2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Fourier-sparse interpolation without a frequency gap

Xue Chen
Department of Computer Science
The University of Texas at Austin
Austin, TX, USA
xchen@cs.utexas.edu

Daniel M. Kane
Departments of CSE and Mathematics
University of California, San Diego
La Jolla, CA, USA
dakane@ucsd.edu

Eric Price
Department of Computer Science
The University of Texas at Austin
Austin, TX, USA
ecprice@cs.utexas.edu

Zhao Song
Department of Computer Science
The University of Texas at Austin
Austin, TX, USA
zhaos@utexas.edu

***Abstract*—We consider the problem of estimating a Fourier-sparse signal from noisy samples, where the sampling is done over some interval $[0, T]$ and the frequencies can be "off-grid". Previous methods for this problem required the gap between frequencies to be above $1/T$, the threshold required to robustly identify individual frequencies. We show the frequency gap is not necessary to estimate the signal as a whole: for arbitrary k-Fourier-sparse signals under ℓ_2 bounded noise, we show how to estimate the signal with a constant factor growth of the noise and sample complexity polynomial in k and logarithmic in the bandwidth and signal-to-noise ratio.**

As a special case, we get an algorithm to interpolate degree d polynomials from noisy measurements, using $O(d)$ samples and increasing the noise by a constant factor in ℓ_2.

***Keywords*-Fourier transform; super-resolution; sparse recovery; polynomial interpolation; compressive sensing;**

I. INTRODUCTION

In an interpolation problem, one can observe $x(t) = x^*(t) + g(t)$, where $x^*(t)$ is a structured signal and $g(t)$ denotes noise, at points t_i of one's choice in some interval $[0, T]$. The goal is to recover an estimate $\widetilde{x}$ of x^* (or of x). Because we can sample over a particular interval, we would like our approximation to be good on that interval, so for any function $y(t)$ we define

$$\|y\|_T^2 = \frac{1}{T}\int_0^T |y(t)|^2 \mathrm{d}t.$$

to be the ℓ_2 error on the sample interval. For some parameters C and δ, we would then like to get

$$\|\widetilde{x} - x^*\|_T \leq C\,\|g\|_T + \delta\,\|x^*\|_T \quad (1)$$

while minimizing the number of samples and running time. Typically, we would like C to be $O(1)$ and to have δ be very small (either zero, or exponentially small). Note that, if we do not care about changing C by $O(1)$, then by the triangle inequality it doesn't matter whether we want to estimate x^* or x (i.e. we could replace the LHS of (1) by $\|\widetilde{x} - x\|_T$).

Of course, to solve an interpolation problem one also needs x^* to have structure. One common form of structure is that x^* have a sparse Fourier representation. We say that a function x^* is k-Fourier-sparse if it can be expressed as a sum of k complex exponentials:

$$x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}.$$

for some $v_j \in \mathbb{C}$ and $f_j \in [-F, F]$, where F is the "bandlimit". Given F, T, and k, how many samples must we take for the interpolation (1)?

If we ignore sparsity and just use the bandlimit, then Nyquist sampling and Shannon-Whittaker interpolation uses $FT + 1/\delta$ samples to achieve (1). Alternatively, in the absence of noise, x^* can be found from $O(k)$ samples by a variety of methods, including Prony's method from 1795 or Reed-Solomon syndrome decoding [1], but these methods are not robust to noise.

If the signal is periodic with period T—i.e., the frequencies are multiples of $1/T$—then we can use sparse discrete Fourier transform methods, which take $O(k \log^c(FT/\delta))$ time and samples (e.g. [2], [3], [4]). If the frequencies are not multiples of $1/T$ (are "off the grid"), then the discrete approximation is only k/δ sparse, making the interpolation less efficient; and even this requires that the frequencies be well separated.

A variety of algorithms have been designed to recover off-grid frequencies directly, but they require the minimum gap among the frequencies to be above some threshold. With frequency gap at least $1/T$, we can achieve a k^c approximation factor using $O(FT)$ samples [5], and with gap above $O(\log^2 k)/T$ we can get a constant approximation using $O(k \log^c(FT/\delta))$ samples and time [6].

Having a dependence on the frequency gap is natural. If two frequencies are very close together—significantly below

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.84

$1/T$—then the corresponding complex exponentials will be close on $[0, T]$, and hard to distinguish in the presence of noise. In fact, from a lower bound in [5], below $1/T$ frequency gap one cannot recover the frequencies in the presence of noise as small as $2^{-\Omega(k)}$. The lower bound proceeds by constructing two signals using significantly different frequencies that are exponentially close over $[0, T]$.

But if two signals are so close, do we need to distinguish them? Such a lower bound doesn't apply to the interpolation problem, it just says that you can't solve it by finding the frequencies. Our question becomes: can we benefit from Fourier sparsity in a regime where we can't recover the individual frequencies?

We answer in the affirmative, giving an algorithm for the interpolation using $O(\mathrm{poly}(k \log(FT/\delta))$ samples. Our main theorem is the following:

Theorem I.1. *Let $x(t) = x^*(t) + g(t)$, where x^* is k-Fourier-sparse signal with frequencies in $[-F, F]$. Given samples of x over $[0, T]$ we can output $\widetilde{x}(t)$ such that with probability at least $1 - 2^{-\Omega(k)}$,*

$$\|\widetilde{x} - x^*\|_T \lesssim \|g\|_T + \delta \|x^*\|_T .$$

Our algorithm uses $\mathrm{poly}(k, \log(1/\delta)) \cdot \log(FT)$ samples and $\mathrm{poly}(k, \log(1/\delta)) \cdot \log^2(FT)$ time. The output $\widetilde{x}$ is $\mathrm{poly}(k, \log(1/\delta))$-Fourier-sparse signal.

Relative to previous work, this result avoids the need for a frequency gap, but loses a polynomial factor in the sample complexity and time. We lose polynomial factors in a number of places; some of these are for ease of exposition, but others are challenging to avoid.

Degree d polynomials are the special case of d-Fourier-sparse functions in the limit of $f_j \to 0$, by a Taylor expansion. This is a regime with no frequency gap, so previous sparse Fourier results would not apply but Theorem I.1 shows that $\mathrm{poly}(d \log(1/\delta))$ samples suffices. In fact, in this special case we can get a better polynomial bound:

Theorem I.2. *For any degree d polynomial $P(t)$ and an arbitrary function $g(t)$, Procedure* ROBUSTPOLYNOMIALLEARNING *takes $O(d)$ samples from $x(t) = P(t) + g(t)$ over $[0, T]$ and reports a degree d polynomial $Q(t)$ in time $O(d^\omega)$ such that, with probability at least $99/100$,*

$$\|P(t) - Q(t)\|_T^2 \lesssim \|g(t)\|_T^2.$$

where $\omega < 2.373$ is matrix multiplication exponent [7], [8], [9].

