Sample-Optimal Fourier Sampling in Any Constant Dimension

Piotr Indyk
MIT
indyk@mit.edu

Michael Kapralov
MIT
kapralov@mit.edu

***Abstract*—We give an algorithm for ℓ_2/ℓ_2 sparse recovery from Fourier measurements using $O(k \log N)$ samples, matching the lower bound of Do Ba-Indyk-Price-Woodruff'10 for non-adaptive algorithms up to constant factors for any $k \le N^{1-\delta}$. The algorithm runs in $\tilde{O}(N)$ time. Our algorithm extends to higher dimensions, leading to sample complexity of $O_d(k \log N)$, which is optimal up to constant factors for any $d = O(1)$. These are the first sample optimal algorithms for these problems.**

A preliminary experimental evaluation indicates that our algorithm has empirical sampling complexity comparable to that of other recovery methods known in the literature, while providing strong provable guarantees on the recovery quality.

***Keywords*-sparse Fourier Transform, sample complexity, compressed sensing, sparse recovery**

I. INTRODUCTION

The Discrete Fourier Transform (DFT) is a mathematical notion that allows to represent a sampled signal or function as a combination of discrete frequencies. It is a powerful tool used in many areas of science and engineering. Its popularity stems from the fact that signals are typically easier to process and interpret when represented in the frequency domain. As a result, DFT plays a key role in digital signal processing, image processing, communications, partial differential equation solvers, etc. Many of these applications rely on the fact that most of the Fourier coefficients of the signals are small or equal to zero, i.e., the signals are (approximately) *sparse*. For example, sparsity provides the rationale underlying compression schemes for audio, image and video signals, since keeping the top few coefficients often suffices to preserve most of the signal energy.

An attractive property of sparse signals is that they can be acquired from only a small number of samples. Reducing the sample complexity is highly desirable as it implies a reduction in signal acquisition time, measurement overhead and/or communication cost. For example, one of the main goals in medical imaging is to reduce the sample complexity in order to reduce the time the patient spends in the MRI machine [30], or the radiation dose received [35]. Similarly in spectrum sensing, a lower average sampling rate enables the fabrication of efficient analog to digital converters (ADCs) that can acquire very wideband multi-GHz signals [39]. As a result, designing sampling schemes and the associated sparse recovery algorithms has been a subject of extensive research in multiple areas, such as:

- *Compressive sensing:* The area of compressive sensing [12], [8], developed over the last decade, studies the task of recovering (approximately) sparse signals from linear measurements. Although several classes of linear measurements were studied, acquisition of sparse signals using few Fourier measurements (or, equivalently, acquisition of Fourier-sparse signals using few signal samples) has been one of the key problems studied in this area. In particular, the seminal work of [8], [34] has shown that one can recover N-dimensional signals with at most k Fourier coefficients using only $k \log^{O(1)} N$ samples. The recovery algorithms are based on linear programming and run in time polynomial in N. See [13] for an introduction to the area.
- *Sparse Fourier Transform:* A different line of research, with origins in computational complexity and learning theory, has been focused on developing algorithms whose sample complexity *and* running time bounds scale with the sparsity. Many such algorithms have been proposed in the literature, including [18], [28], [31], [15], [2], [16], [26], [1], [21], [20], [29], [6], [19], [32], [22], [25]. These works show that, for a wide range of signals, both the time complexity and the number of signal samples taken can be significantly sub-linear in N.

The best known results obtained in both of those areas are summarized in the following table. For the sake of uniformity we focus on algorithms that work for general signals and recover k-sparse approximations satisfying the so-called ℓ_2/ℓ_2 approximation guarantee[1]. In this case, the goal of an algorithm is as follows: given m samples of the Fourier transform $\widehat{x}$ of a signal x[2], and the sparsity parameter k, output x' satisfying

$$\|x - x'\|_2 \le C \min_{k\text{-sparse } y} \|x - y\|_2, \tag{1}$$

The algorithms are randomized and succeed with constant

[1]Some of the algorithms [8], [34], [10] can in fact be made deterministic, but at the cost of satisfying a somewhat weaker ℓ_2/ℓ_1 guarantee. Also, additional results that hold for exactly sparse signals are known, see e.g., [6] and references therein.

[2]Here and for the rest of this paper, we will consider the *inverse* discrete Fourier transform problem of estimating a sparse x from samples of $\widehat{x}$. This leads to a simpler notation. Note that the the forward and inverse DFTs are equivalent modulo conjugation.

Reference	Time	Samples	Approximation	Signal model
[8], [34]				
[10]	$N \times m$ linear program	$O(k \log^3(k) \log(N))$	$C = O(1)$	worst case
[7]	$N \times m$ linear program	$O(k \log N)$	$C = (\log N)^{O(1)}$	worst case
[20]	$O(k \log(N) \log(N/k))$	$O(k \log(N) \log(N/k))$	any $C > 1$	worst case
[14]	$O(k \log^2 N)$	$O(k \log N)$	$C = O(1)$	average case, $k = \Theta(\sqrt{N})$
[33]	$O(N \log N)$	$O(k \log N)$	$C = O(1)$	average case, $k = O(N^\alpha)$, $\alpha < 1$
[25]	$O(k \log^2(N) \log^{O(1)} \log N)$	$O(k \log(N) \log^{O(1)} \log N)$	any $C > 1$	worst case
[11]		$\Omega(k \log(N/k))$	constant C	lower bound

Figure 1. Bounds for the algorithms that recover k-sparse Fourier approximations . All algorithms produce an output satisfying Equation 1 with probability of success that is at least constant.

probability.

As evident from the table, none of the results obtained so far was able to guarantee sparse recovery from the optimal number of samples, unless either the approximation factor was super-constant or the result held for average-case signals. In fact, it was not even known whether there is an *exponential time* algorithm that uses only $O(k \log N)$ samples in the worst case.

A second limitation, that applied to the sub-linear time algorithms in the last three rows in the table, but not to compressive sensing algorithms in the first two rows of the table, is that those algorithms were designed for *one-dimensional* signals. However, the sparsest signals often occur in applications involving higher-dimensional DFTs, since they involve much larger signal lengths N. Although one can reduce, e.g., the two-dimensional DFT over $p \times q$ grid to the one-dimensional DFT over a signal of length pq [16], [27]), the reduction applies only if p and q are relatively prime. This excludes the most typical case of $m \times m$ grids where m is a power of 2. The only prior algorithm that applies to general $m \times m$ grids, due to [16], has $O(k \log^c N)$ sample and time complexity for a rather large value of c. If N is a power of 2, a two-dimensional adaptation of the [20] algorithm (outlined in [14]) has roughly $O(k \log^3 N)$ time and sample complexity, and an adaptation of [25] has $O(k \log^2 N (\log \log N)^{O(1)})$ sample complexity.

Our results: In this paper we give an algorithm that overcomes both of the aforementioned limitations. Specifically, we present an algorithm for the sparse Fourier transform in any fixed dimension that uses only $O(k \log N)$ samples of the signal. This is the first algorithm that matches the lower bound of [11], for k up to $N^{1-\delta}$ for any constant $\delta > 0$. The recovery algorithm runs in time $O(N \log^{O(1)} N)$.

In addition, we note that the algorithm is in fact quite simple. It is essentially a variant of an iterative thresholding scheme, where the coordinates of the signal are updated sequentially in order to minimize the difference between the current approximation and the underlying signal. In Section VII we discuss a preliminary experimental evaluation of this algorithm, which shows promising results.

The techniques introduced in this paper have already found applications. In particular, in a followup paper [23], we give an algorithm that uses $O(k \log(N) \log^{O(1)} \log N)$ samples of the signal and has the running time of $O(k \log^{O(1)}(N) \log^{O(1)} \log N)$ for any constant dimension d. This generalizes the result of [25] to any constant dimension, at the expense of somewhat larger runtime.

