Improved Time Bounds for Near-Optimal Sparse Fourier Representations

A. C. Gilbert[a], S. Muthukrishnan[b], and M. Strauss[c]

[a]Dept. of Mathematics, Univ. of Michigan. Supported in part by NSF DMS 0354600.
annacg@umich.edu;
[b]Rutgers Univ. supported in part by NSF DMS 0354600 and NSF ITR 0220280.
muthu@cs.rutgers.edu;
[c]Depts. of Math and EECS, Univ. of Michigan. Supported in part by NSF DMS 0354600.
martinjs@umich.edu

ABSTRACT

We study the problem of finding a Fourier representation $\mathbf{R}$ of m terms for a given discrete signal $\mathbf{A}$ of length N. The Fast Fourier Transform (FFT) can find the optimal N-term representation in time $O(N \log N)$ time, but our goal is to get *sublinear* time algorithms when $m \ll N$.

Suppose $\|\mathbf{A}\|_2 \leq M \|\mathbf{A} - \mathbf{R}_{\rm opt}\|_2$, where $\mathbf{R}_{\rm opt}$ is the optimal output. The previously best known algorithms output $\mathbf{R}$ such that $\|\mathbf{A} - \mathbf{R}\|_2^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{R}_{\rm opt}\|_2^2$ with probability at least $1-\delta$ in time* $\mathrm{poly}(m, \log(1/\delta), \log N, \log M, 1/\epsilon)$. Although this is sublinear in the input size, the dominating expression is the polynomial factor in m which, for published algorithms, is greater than or equal to the bottleneck at m^2 that we identify below. Our experience with these algorithms shows that this is serious limitation in theory and in practice. Our algorithm beats this m^2 bottleneck.

Our main result is a significantly improved algorithm for this problem and the d-dimensional analog. Our algorithm outputs an $\mathbf{R}$ with the same approximation guarantees but it runs in time

$$m \cdot \mathrm{poly}(\log(1/\delta), \log N, \log M, 1/\epsilon).$$

A version of the algorithm holds for all N, though the details differ slightly according to the factorization of N. For the d-dimensional problem of size $N_1 \times N_2 \times \cdots \times N_d$, the linear-in-$m$ algorithm extends efficiently to higher dimensions for certain factorizations of the N_i's; we give a quadratic-in-m algorithm that works for any values of N_i's.

This article replaces several earlier, unpublished drafts.

Keywords: Fourier analysis, sparse analysis, sampling, randomized approximation algorithms

1. INTRODUCTION

In many computational applications of Fourier analysis, we are interested only in a small number m of the coefficients. The large coefficients capture the major time-invariant wave-like features of the signal, while the smaller ones contribute little information about the signal. The largest few Fourier coefficients are useful in data compression, feature extraction, finding approximate periods, and data mining. The problem of finding the m largest Fourier coefficients of a signal is a fundamental task in computational Fourier analysis. We address the problem of how to find and estimate these coefficients quickly and accurately.

Let us denote the optimal m-term Fourier representation of a signal $\mathbf{A}$ of length N by $\mathbf{R}_{\rm opt}$ and assume that, for some M, we have $(1/M) \leq \|\mathbf{A} - \mathbf{R}_{\rm opt}\|_2 \leq \|\mathbf{A}\|_2 \leq M$. Our main result in this paper is an algorithm that uses at most $m \cdot (\log(1/\delta), \log N, \log M, 1/\epsilon)^{O(1)}$ space and time and outputs a representation $\mathbf{R}$ such that $\|\mathbf{A} - \mathbf{R}\|_2^2 \leq (1+\epsilon)\|\mathbf{A} - \mathbf{R}_{\rm opt}\|_2^2$, with probability at least $1-\delta$. We now give some remarks.

Here, the probability is over random choices made by the algorithm, not over the signal, $\mathbf{A}$. That is, we present a coin-flipping algorithm. Knowing the algorithm but not coin-flip outcomes, an adversary chooses the worst possible $\mathbf{A}$. Then coins are flipped and the algorithm proceeds deterministically from the coin flips and the given signal. For each signal, with

*The expression $\mathrm{poly}()$ denotes a polynomial function of the input.

high probability, the algorithm succeeds. It is *not* true that, with high probability, the algorithm succeeds simultaneously on all signals.

Note that the promised time is much less than N. So our algorithm does not read all the input; we assume that our algorithm can read $\mathbf{A}[t]$ for a t of its choice in constant time. It turns out that the t's where the algorithm looks at the signal are chosen randomly (from a non-uniform distribution) but do not adapt to the signal, $\mathbf{A}$.

Since time m is needed just to output the coefficients, our cost is optimal in the parameter m. We also give extensions to some multidimensional cases, paying a factor $d^{O(1)}$ for a d-dimensional problem. Previously known results[1–3] give similar bounds except for the dependence on m, which is linear in our algorithm, and at least quadratic in the other algorithms. This represents a much-needed, significant improvement. Other work[4] bounds the number of samples somewhat better than ours, but that algorithm[4] is at least linear in N.

There are three previously published papers on this problem; unfortunately, there are some missing links in citations. In the first, breakthrough result,[1] the author studies a variation of our problem and presents an algorithm to which one can immediately reduce our problem for the case N a power of 2 (but not for other N). Later,[2] the authors, unaware of the previous result,[1] give an algorithm for any N, power-of-2 or not, with cost polynomial in $(m \log(N))$. Recently, independently of that,,[2] another work presents[3] an algorithm with cost polynomial in $(m \log(N))$ for values of N beyond just the power of 2 previously considered.[1] The motivating applications in these three papers were quite different: learning,[1] DFT approximation,[2] and list decoding for producing hard-core predicates in cryptography.[3] Unfortunately, the dependence on m in all of these papers is quite high. In some of these results,[1, 2] the cost is "polynomial in m"; a close look at the cost[†] reveals it to be at least $m^{\geq 2}$. Later,[3] the cost is at least $m^{\geq 5.5}$ (the heart of the procedure is in Section 7.2.4 where this complexity emerges). The polynomial in other factors is reasonable.

Previous papers[1–3] focused on learning, complexity theory and sampling complexity respectively. In contrast, our focus is on practical applications of Fourier methods. There are a number of applications, e.g., pseudospectral methods for differential equations,[5] finding approximate correlation of signals,[6] and deconvolving blurred signals,[7] where the best m Fourier coefficients suffice and currently, the full DFT is used instead. The full DFT, however, is a computational bottleneck in these applications. This motivated us to consider sampling algorithms for estimating the m best Fourier terms more efficiently. We consider three algorithms: FFTW[8]—a popular, optimized FFT package, our near-linear-in-m algorithm, and a simplified quadratic-in-m algorithm that, due to its relative simplicity and low overhead, is faster, for small m, than the near-linear-in-m algorithm. We find in practice that, if $m \approx 30$ and $N \approx 4$ million, all three algorithms take about the same amount of time.[9] This shows that asymptotic performance is reached for reasonable values of m and N. We expect comparable performance for certain instances of the multidimensional algorithm. Experimental analysis of an earlier algorithm can also be found.[10]

Our algorithm, at a high level, proceeds in a greedy fashion. Given a signal $\mathbf{A}$, we set the representation $\mathbf{R}$ to zero and consider the residual $\mathbf{A} - \mathbf{R}$. We make progress on lowering $\|\mathbf{A} - \mathbf{R}\|$ by repeatedly

- SAMPLING from $\mathbf{A} - \mathbf{R}$ in approximately m correlated random positions.
- IDENTIFYING a set of "significant" frequencies in the spectrum of $\mathbf{A} - \mathbf{R}$ and
- ESTIMATING the Fourier coefficients of these "significant" frequencies.

Once we have estimated the significant coefficients, we add their contribution to the representation $\mathbf{R}$ and iteratively analyze the residual signal $\mathbf{A} - \mathbf{R}$. This framework is similar to previous work[1–3] although the specific steps differ substantively, and are the achievements of this paper.

From a technical view, there are several $\Omega(m^2)$ bottlenecks in the overall framework. An m-term superposition $\mathbf{A} - \mathbf{R}$ may have only N/m points with non-zero value, in unknown time positions. It follows that one needs to sample approximately m positions and do work $\Omega(m)$ in order to learn anything. It is then critical that this $\Omega(m)$ work not be repeated to learn information about each of the $\Omega(m)$ frequencies in $\mathbf{A}$ or each $\Omega(m)$ frequencies in an intermediate representation $\mathbf{R}$. We break this $\Omega(m^2)$ bottleneck by showing:

[†]While the analyses of the algorithms[1, 2] are not tight in m, one could tighten their analyses and show that the algorithms do, in fact, take time $\Omega(m^2)$.

- how to obtain $\Omega(m)$ samples from $\mathbf{A}$ and $\mathbf{R}$ in $m\text{polylog}(N) = m\log^{O(1)}(N)$ work (despite the fact that $\mathbf{R}$ may contain, say, $m/2$ frequencies),
- how to identify all (at most $O(m)$) significant frequencies in $\mathbf{A} - \mathbf{R}$ in total work $m\text{polylog}(N)$, and
- how to estimate up to $O(m)$ Fourier coefficients in $\mathbf{A}$ at once from m samples with work $m\text{polylog}(N)$.

Note that it takes work approximately m to obtain a *single* sample from a $(m/2)$-term intermediate representation, to identify a *single* significant frequency, or to estimate a *single* coefficient to sufficent accuracy. It follows that a straightforward algorithm for any of these three m-fold steps would cost $\Omega(m^2)$.