In the full version we also show how to reduce the failure probability to an arbitrary $p > 0$ with $O(\log(1/p))$ independent repetitions.

Although we have not seen such a result stated in the literature, our method is quite similar to one used in [10]. Since d samples are necessary to interpolate a polynomial without noise, the result is within constant factors of optimal.

One could apply Theorem I.2 to approximate other functions that are well approximated by polynomials or piecewise polynomials. For example, a Gaussian of standard deviation at least σ can be approximated by a polynomial of degree $O(\left(\frac{T}{\sigma}\right)^2 + \log(1/\delta))$; hence the same bound applies as the sample complexity of improper interpolation of a positive mixture of Gaussians.

A. Related work

Sparse discrete Fourier transforms: There is a large literature on sparse discrete Fourier transforms. Results generally are divided into two categories: one category of results that carefully choose measurements that allow for sublinear recovery time, including [2], [11], [12], [13], [3], [14], [4], [15]. The other category of results expect randomly chosen measurements and show that a generic recovery algorithm such as ℓ_1 minimization will work with high probability; these results often focus on proving the Restricted Isometry Property [16], [17], [18], [19]. At the moment, the first category of results have better theoretical sample complexity and running time, while results in the second category have better failure probabilities and empirical performance. Our result falls in the first category. The best results here can achieve $O(k \log n)$ samples [14], $O(k \log^2 n)$ time [12], or within $\log \log n$ factors of both [4].

For signals that are not periodic, the discrete Fourier transform will not be sparse: it takes k/δ frequencies to capture a $1 - \delta$ fraction of the energy. To get a better dependence on δ, one has to consider frequencies "off the grid", i.e. that are not multiples of $1/T$.

Off the grid: Finding the frequencies of a signal with sparse Fourier transform off the grid has been a question of extensive study. The first algorithm was by Prony in 1795, which worked in the noiseless setting. This was refined by classical algorithms like MUSIC [20] and ESPRIT [21], which empirically work better with noise. Matrix pencil [22] is a method for computing the maximum likelihood signal under Gaussian noise and evenly spaced samples. The question remained how accurate the maximum likelihood estimate is; [5] showed that it has an $O(k^c)$ approximation factor if the frequency gap is at least $1/T$.

Now, the above results all use FT samples, which is analogous to n in the discrete setting. This can be decreased down till $O(k)$ by only looking at a subset of time, i.e. decreasing T; but doing so increases the frequency gap needed for decent robustness results.

A variety of works have studied how to adapt sparse Fourier techniques from the discrete setting to get sublinear sample complexity; they all rely on the minimum separation among the frequencies to be at least c/T for $c \geq 1$. [23] showed that a convex program can recover the frequencies exactly in the noiseless setting, for $c \geq 4$. This was improved in [24] to $c \geq 2$ for complex signals and $c \geq 1.87$ for real signals. [24] also gave a result for $c \geq 2$ that was stable to noise, but this required the signal frequencies to be placed on a finely spaced grid. [25] gave a different convex relaxation that empirically requires smaller c in the noiseless setting. [26] used model-based compressed sensing when $c = \Omega(1)$, again without theoretical noise stability. Note that, in the noiseless setting, exact recovery can be achieved without any frequency separation using Prony's method or Berlekamp-Massey syndrome decoding [1]; the benefit of the above results is that a convex program might be robust to noise, even if it has not been proven to be so.

In the noisy setting, [27] gave an extension of Orthogonal Matching Pursuit (OMP) that can recover signals when $c = \Omega(k)$, with an approximation factor $O(k)$, and a few other assumptions. Similarly, [28] gave a method that required $c = \Omega(k)$ and was robust to certain kinds of noise. [29] got the threshold down to $c = O(1)$, in multiple dimensions, but with approximation factor $O(FTk^{O(1)})$.

[30] shows that, under Gaussian noise and with separation $c \geq 4$, a semidefinite program can optimally estimate $x^*(t_i)$ at evenly spaced sample points t_i from observations $x^*(t_i) + g(t_i)$. This is somewhat analogous to our setting, the differences being that (a) we want to estimate the signal over the entire interval, not just the sampled points, (b) our noise g is adversarial, so we cannot hope to reduce it—if g is also k-Fourier-sparse, we cannot distinguish x^* and g, and of course (c) we want to avoid requiring frequency separation.

In [6], we gave the first algorithm with $O(1)$ approximation factor, finding the frequencies when $c \gtrsim \log(1/\delta)$, and the signal when $c \gtrsim \log(1/\delta) + \log^2 k$.

Now, all of the above results algorithms are designed to recover the frequencies; some of the ones in the noisy setting then show that this yields a good approximation to the overall signal (in the noiseless setting this is trivial). Such an approach necessitates $c \geq 1$: [5] gave a lower bound, showing that any algorithm finding the frequencies with approximation factor $2^{o(k)}$ must require $c \geq 1$.

Thus, in the current literature, we go from not knowing how to get any approximation for $c < 1$, to getting a polynomial approximation at $c = 1$ and a constant approximation at $c \gtrsim \log^2 k$. In this work, we show how to get a constant factor approximation to the signal regardless of c.

Polynomial interpolation: Our result is a generalization of robust polynomial interpolation, and in Theorem I.2 we construct an optimal method for polynomial interpolation as a first step toward interpolating Fourier-sparse signals.

Our result here can be seen as essentially an extension of a technique shown in [10]. The focus of [10] is on the setting where sample points x_i are chosen independently, so $\Theta(d \log d)$ samples are necessary. One of their examples, however, shows a key building block that leads to our theorem.

The recent work [31] looks at robust polynomial interpolation in a different noise model, featuring ℓ_∞ bounded noise with some outliers. In this setting they can get a stronger ℓ_∞ guarantee on the output than is possible in our setting.

Nyquist sampling: The classical method for learning bandlimited signals uses Nyquist sampling—i.e., samples at rate $1/F$, for FT points—and interpolates them using Shannon-Nyquist interpolation. This doesn't require any frequency gap, but also doesn't benefit from sparsity like sparse Fourier transform-based techniques. As discussed in [6], on the signal $x(t) = 1$ it takes $FT + O(1/\delta)$ samples to get δ error on average. Our dependence is logarithmic on both those terms.

B. *Our techniques*

Previous results on sparse Fourier transforms with robust recovery all required a frequency gap. So consider the opposite situation, where all the frequencies converge to zero and the coefficients are adjusted to keep the overall energy fixed. If we take a Taylor expansion of each complex exponential, then the signal will converge to a degree k polynomial. So robust polynomial interpolation is a necessary subproblem for our algorithm.