Our techniques: The overall outline of our algorithms follows the framework of [16], [20], [25], which adapt the methods of [9], [17] from arbitrary linear measurements to Fourier ones. The idea is to take, multiple times, a set of $B = O(k)$ linear measurements of the form

$$\tilde{u}_j = \sum_{i:h(i)=j} s_i x_i$$

for random hash functions $h : [N] \to [B]$ and random sign changes s_i with $|s_i| = 1$. This denotes *hashing* to B *buckets*. With such ideal linear measurements, $O(\log(N/k))$ hashes suffice for sparse recovery, giving an $O(k \log(N/k))$ sample complexity.

The sparse Fourier transform algorithms approximate $\tilde{u}$ using linear combinations of Fourier samples. Specifically, the coefficients of x are first pseudo-randomly permuted, by re-arranging the access to $\hat{x}$ via a random affine permutation. Then the coefficients are partitioned into buckets. This steps uses the"filtering" process that approximately partitions the range of x into intervals (or, in higher dimension, squares) with N/B coefficients each, and collapses each interval into one bucket. To minimize the number of samples taken, the filtering process is approximate. In particular the coefficients contribute ("leak"') to buckets other than the one they are nominally mapped into, although that contribution is limited and controlled by the quality of the filter. The details are described in Section III, see also [21] for further overview.

Overall, this probabilistic process ensures that most of the large coefficients are "isolated", i.e., are hashed to unique buckets, as well as that the contributions from the "tail" of the signal x to those buckets is not much greater than the average; the tail of the signal is defined as $\mathrm{Err}_k(x) = \min_{k-\text{sparse } y} ||x-y||_2$. This enables the algorithm to identify the positions of the large coefficients, as well as estimate their values, producing a sparse estimate χ of x. To improve this estimate, we repeat the process on $x - \chi$ by subtracting the influence of χ during hashing. The repetition will yield a good sparse approximation χ of x.

To achieve the optimal number of measurements, however, our algorithm departs from the above scheme in a crucial way: the algorithm does *not* use fresh hash functions in every repetition. Instead, $O(\log N)$ hash functions are chosen at the beginning of the process, such that each large coefficient is isolated by most of those functions with high probability. The same hash functions are then used throughout the duration of the algorithm. Note that each hash function requires a separate set of samples to construct the buckets, so reusing the hash functions means that the number of samples does not grow with the number of iterations. This enables us to achieve the optimal measurement bound.

At the same time reusing the hash functions creates a major difficulty: if the algorithm identifies a non-existing large coefficient by mistake and adds it to χ, this coefficient will be present in the difference vector $x - \chi$ and will need to be corrected later. And unlike the earlier guarantee for the large coefficients of the *original* signal x, we do not have any guarantees that large *erroneous* coefficients will be isolated by the hash functions, since the positions of those coefficients are determined by those functions. Because of these difficulties, almost all prior works[3] either used a fresh set of measurements in each iteration (almost all sparse Fourier transform algorithms fall into this category) or provided stronger deterministic guarantees for the sampling pattern (such as the restricted isometry property [8]). However, the latter option required a larger number of measurements to ensure the desired properties. Our algorithm circumvents this difficulty by ensuring that no large coefficients are created erroneously. This is non-trivial, since the hashing process is quite noisy (e.g, the bucketing process suffers from leakage). Our solution is to recover the large coefficients in the decreasing order of their magnitude. Specifically, in each step, we recover coefficients with magnitude that exceeds a specific threshold (that decreases exponentially). The process is designed to ensure that (i) all coefficients above the threshold are recovered and (ii) all recovered coefficients have magnitudes close to the threshold. In this way the set of locations of large coefficients stays fixed (or monotonically decreases) over the duration of the algorithms, and we can ensure the isolation properties of those coefficients during the initial choice of the hash functions.

Overall, our algorithm has two key properties (i) it is iterative, and therefore the values of the coefficients estimated in one stage can be corrected in the second stage and (ii) does not require fresh hash functions (and therefore new measurements) in each iteration. Property (ii) implies that the number of measurements is determined by only a single (first) stage, and does not increase beyond that. Property (i) implies that the bucketing and estimation process can be achieved using rather "crude" filters[4], since the estimated values can be corrected in the future stages. As a result each of the hash function require only $O(k)$ samples; since we use $O(\log N)$ hash functions, the $O(k \log N)$ bound follows. This stands in contrast with the algorithm of [16] (which used crude filters of similar complexity but required new measurements per each iteration) or [20] (which used much stronger filters with $O(k \log N)$ sample complexity) or [25] (which used filters of varying quality and sample complexity). The advantage of our approach is amplified in higher dimension, as the ratio of the number of samples required by the filter to the value k grows exponentially in the dimension. Thus, our filters still require $O(k)$ samples in any fixed dimension d, while for [20], [25] this bound increases to $O(k \log^d N)$.

Organization: We give definitions and basic results relevant to sparse recovery from Fourier measurements in section II. Filters that our algorithm uses are constructed in section III. Section IV states the algorithm and provides intuition behind the analysis. The main lemmas of the analysis are proved in section V, and full analysis of the algorithm is provided in section VI. Results of an experimental evaluation are presented in section VII.

II. Preliminaries

For a positive even integer a we will use the notation $[a] = \{-\frac{a}{2}, -\frac{a}{2}+1, \ldots, -1, 0, 1, \ldots, \frac{a}{2}-1\}$. We will consider signals of length $N = n^d$, where n is a power of 2 and $d \geq 1$ is the dimension. We use the notation $\omega = e^{2\pi i/n}$ for the root of unity of order n. The d-dimensional forward and inverse Fourier transforms are given by

$$\hat{x}_j = \frac{1}{\sqrt{N}} \sum_{i \in [n]^d} \omega^{-i^T j} x_i \quad \text{and} \quad x_j = \frac{1}{\sqrt{N}} \sum_{i \in [n]^d} \omega^{i^T j} \hat{x}_i \tag{2}$$

respectively, where $j \in [n]^d$. We will denote the forward Fourier transform by $\mathcal{F}$. Note that we use the orthonormal version of the Fourier transform. Thus, we have $||\hat{x}||_2 = ||x||_2$ for all $x \in \mathbb{C}^N$ (Parseval's identity). We recover a

[3]We are only aware of two exceptions: the algorithms of [14], [32] (which were analyzed only for the easier case where the large coefficients themselves were randomly distributed) and the analysis of iterative thresholding schemes due to [3] (which relied on the fact that the measurements were induced by Gaussian or Gaussian-like matrices).

[4]In fact, our filters are only slightly more accurate than the filters introduced in [16], and require the same number of samples.

signal z such that

$$||x-z||_2 \le (1+\epsilon) \min_{k-\text{ sparse } y} ||x-y||_2$$

from samples of $\widehat{x}$.

We will use pseudorandom spectrum permutations, which we now define. We write $\mathcal{M}_{d\times d}$ for the set of $d\times d$ matrices over $\mathbb{Z}_n$ with odd determinant. For $\Sigma \in \mathcal{M}_{d\times d}, q \in [n]^d$ and $i \in [n]^d$ let $\pi_{\Sigma,q}(i) = \Sigma(i-q) \mod n$. Since $\Sigma \in \mathcal{M}_{d\times d}$, this is a permutation. Our algorithm will use π to hash heavy hitters into B buckets, where we will choose $B \approx k/\epsilon$. It should be noted that unlike many sublinear time algorithms for the problem, our algorithm does not use $O(k)$ buckets with centers equispaced in the time domain. Instead, we think of each point in time domain as having a bucket around it. This improves the dependence of the number of samples on the dimension d. We will often omit the subscript Σ, q and simply write $\pi(i)$ when Σ, q is fixed or clear from context. For $i,j \in [n]^d$ we let $o_i(j) = \pi(j)-\pi(i)$ to be the "offset" of $j \in [n]^d$ relative to $i \in [n]^d$. We will always have $B = b^d$, where b is a power of 2.

