Sublinear Time Low-Rank Approximation of Toeplitz Matrices

Cameron Musco[*] Kshiteej Sheth[†]

Abstract

We present a sublinear time algorithm for computing a near optimal low-rank approximation to any positive semidefinite (PSD) Toeplitz matrix $T \in \mathbb{R}^{d\times d}$, given noisy access to its entries. In particular, given entrywise query access to $T+E$ for an arbitrary noise matrix $E \in \mathbb{R}^{d\times d}$, integer rank $k \leq d$, and error parameter $\delta > 0$, our algorithm runs in time $\mathsf{poly}(k, \log(d/\delta))$ and outputs (in factored form) a Toeplitz matrix $\widetilde{T} \in \mathbb{R}^{d\times d}$ with rank $\mathsf{poly}(k, \log(d/\delta))$ satisfying, for some fixed constant C,

$$\|T - \widetilde{T}\|_F \leq C \cdot \max\{\|E\|_F, \|T - T_k\|_F\} + \delta \cdot \|T\|_F.$$

Here $\|\cdot\|_F$ is the Frobenius norm and T_k is the best (not necessarily Toeplitz) rank-k approximation to T in the Frobenius norm, given by projecting T onto its top k eigenvectors.

Our robust low-rank approximation primitive can be applied in several settings. When $E = 0$, we obtain the first sublinear time near-relative-error low-rank approximation algorithm for PSD Toeplitz matrices, resolving the main open problem of Kapralov et al. SODA '23, which gave an algorithm with sublinear query complexity but exponential runtime. Our algorithm can also be applied to approximate the unknown Toeplitz covariance matrix of a multivariate Gaussian distribution, given sample access to this distribution. By doing so, we resolve an open question of Eldar et al. SODA '20, improving the state-of-the-art error bounds and achieving a polynomial rather than exponential (in the sample size) runtime.

Our algorithm is based on applying sparse Fourier transform techniques to recover a low-rank Toeplitz matrix using its Fourier structure. Our key technical contribution is the first polynomial time algorithm for *discrete time off-grid* sparse Fourier recovery, which may be of independent interest. We also contribute a structural heavy-light decomposition result for PSD Toeplitz matrices, which allows us to apply this primitive to low-rank Toeplitz matrix recovery.

1 Introduction

A Toeplitz matrix $T \in \mathbb{R}^{d\times d}$ is constant along each of its diagonals. I.e., $T_{i,j} = T_{k,l}$ for all i, j, k, l with $i-j = k-l$. These highly structured matrices arise in many fields, including signal processing, scientific computing, control theory, approximation theory, and machine learning – see [9] for a survey. In particular, Toeplitz matrices often arise as the covariance matrices of stationary signals, when the covariance structure is shift invariant. I.e., when the covariance between measurements only depends on their distance in space or time [25]. A row-reversed Toeplitz matrix is known as a Hankel matrix. Such matrices also find broad applications [20, 22, 42].

Given their importance, significant work has studied fast algorithms for basic linear algebraic tasks on Toeplitz matrices. A $d \times d$ Toeplitz matrix can be multiplied by a vector in just $O(d \log d)$ time using the fast Fourier transform. Toeplitz linear systems can be solved in $O(d^2)$ time exactly using Levinson recursion [24], and to high-precision in $O(d \cdot \operatorname{polylog} d)$ time using randomization [56, 57]. A full eigendecomposition of a symmetrix Toeplitz matrix can be computed in $O(d^2 \cdot \operatorname{polylog} d)$ time [45].

1.1 Sublinear query algorithms for Toeplitz matrices.

Recent work has focused on algorithms for Toeplitz matrices with complexity scaling *sublinearly* in the dimension d [1, 12, 19, 34, 39, 47, 55]. Kapralov et al. [34] study low-rank approximation of positive semidefinite (PSD) Toeplitz matrices. They show that by accessing just $\mathsf{poly}(k, \log(d/\delta), 1/\epsilon)$ entries of a PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$, one can compute (in factored form) a symmetric Toeplitz matrix $\widetilde{T}$ with rank $\tilde{O}(k \log(1/\delta)/\epsilon)$ such that:[1]

$$\|T - \widetilde{T}\|_F \leq (1+\epsilon)\|T - T_k\|_F + \delta\|T\|_F, \tag{1.1}$$

[*] Manning College of Information and Computer Sciences, University of Massachusetts Amherst.

[†] School of Computer and Communication Sciences, EPFL.

[1] Throughout we use $\tilde{O}(\cdot)$ to hide polylogarithmic factors in the dimension d and in the argument.

where $\|M\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d M_{ij}^2}$ is the Frobenius norm and $T_k = \underset{B:\operatorname{rank}(B)\le k}{\arg\min} \|T - B\|_F$ is the best rank-k approximation to T in the Frobenius norm, given by projecting T_k onto its top k eigenvectors.

Observe that T_k may not itself be Toeplitz and it is not a priori clear that Toeplitz $\widetilde{T}$ satisfying (1.1) even exists. The key technical contribution of [34] is to prove that it does, and further that $\widetilde{T}$ can be recovered using sample efficient off-grid sparse Fourier transform techniques. Unfortunately, despite its sample efficiency, the algorithm of [34] is computationally inefficient, with runtime that is exponential in $\tilde{O}(k \operatorname{polylog}(d)/\epsilon)$ – i.e., at least $d^{\tilde{O}(k/\epsilon)}$. The main open question left by their work is if this runtime can be improved, giving a sublinear time, not just a sublinear query, algorithm.

Eldar et al. [19] study the related problem of recovering an (approximately) low-rank PSD Toeplitz covariance matrix $T \in \mathbb{R}^{d\times d}$ given independent samples from the d-dimensional Gaussian $\mathcal{N}(0, T)$. Low-rank Toeplitz covariance estimation is widely applied in signal processing, including in direction of arrival (DOA) estimation [36], spectrum sensing for cognitive radio [14,41], and medical and radar image processing [3,8,16,21,49,52]. Motivated by these applications, Eldar et al. [19] consider two relevant sample complexity measures: the vector sample complexity (VSC), which is the number of samples (each a vector in $\mathbb{R}^d$) taken from the distribution $\mathcal{N}(0, T)$, and the entry sample complexity (ESC), which is the number of entries read from each vector sample. They show that with $\mathsf{poly}(k, \log d, 1/\epsilon)$ VSC and ESC, it is possible to return $\widetilde{T}$ satisfying with high probability:[2]

$$\|T - \widetilde{T}\|_2 \lesssim \sqrt{\|T - T_k\|_2 \cdot \operatorname{tr}(T) + \frac{\|T - T_k\|_F \cdot \operatorname{tr}(T)}{k}} + \epsilon\|T\|_2, \tag{1.2}$$

where $\operatorname{tr}(T)$ is the trace and $\|T\|_2$ is the spectral norm. While the above error bound may seem non-standard, Eldar et al. show that it implies fairly strong bounds when T has low stable-rank. Observe that when T is exactly rank-k, the error bound becomes $\epsilon\|T\|_2$. As in [34], a key drawback is that algorithm of [19] has runtime scaling exponentially in $\tilde{O}(k \operatorname{polylog} d)$. Further, the output matrix $\widetilde{T}$ is not itself guaranteed to be low-rank.

1.2 Our contributions.

In this work, we give a *sublinear time* algorithm for computing a near-optimal low-rank approximation of a PSD Toeplitz matrix given noisy access to its entries. Our robust low-rank approximation primitive can be applied in the settings of both [34] and [19], yielding the first sublinear time algorithms for standard PSD Toeplitz low-rank approximation and low-rank Toeplitz covariance estimation. Our main result is:

Theorem 1.1. (Robust Sublinear Time Toeplitz Low-Rank Approximation) *Let $T \in \mathbb{R}^{d\times d}$ be a PSD Toeplitz matrix, $E \in \mathbb{R}^{d\times d}$ be an arbitrary noise matrix, $\delta > 0$ be an error parameter, and k be an integer rank parameter. There exists an algorithm that, given query access to the entries of $T + E$, runs in $\mathsf{poly}(k, \log(d/\delta))$ time and outputs a representation of symmetric Toeplitz matrix $\widetilde{T}$ with rank $\mathsf{poly}(k, \log(d/\delta))$ that satisfies, with probability at least* 0.9,

$$\|T - \widetilde{T}\|_F \lesssim \max\{\|E\|_F, \|T - T_k\|_F\} + \delta\|T\|_F,$$

where $T_k = \arg\min_{B:\operatorname{rank}(B)\le k} \|T - B\|_F$ is the best rank-k approximation to T.