Polynomial interpolation: Let $P(x)$ be a degree d polynomial, and suppose that we can query $f(x) = P(x) + g(x)$ over the interval $[-1, 1]$, where g represents adversarial noise. We would like to query f at $O(d)$ points and output a degree d polynomial $Q(x)$ such that $\|P - Q\| \lesssim \|g\|$, where we define $\|h\|^2 := \int_{-1}^{1} |h(x)|^2 \mathrm{d}x$.

One way to do this would be to sample points $S \subset [-1, 1]$ uniformly, then output the degree d polynomial Q with the smallest empirical error

$$\|P + g - Q\|_S^2 := \frac{1}{|S|} \sum_{x \in S} |(P + g - Q)(x)|^2$$

on the observed points. If $\|R\|_S \approx \|R\|$ for all degree d polynomials R, in particular for $P - Q$, then since usually $\|g\|_S \lesssim \|g\|$ by Markov's inequality, the result follows.

This has two problems: first, uniform sampling is poor because polynomials like Chebyshev polynomials can have most of their energy within $O(1/d^2)$ of the edges of the interval. This necessitates $\Omega(d^2)$ uniform samples before $\|R\|_S \approx \|R\|$ with good probability on a single polynomial. Second, the easiest method to extend from approximating one polynomial to approximating all polynomials uses a union bound over a net exponential in d, which would give an $O(d^3)$ bound.

To fix this, we need to bias our sampling toward the edges of the interval and we need our sampling to not be iid. We partition $[-1, 1]$ into $O(d)$ intervals $I_1, \ldots, I_n$ so that the interval containing each x has width at most $O(\sqrt{1 - x^2})$, except for the $O(1/d^2)$ size regions at the edges. For any degree d polynomial R and any choice of n points $x_i \in I_i$, the appropriately weighted empirical energy is close to $\|R\|$. This takes care of both issues with uniform sampling. If the points are chosen uniformly at random from within their intervals, then $\|g\|$ is probably bounded as well, and the empirically closest degree d polynomial Q will satisfy our requirements.

This result is shown in the full version.

Clusters: Many previous sparse Fourier transform algorithms start with a one-sparse recovery algorithm, then show how to separate frequencies to get a k-sparse algorithm by reducing to the one-sparse case. Without a frequency gap, we cannot hope to reduce to the one-sparse case; instead, we reduce to individual clusters of nearby frequencies.

Essentially the problem is that one *cannot* determine all of the high-energy frequencies of a function x only by sampling it on a bounded interval, as some of the frequencies might cancel each other out on this interval. We also cannot afford to work merely with the frequencies of the truncation of x to the interval $[0, T]$, as the truncation operation will spread the frequencies of x over too wide a range. To fix this problem, we must do something in between the two. In particular, we instead study $x \cdot H$ for a judiciously chosen function H. We want H to approximate the indicator function of the interval $[0, T]$ and have small Fourier-support, $\text{supp}(\widehat{H}) \subset [-k^c/T, k^c/T]$. By using some non-trivial lemmas about the growth rate of x^*, we can show that the difference between $x \cdot H$ on $\mathbb{R}$ and the truncation of x to $[0, T]$ has small ℓ_2 mass, so that we can use the former as a substitute for the latter.

On the other hand, the Fourier transform of $x \cdot H$ is the convolution $\widehat{x} * \widehat{H}$, which has most of its mass within $\text{poly}(k)/T$ of the frequencies of x^*. Although it is impossible to determine the individual frequencies of x^*, we can hope to identify $O(k)$ intervals each of length $\text{poly}(k)/T$ so that all but a small fraction of the energy of $\widehat{x}$ is contained within these intervals.

Note that many of these intervals will represent not individual frequencies of x^*, but small clusters of such frequencies. Furthermore, some frequencies of x^* might not show up in these intervals either because they are too small, or because they cancel out other frequencies when convolved with $\widehat{H}$.

One-cluster recovery: Given our notion of clusters, we start looking at Fourier-sparse interpolation in the special case of *one-cluster recovery*. This is a generalization of one-sparse recovery where we can have multiple frequencies, but they all lie in $[f - \Delta, f + \Delta]$ for some base frequency f and bandwidth $\Delta = k^c/T$. Because all the frequencies are close to each other, values $x(a)$ and $x(a + \beta)$ will tend to have ratio close to $e^{2\pi \mathbf{i} f \beta}$ when β is small enough. We find that $\beta < \frac{1}{\Delta\sqrt{T\Delta}}$ is sufficient, which lets us figure out a frequency $\widetilde{f}$ with $|\widetilde{f} - f| \le \Delta\sqrt{T\Delta} = k^{O(1)}/T$.

Once we have the frequency $\widetilde{f}$, we can consider $x'(t) = x(t)e^{-2\pi \mathbf{i} \widetilde{f}}$. This signal is k-Fourier-sparse with frequencies bounded by $k^{O(1)}/T$. By taking a Taylor approximation to each complex exponential[1], can show x^* is δ-close to $P(t)e^{2\pi \mathbf{i} \widetilde{f}}$ for a degree $d = O(k^c + k \log(1/\delta))$ polynomial P. Thus we could apply our polynomial interpolation algorithm to recover the signal.

k-cluster frequency estimation: Reminiscent of algorithms such as [3], [6], we choose random variables $\sigma \approx T/k^c$, $a \in [0, 1]$, and $b \in [0, 1/\sigma]$ and look at $v \in \mathbb{C}^{k^c}$ given by

$$v_i = (x \cdot H)(\sigma(i - a))e^{-2\pi \mathbf{i}\sigma b i}G(i)$$

where G is a filter function. That is, G has compact support ($\text{supp}(G) \subset [-k^c, k^c]$), and $\widehat{G}$ approximates an interval of length $\Theta(\frac{2\pi}{k})$. In other words, G is the same as $\widehat{H}$ with different parameters: an interval convolved with itself k^c times, multiplied by a sinc function.

We alias v down to $O(k)$ dimensions and take the discrete Fourier transform, getting $\widehat{u}$. It has been implicit in previous work—and we make it explicit—that $\widehat{u}_j$ is equal to $z_{\sigma a}$ for a vector z defined by

$$\widehat{z} = (\widehat{x} * \widehat{H}) \cdot \widehat{G}^{(j)}_{\sigma,b}$$

where $\widehat{G}^{(j)}_{\sigma,b}$ is a particular permutation of $\widehat{G}$. In particular, $\widehat{G}^{(j)}_{\sigma,b}$ has period $1/\sigma$, and approximates an interval of size $\frac{1}{\sigma B}$ within each period.