The task of sampling m times from an intermediate m-term representation and the task of computing the natural estimates for m Fourier coefficients from m random samples are both forms of the *unequally spaced discrete Fourier transform* problem; *i.e.*; multiplying some $k \times k$ submatrix of the $N \times N$ Fourier matrix by a length-k vector, where $k \approx m$. Many time-($k\text{polylog}(N)$) algorithms are known[11] for this. As for identification, previous work has shown how to identify one significant ω out of m with probability at least $1/m$ by using an m-tap random filter with bandwidth N/m; m repetitions of independently chosen m-tap filters will succeed with reasonable probability but with work m^2. Instead, we use a random filterbank of m filters that share a collection of m taps. Computing the m outputs of the filterbank from m inputs turns out to require just an ordinary DFT, which can be done with work $m\log(m)$. The m outputs of the filterbank replace m independent instantiations of a single random filter.

The three tasks above are iterated as we find more and more frequencies and get better and better approximations to the coefficients for frequencies we have already found. Note that each iteration may find just a single significant frequency; a naive overall upper bound would then be m iterations of work $O(m)$ each, for a total of $O(m^2)$ work. Without substantially modifying the algorithm, we give a new bound of approximately $\log(MN/\epsilon)$ iterations. Our new bound analyzes the decrease in $\|\mathbf{A} - \mathbf{R}\|$ rather than the increase in the number of recovered terms.

1.1. Linear versus Quadratic Algorithms

Certainly a near-linear algorithm is quantitatively better than a quadratic algorithm. In this section, we briefly argue that a near-linear-in-m algorithm is *structurally* better than a quadratic-in-m algorithm. We consider the many DFT applications in which time linear in N is needed, *e.g.*, for data acquisition. Then the ($N\log(N)$)-time FFT algorithm becomes a bottleneck, at least in theory, and we want to consider an approximate algorithm that takes at most time N. Thus, for a near-linear-in-m algorithm, we can make m as large as $N/\text{polylog}(N)$; for a quadratic-in-m algorithm, we can only make m as large as $\sqrt{N}/\text{polylog}(N)$. We now consider the structural effects in three applications: convolutions, coding rate, and denoising.

To compute the convolution of two vectors x and y, we need to multiply their spectra. In the worst case, the non-zeros in $\widehat{x}$ correspond to (approximate) zeros in $\widehat{y}$ and vice versa, in which case our approximate algorithms give no useful results. If the spectra are random, however, then we might hope to get non-zeros in $\widehat{x}$ to correspond with non-zeros in $\widehat{y}$ with high probability. If $m = \sqrt{N}/\text{polylog}(N)$, then we are unlikely to find any collisions, and we get no information about the convolution. On the other hand, if $m = N/\text{polylog}(N)$, then we expect to get $N/\text{polylog}(N)$ collisions, which, depending on the context, might result in a useful approximation to the convolution.

Next, consider coding by choosing Fourier basis functions to have non-zero coefficients, where we require a decoding algorithm that runs in time linear in N. A near-linear-in-m algorithm lets us code $\log\binom{N}{m} \approx N/\text{polylog}(N)$ bits; a quadratic-in-m algorithm only lets us code $\log\binom{N}{m} \approx \sqrt{N}/\text{polylog}(N)$ bits—quadratically worse, as expected. But, in coding applications, it is more natural to measure the *rate* of the code—the number of coded bits divided by the length of the codeword, N—and constant rate is often desired. Thus, by improving a quadratic-in-m algorithm to a linear one, the acheivable rate improves from around $1/\sqrt{N}$ to $1/\text{polylog}(N)$, an exponential improvement.

Finally, consider the following *denoising* problem. There is a true signal consisting of a single Fourier mode, ψ_ω. We observe the signal corrupted by additive Gaussian white noise with expected magnitude σ. What is the threshold for σ below which we can determine ω reliably? With high probability, the largest Fourier coefficient of the noise has square magnitude around $\log(N)\sigma^2$, so, even if we had unlimited time, we can recover ω iff $1 \geq \log(N)\sigma^2$, or $\sigma^2 \leq 1/\log(N)$. A dual formulation of our sampling algorithms can recover coefficients with approximately $1/m$ of the total energy (*i.e.*, square of L^2 norm), and the total energy is dominated by the noise energy, $\sigma^2 N$. Thus we need $1 \geq (1/m)\sigma^2 N$ in order

to find a coefficient with energy 1 from a signal plus noise with energy $\sigma^2 N$; this means $\sigma^2 \leq m/N$. In a quadratic-in-m algorithm, $m \approx \sqrt{N}$, so we can only tolerate $\sigma^2 \approx 1/\sqrt{N}$. In a near-linear-in-$m$ algorithm, $m \approx N/\mathrm{polylog}(N)$, so we can tolerate $\sigma^2 \approx 1/\mathrm{polylog}(N)$, exponentially better, and much closer to the information-theoretic limit of $1/\log(N)$.

1.2. Organization

In Section 2, we provide preliminaries on the Fourier transform. In Section 3, we give some technical lemmas. In Section 4, we present our algorithm. In Section 5, we give higher-dimensional variations. In Section 6 we conclude.

2. PRELIMINARIES

Notation. Let $\mathbf{A} = (\mathbf{A}(0), \ldots, \mathbf{A}(N-1))$ be a signal indexed by t, regarded as an integer mod N. The complex conjugate of a number x is denoted by $\overline{x}$ and the dot product $\langle \mathbf{A}, \mathbf{B} \rangle$ of vectors $\mathbf{A}$ and $\mathbf{B}$ is $\sum_t \mathbf{A}(t)\overline{\mathbf{B}(t)}$. The functions $\psi_\omega(t) = \frac{1}{\sqrt{N}} e^{2\pi i \omega t/N}$, $\omega \in \mathbb{Z}_N$, form an orthonormal basis for $\mathbb{Z}/N$. We can represent $\mathbf{A}$ as a linear combination of basis functions

$$\mathbf{A}(t) = \frac{1}{\sqrt{N}} \sum_{\omega=0}^{N-1} \widehat{\mathbf{A}}(\omega) e^{2\pi i \omega t/N}$$

where

$$\widehat{\mathbf{A}}(\omega) = \langle \mathbf{A}, \psi_\omega \rangle = \frac{1}{\sqrt{N}} \sum_t \mathbf{A}(t) e^{-2\pi i \omega t/N}$$

is the ωth Fourier coefficient of $\mathbf{A}$. The vector $\widehat{\mathbf{A}}$ is the *spectrum* of $\mathbf{A}$. The *energy* of $\mathbf{A}$ is $\|\mathbf{A}\|_2^2$, defined by $\|\mathbf{A}\|_2^2 = \sum_t |\mathbf{A}(t)|^2$. Parseval's equality says that $\sum_t |\mathbf{F}(t)|^2 = \sum_\omega |\widehat{\mathbf{F}}(\omega)|^2$. We refer to $|\widehat{\mathbf{A}}(\omega)|^2$ as the energy of the Fourier coefficient $\widehat{\mathbf{A}}(\omega)$ (or the energy of ω) and, similarly, the energy of a set of Fourier coefficients is the sum of the squares of their magnitudes. We write $\mathbf{F} \star \mathbf{G}$ to denote the convolution, $(\mathbf{F} \star \mathbf{G})(t) = \sum_s \mathbf{F}(s)\mathbf{G}(t-s)$. It follows that $\widehat{\mathbf{F} \star \mathbf{G}} = \sqrt{N}\widehat{\mathbf{F}}\widehat{\mathbf{G}}$. We denote by χ_S the signal that equals 1 on a set S and zero elsewhere. The index to χ_S may be time or frequency; this is made clear from context. The *support* $\mathrm{supp}(\mathbf{F})$ of a vector $\mathbf{F}$ is the set of t for which $\mathbf{F}(t) \neq 0$. For a sparse representation $\mathbf{R}$, We will also write $\widehat{\mathrm{supp}}(\mathbf{R})$ to be the set of ω for which $\widehat{\mathbf{R}}(\omega) \neq 0$. More background on Fourier analysis is available in the literature.[12]

A *formal term* is a pair of frequency and coefficient, but will sometimes be written as $c\psi_\omega$ instead of (c, ω). Similarly, a formal representation is a set of formal terms, but will sometimes be written $\sum_{\omega \in \Lambda} c_\omega \psi_\omega$ instead of $\{(c_\omega, \omega) : \omega \in \Lambda\}$. We say that a formal term $c\psi_\omega$ is *bigger* than another term $c'\phi_{\omega'}$ if $|c| > |c'|$. A frequency ω is η-significant in $\mathbf{A}$, for $\eta > 0$, if $|\widehat{\mathbf{A}}(\omega)|^2 \geq \eta \|\mathbf{A}\|^2$.

Define $\mathbf{H}_{K,N}$ by $\mathbf{H}_{K,N}(t) = \frac{\sqrt{N}}{K}\chi_{[0,K)}$. Then $\widehat{\mathbf{H}}_{K,N}(\omega) = \frac{\sin(K\pi\omega/N)}{K\sin(\pi\omega/N)}$, for $\omega \neq 0$ and $\widehat{\mathbf{H}}_{K,N}(0) = 1$. If $K \leq N$, this is called the "Dirichlet kernel" or "Boxcar Filter" and its energy $\|\mathbf{H}_{K,N}\|^2$ is N/K. We will sometimes write $\mathbf{H}_K$ or $\mathbf{H}$ if N or both K and N are clear from context.

Permutation of Spectra. We define a transformation $\mathcal{P}_{\theta,\sigma}$ as follows. For a given signal $\mathbf{A}$ and number σ that is invertible mod N with inverse equal to σ^*, define $(\mathcal{P}_{\theta,\sigma}\mathbf{A})(t)$ by $(\mathcal{P}_{\theta,\sigma}\mathbf{A})(t) = e^{-2\pi i \theta \sigma^* t/N}\mathbf{A}(t\sigma^*)$. First note that one can sample from $\mathcal{P}_{\theta,\sigma}\mathbf{A}$ with approximately the same cost as sampling from $\mathbf{A}$. Next, by elementary facts from Fourier analysis, we have $\widehat{\mathcal{P}_{\theta,\sigma}\mathbf{A}}(\omega) = \widehat{\mathbf{A}}(\theta + \sigma t)$. Since the map $t \mapsto \theta + \sigma t \bmod N$ is invertible iff σ is, $\mathcal{P}_{\theta,\sigma}$ is a spectral permutation. (A number is invertible mod N if and only if it is relatively prime to N.)