Observe that given our runtime, which depends just poly-logarithmically on the input dimension d, it is not possible to output $\widetilde{T} \in \mathbb{R}^{d\times d}$ explicitly. Thus, our algorithm outputs a compressed representation of $\widetilde{T}$. To do so, we use the well known Vandermonde decomposition theorem, which states that any Toeplitz matrix can be diagonalized by a Fourier matrix [17]. We can show that $\widetilde{T}$ in particular, which has rank $r = \mathsf{poly}(k, \log(d/\delta))$, can be written as FDF^* where $F \in \mathbb{C}^{d\times r}$ is a Fourier matrix, with j^{th} column given by $[1, e^{2\pi i\cdot f_j}, e^{2\pi i\cdot 2f_j}, \ldots, e^{2\pi i\cdot(d-1)f_j}]$ for some frequency f_j, and $D \in \mathbb{R}^{r\times r}$ is diagonal. Our algorithm outputs the frequencies $f_1, \ldots, f_r$ and diagonal entries $D_{1,1}, \ldots, D_{r,r}$, which fully determine $\widetilde{T}$.

As an immediate consequence of Theorem 1.1 applied with $E = 0$, we obtain the following sublinear time constant factor low-rank approximation algorithm for PSD Toeplitz matrices, which resolves the main open problem of [34] for the case of constant factor approximation.

Theorem 1.2. (Sublinear Time Toeplitz Low-Rank Approximation) *Let $T \in \mathbb{R}^{d\times d}$ be a PSD Toeplitz matrix, k be an integer rank parameter, and $\delta > 0$ be an error parameter. There exists an algorithm that, given*

[2] Throughout, we let $f(\cdot) \lesssim g(\cdot)$ denote that $f(\cdot) \le c \cdot g(\cdot)$ for some fixed constant c.

query access to entries of T, runs in $\mathsf{poly}(k, \log(d/\delta))$ time and outputs a representation of symmetric Toeplitz matrix $\widetilde{T}$ with rank $\mathsf{poly}(k, \log(d/\delta))$ that satisfies with probability at least 0.9,

$$\|T - \widetilde{T}\|_F \lesssim \|T - T_k\|_F + \delta\|T\|_F.$$

We can further apply Theorem 1.1 to the Toeplitz covariance estimation setting of [19]. Here, we set $E = \frac{1}{s} \cdot XX^T - T$ where $X \in \mathbb{R}^{d \times s}$ has s columns sampled i.i.d. from the Gaussian distribution $\mathcal{N}(0, T)$. That is, E is the error between the true covariance matrix T and a sample covariance matrix $\frac{1}{s}XX^T$ that our algorithm can access. Using similar ideas to [19], we show that for $s = \tilde{O}(k^4/\epsilon^2)$, $\|E\|_F \lesssim \sqrt{\|T - T_k\|_2 \operatorname{tr}(T) + \frac{\|T-T_k\|_F \operatorname{tr}(T)}{k}} + \epsilon\|T\|_2$. Combined with Theorem 1.1 applied with $\delta = \epsilon/\sqrt{d}$ so that $\delta\|T\|_F \le \epsilon\|T\|_2$, this gives:

Theorem 1.3. (Sublinear Time Toeplitz Covariance Matrix Estimation) *Let $T \in \mathbb{R}^{d \times d}$ be a PSD Toeplitz matrix, k be an integer rank parameter, and $\epsilon > 0$ be an error parameter. There is an algorithm that given i.i.d. samples $x^1, \ldots, x^s \sim \mathcal{N}(0, T)$ for $s = \tilde{O}(k^4/\epsilon^2)$, runs in $\mathsf{poly}(k, \log(d/\epsilon), 1/\epsilon)$ time and outputs a representation of symmetric Toeplitz $\widetilde{T}$with rank $\mathsf{poly}(k, \log(d/\epsilon))$ satisfying, with probability at least 0.9,*

$$\|T - \widetilde{T}\|_F \lesssim \sqrt{\|T - T_k\|_2 \operatorname{tr}(T) + \frac{\|T - T_k\|_F \operatorname{tr}(T)}{k}} + \epsilon\|T\|_2.$$

Further, the algorithm has entry sample complexity (ESC) $\mathsf{poly}(k, \log(d/\epsilon))$ – it reads just $\mathsf{poly}(k, \log(d/\epsilon))$ entries of each vector sample x^i.

Observe that the error guarantee of Theorem 1.3 is at least as strong, and can potentially be much stronger, than that of (1.2), since the bound is on $\|T - \widetilde{T}\|_F$ rather than $\|T - \widetilde{T}\|_2$. At the same time, our algorithm achieves the same vector and entry sample complexities as [19] up to $\mathsf{poly}(k, \log(d/\epsilon), 1/\epsilon)$ factors, while running in sublinear time. Moreover, our output $\widetilde{T}$ is guaranteed to be low-rank, while the output of [19] is not.

2 Technical overview

In this section, we sketch the ideas behind the proof of our main result, the sublinear time robust Toeplitz low-rank approximation algorithm of Theorem 1.1.

2.1 Recovering Low-Rank Toeplitz Matrices using Fourier Structure. Our starting point is the main result of [34], which shows that any PSD Toeplitz matrix has a near optimal low-rank approximation $\widetilde{T}$ which is itself Toeplitz. Armed with this existence result, the key idea behind the algorithm of [34] is to leverage the well-known sparse Fourier structure of low-rank Toeplitz matrices [17] to recover $\widetilde{T}$ using a sample efficient (but computationally inefficient) algorithm. The idea behind [19] is similar, except that they rely on a weaker existence statement, where the low-rank approximation $\widetilde{T}$ is guaranteed to have sparse Fourier structure, but not necessarily be Toeplitz. Formally, after defining the notion of a Fourier matrix, we state the main result of [34].

Definition 2.1. (Fourier matrix) *For any set of frequencies $S = \{f_1, f_2, \ldots, f_s\} \subset [0, 1]$, let the Fourier matrix $F_S \in \mathbb{C}^{d \times s}$ have j^{th} column equal to $v(f_j) := [1, e^{2\pi i f}, e^{2\pi i(2f)} \ldots, e^{2\pi i(d-1)f}]$.*

Theorem 2.1. (Theorem 2 of [34]) *For any PSD Toeplitz matrix $T \in \mathbb{R}^{d \times d}$, $0 < \epsilon, \delta < 1$, and $k \le d$, there exists symmetric Toeplitz $\widetilde{T}$ with rank $r = \tilde{O}(k \log(1/\delta)/\epsilon))$ such that*

$$\|T - \widetilde{T}\|_F \le (1 + \epsilon)\|T - T_k\|_F + \delta\|T\|_F.$$

Further, $\widetilde{T}$ can be written as $\widetilde{T} = F_S D F_S^*$ where $F_S \in \mathbb{C}^{d \times r}$ is a Fourier matrix (Def. 2.1) and $D \in \mathbb{R}^{r \times r}$ is diagonal.[3]

[3] Our algorithms will rely on several other properties of the frequency set S and diagonal matrix D guaranteed to exist by Theorem 2.1. See Section 3 for a more complete statement of the theorem, which details these properties.

The first step in our proof of Theorem 1.1 is to apply Theorem 2.1 to the input PSD Toeplitz matrix T with $\epsilon = \Theta(1)$. Our algorithm is given access to the entries of the noisy matrix $T + E$, which we will write as $\widetilde{T} + \widetilde{E}$ for $\widetilde{E} = E + T - \widetilde{T}$. Observe that by triangle inequality, since $\|T - \widetilde{T}\|_F \lesssim \|T - T_k\|_F + \delta\|T\|_F$,

$$\|\widetilde{E}\|_F \lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F \lesssim \max\{\|E\|_F, \|T - T_k\|_F\} + \delta\|T\|_F.$$

Thus, to prove Theorem 1.1, it suffices to show that we can recover $\widetilde{T}$ to within a constant factor of the noise level $\|\widetilde{E}\|_F$. Our key contribution is to improve the runtime of this step from being exponential in the rank $r = \tilde{O}(k \log(1/\delta))$ to being polynomial, and hence sublinear in d (recall that throughout $\tilde{O}(\cdot)$ hides poly logarithmic factors in d and the argument).

To approximately recover $\widetilde{T}$, as in [34] and [19], we will leverage its Fourier structure. In particular, Theorem 2.1 guarantees that $\widetilde{T}$ can be written as $\widetilde{T} = F_S D F_S^*$, where $F_S \in \mathbb{C}^{d \times r}$ is a Fourier matrix (Def. 2.1) and $D \in \mathbb{R}^{r \times r}$ is diagonal. To find $\widetilde{T}$, the algorithm of [34] brute force searches for a good set of frequencies S and corresponding diagonal matrix D. In particular, they consider a large pool of candidate frequency sets drawn from a net over $[0, 1]^r$. For each set S, they solve a regression problem to find a diagonal matrix D that (approximately) minimizes $\|T - F_S D F_S^*\|_F$. These regression problems can be solved using a sublinear number of queries to T using leverage score based random sampling [50, 54]. One can then return the best approximation given by any of the candidate frequency sets as the final output.