In previous work, when σ and b were chosen randomly, each individual frequency would have a good chance of being the only frequency preserved in $\widehat{z}$, and we could apply one-sparse recovery by choosing a variety of a. Without a frequency gap we can't quite say that: we pick $1/\sigma \gg \Delta$ so that the entire cluster usually lands in the same bin, but then nearby clusters can also often land in the same bin. Fortunately, it is still usually true that only nearby clusters will collide. Since our 1-cluster algorithm works when the signal frequencies are nearby, we apply it to find a frequency approximation within $\frac{\sqrt{T/\sigma}}{\sigma} = k^{O(1)}/T$ of the cluster.

The above algorithm recovers each individual frequency with constant probability. By repeating it $O(\log k)$ times, with high probability we find a list L of $O(k)$ frequencies within $k^{O(1)}/T$ of each significant cluster.

k-sparse recovery: Because different clusters aren't anywhere close to orthogonal, we can't simply approximate each cluster separately and add them up. Instead, given the list L of candidate frequencies, we consider the $O(kd)$-dimensional space of functions

$$\widetilde{x}(t) := \sum_{\widetilde{f} \in L} \sum_{i=0}^{d} \alpha_{\widetilde{f},i} t^i e^{2\pi \mathbf{i} \widetilde{f} t}$$

where $d = O(k^{O(1)} + \log(1/\delta))$. We then take a bunch of random samples of x, and choose the $\widetilde{x}(t)$ minimizing the

[1]There is a catch here, that the coefficients of the exponentials are potentially unbounded, if the frequencies are arbitrarily close together. We first use Gram determinants to show that the signal is δ-close to one with frequency gap $\delta 2^{-k}$, and coefficients at most $2^k/\delta$.

empirical error using linear regression. This regression can be made slightly faster using oblivious subspace embeddings [32], [33], [34], [35].

Our argument to show this works is analogous to the naive method we considered for polynomial recovery. Similarly to the one-cluster setting, using Taylor approximations and Gram determinants, we can show that this space includes a sufficiently close approximation to x. Since polynomials are the limit of sparse Fourier as frequencies tend to zero, these functions are arbitrarily close to $O(kd)$-Fourier-sparse functions. Hence we know that the maximum of $|\widetilde{x}(t)|$ is at most a $\mathrm{poly}(kd)$ factor larger than its average over $[0, T]$. Using a net argument, this shows $\mathrm{poly}(kd)$ samples are sufficient to find a good approximation to the nearest function in our space.

Growth rate of Fourier-sparse signals: We need that $\frac{1}{\sqrt{T}} \|x^* \cdot H\|_2 \approx \|x^*\|_T$, where H approximates the interval $1_{[0,T]}$. Because H has support size k^c/T, it has a transition region of size $T/k^{c'}$ at the edges, and it decays as $(t/T)^{-k^{c''}}$ for $t \gg T$. The difference between $\frac{1}{\sqrt{T}} \|x^* \cdot H\|_2$ and $\|x^*\|_T$ involves two main components: mass in the transition region that is lost, and mass outside the sampling interval that is gained. To show the approximation, we need that $|x^*(t)| \lesssim \widetilde{O}(k^2) \|x^*\|_T$ within the interval and $|x^*(t)| \lesssim (kt/T)^{\widetilde{O}(k)} \|x^*\|_T$ outside.

We outline the bound of $\max\limits_{t\in[0,T]} |x^*(t)|$ in terms of its average $\|x^*\|_T$ to bound $|x^*(t)|$ within the interval. Notice that we can assume $|x^*(0)| = \max\limits_{t\in[0,T]} |x^*(t)|$: if $t^* = \arg\max\limits_{t\in[0,T]} |x^*(t)|^2$ is not 0 or T, we can rescale the two intervals $[0, t^*]$ and $[t^*, T]$ to $[0, T]$ separately. Then we show that for any t', there exist $m = \widetilde{O}(k^2)$ and constants $C_1, \cdots, C_m$ such that $x^*(0) = \sum\limits_{j\in[m]} C_j \cdot x^*(j \cdot t')$. Then we take the integration of t' over $[0, T/m]$ to bound $|x^*(0)|^2$ by its average. For any outside $t > T$, we follow this approach to show $x^*(t) = \sum\limits_{j\in[k]} C_j \cdot x^*(t_j)$ where $t_j \in [0, T]$ and $|C_j| \le \mathrm{poly}(k) \cdot (kt/T)^{O(k)}$ for each $j \in [k]$. These results are shown in Section IV.

C. Organization

This paper is organized as follows. We define several notations in Section II. We provide a brief overview about signal recovery in Section III. In Section IV, we show two bounds for signals with k-sparse Fourier transform. We defer the rest statements and proofs to the full version.

II. Notations

For any function f, we define $\widetilde{O}(f)$ to be $f \cdot \log^{O(1)}(f)$. We use $f \lesssim g$ to denote that $f \le Cg$ for some universal constant C. We use $[n]$ to denote $\{1, 2, \cdots, n\}$. We use k, d, T, F for sparsity of signal, degree of polynomial, sample duration, frequency range. For a fixed $T > 0$, we define the inner product of two functions $x, y : [0, T] \to \mathbb{C}$ as

$$\langle x, y\rangle_T = \frac{1}{T}\int_0^T x(t)\overline{y}(t)\mathrm{d}t.$$

We define the $\|\cdot\|_T$ norm as

$$\|x(t)\|_T = \sqrt{\langle x(t), x(t)\rangle_T} = \sqrt{\frac{1}{T}\int_0^T |x(t)|^2\mathrm{d}t}.$$

III. Proof Sketch

In this section we present the key lemmas on the path to producing the algorithm. Proofs are deferred to the full version.

We first consider one-cluster recovery centered at zero, i.e., $x^*(t) = \sum\limits_{j=1}^k v_j \cdot e^{2\pi\mathbf{i} f_j t}$ where every f_j is in $[-\Delta, \Delta]$ for some small $\Delta > 0$. The road map is to replace x^* by a low degree polynomial P such that $\|x^*(t) - P(t)\|_T^2 \lesssim \delta\|x^*\|_T^2$ then recover a polynomial Q to approximate P through the observation $x(t) = P(t) + g'(t)$ where $g'(t) = g(t) + \big(x^*(t) - P(t)\big)$.

A natural way to replace $x^*(t) = \sum\limits_{j=1}^k v_j e^{2\pi\mathbf{i} f_j t}$ by a low degree polynomial $P(t)$ is the Taylor expansion. To bound the error after taking the low degree terms in the expansion by $\delta\|x^*\|_T$, we show the existence of $x'(t) = \sum\limits_{j=1}^k v'_j e^{2\pi\mathbf{i} f'_j t}$ approximating x^* on $[0, T]$ with an extra property—any coefficient v'_j in $x'(t)$ has an upper bound in terms of $\|x'\|_T^2 = \frac{1}{T}\int_0^T |x'(t)|^2\mathrm{d}t$. We prove the existence of $x'(t)$ via two more steps, both of which rely on the estimation of some Gram matrix constituted by these k signals.