Precision. We assume that all signal entries are bounded by some number, M. Similarly, our output will be accurate only to $\pm 1/M$, additively. Thus we need $2\log(M)$ bits to process inputs and outputs, and our algorithm will be allowed time polylogarithmic in M in the bit model. For the sake of clarity, we omit a thorough discussion of precision; we merely point out potential pitfalls. Certain precision issues are actually critical. For example, a classical algorithm[13] to multiply an m-by-m Vandermonde matrix by a vector of length m (a generalization of the unequally-spaced discrete Fourier transform problem) requires just $O(m\log^2(m))$ multiplications, but arithmetic needs to be carried out to $O(m)$ bits, giving only a quadratic bound in the total work in bits.

Asymptotic Notation. We use $O(f)$ to denote the set of functions that grow at most as fast as f and $\Omega(f)$ to denote the set of functions that grow at least as fast as f. We write $\Theta(f) = O(f) \cap \Omega(f)$. We also write $O(f)$, etc., for an anonymous function in the set $O(f)$.

Randomization. Our algorithms are randomized. That is, for each input $\mathbf{A}$ and $3/4$ of the random choices of our algorithm, the algorithm succeeds. The success probability $1/4$ ("significant") or $3/4$ ("high") can be boosted to as close to 1 as desired ("overwhelming") using standard inexpensive techniques. We sometimes omit details.

Non-adaptivity. We have the following non-adaptive access to $\mathbf{A}$. We toss coins, then, based on the coins, compute a sequence $T \subseteq [0, N)$ of indices, learn $\mathbf{A}(T) = \{(t, \mathbf{A}(t)) : t \in T\}$ and run a deterministic algorithm on T, $\mathbf{A}(T)$, and the coin flip outcomes without further access to $\mathbf{A}$. Thus we say that the sampling is *non-adaptive*. For convenience, however, we will present the algorithm in an adaptive way. More specifically, we will present an algorithm that flips coins as it needs them, and such that the algorithm's actions (but not sample locations) depends on previously-computed values, including coin flips and sample values. Our goal is to bound the time used by the algorithm, which implies a bound on the number of samples made. In practice, one can save factors of $\log(N)$ in time by sampling adaptively or by adaptively deciding to make fewer samples, but this is not the focus of this paper.

3. TECHNICAL LEMMAS

3.1. More on the Dirichlet Kernel

In this section we give a useful technical property of the Dirichlet kernel. The lemma says that if we sample $|\widehat{\mathbf{H}}_k(\omega)|^2$ from ω uniformly picked from a certain type of easily-constructible set we get not much more than the value, $1/K$, that we would get if we sampled ω uniformly from all $\mathbb{Z}_N$. This will be motivated below, where it is used.

LEMMA 3.1. *Let N be any number and fix $K \leq N$. For some constant c, let p_0 be any number with $|p_0| \leq cN/K$ and fix $p \neq p_0$ modulo N. Let b be a random invertible number mod N. Then*

$$E_b\left[|\widehat{\mathbf{H}}_K(b(p-p_0)+p_0)|^2 \,\middle|\, \gcd(b,N)=1\right] \leq O\left(\frac{\log(N)}{K}\right).$$

Proof. Let $\phi(N)$ be the Euler totient function of N—the number of positive integers less than N that are relatively prime to N. Then $\phi(N) \geq \Omega(N/\log(N))$. (Stronger statements hold, especially if N is prime or a power of 2. We omit the proof.)

By symmetry of $\widehat{\mathbf{H}}_K$, we may assume $p_0 \geq 0$. Let $g = \gcd(p-p_0, N)$; it follows that the distribution on $b(p-p_0)+p_0$ is the same as the distribution on $bg + p_0$, where b is a random invertible element. It is easy to see that $|\widehat{\mathbf{H}}_K(\omega)| \leq h(\omega)$, where the envelope $h : \mathbb{Z}_N \to \mathbb{R}$ is

$$h(\omega) = \begin{cases} 1, & |\omega| < \frac{N}{2K} \\ \frac{4N}{K|\omega|}, & \frac{N}{2K} \leq |\omega| \leq N/2. \end{cases}$$

Thus it suffices to show that

$$E_b\left[|h(b(p-p_0)+p_0)|^2 \,\middle|\, \gcd(b,N)=1\right] \leq O\left(\frac{N}{\phi(N)K}\right).$$

Note that $E_b[|h(b)|^2] \leq O(1/K)$.

We now show that $h(bg)^2 \leq O(h(bg+p_0)^2)$ for invertible b, so that we may assume $p_0 = 0$. We may assume that $0 \leq bg \leq bg+p_0 < N/2+N/K$, since, otherwise, by the unimodularity of h, it is easy to see that $h(bg) \leq h(bg+p_0)$. We consider two cases. First suppose $bg < N/K$. Then $bg+p_0 \leq O(N/K)$, so, from the definition of $h()$, $h(bg+p_0)^2 \geq \Omega(1)$. Since $h()$ is always at most 1, it follows that $h(bg) \leq 1 \leq O(h(bg+p_0))$. Now suppose $bg \geq N/K$. Then $bg+p_0 \leq O(bg)$. Again, it follows from the definition of $h()$ that $h(bg) \leq O(h(bg+p_0))$.

To bound $E\left[h(bg)^2 \middle| \gcd(b,N)=1\right]$, proceed as follows. We have

$$\begin{aligned} E\left[h(bg)^2 \middle| \gcd(b,N)=1\right] &= \frac{1}{\phi(N)} \sum_{b\in\mathbb{Z}_N^*} h(bg)^2 \\ &\leq \frac{1}{\phi(N)} \sum_{b\in\mathbb{Z}_N\setminus(N/g)\mathbb{Z}_N} h(bg)^2 \\ &= \frac{g}{\phi(N)} \sum_{0<b<N/g} h(bg)^2 \\ &= \frac{g}{\phi(N)} \sum_{-N/(2g)\leq b<N/(2g), b\neq 0} h(bg)^2, \end{aligned}$$

since, by periodicity, $h((b+jN/g)g) = h(bg)$, for integers j. Since $h(x)^2$ decreases as $|x|$ increases, we have $h(bg)^2 \leq \frac{1}{g}\sum_{0\leq x<g} h(bg-x)^2$, if $0 < bg \leq N/2$. Thus we have, using symmetry of h and double counting $h(N/2)$ when 2 divides N/g,

$$\begin{aligned} E\left[h(bg)^2 \middle| \gcd(b,N)=1\right] &\leq \frac{g}{\phi(N)} \sum_{-N/(2g)\leq b<N/(2g), b\neq 0} h(bg)^2 \\ &\leq \frac{2g}{\phi(N)} \left(\sum_{0<b\leq N/(2g)} h(bg)^2 \right) \\ &\leq \frac{2g}{\phi(N)} \left(\sum_{0<b\leq N/(2g)} \frac{1}{g} \sum_{0\leq x<g} h(bg-x)^2 \right) \\ &\leq \frac{1}{\phi(N)} \left(\left(\sum_{0\leq b<N} h(b)^2 \right) - h(0)^2 + h(\lfloor N/2 \rfloor)^2 \right), \end{aligned}$$

where the last term, $h\left(\lfloor N/2 \rfloor\right)^2$, is only needed if 2 divides N/g. By definition of h, we have $h(0)^2 \geq h\left(\lfloor N/2 \rfloor\right)^2$, so that

$$\begin{aligned} E\left[h(bg)^2 \middle| \gcd(b,N)=1\right] &\leq \frac{1}{\phi(N)} \left(\left(\sum_{0\leq b<N} h(b)^2 \right) - h(0)^2 + h(\lfloor N/2 \rfloor)^2 \right) \\ &\leq \frac{1}{\phi(N)} \sum_{0\leq b<N} h(b)^2 \\ &\leq O\left(\frac{N}{\phi(N)} \cdot \frac{1}{K} \right). \end{aligned}$$

□

3.2. AP-Independence

To estimate the expectation $E[Y]$ of a random variable Y, we can form a collection $\{X_k\}$ of independent and identically-distributed copies of Y. We can then define $X = \frac{1}{K}\sum_{k=1}^{K} X_k$. Then $E[X] = E[Y]$ and $\mathrm{var}(X) = \frac{1}{K}\mathrm{var}(Y)$; the reduced variance will allow us to estimate $E[X]$ from a sample of X. It is known that the X_k's need only be pairwise independent for this variance argument to work. We will now claim that the argument holds for a particular construction of X_k's with somewhat less than pairwise independence.

DEFINITION 3.2. *Fix N and let ϕ be a complex-valued function on $\mathbb{Z}_N$. Fix $K \leq N$. Then a sequence $(X_k : 0 \leq k < K)$ of random variables is called* N-arithmetic-progression-independent *(briefly,* ap-independent*) ϕ^{-1}-distributed, if the joint distribution on the sequence can be obtained by writing $X_k = \phi(t_k)$, where a is a random integer mod N, b is a random invertible integer mod N, and $t_k = a + bk$.*

Note that, if N is a prime, then the family $\{a+bk : 0 \le k < K\}$ for random a and random b, is pairwise independent, so the family $X_k = \phi(a+bk)$ of random variables is pairwise independent, as usual. Our construction differs in some important respects. For example, b is restricted to be invertible mod N, so the values $a+bk$ and $a+bk'$ are guaranteed to be different for $k \neq k'$. It follows that $a+bk$ and $a+bk'$ are guaranteed to be *dependent*, and so X_k and $X_{k'}$ are dependent, except for trivial ϕ. Nevertheless, we show that ap-independent random variables are sufficiently independent for the desired variance reduction.