Unfortunately, the above approach incurs exponential runtime, as the number of frequency sets considered grows exponentially in the rank $r = \tilde{O}(k \log(1/\delta))$. A similar issue arises in [19], where a brute force search over frequency sets is also performed, but with a different regression step, where D is relaxed to be any $r \times r$ matrix.

2.2 From Column Recovery to Matrix Recovery. To avoid the exponential runtime, of [34] and [19], we apply tools from the extensive literature on efficient recovery of Fourier-sparse functions [7, 11, 26, 28, 29, 32, 33]. Observe that if we expand out the decomposition $\widetilde{T} = F_S D F_S^*$ from Theorem 2.1, each column of $\widetilde{T}$ is an r-Fourier sparse function from $\{0, \dots, d-1\} \to \mathbb{R}$. In particular, letting $\widetilde{T}_j$ denote the j^{th} column, for any $t \in \{0, \dots, d-1\}$, letting $\{a_f\}_f \in S$ be the diagonal entries of D,

$$\widetilde{T}_j(t) = \sum_{f \in S} a_f e^{2\pi i f (t-j)}.$$

Our algorithm will recover a Fourier-sparse approximation to a single column of $\widetilde{T}$ using samples of $T + E = \widetilde{T} + \widetilde{E}$ in just $\mathsf{poly}(r, \log d)$ time using a sparse Fourier transform algorithm. This Fourier-sparse approximation will give us approximations to the frequencies in F_S, which in turn can be used to form an approximation to $\widetilde{T}$. This approach presents several challenges, which we discuss below.

Column Signal-to-Noise ratio. First, we must ensure that the column that we apply sparse Fourier transform to does not have too much noise on it, compared to its norm. To do so, we sample a column $\widetilde{T}_j$ uniformly at random. By Markov's inequality, $\widetilde{T}_j$ is corrupted with noise whose ℓ_2 norm is bounded as $\|\widetilde{E}_j\|_2 \lesssim \|\widetilde{E}\|_F/\sqrt{d}$ with good probability. Further, using that $\widetilde{T}$ is symmetric Toeplitz and that it is close to T and hence nearly PSD, we can apply a norm compression inequality of Audenaert [4] (see Claim 5.2) to show that each column of $\widetilde{T}$ must have relatively large norm. In particular, for any j we show that $\|\widetilde{T}_j\|_2 \gtrsim \|\widetilde{T}\|_F/\sqrt{d}$ (see Lemma 5.3). In combination, these facts ensure that a random column $\widetilde{T}_j$ has a similar signal-to-noise ratio as the full matrix $\widetilde{T}$ with good probability. Thus, we can expect to recover a good approximation to $\widetilde{T}$ from just this column.

Heavy Frequency Recovery. Of course, given the noise $\widetilde{E}$, we cannot expect to recover approximations to all frequencies in S given sample access to $\widetilde{T} + \widetilde{E}$. For example, if a frequency f corresponds to a very small entry a_f in D, we may not recover it. We must argue that omitting such frequencies from our approximation of $\widetilde{T}$ does not introduce significant error. More formally, our sparse Fourier transform will recover a set of frequencies that well approximates some subset of the input frequencies $S_{heavy} \subset S$, but may not approximate the set frequencies $S_{light} := S \setminus S_{heavy}$. The algorithm guarantees that S_{heavy} suffices to approximate our random column $\widetilde{T}_j$ up to the noise level $\|\widetilde{E}_j\|_2$. Let $Z_{heavy} : \{0, \dots, d-1\} \to \mathbb{R}$ be given by restricting $\widetilde{T}_j$ to the frequencies in S_{heavy}, i.e., $Z_{heavy}(t) = \sum_{f \in S_{heavy}} a_f e^{2\pi i f t}$. Let $Z_{light} = T_j - Z_{heavy}$. Then we can show that

$$\|Z_{light}\|_2 = \|\widetilde{T}_j - Z\|_2 \lesssim \|\widetilde{E}_j\|_2 + \delta\|\widetilde{T}_j\|_2.$$

We have ensured that $\|\widetilde{E}_j\|_2 \lesssim \|\widetilde{E}\|_F/\sqrt{d}$ by choosing a random column. Further, by applying our recovery algorithm with error $\delta' = \delta/\sqrt{d}$, which introduces only additional $\log d$ dependences, we have that $\delta'\|\widetilde{T}_j\|_2 \leq \delta\|\widetilde{T}\|_F/\sqrt{d}$. Thus, we have overall that $\|Z_{light}\|_2 \lesssim \frac{\|\widetilde{E}\|_F + \delta\|\widetilde{T}\|_F}{\sqrt{d}}$.

It remains to argue that this error being small on just column $\widetilde{T}_j$ ensures that it is small on the full matrix. In particular, if we let $\widetilde{T}_{light}$ denote our full Toeplitz matrix restricted to the unrecovered frequencies, we must show that

$$\|\widetilde{T}_{light}\|_F \lesssim \sqrt{d} \cdot \|Z_{light}\|_2 \lesssim \|\widetilde{E}\|_F + \delta\|\widetilde{T}\|_F.$$

Again, we show this by arguing that $\widetilde{T}_{light}$ is itself near PSD and applying the norm compression inequality of [4]. We interpret the existence of $\widetilde{T}_{light}$ and $\widetilde{T}_{heavy} := \widetilde{T} - \widetilde{T}_{light}$ as a heavy-light decomposition of $\widetilde{T}$ into $\widetilde{T} = \widetilde{T}_{heavy} + \widetilde{T}_{light}$, see Section 5.1 for the detailed proof.

Sublinear time Approximate Regression. With the above bound in hand, the proof of Theorem 1.1 is essentially complete, in the last step we use a leverage score based fast approximate regression primitive [50, 54] to regress onto the approximate frequencies that we recover to give an approximation to the full $\widetilde{T}$. We have argued that the noise incurred by not recovering all frequencies is bounded by $\|\widetilde{E}\|_F + \delta\|\widetilde{T}\|_F$, and this bounds the optimum of the regression problem, which we solve approximately, giving the final guarantee of Theorem 1.1. To solve the regression problem efficiently, we crucially rely on the special structure of the Fourier spectrum S of $\widetilde{T}$, stated in the full version of Theorem 2.1 in Section 3.4. Full details of this step are presented in Section 5.2.

2.3 Sublinear Time Discrete-Time Off-Grid Sparse Fourier Transform.

We have argued that applying a sparse Fourier transform to a random column of $\widetilde{T}$ suffices to recover a good approximation to the full matrix. It remains to show that we can implement such a sparse Fourier transform efficiently. We sketch the key ideas behind this efficient sparse Fourier transform here, with the complete proof appearing in Section 4.

If the frequencies in the set S guaranteed to exist by Theorem 2.1 were integer multiples of $1/d$, i.e., they were "on-grid", then we could apply a discrete sparse Fourier transform algorithm to recover a Fourier-sparse approximation to $\widetilde{T}_j$ given samples from the column $[T + E]_j = [\widetilde{T} + \widetilde{E}]_j$. Such algorithms have been studied extensively [23, 26, 27]. The main difficulty in our case is that the frequencies in S may be "off-grid", i.e., arbitrary real numbers in $[0, 1]$. Recent work of [11] solves the off-grid sparse Fourier transform problem given sample access to the function $[\widetilde{T} + \widetilde{E}]_j$ on the *continuous range* $[0, d]$. However, in our setting, we can only access $[\widetilde{T} + \widetilde{E}]_j$ at the integer points $\{0, 1, \ldots, d-1\}$ – i.e., at the entries in the j^{th} column of our matrix. Thus, we must modify the approach of [11] to show that, nevertheless, we can well approximate $\widetilde{T}_j$ from samples at these points. In doing so, we give to the best of our knowledge, the first efficient *discrete time off-grid* sparse Fourier transform algorithm, which may be applicable in other settings that involve off-grid frequencies but on grid time samples.