The first step is to show the existence of a k-Fourier-sparse signal $x'(t)$ with frequency gap $\eta \ge \frac{\exp(-\mathrm{poly}(k))\cdot\delta}{T}$ that is sufficiently close to $x^*(t)$.

Lemma III.1. *There is a universal constant $C_1 > 0$ such that, for any $x^*(t) = \sum\limits_{j=1}^k v_j e^{2\pi\mathbf{i} f_j t}$ and any $\delta > 0$, there always exist $\eta \ge \frac{\delta}{T} \cdot k^{-C_1 k^2}$ and $x'(t) = \sum\limits_{j=1}^k v'_j e^{2\pi\mathbf{i} f'_j t}$ satisfying*

$$\|x'(t) - x^*(t)\|_T \le \delta\|x^*(t)\|_T,$$

with $\min\limits_{i\neq j} |f'_i - f'_j| \ge \eta$ and $\max\limits_{j\in[k]} \{|f'_j - f_j|\} \le k\eta$.

We outline our approach and defer the proof to the full version. We focus on the replacement of one frequency f_k in $x^* = \sum\limits_{j=1}^k v_j e^{2\pi\mathbf{i} f_j t}$ by a new frequency $f_{k+1} \neq f_k$ and its error. The idea is to consider every signal $e^{2\pi\mathbf{i} f_j t}$ as a vector and prove that for any vector x^* in the linear subspace $\mathrm{span}\{e^{2\pi\mathbf{i} f_j t} | j \in [k]\}$, there exists a vector in the linear subspace $\text{span}\{e^{2\pi \mathbf{i} f_{k+1} t}, e^{2\pi \mathbf{i} f_j t} | j \in [k-1]\}$ with distance at most $\exp(k^2) \cdot (|f_k - f_{k+1}| T) \cdot \|x^*\|_T$ to x^*.

The second step is to lower bound $\|x'\|_T^2$ by its coefficients through the frequency gap η in x'.

Lemma III.2. *There exists a universal constant $c > 0$ such that for any $x(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}$ with frequency gap $\eta = \min_{i \neq j} |f_i - f_j|$,*

$$\|x(t)\|_T^2 \geq k^{-ck^2} \min\left((\eta T)^{2k}, 1\right) \sum_{j=1}^{k} |v_j|^2.$$

Combining Lemma III.1 and Lemma III.2, we bound $|v_j'|$ by $\exp(\text{poly}(k)) \cdot \delta^{-O(k)} \cdot \|x'\|_T$ for any coefficient v_j' in x'. Now we apply the Taylor expansion on $x'(t)$ and keep the first $d = O(\Delta T + \text{poly}(k) + k \log \frac{1}{\delta})$ terms of every signal $v_j' \cdot e^{2\pi \mathbf{i} f_j' t}$ in the expansion to obtain a polynomial $P(t)$ of degree at most d. To bound the distance between $P(t)$ and $x'(t)$, we observe that the error of every point $t \in [0, T]$ is at most $(\frac{2\pi \Delta \cdot T}{d})^d \sum_j |v_j'|$, which can be upper bounded by $\delta \|x'(t)\|_T$ via the above connection. We summarize all discussion above as follows.

Lemma III.3. *For any $\Delta > 0$ and any $\delta > 0$, let $x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}$ where $|f_j| \leq \Delta$ for each $j \in [k]$. There exists a polynomial $P(t)$ of degree at most*

$$d = O(T\Delta + k^3 \log k + k \log 1/\delta)$$

such that

$$\|P(t) - x^*(t)\|_T^2 \leq \delta \|x^*\|_T^2.$$

To recover $x^*(t)$, we observe $x(t)$ as a degree d polynomial $P(t)$ with noise. We use properties of the Legendre polynomials to design a method of random sampling such that we only need $O(d)$ random samples to find a polynomial $Q(t)$ approximating $P(t)$.

Theorem I.2. *For any degree d polynomial $P(t)$ and an arbitrary function $g(t)$, Procedure* ROBUSTPOLYNOMIALLEARNING *takes $O(d)$ samples from $x(t) = P(t) + g(t)$ over $[0, T]$ and reports a degree d polynomial $Q(t)$ in time $O(d^{\omega})$ such that, with probability at least 99/100,*

$$\|P(t) - Q(t)\|_T^2 \lesssim \|g(t)\|_T^2.$$

where $\omega < 2.373$ is matrix multiplication exponent [7], [8], [9].

We can either report the polynomial $Q(t)$ or transfer $Q(t)$ to a signal with d-sparse Fourier transform. We defer the technical proofs and the formal statements to the full version and discuss the recovery of k clusters from now on.

As mentioned before, we apply the filter function $(H(t), \widehat{H}(f))$ on x^* such that $\widehat{x^* \cdot H}$ has at most k clusters given $\widehat{x^*}$ with k-sparse Fourier transform. First, we show that all frequencies in the "heavy" clusters of $\widehat{x^* \cdot H}$ constitute a good approximation of x^* in the full version.

Definition III.4. *Given $x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}$, any $\mathcal{N} > 0$, and a filter function $(H, \widehat{H})$ with bounded support in frequency domain. Let L_j denote the interval of $\text{supp}(\widehat{e^{2\pi \mathbf{i} f_j t} \cdot H})$ for each $j \in [k]$.*

Define an equivalence relation $\sim$ on the frequencies f_i by the transitive closure of the relation $f_i \sim f_j$ if $L_i \cap L_j \neq \emptyset$. Let $S_1, \ldots, S_n$ be the equivalence classes under this relation.

Define $C_i = \bigcup_{f \in S_i} L_i$ for each $i \in [n]$. We say C_i is a "heavy" cluster iff $\int_{C_i} |\widehat{H \cdot x^*}(f)|^2 \mathrm{d}f \geq T \cdot \mathcal{N}^2 / k$.

Claim III.5. *Given $x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}$ and any $\mathcal{N} > 0$, let H be the filter function defined in the full version and $C_1, \cdots, C_l$ be the heavy clusters from Definition III.4. For*

$$S = \left\{ j \in [k] \middle| f_j \in C_1 \cup \cdots C_l \right\},$$

we have $x^{(S)}(t) = \sum_{j \in S} v_j e^{2\pi \mathbf{i} f_j t}$ approximates x^* within distance $\|x^{(S)}(t) - x^*(t)\|_T^2 \lesssim \mathcal{N}^2$.

Hence it is enough to recover $x^{(S)}$ for the recovery of x^*. Let Δ_h denote the bandwidth of $\widehat{H}$. We choose $\Delta > k \cdot \Delta_h$ such that for any $j \in S$, $\int_{f_j - \Delta}^{f_j + \Delta} |\widehat{H \cdot x^*}(f)|^2 \mathrm{d}f \geq T \cdot \mathcal{N}^2 / k$ from the fact $|C_i| \leq k \cdot \Delta_h$. Then we prove Theorem III.6 which finds $O(k)$ frequencies to cover all heavy clusters of $\widehat{x^* \cdot H}$.