LEMMA 3.3. *Fix K and N with $K \le N$. Let ϕ be a complex-valued function of $\mathbb{Z}_N$ and let (X_k) be a sequence of K ap-independent ϕ^{-1}-distributed random variables. Let $X = \frac{1}{K}\sum_k X_k$ and let Y be distributed independently and identical to the X_k's. Then $E[X] = E[Y]$, and* $\mathrm{var}(X) \le O\left(\frac{\log(N)}{K}\right)\mathrm{var}(Y)$.

Proof. Note that each t_k is uniformly distributed, so the statement about expectation follows from the linearity of expectation. Note that $E[Y] = \frac{1}{\sqrt{N}}\widehat{\phi}(0)$. From the definition of variance and Parseval's equality,

$$\mathrm{var}(Y) = \frac{1}{N}\left(\sum_t |\phi(t)|^2 - |\widehat{\phi}(0)|^2\right) = \frac{1}{N}\sum_{\omega\neq 0}|\widehat{\phi}(\omega)|^2.$$

Suppose (X_k) is defined on the arithmetic progression $a+bk$, so that $X_k = \phi(a+bk)$. Then $\mathrm{supp}(\mathbf{H}_K) = \{0,1,2\ldots,K-1\}$, so $\mathrm{supp}(\mathcal{P}_{0,-1/b}\mathbf{H}_K) = \{0,-b,-2b,\ldots,-(K-1)b\}$, and so

$$\begin{aligned}(\phi \star \mathcal{P}_{0,-1/b}\mathbf{H}_K)(a) &= \sum_{r+s=a}\phi(r)\left(\mathcal{P}_{0,-1/b}\mathbf{H}_K\right)(s)\\ &= \phi(a)\mathbf{H}_K(0) + \phi(a+b)\mathbf{H}_K(-b) + \phi(a+2b)\mathbf{H}_K(-2b) + \cdots\\ &\qquad +\phi(a+(K-1)b)\mathbf{H}_K(-(K-1)b)\\ &= \frac{\sqrt{N}}{K}\sum_{0\le k<K} X_k,\end{aligned}$$

since $\mathbf{H}_K$ is equal to $\frac{\sqrt{N}}{K}$ on its support. Thus $X = (\phi \star \mathcal{P}_{0,-1/b}\mathbf{H}_K)(a)/\sqrt{N}$. So, conditioned on b and taking variance in a,

$$\mathrm{var}_a(X|b) = \frac{1}{N^2}\sum_{\omega\neq 0}\left|(\phi \star \mathcal{P}_{0,-1/b}\mathbf{H}_K)\widehat{\ }(\omega)\right|^2 = \frac{1}{N}\sum_{\omega\neq 0}\left|\widehat{\phi}(\omega)\widehat{\mathbf{H}}_K(b\omega)\right|^2.$$

We now take expectation of $\mathrm{var}_a(X|b)$ with respect to b. Using Lemma 3.1 with $p=\omega$ and $p_0 = 0$, we have as desired:

$$\mathrm{var}(X) \le \frac{1}{N}\sum_{\omega\neq 0}\left|\widehat{\phi}(\omega)\right|^2 \max_{\omega\neq 0} E_b\left[\left|\widehat{\mathbf{H}}_K(b\omega)\right|^2\right] \le O\left(\frac{\log(N)}{KN}\right)\sum_{\omega\neq 0}\left|\widehat{\phi}(\omega)\right|^2.$$

□

In the sequel, we will substitute AP-independent random variables for pairwise independent random variables. When we do, Lemma 3.1, even for the special case of $p_0 = 0$, suffices to insure that the variance reduction succeeds. The more general Lemma 3.1 for $p_0 \neq 0$ will be used in a different context.

4. ALGORITHM

4.1. Overview

In a bit more detail than above, each iteration of our algorithm proceeds as follows:

- SAMPLE from $\mathbf{A}-\mathbf{R}$ in $K \approx m$ correlated random positions, where $\mathbf{R}$ has L terms, in total time $(K+L)\log^{O(1)}(N)$. (The locations of the samples are defined by the following operations.)
  - Sample from the given signal $\mathbf{A}$ in time $O(1)$ per sample, as hypothesized.

– Sample from $\mathbf{R}$ by performing an unequally-spaced fast Fourier transform.

- IDENTIFY a set of "significant" frequencies in the spectrum of $\mathbf{A} - \mathbf{R}$.
  - ISOLATE one or more modes of $\mathbf{A} - \mathbf{R}$. Generate a set $\{\mathbf{F}_k : k < K\}$ of K new signals from $\mathbf{A} - \mathbf{R}$ where $K \leq O(1/\eta)$ is sufficiently large, so that each ω that is η-significant in $\mathbf{A} - \mathbf{R}$ is likely to be $(1-\gamma)$-significant in some $\mathbf{F}_k$ for some small constant γ.
    * PERMUTE the spectrum of $\mathbf{A} - \mathbf{R}$ by a random dilation σ, getting $\mathcal{P}_{\sigma,0}(\mathbf{A} - \mathbf{R})$
    * FILTER $\mathcal{P}_{\sigma,0}(\mathbf{A} - \mathbf{R})$ by a filterbank of approximately m equally-spaced frequency-domain translations of the Boxcar Filter with bandwidth approximately N/m and approximately m common taps. Get m new signals, some of which have a single overwhelming Fourier mode, $\sigma\omega$, corresponding to significant mode ω in $\mathbf{A} - \mathbf{R}$.
    * Undo the above permutation, thereby making ω overwhelming instead of $\sigma\omega$.
  - GROUP-TEST each new signal to locate the one overwhelming mode, ω. Learn the bits of ω one at a time, least to most significant. *E.g.*, for the least significant bit:
    * PROJECTION. Project each $\mathbf{F}_k$ onto the space of even frequencies and the space of odd frequencies.
    * NORM ESTIMATION. Estimate the norm of each projection, thereby learning the least significant bit of ω.
- ESTIMATE the Fourier coefficients of these "significant" frequencies by computing the Fourier coefficients of the sampled residual using an unequally-spaced fast Fourier transform algorithm and normalizing appropriately.
- ITERATE in a greedy pursult.

We now consider the pieces, one at a time.

4.2. Sampling from a Representation

Given a formal representation $\mathbf{R} = \sum_{\omega \in \Lambda} c_\omega \psi_\omega$ and a set T of times, we want $\{\mathbf{R}(t) : t \in T\}$. Then our problem is a form of unequally-spaced discrete Fourier transform; that is, multiply the Fourier submatrix $F_{\Lambda,T}$ by the vector c_Λ:

$$\begin{pmatrix} \varphi^{t_1\omega_1} & \varphi^{t_1\omega_2} & \dots & \varphi^{t_1\omega_{|\Lambda|}} \\ \varphi^{t_2\omega_1} & \varphi^{t_2\omega_2} & \dots & \varphi^{t_2\omega_{|\Lambda|}} \\ \vdots & \ddots & & \\ \varphi^{t_{|T|}\omega_1} & \varphi^{t_{|T|}\omega_2} & \dots & \varphi^{t_{|T|}\omega_{|\Lambda|}} \end{pmatrix} \cdot \begin{pmatrix} c_{\omega_1} \\ c_{\omega_2} \\ \vdots \\ c_{\omega_{|\Lambda|}} \end{pmatrix},$$

where $\varphi = \frac{1}{\sqrt{N}} e^{2\pi i/N}$. For simplicity, let $K = |\Lambda| + |T|$. There are algorithms[11, 14] to compute this matrix vector multiplication in time $K\text{polylog}(K)$, with only $O(\log(K))$ overhead in the number of bits of precision. This problem is similar to a problem below involving estimating coefficients, where the matrix is transposed. For completeness, we sketch an algorithm for that problem; a similar algorithm works for this problem.

4.3. IDENTIFICATION of Significant Modes

Recall that the IDENTIFICATION step consists of ISOLATION and GROUP TESTING. We address isolation first.

4.3.1. ISOLATION

We use the Dirichlet kernel $\mathbf{H}_K$, for appropriate K, to isolate frequencies. More precisely, we will use a filterbank of K modulations of the Dirichlet kernel (*i.e.*, K translations in the frequency domain). We will first permute the spectrum of a given signal by a random dilation, then filter with $\mathbf{H}_K$, and then undo the dilation. Equivalently (and more conveniently), we will apply a random dilation to the filters in the filterbank, as follows:

DEFINITION 4.1. *Given signal* $\mathbf{A}$ *of length* N*, and given parameter* K *a power of 2, a* K*-shattering of* $\mathbf{A}$ *is a collection* $\left\{ \frac{1}{\sqrt{N}} (\mathcal{P}_{kN/K,\sigma} \mathbf{H}_K) \star \mathbf{A} \right\}_k$*, where* σ *is a random number invertible mod* N*.*

The next lemma guarantees that a K-shattering isolates significant frequencies. That is, each desired frequency is isolated in some element of the shattering with probability close to 1.