Informally, we consider the following sparse Fourier recovery problem:

Problem 1. (Off-Grid Sparse DTFT – Informal) *Let $x^*(t) = \sum_{f \in S} a_f \cdot e^{2\pi i f t}$ for $t \in [d] := \{0, 1, \ldots, d-1\}$ and $S \subset [0, 1]$ with $|S| = k$. Let $x(t) = x^*(t) + g(t)$, where g is arbitrary noise. Given sample access to $x(t)$ for any $t \in [d]$, approximately recover all $f \in S$ that contribute significantly to x.*

In principal, without noise, one can recover x^* from $2k$ samples at any time domain points – including at the integers in $[d]$ – via Prony's method [18]. The key challenge is to recover x^* approximately in the presence of noise. Observe that with noise, it is generally not possible to identify all frequencies in S given sample access to $x(t)$ only on $[d]$. For example, if two frequencies f and f' are extremely close to each other, their contributions to the function x on $[d]$ could nearly cancel out, making it impossible to identify them. This is true even if the coefficients a_f and $a_{f'}$ are arbitrarily large.

Thus, we will settle for approximate frequency recovery. In particular, we will output a list of frequencies L such that it well approximates a subset of frequencies $S_{heavy} \subseteq S$. Denoting for any function $f : [d] \to \mathbb{C}$ let $\|f\|_d^2 = \sum_{i \in [d]} |f(i)|^2$, this subset S_{heavy} spans an approximation to the Fourier sparse function x^* up to error $\lesssim \|g\|_d^2 + \delta\|x\|_d^2$. To find this approximation algorithmically, we will have to regress onto the list of frequencies L, and the final approximation itself will use $\mathsf{poly}(k, \log(d/\delta))$ frequencies. In our Toeplitz setting (Theorem 1.1), this translates to $\mathsf{poly}(k, \log(d/\delta))$ rank of the final Toeplitz matrix that we output. We state our main approximate frequency recovery primitive below. See Section 4 for a more formal statement.

Lemma 2.1. (Approximate Frequency Recovery – Informal) *Consider the setting of Problem 1. Assume that $\|g\|_d^2 \le c\|x^*\|_d^2$ for a small absolute constant c. Then for any given $\delta > 0$, there exists an algorithm that in time and sample complexity $\mathsf{poly}(k, \log(d/\delta))$ outputs a list L of $\mathsf{poly}(k, \log(d/\delta))$ frequencies in $[0,1]$ such that, with probability at least 0.99, letting*

$$S_{heavy} = \left\{ f \in S : \exists f' \in L \ s.t.\ |f - f'|_\circ \le \frac{\mathsf{poly}(k, \log(d/\delta))}{d} \right\},$$

and letting $x^*_{S_{heavy}}(t) := \sum_{f \in S_{heavy}} a_f e^{2\pi i f t}$, we have

$$\|x^* - x^*_{S_{heavy}}\|_d^2 \lesssim \|g\|_d^2 + \delta\|x^*\|_d^2.$$

Here $|\cdot|_\circ$ is the wrap-around distance. For any $f_1, f_2 \in [0,1]$, $|f_1 - f_2|_\circ := \min\{|f_1 - f_2|, |1 - (f_1 - f_2)|\}$.

Lemma 2.1 mirrors Lemma 7.23 in [11], but only requires accessing $x(t)$ at integer $t \in [d]$ rather than at real $t \in [0,d]$. We now discuss how we adapt the approach of [11] to work in this more restricted sampling setting.

One cluster recovery. The approach of [11] first considers the *one cluster* case (see Section 4.1 for a formal definition), where most of the energy of x is concentrated around a single frequency f_0. They show how to approximately recover this central frequency of the cluster up to $\mathsf{poly}(k, \log(d/\delta))/d$ error in $\mathsf{poly}(k, \log(d/\delta))$ time. Roughly, since the signal is close to a pure frequency, by considering the ratio $x(\alpha+\beta)/x(\alpha)$ for carefully chosen sample points α, β, they show that one can approximate $e^{2\pi i f_0 \beta}$, and in turn f_0. In their work, α and β may be real valued. We need to modify the approach to restrict them to be integers. We formally do this in Section 4.1. The key idea follows similar ideas to our other modifications of their approach, discussed below.

Multi-cluster recovery with bounded support. After handling the single cluster case, [11] reduces the general case to it via hashing techniques. In particular, by applying an efficient transformation in time domain, they access a signal whose frequencies are hashed versions of the frequencies in S. This approach is standard in the literature on sparse Fourier transform. The hash function spreads out the frequencies in S so that they lie in different frequency ranges (called 'buckets') with good probability, and can be recovered using the single cluster recovery primitive. The hash function used in [11] is as follows: let $\sigma \in \mathbb{R}, b \in [0,1]$ and B be the number of buckets. Then define:

$$\begin{aligned} \pi_{\sigma,b}(f) &= B\sigma(f-b) \mod B, \\ h_{\sigma,b}(f) &= \mathrm{round}(\pi_{\sigma,b}(f)), \end{aligned} \tag{2.3}$$

where the round(.) function rounds real numbers to the nearest integer. The crucial claim that [11] shows is that if we let σ be a uniformly random real number in $\left[\frac{d}{\mathsf{poly}(k,\log(d/\delta))}, \frac{2d}{\mathsf{poly}(k,\log(d/\delta))}\right]$ then we have that for any f_1, f_2 such that $|f_1 - f_2| \ge \frac{\mathsf{poly}(k,\log(d/\delta))}{d}$,

$$\Pr[h_{\sigma,b}(f_1) = h_{\sigma,b}(f_2)] = \Pr[B\sigma|f_1 - f_2| \in (-1,1) \mod B] \lesssim 1/B. \tag{2.4}$$

This ensures that f_1, f_2 land in different hash buckets with good probability. I.e., after hashing, these frequencies can be separated in the Fourier domain. In particular, [11] applies an efficient filtering approach (see Lemma 4.7) to isolate a cluster of frequencies of width $\mathsf{poly}(k, \log(d/\delta))/d$ around each frequency. If the hash bucket containing this frequency has high SNR, then the frequency can be recovered via single-cluster recovery.

To implement the above approach in our setting, where we can only access the input at integer time points, we need the random seed σ to be a random *integer*. In particular, we let σ to be a uniformly random integer in $\left[\frac{d}{\mathsf{poly}(k,\log(d/\delta))}, \frac{2d}{\mathsf{poly}(k,\log(d/\delta))}\right]$. However this restriction severely affects the collision behavior of the hash function $h_{\sigma,b}$. For example, if $|f_1 - f_2|_\circ = 1/2$, then $B\sigma|f_1 - f_2|_\circ \mod B = 0$ for *every even* σ and thus (2.4) cannot hold with any probability less than 1/2.

To handle this issue, we observe that if an instance only contained frequencies within an interval of width $1/B$ then we would have $|f_1 - f_2|_\circ \le 1/B$ for any f_1, f_2 in the function's support. Then, if σ is a uniformly random integer in $\left[\frac{d}{\mathsf{poly}(k,\log(d/\delta))}, \frac{2d}{\mathsf{poly}(k,\log(d/\delta))}\right]$, $B\sigma|f_1 - f_2|_\circ$ is uniformly distributed on a grid of spacing at most $B|f_1 - f_2|_\circ \le 1$. Thus for every σ such that $B\sigma|f_1 - f_2|_\circ \in (-1, 1) \mod B$, there are roughly at least B other values of σ for which $B\sigma|f_1 - f_2|_\circ \notin (-1, 1) \mod B$. This is enough to show that (2.4) holds.

General multi-cluster recovery. Of course, a general input instance may have frequencies that do not lie in an interval of width $1/B$. To reduce to this setting, we apply an additional filtering step in the Fourier domain. Our filter splits the interval $[0, 1]$ into B intervals of width $1/B$, based on the same filtering approach of Lemma 4.7 as discussed before. For the complete proof, see Section 4.2. This completes our proof sketch for our discrete time off-grid sparse Fourier transform primitive (Lemma 2.1).

2.4 Other related work.

A large body of work in numerical linear algebra, applied mathematics, theoretical computer science, and signal processing has studied the problem of computing low-rank approximations of Toeplitz and Hankel matrices [10, 31, 37, 40, 44, 46, 51]. In many applied signal processing settings, Toeplitz matrices arise as PSD covariance matrices, motivating our focus on the PSD case. Significant prior work has also studied the problem of computing an optimal *structure-preserving* Toeplitz low-rank approximation, where the low-rank approximation $\widetilde{T}$ is itself required to be Toeplitz. However, no simple characterization of the optimal solution to this problem is known [13] and polynomial time algorithms are only known in the special cases of $k = 1$ and $k = d - 1$ [13, 35]. On the practical side a range of heuristics are known, based on techniques such as convex relaxation [10, 20, 44], alternating minimization [13, 53], and sparse Fourier transform [37].