Theorem III.6. *Let $x^*(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t}$ and $x(t) = x^*(t) + g(t)$ be our observable signal where $\|g(t)\|_T^2 \leq c \|x^*(t)\|_T^2$ for a sufficiently small constant c. Then Procedure* FREQUENCYRECOVERYKCLUSTER *returns a set L of $O(k)$ frequencies that covers all heavy clusters of x^*, which uses $\text{poly}(k, \log(1/\delta)) \log(FT)$ samples and $\text{poly}(k, \log(1/\delta)) \log^2(FT)$ time. In particular, for $\Delta = \text{poly}(k, \log(1/\delta))/T$ and $\mathcal{N}^2 := \|g(t)\|_T^2 + \delta \|x^*(t)\|_T^2$, with probability $1 - 2^{-\Omega(k)}$, for any f^* with*

$$\int_{f^* - \Delta}^{f^* + \Delta} |\widehat{x \cdot H}(f)|^2 \mathrm{d}f \geq T \mathcal{N}^2 / k, \quad (2)$$

there exists an $\widetilde{f} \in L$ satisfying

$$|f^* - \widetilde{f}| \lesssim \Delta \sqrt{\Delta T}.$$

Let $L = \{\widetilde{f}_1, \cdots, \widetilde{f}_l\}$ be the list of frequencies from the output of Procedure FREQUENCYRECOVERYKCLUSTER in Theorem III.6. The guarantee is that, for any f_j in $x^{(S)}$, there exists some $p_j \in [l]$ such that $|\widetilde{f}_{p_j} - f_j| \lesssim \Delta \sqrt{\Delta T}$ for

$\Delta = \text{poly}(k, \log(1/\delta))/T$. Hence we rewrite

$$x^{(S)}(t) = \sum_{i\in[l]} e^{2\pi \mathbf{i} \widetilde{f}_i t}\Big(\sum_{j\in S: p_j = i} e^{2\pi \mathbf{i}(f_j - \widetilde{f}_i)t}\Big).$$

For each $i \in [l]$, we apply Lemma III.3 of one-cluster recovery on $\sum_{j\in S:p_j=i} e^{2\pi\mathbf{i}(f_j-\widetilde{f}_i)t}$ to approximate it by a degree d polynomial $P_i(t)$.

Now we consider $x(t) = \sum_{i\in[l]} e^{2\pi\mathbf{i}\widetilde{f}_i t} \cdot P_i(t) + g''(t)$ where $\|g''(t)\|_T \lesssim \|g(t)\|_T + \delta\|x^*(t)\|_T$. To recover $\sum_{i\in[l]} e^{2\pi\mathbf{i}\widetilde{f}_i t} \cdot P_i(t)$, we treat it as a vector in the linear subspace

$$V = \text{span}\left\{ e^{2\pi\mathbf{i}\widetilde{f}_i t} \cdot t^j \middle| j \in \{0, \cdots, d\}, i \in [l] \right\}$$

with dimension at most $l(d+1)$ and find a vector in this linear subspace approximating it.

We show that for any $v \in V$, the average of $\text{poly}(kd)$ random samples on v is enough to estimate $\|v\|_T^2$. In particular, any vector in this linear subspace satisfies that the maximum of it in $[0,T]$ has an upper bound in terms of its average in $[0,T]$. Then we apply the Chernoff bound to prove that $\text{poly}(kd)$ random samples are enough for the estimation of one vector $v \in V$.

Claim III.7. *For any*

$$\vec{u} \in \text{span}\left\{ e^{2\pi\mathbf{i}\widetilde{f}_i t} \cdot t^j \middle| j \in \{0, \cdots, d\}, i \in [l] \right\},$$

there exists some universal constants $C_1 \leq 4$ and $C_2 \leq 3$ such that

$$\max_{t\in[0,T]} \{|\vec{u}(t)|^2\} \lesssim (ld)^{C_1} \log^{C_2}(ld) \cdot \|\vec{u}\|_T^2.$$

At last we use an ϵ-net to argue that $\text{poly}(kd)$ random samples from $[0,T]$ are enough to interpolate $x(t)$ by a vector $v \in V$. Because the dimension of this linear subspace is at most $l(d+1) = O(kd)$, there exists an ϵ-net in this linear subspace for unit vectors with size at most $\exp(kd)$. Combining the Chernoff bound on all vectors in the ϵ-net and Claim III.7, we know that $\text{poly}(kd)$ samples are sufficient to estimate $\|v\|_T^2$ for any vector $v \in V$. In the full version, we show that a vector $v \in V$ minimizing the distance on $\text{poly}(kd)$ random samples is a good approximation for $\sum_{i\in[l]} e^{2\pi\mathbf{i}\widetilde{f}_i t} \cdot P_i(t)$, which is a good approximation for $x^*(t)$ from all discussion above.

Theorem I.1. *Let $x(t) = x^*(t) + g(t)$, where x^* is k-Fourier-sparse signal with frequencies in $[-F, F]$. Given samples of x over $[0,T]$ we can output $\widetilde{x}(t)$ such that with probability at least $1 - 2^{-\Omega(k)}$,*

$$\|\widetilde{x} - x^*\|_T \lesssim \|g\|_T + \delta\,\|x^*\|_T\,.$$

Our algorithm uses $\text{poly}(k, \log(1/\delta)) \cdot \log(FT)$ samples and $\text{poly}(k, \log(1/\delta)) \cdot \log^2(FT)$ time. The output $\widetilde{x}$ is $\text{poly}(k, \log(1/\delta))$-Fourier-sparse signal.

IV. Bounds on the magnitude of a Fourier-sparse signal in terms of its average norm

The main results in this section are two upper bounds, Lemma IV.1 on $\max_{t\in[0,T]} |x(t)|^2$ and Lemma IV.5 on $|x(t)|^2$ for $t > T$, in terms of the typical signal value $\|x\|_T^2 = \frac{1}{T}\int_0^T |x(t)|^2 \mathrm{d}t$. We prove Lemma IV.1 in Section IV-A and Lemma IV.5 in Section IV-B.

A. Bounding the maximum inside the interval

The goal of this section is to prove Lemma IV.1.