LEMMA 4.2. *Let γ be any positive constant. Let ω_0 be an η-significant frequency in $\mathbf{A}$, so that $|\widehat{\mathbf{A}}(\omega_0)|^2 \geq \eta \|\mathbf{A}\|^2$. Then, for some $K \leq O(\log(N)/(\gamma\eta))$, in a K-shattering $\{\mathbf{F}_k\}_k$ of $\mathbf{A}$, with probability $\Omega(1)$, there exists $k \leq K$ such that $|\widehat{\mathbf{F}}_k(\omega_0)|^2 \geq (1-\gamma)\|\mathbf{F}_k\|^2$.*

Proof. Note that there is some k such that

$$|\sigma\omega_0 + kN/K| \leq N/(2K). \tag{1}$$

Then, by the properties of the Dirichlet kernel, $|\widehat{(\mathcal{P}_{kN/K,\sigma}\mathbf{H}_K)}(\omega_0)| = |\widehat{\mathbf{H}}_K(\sigma\omega_0 + kN/K)| \geq 2/\pi$, whence, for some constant c, $|\widehat{\mathbf{F}}_k(\omega_0)|^2 = |\widehat{\mathbf{H}}_K(\sigma\omega_0 + kN/K)|^2|\widehat{\mathbf{A}}(\omega_0)|^2 \geq c\eta\|\mathbf{A}\|^2$. Next, for each k,

$$\|\mathbf{F}_k\|^2 - |\widehat{\mathbf{F}}_k(\omega_0)|^2 = \sum_{\omega\neq\omega_0} |\widehat{\mathbf{F}}_k(\omega)|^2 = \sum_{\omega\neq\omega_0} |\widehat{\mathbf{A}}(\omega)|^2|\widehat{\mathbf{H}}_K(\sigma\omega + kN/K)|^2.$$

Lemma 3.1 implies that $E_\sigma\left[|\widehat{\mathbf{H}}_K(\sigma\omega + kN/K)|^2 \middle| (1) \text{ holds}\right] \leq \frac{1}{4}c\gamma\eta$, so that

$$E\left[\|\mathbf{F}_k\|^2 - |\widehat{\mathbf{F}}_k(\omega_0)|^2 \middle| (1) \text{ holds}\right] \leq \frac{1}{4}c\gamma\eta\|\mathbf{A}\|^2.$$

By Markov, it follows that, for each k, with probability at least $3/4$, if (1) holds, then $\|\mathbf{F}_k\|^2 - |\widehat{\mathbf{F}}_k(\omega_0)|^2 \leq c\gamma\eta\|\mathbf{A}\|^2$. In that case, $|\widehat{\mathbf{F}}_k(\omega_0)|^2 \geq (1-\gamma)\|\mathbf{F}_k\|^2$, as desired. □

LEMMA 4.3. *Given sampling access to a signal $\mathbf{A}$, and a dilation σ, one can compute the K-shattering*

$$\left\{\frac{1}{\sqrt{N}}(\mathcal{P}_{kN/K,\sigma}\mathbf{H}_K) \star \mathbf{A}\right\}_k$$

in time $O(K\log(K))$.

Proof. We assume $\sigma = 1$; the general case is similar. From the definitions, $\left\{(\mathcal{P}_{kN/K,1}\mathbf{H}_K) \star \mathbf{A}\right\}_k$ is the DFT of the sequence $\mathbf{A}[0], \mathbf{A}[1], \ldots, \mathbf{A}[K-1]$, which can be computed efficiently by the FFT algorithm. □

4.3.2. GROUP TESTING

Now we show how to project the signal approximately onto the space of even and odd frequencies. We also generalize this to other pairs of subspaces that we will need. Note that we do *not* assume that N is even.

DEFINITION 4.4. *Define the filterbank pair $\mathbf{G}_n^\pm$ by $\mathbf{G}_n^\pm = \left(\frac{\sqrt{N}}{2}\right)\left(\delta_0 \pm \delta_{\lfloor N/2^{n+1}\rfloor}\right)$.*

The next Lemma shows how these filters can be used to find the $(n+1)$'s least significant bit of ω, provided we know the n least significant bits (by induction) and a few of the most significant bits, which we can assume by trying all possibilities exhaustively. If we know the values of any bit positions in an otherwise unknown ω_0, we can modulate the relevant signal (*i.e.*, translate the spectrum), so that all of the known bit positions become zeros, which we assume below. This does not change any of the unknown bit positions. Below, we can assume that $0 \leq \omega_0 < N\gamma/\pi$ by exhaustively trying all π/γ possibilities for the most significant $\log(\pi/\gamma) \leq O(1)$ bits of ω_0.

LEMMA 4.5 (PROJECTION). *For all $\gamma > 0$ and all n, if $0 \leq \omega_0 < N\gamma/\pi$ and $\omega_0 = 0 \bmod 2^n$, then*

$$\begin{cases} 1-\gamma \leq |\widehat{\mathbf{G}_n^+}(\omega_0)|^2 \leq 1 \quad \textit{and} \quad |\widehat{\mathbf{G}_n^-}(\omega_0)| \leq \gamma, \quad \textit{if } \omega_0 = 0 \bmod 2^{n+1} \\ 1-\gamma \leq |\widehat{\mathbf{G}_n^-}(\omega_0)|^2 \leq 1 \quad \textit{and} \quad |\widehat{\mathbf{G}_n^+}(\omega_0)| \leq \gamma, \quad \textit{if } \omega_0 = 2^n \bmod 2^{n+1}. \end{cases}$$

Proof. For N a power of 2, the result is known from earlier.[1] We sketch that algorithm as an ideal case. We then generalize the result to other N.

So assume N is a power of 2. Define $\mathbf{F}_0^\pm = (\sqrt{N}/2)(\delta_0 \pm \delta_{N/2})$. It's easy to see that

$$\widehat{\mathbf{F}_0^+}(\omega) = \begin{cases} 1, & \omega \text{ even} \\ 0, & \omega \text{ odd}, \end{cases}$$

and, similarly, $\widehat{\mathbf{F}_0^-}(\omega)$ is 0 or 1, respectively, depending on whether ω is even or odd. More generally, define $\mathbf{F}_n^\pm = (\sqrt{N}/2)(\delta_0 \pm \delta_{N/2^{n+1}})$. It's easy to see that $\widehat{\mathbf{F}_n^+}(\omega) = 1$ if $\omega = 0 \bmod 2^{n+1}$ and $\widehat{\mathbf{F}_n^+}(\omega) = 0$ if $\omega = 2^n \bmod 2^{n+1}$; $\widehat{\mathbf{F}_n^-}(\omega) = 1 - \widehat{\mathbf{F}_n^+}(\omega)$ if $\omega = 0 \bmod 2^{n+1}$.

Now consider general N. We cannot use the ideal filters $(\sqrt{N}/2)(\delta_0 \pm \delta_{N/2})$, because $N/2$ is not necessarily an integer. Instead, we will define $\mathbf{G}_0^\pm = (\sqrt{N}/2)(\delta_0 \pm \delta_{\lfloor N/2 \rfloor})$. Then

$$\widehat{\mathbf{G}_0^+}(\omega) = \frac{1 + e^{2\pi i \omega \lfloor N/2 \rfloor / N}}{2}.$$

Put $\Delta = N/2 - \lfloor N/2 \rfloor$, so $\Delta = 0$ or $\Delta = 1/2$. Thus

$$\widehat{\mathbf{G}_0^+}(\omega) = \frac{1 + e^{-2\pi i \omega \Delta / N}(-1)^\omega}{2}.$$

Now, suppose $0 \le \omega_0 < N\gamma/\pi$. Then $\pi\omega_0\Delta/N \le \gamma/2$. Thus, if ω_0 is even, $\widehat{\mathbf{G}_0^+}(\omega_0) = (1 + e^{-2\pi i \omega_0 \Delta/N})/2 \approx 1 + \pi i \omega_0 \Delta/N$ so that $|\widehat{\mathbf{G}_0^+}(\omega_0) - 1| \le \gamma$ and, similarly, $|\widehat{\mathbf{G}_0^-}(\omega_0)| \le \gamma$, provided γ is small enough to force $\pi\omega/N \le \gamma$ to be small enough that $|1 - e^{-2\pi i \omega_0 \Delta/N}| \le 4\pi\omega_0\Delta/N$. Similarly, if ω_0 is odd, then $|\widehat{\mathbf{G}_0^+}(\omega_0)| \le \gamma$ and $|\widehat{\mathbf{G}_0^-}(\omega_0) - 1| \le \gamma$. It follows that, for sufficiently small $\gamma \ge \Omega(1)$, the filters $\mathbf{G}_0^\pm$ behave similarly to $\mathbf{F}_0^\pm$, provided $0 \le \omega_0 < N\gamma/\pi$; *i.e.*, provided the most significant $\log(\pi/\gamma)$ bits of ω_0 are zero. Similar considerations hold for $\mathbf{G}_k^\pm$, $k > 0$. In general, we will perturb $t \approx N/2^n$ by $\Delta \le 1/2$ to get an integer $\lfloor N/2^n \rfloor$ or $\lceil N/2^n \rceil$; we then need $|\omega| \le N\gamma/\pi$ so that the perturbation $2\pi i \omega \Delta/N$ is at most γ, so that $(e^{2\pi i \Delta t/N} - 1)/2 \approx \gamma/2 \le \gamma$. Finally, $|\widehat{\mathbf{G}_n^\pm}(\omega)|^2 \le 1$ for all ω by the triangle inequality. □

4.4. Norm Estimation

Next, we show how to estimate norms (equivalently, energies), by sampling. We cannot estimate norms reliably (for example, if the signal consists of a single spike, our sampling algorithm cannot find the spike to learn its height). But we can get a certain one-sided estimate that suffices for us. Similar lemmas appeared earlier.[1, 2, 10] Below, we sketch a proof.