Many results in the signal processing literature study sublinear query algorithms for Toeplitz matrices, these are referred to as *sparse array methods* which proceed by querying a small principal submatrix to obtain an approximation to the whole matrix [1, 12, 39, 47]. The work of [19] and [34] on sublinear query PSD Toeplitz low-rank approximation and Toeplitz covariance estimation is closely related to this literature.

Beyond Toeplitz matrices, significant work has focused on sublinear time low-rank approximation algorithms for other structured matrix classes. This includes positive semidefinite matrices [5, 43], distance matrices [6, 30], and kernel matrices [2, 43, 58].

2.5 Roadmap.

The remainder of the paper is organized as follows. In Section 3 we define basic notation and import lemmas and theorems from prior work that will be used in our proofs. In Section 4 we give the analysis of our discrete time off-grid sparse Fourier transform. In Section 5 we apply this sparse Fourier transform to the low-rank Toeplitz recovery problem, proving our main result, Theorem 1.1. In Section 5.3, we show how to instantiate this general robust low-rank approximation result to obtain our sublinear time Toeplitz low-rank approximation and covariance estimation results (Theorems 1.2 and 1.3).

3 Notation and preliminaries

In this section, we introduce notation and preliminary concepts that are used throughout this paper.

3.1 General and linear algebraic notation.

Consider functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ for input domain $\mathcal{X}$. We write $f(\cdot) \lesssim g(\cdot)$ if there exists a constant $C > 0$ such that $f(x) \le Cg(x)$ for all $x \in \mathcal{X}$. For any integer $d > 0$, let $[d] = \{0, 1, \ldots, d-1\}$. For any function $x : \mathbb{Z} \to \mathbb{C}$, and integer $d > 0$, we let $\|x\|_d^2 := \sum_{j\in[d]} |x(j)|^2$. For any set N, let N^n denote the set of all subsets of N with n elements.

For a matrix A, let A^T and A^* denote its transpose and Hermitian transpose, respectively. Let $A_{[i,j]}$ denote the i, j entry of A and for $i_1 < j_1, i_2 < j_2$, let $A_{[i_1:j_1, i_2:j_2]}$ denote the submatrix containing entries from rows i_1 to j_1 and columns i_2 to j_2. For any vector $x \in \mathbb{C}^d$, let $\|x\|_2 = \sqrt{x^*x}$ denote its ℓ_2 norm. For a matrix $A \in \mathbb{C}^{d\times d}$, let $\|A\|_2 = \sup_{x\in\mathbb{C}^d} \|Ax\|_2/\|x\|_2$ denote its spectral norm and $\|A\|_F = \sqrt{\sum_{i,j\in[d]} |A_{[i,j]}|^2}$ denote its Frobenius norm.

A Hermitian matrix $A \in \mathbb{C}^{d\times d}$ is positive semidefinite (PSD) if for all $x \in \mathbb{C}^d$, $x^*Ax \ge 0$. Let $\lambda_1(A) \ge \ldots \ge \lambda_d(A) \ge 0$ denote its eigenvalues. Let $\preceq$ denote the Loewner ordering, that is $A \preceq B$ if and only if $B - A$ is PSD. Let $A = U\Sigma V^*$ denote the compact singular value decomposition of A, and when A is PSD note that $U\Sigma U^*$ is its eigenvalue decomposition. In this case, let $A^{1/2} = U\Sigma^{1/2}$ denote its matrix square root, where $\Sigma^{1/2}$ is obtained by taking the elementwise square root of Σ. Let $A_k = U_k\Sigma_k V_k^*$ denote the projection of A onto its top k singular vectors. Here, $\Sigma_k \in \mathbb{R}^{k\times k}$ is the diagonal matrix containing the k largest singular values of A, and $U_k, V_k \in \mathbb{C}^{d\times k}$ denote the corresponding k left and right singular vectors of A. Note that A_k is the optimal rank k approximation to A in the spectral and Frobenius norms. That is, $A_k = \arg\min_{B:\text{rank}(B)\le k} \|A - B\|_2$ and $A_k = \arg\min_{B:\text{rank}(B)\le k} \|A - B\|_F$.

We let $*$ denote the convolution operator – in both the discrete and continous settings. For discrete functions $x, y : \mathbb{Z} \to \mathbb{C}$, we have $[x * y](n) = \sum_{k \in \mathbb{Z}} x(k) y(n-k)$. For continuous functions $x, y : [0,1] \to \mathbb{C}$, we have $[x * y](f) = \int_0^1 x(f') y(f - f') df'$. For any function $f : \mathcal{D} \to \mathbb{C}$ defined over domain $\mathcal{D}$ (e.g., $\mathcal{D}$ can be $\mathbb{Z}, \mathbb{R}$ etc.), we let *supp*(f) denote its support. That is, *supp*$(f) = \{x \in \mathcal{D} : |f(x)| > 0\}$.

3.2 Fourier analytic notation. Throughout, we use the standard discrete-time Fourier transform.

Definition 3.1. (Discrete-time Fourier transform (DTFT)) *For $x : \mathbb{Z} \to \mathbb{C}$, its DTFT $\widehat{x}(f) : [0,1] \to \mathbb{C}$ is defined as*

$$\widehat{x}(f) = \sum_{n \in \mathbb{Z}} x(n) e^{-2\pi i f n}.$$

The inverse DTFT is given by:

$$x(n) = \int_0^1 \widehat{x}(f) e^{2\pi i f n} \, \mathrm{d}f.$$

The discrete time version of Parseval's theorem allows us to relate the energy in time and Fourier domains.

Lemma 3.1. (Parseval's identity) *For $x : \mathbb{Z} \to \mathbb{C}$ with DTFT $\widehat{x} : [0,1] \to \mathbb{C}$, we have:*

$$\sum_{t=-\infty}^{\infty} |x(t)|^2 = \int_0^1 |\widehat{x}(f)|^2 df.$$

We define the wrap around distance between two frequencies as follows.

Definition 3.2. (Wrap around distance) *For any $f_1, f_2 \in [0,1]$, let $|f_1 - f_2|_\circ = \min\{|f_1 - f_2|, |1 - (f_1 - f_2)|\}$.*

3.3 Compact filter functions. Our algorithm will use a function $H : \mathbb{Z} \to \mathbb{R}$ which approximates the indicator function of an interval $[d]$ by preserving energies of k-Fourier sparse functions on this interval and almost killing off their energy outside this interval. At the same time, the support of this function in the Fourier domain is compact. Formally:

Lemma 3.2. (Discrete version of Lemma 6.6 of [11]) *Given positive integers d, k and real $\delta > 0$, let $s_0 = \mathsf{poly}(k, \log(d/\delta)), s_1 = \mathsf{poly}(k, \log(d/\delta)), s_3 = 1 - 1/\mathsf{poly}(k, \log(d/\delta))$, and $l = \Theta(k \log(k/\delta))$, there is a function $H : \mathbb{Z} \to \mathbb{R}$ with DTFT $\widehat{H} : [0,1] \to \mathbb{R}$ having the following properties:*

$$\text{Property I:} \quad H(t) \in [1-\delta, 1] \quad \forall t \in \mathbb{Z} : |t - d/2| \leq d \left(\frac{1}{2} - \frac{2}{s_1} \right) s_3,$$

$$\text{Property II:} \quad H(t) \in [0,1] \quad \forall t \in \mathbb{Z} : d \left(\frac{1}{2} - \frac{2}{s_1} \right) s_3 \leq |t - d/2| \leq \frac{d}{2} s_3,$$

$$\text{Property III:} \quad H(t) \leq s_0 \cdot \left(s_1 \left(\frac{|t - d/2|}{d s_3} - \frac{1}{2} \right) + 2 \right)^{-l} \quad \forall t \in \mathbb{Z} : |t - d/2| \geq \frac{d}{2} s_3,$$

$$\text{Property IV:} \quad supp(\widehat{H}) \subseteq \left[-\frac{s_1 l}{2 d s_3}, \frac{s_1 l}{2 d s_3} \right] \text{ and so } \Delta_h := |supp(\widehat{H})| = s_1 l / (d s_3) = \mathsf{poly}(k, \log(d/\delta))/d.$$

For any exact k-Fourier-sparse signal $x^*(t)$ we additionally have the following:

$$\text{Property V:} \quad \sum_{t \in \mathbb{Z} \setminus [d]} |x^*(t) \cdot H(t)|^2 \leq \delta \|x^*\|_d^2,$$

$$\text{Property VI:} \quad \|x^* \cdot H\|_d^2 \in [0.99 \|x^*\|_d^2, \|x^*\|_d^2].$$

The first four properties guarantee that H approximates the indicator function of the interval $[d]$ and has a compact support in Fourier domain. The final two properties capture the fact that multiplying H by any k-Fourier sparse function almost kills of its energy outside $[d]$ and almost preserves its energy inside $[d]$. Lemma 3.2 follows by modifying the proof of Lemma 6.6 of [11], which gives an analogous filter function in the continuous domain. We state Lemma 6.6 below, followed by the proof of Lemma 3.2.