Lemma IV.1. *For any k-Fourier-sparse signal $x(t) : \mathbb{R} \to \mathbb{C}$ and any duration T, we have*

$$\max_{t\in[0,T]} |x(t)|^2 \lesssim k^4 \log^3 k \cdot \|x\|_T^2.$$

Proof: Without loss of generality, we fix $T = 1$. Then $\|x\|_T^2 = \int_0^1 |x(t)|^2 \mathrm{d}t$. Because $\|x\|_T^2$ is the average over the interval $[0,T]$, if $t^* = \arg\max_{t\in[0,T]} |x(t)|^2$ is not 0 or $T = 1$, we can rescale the two intervals $[0, t^*]$ and $[t^*, T]$ to $[0,1]$ and prove the desired property separately. Hence we assume $|x(0)|^2 = \max_{t\in[0,T]} |x(t)|^2$ in this proof.

Claim IV.2. *For any k, there exists $m = O(k^2 \log k)$ such that for any k-Fourier-sparse signal $x(t)$, any $t_0 \geq 0$ and $\tau > 0$, there always exist $C_1, \cdots, C_m \in \mathbb{C}$ such that the following properties hold,*

$$\begin{aligned} &\text{Property I} && |C_j| \leq 11 \ \textit{ for all } j \in [m], \\ &\text{Property II} && x(t_0) = \sum_{j\in[m]} C_j \cdot x(t_0 + j\cdot\tau). \end{aligned}$$

We first use this claim to finish the proof of Lemma IV.1. We choose $t_0 = 0$ such that $\forall \tau > 0$, there always exist $C_1, \cdots, C_m \in \mathbb{C}$, and

$$x(0) = \sum_{j\in[m]} C_j \cdot x(j\cdot\tau).$$

By the Cauchy-Schwarz inequality, it implies that for any τ,

$$\begin{aligned} |x(0)|^2 &\leq m \sum_{j\in[m]} |C_j|^2 |x(j\cdot\tau)|^2 \\ &\lesssim m \sum_{j\in[m]} |x(j\cdot\tau)|^2. \end{aligned} \tag{3}$$

At last, we obtain

$$\begin{aligned} |x(0)|^2 = &\quad m\int_0^{1/m}|x(0)|^2\mathrm{d}\tau \\ \lesssim &\quad m\cdot\int_0^{1/m}(m\sum_{j=1}^{m}|x(j\cdot\tau)|^2)\mathrm{d}\tau \\ = &\quad m^2\cdot\sum_{j=1}^{m}\int_0^{1/m}|x(j\cdot\tau)|^2\mathrm{d}\tau \\ = &\quad m^2\cdot\sum_{j=1}^{m}\frac{1}{j}\int_0^{j/m}|x(\tau)|^2\mathrm{d}\tau \\ \leq &\quad m^2\cdot\sum_{j=1}^{m}\frac{1}{j}\cdot\int_0^{1}|x(\tau)|^2\mathrm{d}\tau \\ \lesssim &\quad m^2\log m\cdot\|x\|_T^2, \end{aligned}$$

where the first inequality follows by Equation (3), the second inequality follows by $j/m \leq 1$ and the last step follows by $\sum_{i=1}^{m}\frac{1}{i} = O(\log m)$. From $m = O(k^2\log k)$, we obtain $|x(0)|^2 = O(k^4\log^3 k\|x\|_T^2)$. ■

To prove Claim IV.2, we use the following lemmas about polynomials. We defer their proofs to the full version.

Lemma IV.3. *Let $Q(z)$ be a degree k polynomial, all of whose roots are complex numbers with absolute value 1. For any integer n, let $r_{n,k}(z) = \sum_{l=0}^{k-1} r_{n,k}^{(l)}\cdot z^l$ denote the residual polynomial of*

$$r_{n,k}(z) \equiv z^n \pmod{Q(z)}.$$

Then, each coefficient of $r_{n,k}$ is bounded: $|r_{n,k}^{(l)}| \leq 2^k n^{k-1}$ for any l.

Lemma IV.4. *For any $k \in \mathbb{Z}$ and any $z_1, \cdots, z_k$ on the unit circle of $\mathbb{C}$, there always exists a degree $m = O(k^2\log k)$ polynomial $P(z) = \sum_{j=0}^{m} c_j z^j$ with the following properties:*

Property I	$P(z_i) = 0, \forall i \in \{1, \cdots, k\}$,
Property II	$c_0 = 1$,
Property III	$\|c_j\| \leq 11, \forall j \in \{1, \cdots, m\}$.

Proof of Claim IV.2. For $x(t) = \sum_{i=1}^{k} v_i e^{2\pi \mathbf{i} f_i t}$, we fix t_0 and τ then rewrite $x(t_0+j\cdot\tau)$ as a polynomial of $b_i = v_i\cdot e^{2\pi\mathbf{i}f_i t_0}$ and $z_i = e^{2\pi\mathbf{i}f_i\tau}$ for each $i \in [k]$.

$$\begin{aligned} x(t_0+j\cdot\tau) &= \sum_{i=1}^{k} v_i e^{2\pi\mathbf{i}f_i\cdot(t_0+j\cdot\tau)} \\ &= \sum_{i=1}^{k} v_i e^{2\pi\mathbf{i}f_i t_0}\cdot e^{2\pi\mathbf{i}f_i\cdot j\tau} \\ &= \sum_{i=1}^{k} b_i\cdot z_i^j. \end{aligned}$$

Given k and $z_1, \cdots, z_k$, let $P(z) = \sum_{j=0}^{m} c_j z^j$ be the degree m polynomial in Lemma IV.4.

$$\begin{aligned} \sum_{j=0}^{m} c_j x(t_0+j\tau) &= \sum_{j=0}^{m} c_j\sum_{i=1}^{k} b_i\cdot z_i^j \\ &= \sum_{i=1}^{k} b_i\sum_{j=0}^{m} c_j\cdot z_i^j \\ &= \sum_{i=1}^{k} b_i P(z_i), \\ &= 0, \end{aligned} \tag{4}$$

where the last step follows by Property I of $P(z)$ in Lemma IV.4. From the Property II and III of $P(z)$, we obtain $x(t_0) = -\sum_{j=1}^{m} c_j x(t_0+j\tau)$. □

B. Bounding growth outside the interval

Here we show signals with sparse Fourier transform cannot grow too quickly outside the interval.

Lemma IV.5. *Let $x(t)$ be a k-Fourier-sparse signal. For any $T > 0$ and any $t > T$,*

$$|x(t)|^2 \leq k^7\cdot(2kt/T)^{2.5k}\cdot\|x\|_T^2.$$

Proof: For any $t > T$, let $t = t_0 + n\cdot\tau$ such that $t_0 \in [0, T/k], \tau \in [0, T/k]$ and $n \leq \frac{2kt}{T}$. We define $b_i = v_i e^{2\pi\mathbf{i}f_i t_0}$, and $z_i = v_i e^{2\pi\mathbf{i}f_i\tau}$ such that $x(t_0+n\cdot\tau) = \sum_{j=1}^{k} b_j z_j^n$.