Definition II.1. *Suppose that Σ^{-1} exists $\mod n$. For $a, q \in [n]^d$ we define the permutation $P_{\Sigma,a,q}$ by $(P_{\Sigma,a,q}\hat{x})_i = \hat{x}_{\Sigma^T(i-a)}\omega^{i^T\Sigma q}$.*

Lemma II.2. $\mathcal{F}^{-1}(P_{\Sigma,a,q}\hat{x})_{\pi_{\Sigma,q}(i)} = x_i \omega^{a^T\Sigma i}$

The proof is similar to the proof of Claim B.3 in [14] and is deferred to the full version of the paper [24] for completeness. Define

$$\text{Err}_k(x) = \min_{k-\text{sparse } y} ||x-y||_2 \text{ and } \mu^2 = \text{Err}_k^2(x)/k. \quad (3)$$

In this paper, we assume knowledge of μ (a constant factor upper bound on μ suffices). We also assume that the signal to noise ratio is bounded by a polynomial, namely that $R^* := ||x||_\infty/(\sqrt{\epsilon}\mu) \le n^C$ for a constant $C > 0$. We use the notation $\mathbb{B}_r^\infty(x)$ to denote the ℓ_∞ ball of radius r around x:

$$\mathbb{B}_r^\infty(x) = \{y \in [n]^d : ||x-y||_\infty \le r\},$$

where $||x-y||_\infty = \max_{1\le s\le d} ||x_s - y_s||_\circ$, and $||x_s-y_s||_\circ$ is the circular distance on $\mathbb{Z}_n$. We will also use the notation $f \lesssim g$ to denote $f = O(g)$. We sometimes write u.a.r. to denote 'uniformly at random'.

III. Filter construction and properties

For an integer $b > 0$ a power of 2 let

$$\hat{H}_i^1 = \begin{cases} \frac{\sqrt{n}}{b-1}, & \text{if } |i| < b/2 \\ 0 & \text{o.w.} \end{cases} \quad (4)$$

Let $\hat{H}^F$ denote the F-fold convolution of $\hat{H}^1$ with itself, so that supp $\hat{H}^F \subseteq [-F\cdot b, F\cdot b]$. Here and below F is a parameter that we will choose to satisfy $F \ge 2d, F = \Theta(d)$. The Fourier transform of $\hat{H}^1$ is the Dirichlet kernel (see e.g. [36], page 37):

$$H_j^1 = \frac{1}{b-1}\sum_{|i|<b/2}\omega^{ij} = \frac{\sin(\pi(b-1)j/n)}{(b-1)\sin(\pi j/n)} \text{ for } j \ne 0$$

$$H_0^1 = 1.$$

Thus, $H_j^F = \left(\frac{1}{b-1}\sum_{|i|<b/2}\omega^{ij}\right)^F = \left(\frac{\sin(\pi(b-1)j/n)}{(b-1)\sin(\pi j/n)}\right)^F$ for $j \ne 0$, and $H_0^F = 1$. For $i \in [n]^d$ let

$$G_i = \prod_{s=1}^d H_{i_s}^F, \quad (5)$$

so that $\hat{G}_i = \prod_{s=1}^d \hat{H}_{i_s}^F$ and $\text{supp}\, \hat{G} \subseteq [-F\cdot b, F\cdot b]^d$. We will use the following simple properties of G:

Lemma III.1. *For any $F \ge 1$ one has* **(1)** *$G_0 = 1$, and $G_j \in [\frac{1}{(2\pi)^{F\cdot d}}, 1]$ for all $j \in [n]^d$ such that $||j||_\infty \le \frac{n}{2b}$;* **(2)** *$|G_j| \le \left(\frac{2}{1+(b/n)||j||_\infty}\right)^F$ for all $j \in [n]^d$ as long as $b \ge 3$.*

The two properties imply that most of the mass of the filter is concentrated in a box of side $O(n/b)$, approximating the "ideal" filter (whose value would be equal to 1 for entries within the square and equal to 0 outside of it). The proof of the lemma is similar to the analysis of filters in [21], [25] and is deferred to the full version of the paper [24]. We will not use the lower bound on G given in the first claim of Lemma III.1 for our $\tilde{O}(N)$ time algorithm in this paper. We state the Lemma in full form for later use in [23], where we present a sublinear time algorithm.

The following property of pseudorandom permutations $\pi_{\Sigma,q}$ makes hashing using our filters effective (i.e. allows us to bound noise in each bucket, see Lemma III.3, see below):

Lemma III.2. *Let $i, j \in [n]^d$. Let Σ be uniformly random with odd determinant. Then for all $t \ge 0$* **Pr**$[||\Sigma(i-j)||_\infty \le t] \le 2(2t/n)^d$.

A somewhat incomplete proof of this lemma for the case $d = 2$ appeared as Lemma B.4 in [14]. We give a full proof for arbitrary d in the full version of the paper [24].

We access the signal x via random samples of $\hat{x}$, namely by computing the signal $\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{x})\cdot\hat{G})$. As Lemma III.3 below shows, this effectively "hashes" x into $B = b^d$ bins by convolving it with the filter G constructed above. Since our algorithm runs in $\tilde{O}(N)$ as opposed to $\tilde{O}(k)$ time, we can afford to work with bins around any location in time domain (we will be interested in locations of heavy hitters after applying the permutation, see Lemma III.3). This improves the dependence of our sample complexity on d. The properties of the filtering process are summarized in

Lemma III.3. *Let $x \in \mathbb{C}^N$. Choose $\Sigma \in \mathcal{M}_{d\times d}, a, q \in [n]^d$ uniformly at random, independent of x. Let $u = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{x})\cdot\hat{G})$, where G is the filter constructed* *in (5). Let $\pi = \pi_{\Sigma,q}$. For $i \in [n]^d$ let $\mu^2_{\Sigma,q}(i) = \sum_{j\in[n]^d\setminus\{i\}} |x_j G_{o_i(j)}|^2$, where $o_i(j) = \pi(j) - \pi(i)$ as before. Suppose that $F \geq 2d$. Then for any $i \in [n]^d$*

1) *$\mathbf{E}_{\Sigma,q}[\mu^2_{\Sigma,q}(i)] \leq C^d \|x\|_2^2/B$ for a constant $C > 0$.*
2) *for any Σ, q one has $\mathbf{E}_a[|\omega^{-a^T\Sigma i} u_{\pi(i)} - x_i|^2] \lesssim \mu^2_{\Sigma,q}(i) + \delta\|x\|_2^2$, where the last term corresponds to the numerical error incurred from computing FFT with $O(\log 1/\delta)$ bits of machine precision.*

The proof of Lemma III.3 is given in the full version of the paper [24].

Remark III.4. *We assume throughout the paper that arithmetic operations are performed on $C \log N$ bit numbers for a sufficiently large constant $C > 0$ such that $\delta\|x\|_2^2 \leq \delta(R^*)^2 n\mu^2 \leq \mu^2/N$, so that the effect of rounding errors on Lemma III.3 is negligible.*

IV. The algorithm

In this section we present our $\tilde{O}(N)$ time algorithm that achieves $d^{O(d)}\frac{1}{\epsilon}k\log N$ sample complexity and give the main definitions required for its analysis. Our algorithm follows the natural iterative recovery scheme. The main body of the algorithm (Algorithm 1) takes samples of the signal $\widehat{x}$ and repeatedly calls the LocateAndEstimate function (Algorithm 2), improving estimates of the values of dominant elements of x over $O(\log n)$ iterations. Crucially, samples of $\widehat{x}$ are only taken at the beginning of Algorithm 1 and passed to each invocation of LocateAndEstimate. Each invocation of LocateAndEstimate takes samples of $\widehat{x}$ as well as the current approximation χ to x as input, and outputs a constant factor approximation to dominant elements of $x - \chi$ (see section VI for analysis of LocateAndEstimate).