Lemma 3.3. (Lemma 6.6 of [11]) *Given positive integers d, k and real $\delta > 0$, let $s_0 = \mathsf{poly}(k, \log(d/\delta)), s_1 = \mathsf{poly}(k, \log(d/\delta)), s_3 = 1 - 1/\mathsf{poly}(k, \log(d/\delta))$, and $l = \Theta(k \log(k/\delta))$, there is a function $H' : \mathbb{R} \to \mathbb{R}$ with continuous Fourier transform $\widehat{H'} : \mathbb{R} \to \mathbb{R}$ having the following properties:*

$$\text{Property I:} \quad H'(t) \in [1-\delta, 1] \quad \forall t \in \mathbb{R} : |t - d/2| \leq d\left(\frac{1}{2} - \frac{2}{s_1}\right) s_3,$$

$$\text{Property II:} \quad H'(t) \in [0, 1] \quad \forall t \in \mathbb{R} : d\left(\frac{1}{2} - \frac{2}{s_1}\right) s_3 \leq |t - d/2| \leq \frac{d}{2} s_3,$$

$$\text{Property III:} \quad H'(t) \leq s_0 \cdot \left(s_1\left(\frac{|t - d/2|}{ds_3} - \frac{1}{2}\right) + 2\right)^{-l} \quad \forall t \in \mathbb{R} : |t - d/2| \geq \frac{d}{2} s_3,$$

$$\text{Property IV:} \quad supp(\widehat{H'}) \subseteq \left[-\frac{s_1 l}{2ds_3}, \frac{s_1 l}{2ds_3}\right] \text{ and so } \Delta_h := |supp(\widehat{H'})| = s_1 l/(ds_3) = \mathsf{poly}(k, \log(d/\delta))/d.$$

Proof. [Proof of Lemma 3.2] Consider the function $H' : \mathbb{R} \to \mathbb{R}$ given in Lemma 3.3. To obtain our function H, we discretize H' by restricting it to the integers. The first three time domain properties of our function thus follow directly from the first three properties of Lemma 3.3. Discretizing H' in time domain to obtain H results in aliasing of $\widehat{H'}$ in Fourier domain to obtain $\widehat{H}$. However, by Property IV of Lemma 3.3, we know that the support of $\widehat{H'}$ is contained in $[-1/2, 1/2]$ assuming $d \gg \mathsf{poly}(k)$. Thus $\widehat{H}(f) = \widehat{H'}(f)$ for all $f \in [-1/2, 1/2]$, thus implying that property 4 of Lemma 3.2 follows from property 4 of Lemma 3.3.

Finally, to prove properties 5 and 6, we must use discrete versions of Lemmas 5.1 and 5.5 of [11], which are used to prove Properties of 5 and 6 of Lemma 6.6 in [11]. The discrete version of Lemma 5.1 was shown in Lemma C.1 of [19], and establishes that for any k-Fourier sparse $x^* : \mathbb{Z} \to \mathbb{C}$, $\forall i \in [d]$, $|x^*(i)|^2 \lesssim k^6 \log^3(k)(\|x^*\|_d^2/d)$. It is also fairly easy to obtain the following bound from inspecting the proof of Lemma 5.5 in [11]: for any k-Fourier sparse $x^* : \mathbb{Z} \to \mathbb{C}$ $\forall i \in \mathbb{Z} \setminus [d]$, $|x^*(i)|^2 \lesssim k^{13}(ki/d)^{2.2k}(\|x^*\|_d^2/d)$. Using these bounds in the proof of Properties 5 and 6 of Lemma 6.6 of [11] and replacing integrals with sums, we obtain Properties 5 and 6 of Lemma 3.2. ∎

We will also need the following filter function from [11], which when convolved with, allows us to access the input signal whose Fourier transform is restricted to a desired interval.

Lemma 3.4. (Lemma 6.7 of [11]) *Given $B > 1$, $\delta, k, w > 0$, let $l = O(k \log(k/\delta))$. Then there exists a function $G : \mathbb{R} \to \mathbb{C}$ with continous Fourier transform $\widehat{G} : \mathbb{R} \to \mathbb{C}$ satisfying the following*

$$\text{Property I:} \quad \widehat{G}(f) \in [1 - \delta/k, 1] \quad if |f| \leq (1-w)/2B.$$

$$\text{Property II:} \quad \widehat{G}(f) \in [0, 1] \quad if (1-w)/2B \leq |f| \leq 1/2B.$$

$$\text{Property III:} \quad \widehat{G}(f) \in [-\delta/k, \delta/k] \quad if |f| \geq 1/2B.$$

$$\text{Property IV:} \quad supp(G(t)) \subset \left[\frac{-lB}{w}, \frac{lB}{w}\right].$$

$$\text{Property V:} \quad \max(G(t)) \lesssim \mathsf{poly}(B, l).$$

3.4 Structure preserving Toeplitz low-rank approximation. We next formally state the main result of [34] which shows that for any PSD Toeplitz matrix, there exists a near optimal low-rank approximation in the Frobenius norm which itself is Toeplitz. We will use this fact in the proof of Theorem 1.1 by interpreting the input PSD Toeplitz matrix as a noisy version of the near optimal Toeplitz low-rank approximation, further corrupted by noise E.

Theorem 2.1. *Given PSD Toeplitz matrix $T \in \mathbb{R}^{d \times d}$, $\epsilon, \delta \in (0, 1)$, and an integer rank $k \leq d$, let $r_1 = \widetilde{O}(k/\epsilon)$ and $r_2 = \widetilde{O}(\log(1/\delta))$. There exists a symmetric Toeplitz matrix $\widetilde{T} = F_S D F_S^*$ of rank $r = 2r_1 r_2 = \widetilde{O}(k \log(1/\delta)/\epsilon)$ such that,*

1. $\|T - \widetilde{T}\|_F \leq (1+\epsilon)\|T - T_k\|_F + \delta\|T\|_F$.

2. *$F_S \in \mathbb{R}^{d\times r}$ and $D \in \mathbb{R}^{r\times r}$ are Fourier (Def. 2.1) and diagonal matrices respectively. The set of frequencies S can be partitioned into r_1 sets $\widetilde{S}_1, \ldots, \widetilde{S}_{r_1}$ where each $\widetilde{S}_i$ is as follows:*

$$\widetilde{S}_i = \bigcup_{1\le j\le r_2} \{f_i + \gamma j, f_i - \gamma j\},$$

where $f_i \in \{1/2d, 3/2d, \ldots, 1-1/2d\}$ for all $i \in [r_1]$, and $\gamma = \delta/(2^{C\log^7 d})$ for fixed constant $C > 0$.

3. *Let $-\widetilde{S}_i = \{1-f | \forall f \in \widetilde{S}_i\}$, then $-\widetilde{S}_i \subseteq S$ for all $i \in [r_1]$.*

4. *Let $S_i = -\widetilde{S}_i \cup \widetilde{S}_i$ and D_i contain the corresponding entries of D for every $i \in [r_1]$. We have that for any subset S' of $\cup_{i\in[r_1]}\{S_i\}$, if we let $\widetilde{T}' = \sum_{S_i\in S'} F_{S_i} D_i F^*_{S_i}$, then there exists a PSD Toeplitz matrix T' such that $\|\widetilde{T}' - T'\|_F \le \delta\|T'\|_F$.*

Points 2,3 and 4 of the previous Lemma provide essential structural properties of the near optimal Toeplitz low-rank approximation crucial for our approach. Point 2 will be used in our sublinear time approximate regression primitive by essentially providing a finite set inside $[0,1]$ where possible frequencies of the near optimal Toeplitz low-rank approximation could lie. Point 3,4 implies in particular that $\widetilde{T}$ is almost PSD, which is an important property used to show that a random column of $\widetilde{T}$ must have similar signal-to-noise ratio as the full matrix and thus a sparse Fourier transform of a random column of $\widetilde{T}$ yields good frequencies for approximating $\widetilde{T}$. Point 4 in fact implies a more fine grained notion, which is useful for our proof that the recovered heavy frequencies suffice to approximate $\widetilde{T}$.

4 Discrete Time Off-Grid Sparse Fourier Recovery

In this section we prove our main discrete time off-grid Fourier recovery result of Lemma 2.1, which was outlined in Section 2.3. We re-state the complete version of the lemma here.