By Lemma IV.3, we have for any $z_1, z_2, \cdots, z_k$ and any n,

$$z^n \equiv \sum_{i=0}^{k-1} a_i z^i \pmod{\prod_{i=1}^{k}(z - z_i)},$$

where $|a_i| \leq 2^k\cdot n^k, \forall i \in \{0, 1, \cdots, k-1\}$. Thus, we obtain

$$x(t_0+n\tau) = \sum_{j=1}^{k} b_j z_j^n = \sum_{j=1}^{k} b_j(\sum_{i=0}^{k-1} a_i z_j^i).$$

From the fact that $x(t_0 + i \cdot \tau) = \sum_{j=1}^{k} b_j z_j^i$, we simplify it to be

$$x(t_0 + n\tau) = \sum_{i=0}^{k-1} a_i \sum_{j=1}^{k} b_j z_j^i = \sum_{i=0}^{k-1} a_i x(t_0 + i \cdot \tau).$$

Because $(t_0 + i \cdot \tau) \in [0, T]$ for any $i = 0, \cdots, k-1$, we have $|x(t_0 + i\tau)|^2 \leq \max_{t \in [0,T]} |x(t)|^2 \lesssim k^4 \log^3 k \|x\|_T^2$ from Lemma IV.1. Hence

$$\begin{aligned} |x(t_0 + n \cdot \tau)|^2 \leq & \quad k \sum_{i=0}^{k-1} |a_i|^2 \cdot |x(t_0 + i \cdot \tau)|^2 \\ \leq & \quad k \sum_{i=0}^{k-1} n^{2.2k} \cdot \max_{t \in [0,T]} |x(t)|^2 \\ \leq & \quad k^7 \cdot (2kt/T)^{2.2k} \|x\|_T^2. \end{aligned}$$

Thus, we complete the proof. ■

ACKNOWLEDGMENT

The authors would to like thank Aaron Sidford and David Woodruff for useful discussions.

REFERENCES

[1] J. L. Massey, "Shift-register synthesis and BCH decoding," *Information Theory, IEEE Transactions on*, vol. 15, no. 1, pp. 122–127, 1969.

[2] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, "Near-optimal sparse Fourier representations via sampling," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 152–161.

[3] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse Fourier transform," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012.

[4] P. Indyk, M. Kapralov, and E. Price, "(Nearly) Sample-optimal sparse Fourier transform," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2014, pp. 480–499.

[5] A. Moitra, "The threshold for super-resolution via extremal functions," in *STOC*, 2015.

[6] E. Price and Z. Song, "A robust sparse Fourier transform in the continuous setting," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 583–600.

[7] V. Strassen, "Gaussian elimination is not optimal," *Numerische Mathematik*, vol. 13, no. 4, pp. 354–356, 1969.

[8] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," in *Proceedings of the nineteenth annual ACM symposium on Theory of computing*. ACM, 1987, pp. 1–6.

[9] V. V. Williams, "Multiplying matrices faster than coppersmith-winograd," in *STOC*. ACM, 2012, pp. 887–898.

[10] A. Cohen, M. A. Davenport, and D. Leviatan, "On the stability and accuracy of least squares approximations," *Foundations of computational mathematics*, vol. 13, no. 5, pp. 819–834, 2013.

[11] A. C. Gilbert, S. Muthukrishnan, and M. Strauss, "Improved time bounds for near-optimal sparse Fourier representations," in *Optics & Photonics 2005*. International Society for Optics and Photonics, 2005, pp. 59 141A–59 141A.

[12] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Simple and practical algorithm for sparse Fourier transform," in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2012, pp. 1183–1194.

[13] M. A. Iwen, "Improved approximation guarantees for sublinear-time Fourier algorithms," *Applied And Computational Harmonic Analysis*, vol. 34, no. 1, pp. 57–82, 2013.

[14] P. Indyk and M. Kapralov, "Sample-optimal Fourier sampling in any constant dimension," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 514–523.

[15] M. Kapralov, "Sparse Fourier transform in any constant dimension with nearly-optimal sample complexity in sublinear time," in *Symposium on Theory of Computing Conference, STOC'16, Cambridge, MA, USA, June 19-21, 2016*, 2016.

[16] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[17] M. Rudelson and R. Vershynin, "On sparse reconstruction from fourier and gaussian measurements," *Communications on Pure and Applied Mathematics*, vol. 61, no. 8, pp. 1025–1045, 2008.

[18] J. Bourgain, "An improved estimate in the restricted isometry problem," in *Geometric Aspects of Functional Analysis*. Springer, 2014, pp. 65–70.

[19] I. Haviv and O. Regev, "The restricted isometry property of subsampled Fourier matrices," *arXiv preprint arXiv:1507.01768*, 2015.

[20] R. O. Schmidt, "A signal subspace approach to multiple emitter location spectral estimation," *Ph. D. Thesis, Stanford University*, 1981.

[21] R. Roy, A. Paulraj, and T. Kailath, "Esprit–a subspace rotation approach to estimation of parameters of cisoids in noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 5, pp. 1340–1342, 1986.

[22] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1081–1089, Oct 1986.

[23] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *Information Theory, IEEE Transactions on*, vol. 59, no. 11, pp. 7465–7490, 2013.

[24] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[25] Z. Yang and L. Xie, "Achieving high resolution for super-resolution via reweighted atomic norm minimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3646–3650.

[26] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 111–129, 2013.

[27] A. Fannjiang and W. Liao, "Coherence pattern-guided compressive sensing with unresolved grids," *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 179–202, 2012.

[28] P. Boufounos, V. Cevher, A. C. Gilbert, Y. Li, and M. J. Strauss, "What's the frequency, Kenneth?: Sublinear Fourier sampling off the grid," in *Algorithmica(A preliminary version of this paper appeared in the Proceedings of RANDOM/APPROX 2012, LNCS 7408, pp. 61-72)*. Springer, 2014, pp. 1–28.

[29] Q. Huang and S. M. Kakade, "Super-resolution off the grid," in *Advances in Neural Information Processing Systems*, 2015, pp. 2647–2655.

[30] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," *Information Theory, IEEE Transactions on*, vol. 61, no. 1, pp. 499–512, 2015.

[31] V. Guruswami and D. Zuckerman, "Robust Fourier and polynomial curve fitting," in *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE, 2016.

[32] K. L. Clarkson and D. P. Woodruff, "Low rank approximation and regression in input sparsity time," in *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, 2013, pp. 81–90.

[33] J. Nelson and H. L. Nguyên, "OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 117–126.

[34] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *arXiv preprint arXiv:1411.4357*, 2014.

[35] M. B. Cohen, J. Nelson, and D. P. Woodruff, "Optimal approximate matrix product in terms of stable rank," *arXiv preprint arXiv:1507.02268*, 2015.