We first give intuition behind the algorithm and the analysis. We define the set $S \subseteq [n]^d$ to contain elements $i \in [n]^d$ such that $|x_i|^2 \geq \epsilon\mu^2$ (i.e. S is the set of *head elements* of x). As we show later (see section VI) it is sufficient to locate and estimate all elements in S up to $O(\epsilon\mu^2)$ error term in order obtain ℓ_2/ℓ_2 guarantees that we need[5]. Algorithm 1 performs $O(\log N)$ rounds of location and estimation, where in each round the located elements are estimated up to a constant factor. The crucial fact that allows us to obtain an optimal sampling bound is that the algorithm uses the same samples during these $O(\log N)$ rounds. Thus, our main goal is to show that elements of S will be successfully located and estimated throughout the process, *despite the dependencies* between the sampling pattern and the residual signal $x - \chi^{(t)}$ that arise due to reuse of randomness in the main loop of Algorithm 1.

We now give an overview of the main ideas that allow us to circumvent lack of independence. Recall that our

[5]In fact, one can see that our algorithm gives the stronger ℓ_∞/ℓ_2 guarantee

Algorithm 1 Overall algorithm: perform Sparse Fourier Transform

```
1: procedure SparseFFT($\hat{x}, k, \epsilon, R^*, \mu$)  ▷ $R^*$ is a bound on $||x||_\infty/(\sqrt{\epsilon}\mu)$
2:     $\chi^{(0)} \leftarrow 0$                                  ▷ in $\mathbb{C}^n$.
3:     $T \leftarrow \log_2 R^*$
4:     $B \leftarrow k/(\epsilon\alpha^d)$              ▷ $B = b^d$ for $b$ a power of 2
5:     $G, \widehat{G} \leftarrow$ filter as in (5)
6:     $r_{max} \leftarrow \Theta(\log N)$
7:     for $r = 0$ to $r_{max}$ do
8:         Choose $\Sigma_r \in \mathcal{M}_{d\times d}, a_r, q_r \in [n]^d$ u.a.r.
9:         For $r = 1, \ldots, r_{max}$,
10:            $u^r \leftarrow \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{x}) \cdot \hat{G})$
11:        ▷ Note that $u^r \in \mathbb{C}^{[n]^d}$ for all $r$
12:    end for
13:    for $t = 0, 1, \ldots, T-1$ do
14:        $\chi' \leftarrow$ LocateAndEstimate($\hat{x}, \chi^{(t)}$,
15:            $\{(\Sigma_r, a_r, b_r), u_r\}_{r=1}^{r_{max}}, r_{max}, \widehat{G}, 4\sqrt{\epsilon}\mu 2^{T-(t+1)}$)
16:        $\chi^{(t+1)} \leftarrow \chi^{(t)} + \chi'$
17:    end for
18:    return $\chi^{(T)}$
19: end procedure
```

Algorithm 2 LocateAndEstimate($\hat{x}, \chi, \{(\Sigma_r, a_r, q_r), u_r\}_{r=1}^{r_{max}}, r_{max}, \widehat{G}, \nu$)

```
1: procedure LocateAndEstimate($\hat{x}, \chi, \{(\Sigma_r, a_r, q_r), u_r\}_{r=1}^{r_{max}}, r_{max}, \widehat{G}, \nu$)
2:     Requires that $||x - \chi||_\infty \leq 2\nu$
3:     Guarantees that $||x - \chi - \chi'||_\infty \leq \nu$
4:     $L \leftarrow \emptyset$
5:     $w \leftarrow 0$
6:     for $r = 0$ to $r_{max}$ do
7:         $v^r \leftarrow u^r - \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{\chi}) \cdot \hat{G})$
8:         ▷ Update signal: note this does not use any new samples
9:     end for
10:    for $f \in [n]^d$ do
11:        $S \leftarrow \emptyset$
12:        for $r = 0$ to $r_{max}$ do
13:            Denote permutation $\pi_{\Sigma_r,q_r}$ by $\pi$
14:            $S \leftarrow S \cup \{v^r_{\pi(f)} \cdot \omega^{-a^T\Sigma f}\}$
15:        end for
16:        $\eta \leftarrow$ median($S$)
17:        If $|\eta| \leq \nu/2$ then continue
18:        $L \leftarrow L \cup \{f\}$
19:        $w_f \leftarrow \eta$
20:    end for
21:    return $w$
22: end procedure
```

algorithm needs to estimate all *head elements*, i.e. elements $i \in S$, up to $O(\epsilon\mu^2)$ additive error. Fix an element $i \in S$ and for each permutation π consider balls $\mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})$ around the position that i occupies in the permuted signal. For simplicity, we assume that $d = 1$, in which case the balls $\mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})$ are just intervals:

$$\mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2}) = \pi(i) + \left[-\frac{n}{b}\cdot 2^{t+2}, +\frac{n}{b}\cdot 2^{t+2}\right], \quad (6)$$

where addition is modulo n. Since our filtering scheme is essentially "hashing" elements of x into $B = \Omega(|S|/\alpha)$ "buckets" for a small constant $\alpha > 0$, we expect at most $O(\alpha)2^{t+2}$ elements of S to land in a ball (6) (i.e. the expected number of elements that land in this ball is proportional to its volume).

First suppose that this expected case occurs for any permutation, and assume that all head elements (elements of S) have the same magnitude (equal to 1 to simplify notation). It is now easy to see that the number of elements of S that are mapped to (6) *for any* $t \geq 0$ does not exceed its expectation (we call element i "isolated" with respect to π *at scale* t in that case), then the contribution of S to i's estimation error is $O(\alpha)$. Indeed, recall that the contribution of an element $j \in [n]$ to the estimation error of i is about $(1 + (b/n)|\pi(i) - \pi(j)|)^{-F}$ by Lemma III.1, (2), where we can choose F to be any constant without affecting the asymptotic sample complexity. Thus, even if $F = 2$, corresponding to the boxcar filter, the contribution to i's estimation error is bounded by

$$\begin{aligned}&\sum_{t\geq 0,(n/b)\cdot 2^{t+2}<n/2} \left|\pi(S)\cap \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})\right| \\ &\cdot \max_{y\in \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})\setminus \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+1})} |G_{\pi(i)-y}| \\ &= \sum_{t\geq 0,(n/b)\cdot 2^{t+2}<n/2} O(\alpha 2^{t+2})\cdot (1+2^{t+1})^{-F} = O(\alpha).\end{aligned}$$

Thus, if not too many elements of S land in intervals around $\pi(i)$, then the error in estimating i is at most $O(\alpha)$ times the maximum head element in the current residual signal (plus noise, which can be handled separately). This means that the median in line 15 of Algorithm 2 is an additive $\pm O(\alpha)||x-\chi||_\infty$ approximation to element f. Since Algorithm 2 only updates elements that pass the magnitude test in line 16, we can conclude that whenever we update an element, we have a $(1\pm O(\alpha))$ multiplicative estimate of its value, which is sufficient to conclude that we decrease the ℓ_∞ norm of $x-\chi$ in each iteration. Finally, we crucially ensure that the signal is never updated outside of the set S. This means that the set of head elements is fixed in advance and does not depend on the execution path of the algorithm! This allows us to formulate a notion of isolation with respect to the set S of head elements fixed in advance, and hence avoid issues arising from the lack of independence of the signal $x - \chi$ and the permutations we choose.