Lemma 2.1. *Consider $x, x^*, g : \mathbb{Z} \to \mathbb{R}$, where $x^*(t) = \sum_{f\in S} a_f e^{2\pi i f t}$ for $S \subseteq [0,1]$ with $|S| \le k$, $g(t)$ is arbitrary noise, and $x(t) = x^*(t) + g(t)$. Assume that we can access $x(t)$ for $t \in [d]$ and that $\|g\|_d^2 \le c\|x^*\|_d^2$ for a small absolute constant $c > 0$. Let $\delta > 0$ be an error parameter, and let $\mathcal{N}^2 = \frac{1}{d}(\|g\|_d^2 + \delta\|x^*\|_d^2)$ be the noise threshold. Let $\Delta = \mathsf{poly}(k, \log(d/\delta))/d$ such that $\Delta \ge k \cdot |supp(\widehat{H}(f))|$, where $H(t)$ (with Fourier transform $\widehat{H}(f)$) is the function guaranteed to exist by Lemma 3.2 for parameters k, δ. Then in time and sample complexity $\mathsf{poly}(k, \log(d/\delta))$ one can find a list L of $\mathsf{poly}(k, \log(d/\delta))$ frequencies that satisfies the following with probability at least 0.99: Let*

$$S_{heavy} = \{f \in S : \exists f' \in L \text{ s.t. } |f - f'|_\circ \lesssim k\Delta\sqrt{k\Delta d}\},$$

and let $x^*_{S_{heavy}}(t) = \sum_{f\in S_{heavy}} a_f e^{2\pi i f t}$. Then

$$\|x^* - x^*_{S_{heavy}}\|_d^2 \lesssim d\mathcal{N}^2.$$

In the subsequent subsections we state intermediary claims and their proofs building up to the proof of Lemma 2.1. In Section 4.1 we first consider the case when all frequencies of the Fourier sparse function x^* are very close to each other – i.e., the one-cluster case. In Section 4.1 we describe how to approximately recover the central frequency of the one cluster. Then in Section 4.2 we reduce the general case to the one-cluster case. In Section 4.2.1, we first give the reduction in the setting when the frequencies lie in a relatively small interval. We call such instances ‘bounded’. We then give a reduction from fully general instances to bounded instances in Section 4.2.3.

4.1 One cluster case. In this section, we consider discrete time signals that are *clustered* in Fourier domain with approximately bounded support in time domain, as formalized in the following definition.

Definition 4.1. ((ϵ, Δ)-one clustered signal) *$z : \mathbb{Z} \to \mathbb{C}$ is (ϵ, Δ)-one clustered around $f_0 \in [0,1]$ if*

$$\text{Property I:} \quad \int_{f_0-\Delta}^{f_0+\Delta} |\widehat{z}(f)|^2 df \ge (1-\epsilon)\int_0^1 |\widehat{z}(f)|^2 df,$$

$$\text{Property II:} \quad \|z\|_d^2 \ge (1-\epsilon)\sum_{t\in\mathbb{Z}} |z(t)|^2.$$

The main result of this subsection is that we can approximately recover the central frequency f_0 of a clustered signal in sublinear time only using samples of the function z at on-grid time domain points $\{0, \ldots, d-1\}$.

LEMMA 4.1. (VARIANT OF LEMMAS 7.3, 7.17 OF [11]) *Let z be a (ϵ, Δ')-clustered signal around f_0 (Def. 4.1) for any ϵ smaller than an absolute constant. Suppose $f_0 \in \mathcal{I}$ for an interval $\mathcal{I} \subseteq [0,1]$ of size $|\mathcal{I}|$ smaller than an absolute constant with $\mathcal{I}$ known to the algorithm a priori. Let $\Delta' = O(k\Delta)$for k, Δ as defined in Lemma 2.1.Then procedure* FREQUENCYRECOVERY1CLUSTER *returns an $\widetilde{f}_0$ such that, with probability at least $1 - 2^{-\Omega(k)}$,*

$$|\widetilde{f}_0 - f_0|_\circ \lesssim \Delta' \sqrt{\Delta' d}.$$

Moreover, FREQUENCYRECOVERY1CLUSTER *has time and sample complexity of* $\mathsf{poly}(k, \log(d/\delta))$.

We now present the main subroutines and their proofs of correctness that lead the proof of Lemma 4.1.

4.1.1 Sampling the signal. In this section, we present algorithms GETEMPIRICALENERGY and GETLEGALSAMPLE and show their correctness, with the goal of proving Lemma 4.3, which is a key primitive to obtain a weak estimate of the central frequency of a clustered signal. This primitive will then be used in the next section, Section 4.1.2, repeatedly to obtain a good estimate for the central frequency of the clustered signal thus proving Lemma 4.1. This section mirrors Section 7.2 from [11], with algorithms GETEMPIRICALENERGY and GETLEGALSAMPLE being analogous to GETEMPIRICAL1ENERGY and GETLEGAL1SAMPLE but operating on a discrete time signal, rather than a continuous time one. We refer the reader to [11] for the detailed proofs (we only state the why the proofs carry over here), since they are the same up to the replacing the continuous Fourier transform with the DTFT.

Algorithm 4.1 GETEMPIRICALENERGY(z, d, Δ')

```
Input: Query access to $z$ on $[d]$, $d$, $\Delta'$.
$R_{est} \leftarrow O((d\Delta')^2)$.
$z_{emp} \leftarrow 0$.
for $i \in [R_{est}]$ do
    Choose $\alpha_i \in [d]$ uniformly at random.
    $z_{emp} \leftarrow z_{emp} + |z(\alpha_i)|^2$.
end for
$z_{emp} \leftarrow \sqrt{z_{emp}/R_{est}}$.
Return $z_{emp}$.
```

LEMMA 4.2. (COUNTERPART OF CLAIM 7.11 IN [11]) *Let z be an (ϵ, Δ')-clustered signal as in Lemma 4.1.* GETEMPIRICALENERGY *(Algorithm 4.1) takes $O((d\Delta')^2)$ samples to output z_{emp} such that $z_{emp} \in (1 \pm 0.2)||z||_d/\sqrt{d}$ with probability at least* 0.9.

Proof. The proof follows that of Claim 7.11 of [11], up to replacing the continuous Fourier transform with the DTFT. The main reason why the proof carries over is that since z has approximately compact support in Fourier domain, it is a smooth function. Thus its energy in time domain that is concentrated on $[d]$ can be estimated using uniform sampling with few samples. □

Algorithm 4.2 GETLEGALSAMPLE$(z, d, \Delta', \beta, z_{emp})$

```
**Input:**Query access to $z$ on $[d]$, $d$, $\Delta'$, $\beta$, $z_{emp}$.
$R_{repeat} \leftarrow O((d\Delta')^3)$, $S_{heavy} \leftarrow \emptyset$.
**for** $i \in [R_{repeat}]$ **do**
    Choose $\alpha_i \in [d]$ uniformly at random.

    **if** $|z(\alpha_i)| \geq 0.5z_{emp}$ **then**
        $S_{heavy} \leftarrow S_{heavy} \cup \{i\}$.
    **end if**
**end for**
**for** $i \in S_{heavy}$ **do**
    $w(i) \leftarrow |z(\alpha_i)|^2 + |z(\alpha_i + \beta)|^2$.
**end for**
Choose $\alpha$ among $\alpha_i$, $i \in S_{heavy}$ with probability $w(i)/\sum_{j \in S_{heavy}} w(j)$.
**Return:** $\alpha$.
```

LEMMA 4.3. (COUNTERPART OF LEMMA 7.2 IN [11]) *Let* $\widehat{\beta} = \frac{C_\beta}{\Delta'\sqrt{\Delta' d}}$ *for a sufficiently small constant* C_β*. Let* z *be a* (ϵ, Δ')*-one clustered signal around* f_0 *as in Lemma 4.1. Then Algorithm* GETLEGALSAMPLE*, with any integer* $\beta \leq 2\widehat{\beta}$ *and* z_{emp} *satisfying* $z_{emp} \in (1 \pm 0.2)\|z\|_d/\sqrt{d}$*, takes* $O((d\Delta')^3)$ *samples to output* $\alpha \in [d]$ *such that* $\alpha + \beta \in [d]$ *and, with probability at least* 0.6*,*

$$|z(\alpha + \beta) - z(\alpha)e^{2\pi i f_0 \beta}| \leq 0.08(|z(\alpha)| + |z(\alpha + \beta)|).$$

Proof. The proof follows that of Lemma 7.2 of [11] upto the replacement of the continuous Fourier transform with the DTFT. □

The above lemma, Lemma 4.3 obtains a weak estimate for f_0 by taking the logarithm of the ratio of the input evaluated at α and $\alpha + \beta$. This lemma is then used repeatedly in Section 4.1.2

4.1.2 Frequency recovery.

The main goal of this subsection is to present algorithms that can estimate the central frequency of a clustered signal approximately, leading to the proof of Lemma 4.1.These algorithms use the primitives presented in the previous section.