DEFINITION 4.6. *For integer J and signal $\mathbf{A}$, define the estimator $\|\mathbf{A}\|_{\sim J}^2$ (or $\|\mathbf{A}\|_\sim^2$, if J is understood) as follows. Choose J points t at random. Take the median of $\frac{N}{8}|\mathbf{A}(t)|^2$ over the points t.*

LEMMA 4.7 (NORM ESTIMATION). *There exists constants $\alpha > 0$ and $\beta > 0$, and, for all $\delta > 0$, there exists $J \le O(\log(1/\delta))$ such that the following hold with probability at least $1 - \delta$:*

- $\|\mathbf{A}\|_{\sim J}^2 \le \|\mathbf{A}\|^2$ *(for any $\mathbf{A}$);*
- *if $|\widehat{\mathbf{A}}(\omega)|^2 \ge (1-\alpha)\|\mathbf{A}\|^2$ for some ω, then $\|\mathbf{A}\|_{\sim J}^2 \ge \beta \|\mathbf{A}\|^2$.*

Proof. Let $X = N|\mathbf{A}(t)|^2$ for random t Then $E[X] = \|\mathbf{A}\|^2$. First consider the upper bound. Since $X \ge 0$, by the Markov inequality, it follows that $\Pr(X > 8E[X]) \le 1/8$.

Now consider the lower bound. Write $\psi = \widehat{\mathbf{A}}(\omega)\psi_\omega$ and $\nu = \mathbf{A} - \psi$, where $\langle \psi, \nu \rangle = 0$. Let $T = \{t : \frac{1}{8}N|\mathbf{A}(t)|^2 < \beta\|\mathbf{A}\|^2\}$ for some constant β to be determined; *i.e.*, T is the set of positions t that would cause $\frac{1}{8}X(t)$ to be a severe underestimate of the energy. For $t \in T$, we have $|\nu(t)|^2 \ge (|\psi(t)| - |\mathbf{A}(t)|)^2 \ge \frac{(\sqrt{1-\alpha} - \sqrt{8\beta})^2}{N}\|\mathbf{A}\|^2 = \frac{\gamma}{N}\|\mathbf{A}\|^2$. Thus $\alpha\|\mathbf{A}\|^2 \ge \|\nu\|^2 \ge |T| \cdot \frac{\gamma}{N}\|\mathbf{A}\|^2$, so $|T|/N < 1/8$ for appropriate α, β, and γ. That is, $\Pr\left(\frac{1}{8}X < \beta\|\mathbf{A}\|^2\right) = |T|/N < 1/8$. Thus, with probability at least 3/4, we have $\beta\|\mathbf{A}\|^2 \le \frac{1}{8}X \le \|\mathbf{A}\|^2$ ("success"). If the median of J repetitions fails then at least $J/2$ of the repetitions fail, which has probability at most $e^{-\Omega(J)} = \delta$ by the Chernoff equality. □

4.5. ESTIMATION of Coefficients for Significant Frequencies

In the previous step we generate a set of L significant frequencies. In the ESTIMATION step, we estimate the contribution each significant frequency makes to the signal; *i.e.*, we estimate its Fourier coefficient. Our technique allows us to estimate several coefficients (not just one) with bulk sampling. In what follows, we define an estimator, show that the estimator approximates Fourier coefficients, then show how to compute the estimator efficiently in bulk.

DEFINITION 4.8. *Given frequency ω and parameters J and K, define the random variable $\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega)$ as:*

$$\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega) = \text{median}_{j\leq J}\text{mean}_{k\leq K}\sqrt{N}\mathbf{A}(t_{j,k})e^{-2\pi i\omega t_{j,k}/N},$$

where, independently for each j and each k, we pick a random $t_{j,k}$. The median of the set of complex numbers is performed by taking medians separately in the real and imaginary directions.

LEMMA 4.9. *For any signal $\mathbf{A}$, any frequency ω, and any parameters $\eta > 0$ and $\delta > 0$, there are a $J \leq O(\log(1/\delta))$ and a $K \leq O(1/\eta)$, such that $\left\|\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega) - \widehat{\mathbf{A}}(\omega)\right\|^2 \leq \eta \|\mathbf{A}\|^2$ with probability at least $1-\delta$.*

Proof. (Sketch.) Let $X = N\mathbf{A}(t)\overline{\psi_\omega(t)}$. Then $E[X] = \widehat{\mathbf{A}}(\omega)$ and $\text{var}(X) \leq \|\mathbf{A}\|^2$. Then the mean of K copies has variance at most $O(\eta\|\mathbf{A}\|^2)$. Then take a median of $J \leq O(\log 1/\delta)$ trials of this random variable. The standard Chernoff inequality guarantees our result. □

Henceforth, assume $J = 1$. At this point, we have not discussed the computation of $\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega)$, though it is straightforward to compute each $\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega)$ in time $JK = \Theta(J/\eta)$. One can see from the definitions, however, that computation of all K estimates (where, wlog, $J = 1$), is an unequally-spaced discrete Fourier transform problem. Thus we have

LEMMA 4.10 (BULK ESTIMATION). *There exists a sampling algorithm that takes oracle $\mathbf{A}$ and takes inputs K, L, M, N and L frequencies ω_l, such that the algorithm has time cost $(L+K)\log^{O(1)}(N)\log(M)$, and the algorithm outputs a sample from the distribution on the vector $\left(\widetilde{\widehat{\mathbf{A}}}_{1,K}(\omega_\ell)\right)_\ell$ of specified Fourier coefficient estimates to within precision $\pm\frac{1}{M}\|\mathbf{A}\|$, additively.*

In fact, only pairwise independent set of t's will suffice for the variance reduction above. Furthermore, by Lemma 3.3, we can take the t_k's to be points on a random arithmetic $a + bk$ progression with k invertible, which results in a pairwise independent family if N is prime, but generally not otherwise. If the t_k's are on a random arithmetic progression, then the computation can become simpler, in theory and practice. For completeness, we sketch a simple algorithm, a variation of which appeared earlier.[15] A similar algorithm works for bulk sampling, Section 4.2, above.

Our goal is to compute $\sum_k \mathbf{A}(t_k)e^{-2\pi i\omega t_k/N} = \sum_k \mathbf{A}(a+bk)e^{-2\pi i\omega(a+bk)/N}$ for each ω in some set Λ. We assume $a = 0$ and $b = 1$; the general case is similar. Then $\sum_k \mathbf{A}(k)(e^{-2\pi i\omega/N})^k$ is the evaluation of a degree-K polynomial p on each of $L = |\Lambda|$ complex points of unit norm. Note that if Λ is a subset of the set Λ^* of equally-spaced ("cyclotomic") points around the unit circle, then we would be asking for the (ordinary) discrete Fourier transform of the sequence $\mathbf{A}(0),\ldots,\mathbf{A}(K-1)$, for which there are efficient algorithms. Instead, compute $p, p', p'', \ldots$ on Λ^*, where p', p'', etc., are the first few derivatives of p, computed termwise. Then approximate $p(\omega)$ for each $\omega \in \Lambda$ by expanding p in a Taylor polynomial at a cyclotomic point near ω. One can verify that the convergence of the Taylor series is exponential, so just a small number of terms are needed. We can call this lemma J times and take a median to compute $\widetilde{\widehat{\mathbf{A}}}_{J,K}(\omega_k)$ for $J > 1$.

Although Lemma 4.10 suffices in theory, a better estimator will be useful both for our analysis and in practice. We want to bound the coefficient error in terms of the possibly much smaller quantity $\|\mathbf{A}-\mathbf{R}^*\|^2 \leq \|\mathbf{A}\|^2$, where $\mathbf{R}^*$ is the representation over Λ with optimal coefficients. This is done iteratively. A proof of the following appeared earlier.[10]

LEMMA 4.11 (ITERATIVE BULK ESTIMATION). *There exists a sampling algorithm that takes oracle $\mathbf{A}$ and takes inputs (δ,ϵ,L,M,N) and L frequencies ω_l, such that, with probability at least $1-\delta$, the algorithm outputs an L-term representation $\mathbf{R}$ such that, for each $\ell < L$, $|\widehat{\mathbf{A}-\mathbf{R}}(\omega_\ell)|^2 \leq (\epsilon/L)\|\mathbf{A}-\mathbf{R}^*\|^2$, where $\mathbf{R}^*$ is the optimal L-term representation, the algorithm has time cost bounded by $L\log(1/\delta)\log^{O(1)}(N)\log(M)/\epsilon$, and where we assume $\|\mathbf{A}-\mathbf{R}^*\|^2 \geq \frac{1}{M}\|\mathbf{A}\|^2$.*

4.6. ITERATION upon IDENTIFICATION and ESTIMATION

Above we showed, for any η and L, given a signal $\mathbf{A}$ as oracle, and a representation $\mathbf{R}$ of at most L terms we can, in time $(L/\epsilon+1/\eta)(\log(N)+\log(1/\eta))^{O(1)}\log(1/\delta)$, find a list Λ containing all frequencies ω with $|\widehat{\mathbf{A}-\mathbf{R}}(\omega)|^2 \geq \eta \|\mathbf{A}-\mathbf{R}\|^2$ and, for each $\omega \in \Lambda$, estimate $\widehat{\mathbf{A}-\mathbf{R}}(\omega)$ as $\widetilde{\widehat{\mathbf{A}-\mathbf{R}}}(\omega)$ with

$$\left|\widetilde{\widehat{\mathbf{A}-\mathbf{R}}}(\omega) - \widehat{\mathbf{A}-\mathbf{R}}(\omega)\right|^2 \leq (\epsilon/L)\|\mathbf{A}-\mathbf{R}^*\|^2,$$

where $\mathbf{R}^*$ is the projection of $\mathbf{A}$ onto the frequencies in Λ. We now show how to use this repeatedly to recover an approximate representation for the signal.

In the sequel, L and $1/\eta$ can be as large as m, so a single round of identification and estimation may already take time linear in m. On the other hand, it is possible that, for some intermediate representation $\mathbf{R}$ with $\|\mathbf{A}-\mathbf{R}\|^2 \gg \|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$, a single iteration of greedy pursuit only produces a single new frequency. So a straightforward bound on the number of iterations would be m, giving a runtime bound of m^2. Instead, we will bound the number of interations independently of m (ignoring log factors), by considering the decrease in $\|\mathbf{A}-\mathbf{R}\|^2$ rather than counting the number of new frequencies.