We formalize this notion in Definition V.2, where we define what it means for $i \in S$ to be isolated under π. Note that the definition is essentially the same as asking that the balls in (6) do not contain more than the expected number of elements of S. However, we need to relax the condition somewhat in order to argue that it is satisfied with good enough probability *simultaneously for all* $t \geq 0$. A adverse effect of this relaxation is that our bound on the number of elements of S that are mapped to a balls around $\pi(i)$ are weaker than what one would have in expectation. This, however, is easily countered by choosing a filter with stronger, but still polynomial, decay (i.e. setting the parameter F in the definition of our filter G in (5) sufficiently large).

As noted before, the definition of being isolated crucially only depends on the **locations** of heavy hitters as opposed to their **values**. This allows us to avoid an (intractable) union bound over all signals that appear during the execution of our algorithm. We give formal definitions of isolationin section V, and then use them to analyze the algorithm in section VI.

V. Isolated elements and main technical lemmas

We now give the technical details for the outline above.

Definition V.1. *For a permutation π and a set $S \subseteq [n]^d$ we denote $S^\pi := \{\pi(x) : x \in S\}$.*

Definition V.2. *Let $\Sigma \in \mathcal{M}_{d\times d}, q \in [n]^d$, $S \subseteq [n]^d$, and let $\pi = \pi_{\Sigma,q}$. We say that an element i is* isolated *under permutation* π at scale t *if*

$$|(S\setminus\{i\})^\pi \cap \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})| \leq \alpha^{d/2}2^{(t+3)d}\cdot 2^t.$$

We say that i is simply isolated *under permutation $\pi_{\Sigma,q}$ if it is isolated under $\pi_{\Sigma,q}$ at all scales $t \geq 0$.*

Remark V.3. *We will use the definition of isolated elements for a set S with $|S| \approx k/\epsilon$.*

The following lemma shows that every $i \in [n]^d$ is likely to be isolated under a randomly chosen permutation π:

Lemma V.4. *Let $S \subseteq [n]^d, |S| \leq 2k/\epsilon$. Let $B \geq k/(\epsilon\alpha^d)$. Let $\Sigma \in \mathcal{M}_{d\times d}, q \in [n]^d$ be chosen uniformly at random, and let $\pi = \pi_{\Sigma,q}$. Then each $i \in [n]^d$ is* isolated *under permutation π with probability at least $1 - O(\alpha^{d/2})$.*

Proof: By Lemma III.2, for fixed i, $j \neq i$, and any $r \geq 0$,

$$\mathbf{Pr}_\Sigma[\|\Sigma(i-j)\|_\infty \leq r] \leq 2(2r/n)^d. \quad (7)$$

Setting $r = (n/b)\cdot 2^{t+2}$, we get

$$\begin{aligned}&\mathbf{E}_{\Sigma,q}[|(S\setminus\{i\})^\pi \cap \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})|] \\ &= \sum_{j\in S\setminus\{i\}} \mathbf{Pr}_{\Sigma,q}[\pi(j) \in \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})] \quad (8)\end{aligned}$$

Since $\pi_{\Sigma,q}(i) = \Sigma(i-q)$ for all $i \in [n]^d$, we have

$$\begin{aligned}&\mathbf{Pr}_{\Sigma,q}[\pi(j) \in \mathbb{B}^{\infty}_{\pi(i)}((n/b)\cdot 2^{t+2})]\\ &= \mathbf{Pr}_{\Sigma,q}[||\pi(j)-\pi(i)||_\infty \le (n/b)\cdot 2^{t+2}]\\ &= \mathbf{Pr}_{\Sigma,q}[||\Sigma(j-i)||_\infty \le (n/b)\cdot 2^{t+2}] \le 2(2^{t+3}/b)^d,\end{aligned}$$

where we used (7) in the last step. Using this in (8), we get

$$\begin{aligned}&\mathbf{E}_{\Sigma,q}[|(S\setminus\{i\})^\pi \cap \mathbb{B}^{\infty}_{\pi(i)}((n/b)\cdot 2^{t+2})|]\\ &\le |S|\cdot(2^{t+3}/b)^d \le (|S|/B)\cdot 2^{(t+3)d} \lesssim \epsilon\alpha^d 2^{(t+3)d}.\end{aligned}$$

Now by Markov's inequality we have that i fails to be isolated at scale t with probability at most

$$\begin{aligned}&\mathbf{Pr}_{\Sigma,q}\left[|(S\setminus\{i\})^\pi \cap \mathbb{B}^{\infty}_{\pi(i)}(\frac{n}{b}\cdot 2^{t+2})| > \alpha^{d/2}2^{(t+3)d+t}\right]\\ &\lesssim 2^{-t}\alpha^{d/2}.\end{aligned}$$

Taking the union bound over all $t \ge 0$, we get $\mathbf{Pr}_{\Sigma,q}[i$ is not isolated$] \lesssim \sum_{t\ge 0} 2^{-t}\alpha^{d/2} \lesssim \alpha^{d/2}$ as required. ■

The contribution of tail noise to an element $i \in [n]^d$ is captured by the following

Definition V.5. *Let $x \in \mathbb{C}^N$. Let $S \subseteq [n]^d, |S| \le 2k/\epsilon$. Let $B \ge k/(\epsilon\alpha^d)$. Let $u = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{x}_{[n]^d\setminus S})\cdot \hat{G})$. We say that an element $i \in [n]^d$ is* well-hashed with respect to noise under $(\pi_{\Sigma,q}, a)$ *if $|u_{\pi(i)}\omega^{-a^T\Sigma i} - x_i|^2 = O(\sqrt{\alpha})\epsilon\mu^2$, where we let $\pi = \pi_{\Sigma,q}$ to simplify notation.*

Lemma V.6. *Let $S \subseteq [n]^d, |S| \le 2k/\epsilon$, be such that $||x_{[n]^d\setminus S}||_\infty \le \mu$. Let $B \ge k/(\epsilon\alpha^d)$. Let $\Sigma_r \in \mathcal{M}_{d\times d}, q_r, a_r \in [n]^d, r = 1,\ldots,r_{max}, r_{max} \ge (C/\sqrt{\alpha})\log N$ be chosen uniformly at random, where $\alpha > 0$ is a constant and $C > 0$ is a sufficiently large constant that depends on α. Then with probability at least $1 - N^{-\Omega(C)}$* **1)** *each $i \in [n]^d$ is isolated with respect to S under at least $(1 - O(\sqrt{\alpha}))r_{max}$ permutations $\pi_r, r = 1,\ldots,r_{max}$; and* **2)** *each $i \in [n]^d$ is well-hashed with respect to noise under at least $(1-O(\sqrt{\alpha}))r_{max}$ pairs $(\pi_r, a_r), r = 1,\ldots,r_{max}$.*

Proof: The first claim follows by an application of Chernoff bounds and Lemma V.4. For the second claim, let $u = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{x}_{[n]^d\setminus S})\cdot\hat{G})$, where $(\Sigma,a,q) = (\Sigma_r, a_r, q_r)$ for some $r = 1,\ldots,r_{max}$. Letting $\pi = \pi_{\Sigma,q}$, by Lemma III.3, (1) and (2) we have

$$\begin{aligned}&\mathbf{E}_{\Sigma,q,a}[|u_{\pi(i)}\omega^{-a^T\Sigma i} - (x_{[n]^d\setminus S})_i|^2]\\ &\le (C')^d||x_{[n]^d\setminus S}||^2/B + ||x||^2\cdot N^{-\Omega(c)}\end{aligned}$$

for a constant $C' > 0$, where we asssume that arithmetic operations are performed on $c\log N$-bit numbers for some constant $c > 0$. Since we assume that $R^* \le \mathrm{poly}(N)$, we have

$$\begin{aligned}&\mathbf{E}_{\Sigma,q,a}[|u_{\pi(i)}\omega^{-a^T\Sigma i} - (x_{[n]^d\setminus S})_i|^2]\\ &\le (C'')^d||x_{[n]^d\setminus S}||^2/B + \mu^2\cdot N^{-\Omega(c)}.\end{aligned}$$

Let $S^* \subset [n]^d$ denote a set of top k coefficients of x (with ties broken arbitrarily). We have

$$\begin{aligned}&||x_{[n]^d\setminus S}||^2 \le ||x_{[n]^d\setminus(S\cup S^*)}||^2 + ||x_{S^*\setminus S}||^2\\ &\le ||x_{[n]^d\setminus S}||^2 \le ||x_{[n]^d\setminus S^*}||^2 + k\cdot||x_{[n]^d\setminus S}||^2_\infty \le 2k\mu^2.\end{aligned}$$

Since $B \ge k/(\epsilon\alpha^d)$, we thus have

$$\mathbf{E}_{\Sigma,q,a}[|u_{\pi(i)}\omega^{-a^T\Sigma i} - (x_{[n]^d\setminus S})_i|^2] \le (C'''\alpha)^d\epsilon\mu^2$$

for a constant $C''' > 0$. By Markov's inequality

$$\mathbf{Pr}_{\Sigma,q,a}[|u_{h(i)}\omega^{-a^T\Sigma i} - (x_{[n]^d\setminus S})_i|^2 > (C'''\sqrt{\alpha})^d\epsilon\mu^2] < \alpha^{d/2}.$$

As before, an application of Chernoff bounds now shows that each $i \in [n]^d$ is well-hashed with respect to noise with probability at least $1 - N^{-10}$, and hence all $i \in [n]^d$ are well-hashed with respect to noise with probability at least $1 - N^{-\Omega(C)}$ as long as α is smaller than an absolute constant. ■

We now combine Lemma V.4 with Lemma V.6 to derive a bound on the noise in the "bucket" of an element $i \in [n]^d$ due to both heavy hitters and tail noise. Note that crucially, the bound only depends on the ℓ_∞ norm of the head elements (i.e. the set S), and in particular, works for *any signal* that coincides with x on the complement of S. Lemma V.7 will be the main tool in the analysis of our algorithm in the next section.

Lemma V.7. *Let $x \in \mathbb{C}^N$. Let $S \subseteq [n]^d, |S| \le 2k/\epsilon$. Let $B \ge k/(\epsilon\alpha^d)$. Let $y \in \mathbb{C}^N$ be such that $y_{[n]^d\setminus S} = x_{[n]\setminus S}$ and $||y_S||_\infty \le 4\sqrt{\epsilon}\mu 2^w$ for some $t \ge 0$. Let $u = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{y})\cdot\hat{G})$. Then for each $i \in [n]^d$ that is isolated and well-hashed with respect to noise under (Σ_r, q_r, a_r) one has $|u_j\omega^{-a^T\Sigma i} - y_i|^2 \lesssim \sqrt{\alpha}\epsilon((4\mu 2^w)^2 + \mu^2)$.*

Proof: We have $\left|u_j\omega^{-a^T\Sigma i} - x_i\right|^2 \le 2|A^H|^2 + 2|A^T|^2$, where $A^H = u^H_j\omega^{-a^T\Sigma i} - (y_S)_i$ for $u^H = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{y}_S)\cdot\hat{G})$ and $A^T = u^T_j\omega^{-a^T\Sigma i} - (y_{[n]^d\setminus S})_i$ for $u^T = \sqrt{N}\mathcal{F}^{-1}((P_{\Sigma,a,q}\hat{y}_{[n]^d\setminus S})\cdot\hat{G})$

We first bound A^H. Fix $i \in [n]^d$. If i is isolated, we have

$$|(S\setminus\{i\})^\pi \cap \mathbb{B}^\infty_{\pi(i)}((n/b)\cdot 2^{t+2})| \le \alpha^{d/2}2^{(t+3)d}\cdot 2^t.$$

for all $t \ge 0$. We have

$$\begin{aligned}|A^H| &= |u^H_j\omega^{-a^T\Sigma i} - y_i| = \left|\sum_{j\in S\setminus\{i\}} y_j G_{o_i(j)}\omega^{-a^T\Sigma j}\right|\\ &\le \sum_{j\in S\setminus\{i\}} |y_j G_{o_i(j)}|\\ &\le ||y_S||_\infty \cdot \sum_{t\ge 0}\left(\frac{2}{2^{t+2}}\right)^F |(S\setminus\{i\})^\pi \cap \mathbb{B}^\infty_{\pi(i)}\left(\frac{n}{b}\cdot 2^{t+2}\right)|\\ &\le ||y_S||_\infty \cdot \sum_{t\ge 0}\left(\frac{2}{2^{t+2}}\right)^F \alpha^{d/2}2^{(t+3)d}\cdot 2^t\end{aligned}$$

$$= ||y_S||_\infty \cdot O(\alpha^{d/2}) = O(\alpha^{d/2}\sqrt{\epsilon}\mu 2^w)$$

as long as $F \geq 2d, F = \Theta(d)$. Further, if i is well-hashed with respect to noise, we have $|A^T| \lesssim \sqrt{\alpha}\epsilon\mu^2$. Putting these estimates together yields the result. ■

VI. Main result

In this section we use Lemma V.7 to prove that our algorithm satisfies the stated ℓ_2/ℓ_2 sparse recovery guarantees. The proof consists of two main steps: Lemma VI.1 proves that one iteration of the peeling process (i.e. one call to LocateAndEstimate) outputs a list containing all elements whose values are close to the current ℓ_∞ norm of the residual signal. Furthermore, approximations that LocateAndEstimate returns for elements in its output list are correct up to a multiplicative $1 \pm 1/3$ factor. Lemma VI.2 then shows that repeated invocations of LocateAndEstimate reduce the ℓ_∞ norm of the residual signal as claimed.

Lemma VI.1. *Let $x \in \mathbb{C}^N$. Let $S \subseteq [n]^d, |S| \leq 2k/\epsilon$, be such that $||x_{[n]^d\setminus S}||_\infty \leq \mu$. Let $B \geq k/(\epsilon\alpha^d)$. Consider the t-th iteration of the main loop in Algorithm 1. Suppose that $||x-\chi||_\infty \leq 2\nu, \nu > \sqrt{\epsilon}\mu$. Suppose that each element $i \in [n]^d$ is isolated with respect to S and well-hashed with respect to noise under at least $(1-O(\sqrt{\alpha}))r_{max}$ values of $r = 1, \ldots, r_{max}$. Let $y = x - \chi^{(t)}$, and let χ' denote the output of* LocateAndEstimate. *Then one has*

1) $|\chi'_i - y_i| < \frac{1}{3}|y_i|$ *for all $i \in L$;*
2) *all i such that $|y_i| \geq \nu$ are included in L;*
3) *for all $i \in L$ one has $|y_i| \geq \nu/4$*

as long as $\alpha > 0$ is a sufficiently small constant.

Proof: We let $y := x - \chi^{(t)}$ to simplify notation. Fix $i \in [n]^d$. Consider r such that i is isolated under π_r and well-hashed with respect to noise under (π_r, a_r). Then we have by Lemma V.7

$$\begin{aligned} |v^r_j\omega^{-a^T\Sigma i} - y_i|^2 &\lesssim \sqrt{\alpha}(\epsilon(||y_{[S]}||_\infty)^2 + \epsilon\mu^2) \\ &\lesssim \sqrt{\alpha}\epsilon((2\nu)^2 + \mu^2) \leq (\frac{1}{16}\nu)^2 \end{aligned} \tag{9}$$

as long as α is smaller than an absolute constant.

Since each i is well-hashed with respect to at least at $1 - O(\sqrt{\alpha})$ fraction of permutations, we get that $|y_i - \eta| \leq \frac{1}{16}\nu$. Now if $i \in L$, it must be that $|\eta| > \nu/2$, but then

$$|y_i| \geq |\eta| - \nu/16 > \nu/2 - \nu/16 > (1/4)\nu, \tag{10}$$

implying the third claim. This also implies that $|\chi'_i - y_i| \leq \nu/16 < 4|y_i|/16 < |y_i|/3$, so the first claim follows.