Our goal will be a representation $\mathbf{R}$ such that $\|\mathbf{A}-\mathbf{R}\|^2$ is bounded by either $(1+\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$ or $1/M^2$. Intuitively, the bound of $(1+\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$ holds in the noisy case and $1/M^2$ holds when $\mathbf{A}$ is an exact superposition. In the latter case, note that the coefficients in $\mathbf{R}_{\rm opt}$ may be irrational even if the input values in $\mathbf{A}$ are integers, so a scheme that outputs an answer correct to $\pm 1/M$, *i.e.*, $\log(M)$ bits, in time $\log^{O(1)}(MN)$ is considered "exact output."

Next we give our algorithm. The algorithm will use positive constants c_0, c_1, c_2 and our analysis will use additional positive constants c_3, etc., described below.

ALGORITHM 4.12. Input: $M, N, \epsilon, m, \delta$
Oracle: $\mathbf{A}$, representing a signal of length N
Output: m-term representation, $\mathbf{R}$, such that, with probability at least $1-\delta$, we have $\|\mathbf{A}-\mathbf{R}\|^2 \leq (1+\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$ or $\|\mathbf{A}-\mathbf{R}\|^2 \leq 1/M^2$.

$\mathbf{R}_0 = 0$, formally, and $\Lambda_0 = \emptyset$
$T = c_0 \log(M/\epsilon)/\epsilon^2$
for($t = 0$; $t \leq T$; t^{++}){
  Find a list L of ω with $\left|\widehat{\mathbf{A}-\mathbf{R}}_t(\omega)\right|^2 \geq c_1 \frac{\epsilon^2}{m}\|\mathbf{A}-\mathbf{R}_t\|^2$.
  Put $\Lambda_{t+1} = L \cup \Lambda_t$.
  Estimate coefficients for $\omega \in \Lambda_{t+1}$ so that $|\widetilde{\widehat{\mathbf{A}-\mathbf{R}}}_{t+1}(\omega)|^2 \leq \frac{c_2\epsilon^2}{|\Lambda_{t+1}|+m}\left\|\mathbf{A}-\mathbf{A}_{\Lambda_{t+1}}\right\|^2$
}
Output $\mathbf{R}_T^m$, the top m terms in $\mathbf{R}_T$.

□

We now proceed with analysis. First we show that we can assume the signal is noisy. Starting with a signal, $\mathbf{A}$, add a spike (*i.e.*, noise), ν, with $\|\nu\| = \frac{\epsilon}{36M}$, in a random location, and with sign chosen so that, for any $(N/4)$-term representation $\mathbf{R}$, $\|\mathbf{A}+\nu-\mathbf{R}\| \geq \frac{\epsilon}{144M}$. Then find a $(1+\epsilon)$-factor approximation $\mathbf{R}$ to $\mathbf{A}+\nu$ and return it as an representation for $\mathbf{A}$. It is straightforward to check that $\|\mathbf{A}-\mathbf{R}\|^2$ is bounded by either $(1+\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$ or $1/M^2$, as desired. Since our sampling algorithm sees ν only with small probability, our algorithm won't change as a result of adding ν; this is a fiction for analysis only. So, henceforth, we assume that, for any $(N/4)$-term representation $\mathbf{R}$, $\|\mathbf{A}-\mathbf{R}\| \geq \frac{\epsilon}{144M}$; we show how to get a relative-error approximation to $\mathbf{A}$.

We now give an informal discussion of correctness. Similar proofs have appeared,[1–3] but some care is needed to keep the overall time cost approximately linear in m rather than quadratic in m, by showing that the number of rounds is roughly independent of m and that the previous lemmata for finding frequencies and estimating coefficients can be used with parameters leading to time linear in m.

Fix an optimal representation, $\mathbf{R}_{\rm opt} = \sum_{\omega\in\Lambda_{\rm opt}} c_\omega \psi_\omega$. Consider the relative improvement $\frac{\|\mathbf{A}-\mathbf{R}_{t+1}\|}{\|\mathbf{A}-\mathbf{R}_t\|}$. For appropriate constant c_3, this ratio cannot be less than $(1-c_3\epsilon^2)$ for all t, since, otherwise, $\|\mathbf{A}-\mathbf{R}_T\| \le (1-c_3\epsilon^2)^T \|\mathbf{A}-\mathbf{R}_0\| \le \epsilon/(144M)$, a contradiction, if $T \le O(\log(M/\epsilon)/\epsilon^2)$ is sufficiently large. (Note that, crucially, T is independent of m.) So suppose, for some j and appropriate c_3, that $\|\mathbf{A}-\mathbf{R}_{j+1}\| > (1-c_3\epsilon^2)\|\mathbf{A}-\mathbf{R}_j\|$.

Define $\Lambda_{\rm opt'} = \{\omega \in \Lambda_{\rm opt} : |c_\omega|^2 \ge (c_4\epsilon/m)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2\}$, for some constant c_4. Note that, if $\mathbf{R}_{\rm opt'} = \sum_{\omega\in\Lambda_{\rm opt'}} c_\omega\psi_\omega$, then $\|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2 \le (1+c_4\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$. Thus we can bound $\|\mathbf{A}-\mathbf{R}\|^2$ in terms of $\|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2$ instead of $\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$.

We next claim that $\Lambda_{\rm opt'} \subseteq \Lambda_{j+1} \subseteq \Lambda_T$. First, toward a contradiction, suppose $\|\mathbf{A}-\mathbf{R}_j\|^2 > 2\|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2$. Then, ignoring the coefficient approximation error in $\mathbf{R}_j$ (by Lemma 4.11, the total square error is much less than $\|\mathbf{A}-\mathbf{R}_j\|^2$), we conclude that $\Lambda_{\rm opt'} \setminus \Lambda_j$ has energy at least $\|\mathbf{A}-\mathbf{R}_j\|^2 - \|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2 \ge (1/2)\|\mathbf{A}-\mathbf{R}_j\|^2$. An iteration of greedy pursuit finds frequencies ω and good approximations to coefficients c_ω provided that $|c_\omega|^2 \ge (c_1\epsilon^2/m)\|\mathbf{A}-\mathbf{R}_j\|^2$; it follows that most of the energy in $\Lambda_{\rm opt'} \setminus \Lambda_j$ is found—all but $c_1\epsilon^2\|\mathbf{A}-\mathbf{R}_j\|^2$ is found—so that $\|\mathbf{A}-\mathbf{R}_{j+1}\|^2 \le (3/4)\|\mathbf{A}-\mathbf{R}_j\|^2$, contradicting the definitions of j and c_3. So suppose $\|\mathbf{A}-\mathbf{R}_j\|^2 \le 2\|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2$. Then $\omega \in \Lambda_{\rm opt'}$ satisfies $|c_\omega|^2 \ge (c_4\epsilon/m)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2 \ge (1+c_4\epsilon)^{-1}(c_4\epsilon/m)\|\mathbf{A}-\mathbf{R}_{\rm opt'}\|^2 \ge (c_4\epsilon/(2m))\|\mathbf{A}-\mathbf{R}_j\|^2$, so we recover such ω on the next iteration of greedy pursuit.

Since the coefficient estimation error is small enough, it follows that $\mathbf{R}_{j+1}$ is a representation with sufficiently small error that is "constructed" sufficiently quickly. Several problems remain. First, we don't know what j is, in general, so we would have to use $\mathbf{R}_T$. Second, we are required to return just an m-term representation whereas $\mathbf{R}_T$ may have, say, $m\log(N)$ terms. (In some applications, of course, the number of terms is not critical, only the goodness of approximation and construction time. In that case, $\mathbf{R}^T$ is an acceptable output.) Although our estimates of coefficients are good enough to be ignored from the perspective of overall error, tiny errors in coefficient estimation may cause us to choose the wrong set of frequencies as the top m terms in $\mathbf{R}^T$. We now show that, while this can happen, the resulting error is acceptable.

Suppose $\omega \in \Lambda_{\rm opt'}$ with ideal coefficient c_ω is displaced by $\omega_* \in \Lambda_T^m$, and suppose $\widetilde{c}_{\omega_*}$ is the coefficient of ω_* in our output. Then the error attributed to this exchange is $E_x = |\widetilde{c}_{\omega_*} - c_{\omega_*}|^2 + |c_\omega|^2 - |c_{\omega_*}|^2$. By goodness of approximation, $|\widetilde{c}_{\omega_*} - c_{\omega_*}|^2 \le (c_2\epsilon^2/m)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$. Next, because $\omega \in \Lambda_{\rm opt'}$, it follows that $|\widetilde{c}_{\omega_*} - c_{\omega_*}|^2 \le (c_2\epsilon/c_4)|c_\omega|^2$. Thus it follows, if constants are chosen properly, that $E_x \le (1+\epsilon)|c_\omega|^2 - |c_{\omega_*}|^2$, and we need to show that E_x is small compared with the unavoidable error, $|c_{\omega_*}|^2$. Because we chose ω_* over ω, we have $|\widetilde{c}_{\omega_*}|^2 \ge |\widetilde{c}_\omega|^2$, which is at least $(|c_\omega| - |c_\omega - \widetilde{c}_\omega|)^2$ by the triangle inequality. By goodness of approximation and because $\omega \in \Lambda_{\rm opt'}$, this is at least $(|c_\omega| - \sqrt{c_2\epsilon/c_4}|c_\omega|)^2 = (1-c_5\sqrt{\epsilon})|c_\omega|^2$. It follows that $E_x \le c_6\sqrt{\epsilon}|c_{\omega_*}|^2$. By replacing ϵ by ϵ^2/c_6^2, we conclude that each displacement of an optimal frequency by a suboptimal one increases its contribution to the error by the factor $(1+\epsilon)$, so the combined contributions blow up the error by at most the same factor. We conclude that $\|\mathbf{A}-\mathbf{R}_T^m\|^2 \le (1+\epsilon)\|\mathbf{A}-\mathbf{R}_{\rm opt}\|^2$; *i.e.*, the algorithm is correct. We claim without proof that the constants $c_0, c_1, \ldots$ can be set consistently. The main result follows.