For the second claim, it suffices to note that if $|y_i| > \nu$, then we must have $|\eta| \geq |y_i| - \nu/16 > \nu/2$, so i passes the magnitude test and is hence included in L. ■

We can now prove the main lemma required for analysis of Algorithm 1:

Lemma VI.2. *Let $x \in \mathbb{C}^N$. Let $\Sigma_r \in \mathcal{M}_{d\times d}, a_r, q_r \in [n]^d, r = 1, \ldots, r_{max} = C\log N$, where $C > 0$ is a sufficiently large constant, be chosen uniformly at random. Then with probability at least $1 - N^{-\Omega(C)}$ one has $||x - \chi^{(T)}||_\infty \leq 4\sqrt{\epsilon}\mu$.*

Proof: We now fix a specific choice of the set $S \subseteq [n]^d$. Let

$$S = \{i \in [n]^d : |x_i| > \sqrt{\epsilon}\mu\}. \tag{11}$$

First note that $||x_{[n]^d\setminus S}||_\infty \leq \sqrt{\epsilon}\mu$ (in particular, $||x_{[n]^d\setminus S}||_\infty \leq \mu$). Also, we have $|S| \leq 2k/\epsilon$. Indeed, recall that $\mu^2 = \mathrm{Err}^2_k(x)/k$. If $|S| > 2k/\epsilon$, more than k/ϵ elements of S belong to the tail, amounting to at least $\epsilon\mu^2 \cdot (k/\epsilon) > \mathrm{Err}^2_k(x)$ tail mass. Thus, since $r_{max} \geq C\log N$, and by the choice of B in Algorithm 1, we have by Lemma V.6 that with probability at least $1 - N^{-\Omega(C)}$

1) each $i \in [n]^d$ is isolated with respect to S under at least $(1 - O(\sqrt{\alpha}))r_{max}$ permutations $\pi_r, r = 1, \ldots, r_{max}$;
2) each $i \in [n]^d$ is well-hashed with respect to noise under at least $(1-O(\sqrt{\alpha}))r_{max}$ permutations $\pi_r, r = 1, \ldots, r_{max}$.

This ensures that the preconditions of Lemma VI.1 are satisfied. We now prove the following statement for $t \in [0 : T]$ by induction on t:

1) $\chi^{(t)}_{[n]\setminus S} \equiv 0$
2) $||(x - \chi^{(t)})_S||_\infty \leq 4\sqrt{\epsilon}\mu 2^{T-t}$.
3) $|x_i - \chi^{(t)}_i| \leq |x_i|$ for all $i \in [n]^d$.

Base:$t = 0$**.** True by the choice of T.

Inductive step: $t \to t+1$**.** Consider the list L constructed by LocateAndEstimate at iteration t. Let $S^* := \{i \in S : |(x-\chi^{(t)})_i| > 4\sqrt{\epsilon}\mu 2^{T-(t+1)}\}$. We have $S^* \subseteq L$ by Lemma VI.1, (2). Thus,

$$\begin{aligned} ||(x - \chi^{(t+1)})_S||_\infty \leq \max\{&||(x - \chi^{(t+1)})_{S^*}||_\infty, \\ &||(x - \chi^{(t+1)})_{S\setminus S^*}||_\infty, \\ &||(x - \chi^{(t+1)})_{[n]^d\setminus S}||_\infty\}. \end{aligned}$$

By Lemma VI.1, (1) we have $|x_i - \chi^{(t+1)}_i| \leq |x_i - \chi^{(t)}_i|/3$ for all $i \in L$, so (3) follows. Furthemore, this implies that

1) $||(x - \chi^{(t+1)})_{S^*}||_\infty \leq ||x - \chi^{(t)}||_\infty/3 \leq 4\sqrt{\epsilon}\mu 2^{T-(t+1)}$ by the inductive hypothesis;
2) $||(x-\chi^{(t+1)})_{S\setminus S^*}||_\infty \leq 4\sqrt{\epsilon}\mu 2^{T-(t+1)}$ by definition of S^*;
3) $||(x - \chi^{(t+1)})_{[n]^d\setminus S}||_\infty = ||x_{[n]^d\setminus S}||_\infty \leq \sqrt{\epsilon}\mu \leq 4\sqrt{\epsilon}\mu 2^{T-(t+1)}$ by the inductive hypothesis and the fact that $t \leq T - 1$.

This proves (2).

Finally, by Lemma VI.1, (3) only elements i such that $|(x-\chi^{(t)})_i| > \frac{1}{4}4\sqrt{\epsilon}\mu 2^{T-t} > \sqrt{\epsilon}\mu$ are included in L. Since $|(x - \chi^{(t)})_i| \leq |x_i|$, this means that $|x_i| > \sqrt{\epsilon}\mu$, i.e. $i \in S$ and $\chi^{(t+1)}_{[n]\setminus S} = 0$, as required. ■

We can now prove

Theorem VI.3. *Algorithm 1 returns a vector χ such that $||x-\chi||_2 \le (1+O(\epsilon))\,\mathrm{Err}_k(x)$. The number of samples is bounded by $d^{O(d)}\frac{1}{\epsilon}k\log N$, and the runtime is bounded by $O(N\log^3 N)$.*

Proof: By Lemma VI.2 we have that $||x-\chi||_\infty \le 4\sqrt{\epsilon}\mu$, so

$$\begin{aligned}||x-\chi||_2^2 &\le ||(x-\chi)_{[k]}||_\infty^2 \cdot k + ||(x-\chi)_{[n]^d\setminus[k]}||_2^2\\ &\le ||(x-\chi)_{[k]}||_\infty^2 \cdot k + ||x_{[n]^d\setminus[k]}||_2^2\\ &\le (1+O(\epsilon))\,\mathrm{Err}_k^2(x),\end{aligned}$$

where we used Lemma VI.2, (3) to upper bound $||(x-\chi)_{[n]^d\setminus[k]}||_2^2$ with $||x_{[n]^d\setminus[k]}||_2^2$.

We now bound sampling complexity. The support of the filter G is bounded by $O(BF^d) = d^{O(d)}\frac{1}{\epsilon}k$ by construction. We are using $r_{max} = \Theta(\log N)$, amounting to $d^{O(d)}\frac{1}{\epsilon}k\log N$ sampling complexity overall. The location and estimation loop takes $O(N\log^3 N)$ time: each time the vector v^r is calculated in LOCATEANDESTIMATE $O(N\log N)$ time is used by the FFT computation, so since we compute $O(\log N)$ vectors during each of $O(\log N)$ iterations, this results in an $O(N\log^3 N)$ contribution to runtime. ■

VII. EXPERIMENTAL EVALUATION

In this section we describe results of an experimental evaluation of our algorithm from section IV. In order to avoid the issue of numerical precision and make the notion of recovery probability well-defined, we focus the problem of *support recovery*, where the goal is to recover the *positions* of the non-zero coefficients. We first describe the experimental setup that we used to evaluate our algorithm, and follow with evaluation results.

Experimental setup: We present experiments for support recovery from one-dimensional Fourier measurements (i.e. $d=1$). In this problem one is given frequency domain access to a signal $\widehat{x}$ that is *exactly* k-sparse in the time domain, and needs to recovery the support of x exactly. The support of x was chosen to be uniform among subsets of $[N]$ of size k.

We compared our algorithm to ℓ_1-minimization, a state-of-the-art technique for practical sparse recovery using Gaussian and Fourier measurements. We used the implementation from SPGL1 [38], [37], a standard Matlab package for sparse recovery using ℓ_1-minimization. We also present a comparison to SSMP [5], which is a state-of-the art iterative algorithm for sparse recovery using sparse matrices. We compare our results to experiments in [4]. The details of our implementation of Algorithm 1 and the input distributions in the two experiments are described in the full version of the paper [24].

Results: A plot of recovery probability as a function of the number of (complex) measurements and sparsity for SPGL1 and Algorithm 1 is given in Fig. 2. The empirical sample complexity of our algorithm is within a factor of 2 of ℓ_1 minimization if success probability 0.9 is desired. The best known theoretical bounds for the general setting of approximate sparse recovery show that $O(k\log^3 k\log N)$ samples are sufficient. The runtime is bounded by the cost of solving an $N\times m$ linear program. Our algorithm provides comparable empirical performance, while providing optimal measurement bound and $\tilde{O}(N)$ runtime.

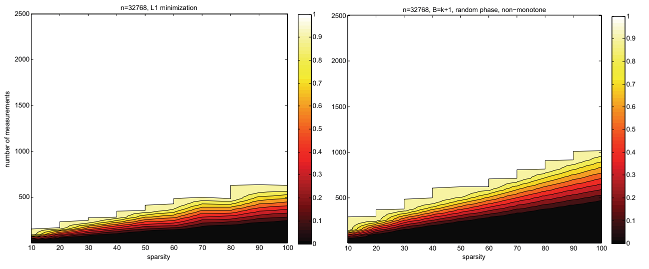

Figure 2. Success probability as a function of sparsity and number of (complex) measurements: SPGL1 (left panel) and Algorithm 1 (right panel).

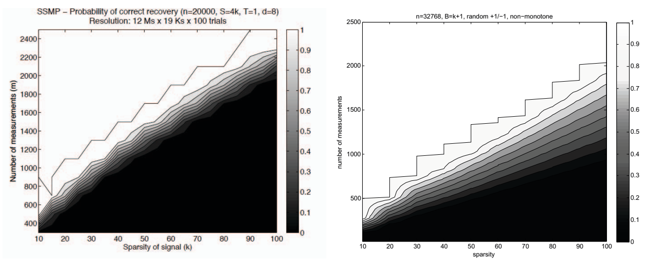

Figure 3. Success probability as a function of sparsity and number of (real) measurements: SSMP (left panel) and Algorithm 1 (right panel).

Since experiments in [4] used real measurements, we multiply the number of our (complex) measurements by 2 for this comparison (note that the right panel of Fig. 3 is the same as the right panel of Fig. 2, up to the factor of 2 in the number of measurements). We observe that our algorithm improves upon SSMP by a factor of about 1.15 when 0.9 success probability is desired.

ACKNOWLEDGMENT

We would like to thank the anonymous FOCS reviewers for comments that helped improve the presentation of the paper. This research was supported by NSF award CCF-1065125, MADALGO center, Packard Foundation and Simons Foundation. We also acknowledge financial support from grant #FA9550-12-1-0411 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).

REFERENCES

[1] A. Akavia. Deterministic sparse Fourier approximation via fooling arithmetic progressions. *COLT*, pages 381–393, 2010.

[2] A. Akavia, S. Goldwasser, and S. Safra. Proving hard-core predicates using list decoding. *FOCS*, 44:146–159, 2003.

[3] M. Bayati, M. Lelarge, and A. Montanari. Universality in polytope phase transitions and message passing algorithms. 2012.

[4] Radu Berinde. Advances in sparse signal recovery methods. *MIT*, 2009.

[5] Radu Berinde and Piotr Indyk. Sequential sparse matching pursuit. *Allerton'09*, pages 36–43, 2009.

[6] P. Boufounos, V. Cevher, A. C. Gilbert, Y. Li, and M. J. Strauss. What's the frequency, Kenneth?: Sublinear Fourier sampling off the grid. *RANDOM/APPROX*, 2012.

[7] E. Candes and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 2010.

[8] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies. *IEEE Trans. on Info.Theory*, 2006.

[9] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. *ICALP*, 2002.

[10] Mahdi Cheraghchi, Venkatesan Guruswami, and Ameya Velingker. Restricted isometry of Fourier matrices and list decodability of random linear codes. *SODA*, 2012.

[11] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower Bounds for Sparse Recovery. *SODA*, 2010.

[12] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[14] Badih Ghazi, Haitham Hassanieh, Piotr Indyk, Dina Katabi, Eric Price, and Lixin Shi. Sample-optimal average-case sparse Fourier transform in two dimensions. *arXiv preprint arXiv:1303.1209*, 2013.

[15] A. Gilbert, S. Guha, P. Indyk, M. Muthukrishnan, and M. Strauss. Near-optimal sparse Fourier representations via sampling. *STOC*, 2002.

[16] A. Gilbert, M. Muthukrishnan, and M. Strauss. Improved time bounds for near-optimal space Fourier representations. *SPIE Conference, Wavelets*, 2005.

[17] A. C. Gilbert, Y. Li, E. Porat, and M. J. Strauss. Approximate sparse recovery: optimizing time and measurements. In *STOC*, pages 475–484, 2010.

[18] O. Goldreich and L. Levin. A hard-corepredicate for alloneway functions. *STOC*, pages 25–32, 1989.

[19] H. Hassanieh, F. Adib, D. Katabi, and P. Indyk. Faster GPS Via the Sparse Fourier Transform. *MOBICOM*, 2012.

[20] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Near-optimal algorithm for sparse Fourier transform. *STOC*, 2012.

[21] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Simple and practical algorithm for sparse Fourier transform. *SODA*, 2012.

[22] Sabine Heider, Stefan Kunis, Daniel Potts, and Michael Veit. A sparse Prony FFT. *SAMPTA*, 2013.

[23] Piotr Indyk and Michael Kapralov. Sample-Optimal Fourier Sampling – II. *http://www.mit.edu/~kapralov/ft-hd-part1.pdf*, 2014.

[24] Piotr Indyk and Michael Kapralov. Sample-Optimal Fourier Sampling in Any Constant Dimension. *http://arxiv.org/abs/1403.5804*, 2014.

[25] Piotr Indyk, Michael Kapralov, and Eric Price. (Nearly) sample-optimal sparse Fourier transform. *SODA*, 2014.

[26] M. A. Iwen. Combinatorial sublinear-time Fourier algorithms. *Foundations of Computational Mathematics*, 10:303–338, 2010.

[27] M.A. Iwen. Improved approximation guarantees for sublinear-time Fourier algorithms. *Applied And Computational Harmonic Analysis*, 2012.

[28] E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *STOC*, 1991.

[29] D. Lawlor, Y. Wang, and A. Christlieb. Adaptive sub-linear time Fourier algorithms. *arXiv:1207.6368*, 2012.

[30] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing MRI. *Signal Processing Magazine, IEEE*, 25(2):72–82, 2008.

[31] Y. Mansour. Randomized interpolation and approximation of sparse polynomials. *ICALP*, 1992.

[32] Sameer Pawar and Kannan Ramchandran. Computing a k-sparse n-length Discrete Fourier Transform using at most $4k$ samples and $O(k \log k)$ complexity. *ISIT*, 2013.

[33] Sameer Pawar and Kannan Ramchandran. A robust FFAST framework for computing a k-sparse n-length DFT in $O(k \log n)$ sample complexity using sparse-graph codes. *Manuscript*, 2014.

[34] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *CPAM*, 2008.

[35] Emil Sidky. What does compressive sensing mean for X-ray CT and comparisons with its MRI application. In *Conference on Mathematics of Medical Imaging*, 2011.

[36] Elias M. Stein and Rami Shakarchi. *Fourier Analysis:An Introduction*. Princeton University Press, 2003.

[37] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[38] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[39] Juhwan Yoo, S. Becker, M. Loh, M. Monge, E. Candès, and A. E-Neyestanak. A 100MHz–2GHz 12.5x subNyquist rate receiver in 90nm CMOS. In *IEEE RFIC*, 2012.