5. HIGHER DIMENSIONAL VERSIONS

In this section, we sketch how to generalize our algorithm to more than one dimension. Suppose that the time and frequency domain is $N_1 \times N_2 \times \cdots \times N_d$, where $N = \prod_j N_j$. In higher dimensions, the basis functions are of the form $\psi_{\omega_1,\omega_2,\ldots,\omega_d}$, defined by

$$\psi_{\omega_1,\omega_2,\ldots,\omega_d}(t_1,t_2,\ldots,t_d) = \psi_{\omega_1}(t_1)\psi_{\omega_2}(t_2)\cdots\psi_{\omega_d}(t_d) = \frac{1}{\sqrt{N}} e^{2\pi i(\omega_1 t_1/N_1 + \omega_2 t_2 + \cdots + \omega_d t_d)/N_1}.$$

Observe that if the original $N_1, N_2, \ldots$ are relatively prime, then the $N_1 \times N_2 \times \cdots \times N_d$ problem in d dimensions is equivalent to the one-dimensional problem of size $N_1 \cdot N_2 \cdots N_d$. The equivalence is efficient to realize, using the Chinese remainder theorem. More generally, by the fundamental theorem of finitely-generated abelian groups, we may assume that

$N_1 \big| N_2 \big| \cdots \big| N_{d'}$, where $d' \leq d$, and where each N_i is greater than 1. By extending the data periodically, we may assume that all the N_i's are equal, since this "only" blows up the size of each dimension by the factor $N_{d'} \leq N$, for a factor of N^d overall, and $\mathrm{poly}(d, \log(N^d))$ is comparable to $\mathrm{poly}(d, \log(N))$. Henceforth, we'll assume that the problem shape is $\overbrace{N_1 \times N_1 \times \cdots \times N_1}^{d}$.

There are two main approaches, depending on whether we use $\mathbf{H}_{K^{1/d}} \otimes \cdots \otimes \mathbf{H}_{K^{1/d}}$ or $\mathbf{H}_K \otimes \mathbf{H}_1 \otimes \mathbf{H}_1 \otimes \cdots \otimes \mathbf{H}_1$ as the d-dimensional analog of the Dirichlet kernel $\mathbf{H}_K$ in one dimension. Both have issues.

If we use $\mathbf{H}_{K^{1/d}} \otimes \cdots \otimes \mathbf{H}_{K^{1/d}}$, then the algorithm above for one-dimensional signals can be generalized to higher dimensions in a straightforward way, but it will not be efficient. First, we pay a factor $O(1)$ in one dimension because the "pass" region $|\omega| \leq \frac{N}{2K}$ of $\mathbf{H}_k$ actually attenuates energy by as much as $(2/\pi)^2$. In higher dimensions, the $O(1)$-factor penalty becomes $2^{O(d)}$ (*i.e.*, something like $(\pi/2)^{2d}$), depending on engineering choices. This may be acceptable for $d \leq 3$. Other computational bottlenecks, however, are much worse than $2^{O(d)}$. For example, a straightforward generalization of an unequally-spaced fast Fourier transform algorithm to d dimensions may cost $(\log(M)\log(N))^d$, which is typically *not* acceptable.

Instead, if we use $\mathbf{H}_K \otimes \mathbf{H}_1 \otimes \mathbf{H}_1 \otimes \cdots \otimes \mathbf{H}_1$, then it is necessary to permute the spectrum pairwise randomly (or "close" to that, in some sense), and the straightforward techniques fail. For any N_1, we can map each *one* spectral position uniformly. But suppose N_1 is a power of 2 and $d = 2$, and (ω_1, ω_2) and (θ_1, θ_2) are two freqencies with difference $(\omega_1 - \theta_1, \omega_2 - \theta_2) = (0, N/2)$. Consider a mapping of the form

$$\mathcal{P} : \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \mapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \tag{2}$$

where the matrix is invertible; this is a spectral permutation. Any such mapping will map (θ_1, θ_2) to $(\omega_1, \omega_2) + (g, h)$, where (g, h) is in the set $\{(0, N_1/2), (N_1/2, 0), (N_1/2, N_1/2)\}$. That is, conditioned on $\mathcal{P}(\omega_1, \omega_2)$, there is just a small orbit of three possibilities for $\mathcal{P}(\theta_1, \theta_2)$. The kernel $\mathbf{H}_K \otimes \mathbf{H}_1 \otimes \mathbf{H}_1 \otimes \cdots \otimes \mathbf{H}_1$ will pass at least one of these. So, conditioned on $\mathbf{H}_K \otimes \mathbf{H}_1 \otimes \mathbf{H}_1 \otimes \cdots \otimes \mathbf{H}_1$ passing $\mathcal{P}(\omega_1, \omega_2)$, there is a $1/3$ chance that $\mathcal{P}(\theta_1, \theta_2)$ will also pass. Recall that, unfortunately, we wanted just a $1/m$ chance (approximately), so the expected contribution of (θ_1, θ_2) would be attenuated by the factor m.

We now discuss partial results.

General Result in Low Dimensions. Note that an algorithm with time cost polynomial in d, m, $\log(1/\delta)$, $\log(N)$, $\log M$, and $1/\epsilon$—polynomial in m instead of linear in m—was presented[1] for the case of N_1 equal to a power of 2. That algorithm can be modified in a straightforward way using tools of this paper to handle all N_1's and to have dependence just quadratic on m, using an ordinary matrix-vector multiplication algorithm instead of an unequally-spaced fast Fourier transform.

Large Square-Free Divisor. For any (ω_1, ω_2) and (θ_1, θ_2) subject to the map (2), conditioned on the kernel $\mathbf{H}_{K^{1/d}} \otimes \cdots \otimes \mathbf{H}_{K^{1/d}}$ passing $\mathcal{P}(\omega_1, \omega_2)$, that kernel will attenuate the energy of $\mathcal{P}(\theta_1, \theta_2)$ by approximately the factor $1/a$ or smaller, where a is the largest square-free divisor of N_1—that is, a is the product of all primes dividing N_1. If a is at least m, then this will suffice for our purposes. More generally, if $a < m$, we can get the algorithm to work at additional cost factor m/a. As a concrete example, if N_1 is itself a prime at least m or so, then this algorithm will work.

Power of 2. It remains open to provide an algorithm with cost $m\mathrm{poly}(d)$ in d dimensions if N_1 is a power of 2.

6. CONCLUSION

We provided a sampling algorithm that yields, with high probability, a m-term Fourier representation $\mathbf{R}$ for any input signal $\mathbf{A}$ of length N, with the guarantee that $\|\mathbf{A} - \mathbf{R}\|_2^2$ is within a factor $(1 + \epsilon)$ of the best possible m-term Fourier representation. The algorithm samples

$$m \cdot \mathrm{poly}(\log N, \log \|\mathbf{A}\|, 1/\epsilon)$$

positions non-adaptively and spends time and space linear in this quantity. Preliminary implementations of our algorithm indicate that, for exact m-term superpositions for small m, our algorithm is more efficient than an optimized, publicly-available FFT package for N approximately 1 million.

The overall structure of this algorithm follows previous work,[1–3] but we have to apply two key ideas: bulk estimation of multipoint polynomial evaluation using an unequally-spaced Fourier tranform, and use of arithmetic-progression independent random variables to enable the iterative algorithm. As a result we improve the $m^{\geq 4}$ factor in previous results to being linear in m.

Acknowledgments

For helpful discussions, we thank: Ingrid Daubechies, Björn Enquist, and Jing Zou.

REFERENCES

1. Y. Mansour, "Randomized interpolation and approximation of sparse polynomials," *SIAM J. Computing* **24**(2), 1995.
2. A. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. J. Strauss, "Near optimal sparse fourier representation via sampling," in *Proc. ACM STOC*, 2002.
3. A. Akavia, S. Goldwasser, and S. Safra, "Proving hard-core predicates by list decoding," in *Proc. IEEE FOCS*, pp. 146–157, 2003.
4. E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information." Submitted, 2004.
5. D. Gottlieb and S. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, no. 26 in CMBS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1977.
6. G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice Hall, 1994.
7. J. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, 1996.
8. "FFTW." `http://www.fftw.org`.
9. J. Zou, 2003. Personal communication.
10. J. Zou, A. Gilbert, M. Strauss, and I. Daubechies, "Theoretical and experimental analysis of a randomized algorithm for sparse fourier transform analysis." Submitted, 2004.
11. A. Nieslony and G. Steidl, "Approximate factorizations of Fourier matrices with nonequispaced knots," *Linear Algebra and its Applications* **366**, pp. 337–51, June 2003.
12. H. J. Weaver, *Applications of Discrete and Continuous Fourier Analysis*, Wiley, 1983.
13. A. Aho, J. Hopcroft, and J. Ullman, *Design and analysis of algorithms*, Addison-Wesley, 1972.
14. A. Dutt and V. Rokhlin, "Fast Fourier transforms for nonequispaced data," *SIAM J. Sci. Comput.* **14**, pp. 1368–1393, 1993.
15. C. Anderson and M. D. Dahleh, "Rapid compuation of the discrete Fourier transform," *SIAM J. Sci. Comput.* **17**, pp. 913–919, 1996.