Algorithm 4.3 LOCATE1INNER$(z, \Delta', d, \widehat{\beta}, z_{emp}, \widehat{L}, R_{loc})$

```
$v_q \leftarrow 0, \forall q \in [t]$.
**while** $r = 1 \rightarrow R_{loc}$ **do**
    Choose $\beta \in [0.5\widehat{\beta}, \widehat{\beta}] \cap \mathbb{Z}$ uniformly at random.
    $\gamma \leftarrow$ GETLEGALSAMPLE$(z, \Delta', d, \beta, z_{emp})$.
    $\theta' \leftarrow \text{phase}(z(\gamma)/z(\gamma + \beta))/2\pi$.
    **for** $q \in [t]$ **do**
        $\theta_q = \widehat{L} - \Delta l/2 + \frac{q-0.5}{t}\Delta l$.
        If $\|2\pi\theta' - 2\pi\beta\theta_q\|_\circ \leq s\pi$ then add vote to $v_q, v_{q-1}, v_{q+1}$     ($\|x - y\|_\circ = \min_{z \in \mathbb{Z}} |x - y + 2\pi z|$ for any
        $x, y \in \mathbb{R}$
    **end for**
**end while**
$q^* \leftarrow \{q | v_q > \frac{R_{loc}}{2}\}$.
**Return:** $L \leftarrow \widehat{L} - \Delta l/2 + \frac{q^* - 0.5}{t}\Delta l$.
```

Algorithm 4.4 LOCATE1SIGNAL$(z, d, F, \Delta', z_{emp}, \mathcal{I})$

```
$t = \log(d)$, $t' = t/4$, $D_{max} = \log_t(d)$, $R_{loc} \approx \log_{1/c}(tc)$ ($c < 1/2$ is some constant), $L^{(1)} \leftarrow$ midpoint of $\mathcal{I}$,
    $i_0 = \log_{t'}(1/|\mathcal{I}|) + 1$.
for $i = i_0 \to D_{max}$ do
    $\Delta l = 1/(t')^{i-1}$, $s \approx c$, $\widehat{\beta} \leftarrow \frac{ts}{2\Delta' l}$.
    if $\widehat{\beta} \gtrsim d/(d\Delta)^{3/2}$ then
        Break.
    else
        $L^{(i)} \leftarrow$ LOCATEINNER$(z, \Delta', d, \widehat{\beta}, z_{emp}, L^{(i-1)}, R_{loc})$.
    end if
end for
Return: $L^{(D_{max})}$.
```

Algorithm 4.5 FREQUENCYRECOVERY1CLUSTER$(z, d, \Delta', \mathcal{I})$

```
$z_{emp} \leftarrow$ GETEMPIRICALENERGY$(z, d, \Delta')$
for $i = 1 \to O(k)$ do
    $L_r \leftarrow$ LOCATE1SIGNAL$(z, d, \Delta', z_{emp}, \mathcal{I})$.
end for
Return: $L^* \leftarrow \text{median}\{L_r | r \in [O(k)]\}$.
```

The following lemma formalizes the guarantees of the LOCATE1INNER primitive which is used iteratively in LOCATE1SIGNAL to refine and narrow-down the estimate for f_0, the frequency around which z is clustered. The final algorithm leading to the proof of Lemma 4.1 FREQUENCYRECOVERY1CLUSTER then runs LOCATE1SIGNAL multiple times and returns the median of all the runs to get an approximation to f_0 with high-probabilityThis lemma is the analogue of Lemma 7.14 of [11], however since β is always restricted to be an integer in Algorithm LOCATE1INNER, an alternate proof of correctness is needed. Another major difference is that since we already know an interval $\mathcal{I}$ of size $o(1)$ such that $f_0 \in \mathcal{I}$, the initialization of the frequency searching primitive LOCATE1SIGNAL uses this information.

LEMMA 4.4. (VARIANT OF LEMMA 7.14 OF [11]) *Consider an invocation of* LOCATE1INNER *on inputs (as per Alg. 4.4) such that there is a* $q' \in [t]$ *with* $f_0 \in [\widehat{L} - \Delta l/2 + \frac{q'-1}{t}\Delta l, \widehat{L} - \Delta l/2 + \frac{q'}{t}\Delta l]$*. Let* β *be sampled uniformly at random from* $[\frac{st}{4\Delta l}, \frac{st}{2\Delta l}] \cap \mathbb{Z}$ *and let* γ *denote the output of procedure* GETLEGALSAMPLE$(z, \Delta', d, \beta, z_{emp}, \widehat{L}, R_{loc})$*. Then the following holds,*

- *with probability at least* $1 - s$*,* $v_{q'}$ *will increase by one,*
- *for any* $|q - q'| > 3$*, with probability at least* $1 - 15s$ v_q *will not increase.*

Proof. The proof follows that of Lemma 7.14 of [11]. In their notation $\theta = f_0$ and $\theta_q = \widehat{L} - \Delta l/2 + \frac{q-1/2}{t}\Delta l$. The major and only difference lies in analyzing the case when $|q - q'| > 3$ and $|\theta - \theta_q|_\circ \geq \frac{\Delta l}{st}$ and showing that in this case v_q will not increase with high constant probability. Here, we adopt the analysis of [26] and instead of using Lemma 6.5 of [11], we will use a corollary of Lemma 4.3 of [26] since our choice of β is a random *integer* rather than a random *real number* in some range. This lemma is as follows.

LEMMA 4.5. (COROLLARY OF LEMMA 4.3 OF [26]) *For some integer number* m*, if we sample* β *uniformly at random from a set* $T \subseteq [m]$ *of* t *consecutive integers, for any* $i \in [d]$ *and a set* $S \subseteq [d]$ *of* l *consecutive integers,*

$$\Pr[\beta i \mod d \in S] \leq \frac{1}{t} + \frac{im}{dt} + \frac{lm}{dt} + \frac{l}{it}.$$

We use Lemma 4.5 with the following values — we set $m = \lceil \frac{st}{2\Delta l} \rceil$, $T = [\frac{st}{4\Delta l}, \frac{st}{2\Delta l}] \cap \mathbb{Z}$, $t \geq \frac{st}{4\Delta l} - 1$, $S = [0, \frac{3s}{4}d] \cap \mathbb{Z}$, $l \leq \frac{3}{4}sd + 1$, $i = d|\theta - \theta_q|_\circ$. Without loss of generality we can assume that i is an integer by rounding $\theta = f_0$ to the nearest integer multiple of $1/d$. This is feasible as the signal will still be clustered around this rounded f_0 since the width of the cluster $\Delta' = \mathsf{poly}(k, \log(d/\delta))/d \gg 1/d$. Recall that $t = \log d$. By assuming that d is large enough, we assume that $\frac{st}{4\Delta l} \geq \max(4, 10/s+1)$. Note that $d\frac{\Delta l}{st} \leq i \leq d\Delta l$. Observe that $\frac{m}{t} \leq (\frac{st}{2\Delta l}+1)/(\frac{st}{4\Delta l}-1) \leq 3$. Hence,

$$\begin{aligned}\Pr[\beta i \mod d \in S] &\leq \frac{s}{10} + \frac{3d\Delta l}{d} + 3\frac{(3/4)sd+1}{d} + \frac{(3/4)sd+1}{d\Delta l(st)^{-1}\cdot(st(4\Delta l)^{-1}+1)}\\ &\leq \frac{s}{10} + o(1) + 9/4s + 3s \leq 7.5s,\end{aligned}$$

where we used the fact that $\Delta l \leq |\mathcal{I}| \leq s/1.5$ since $s = \Theta(1)$ and $|\mathcal{I}|$ is a small enough constant, By the same bound for $S = [-\frac{3s}{4}d, 0]$ and a union bound, we conclude

$$\Pr[\beta d|\theta - \theta_q|_\circ \mod d \in [-\frac{3s}{4}d, \frac{3s}{4}d]] \leq 15s,$$

which is equivalent to

$$\Pr[2\pi\beta|\theta - \theta_q|_\circ \mod 2\pi \in [-\frac{3s}{4}2\pi, \frac{3s}{4}2\pi]] \leq 15s.$$

Recall that we denote $\|x-y\|_\circ = \min_{z\in\mathbb{Z}} |x-y+2\pi z|$ as the circular distance between x, y for any $x, y \in \mathbb{R}$.Thus, overall we get that with probability at least $15s$, $\|2\pi\beta(\theta_q - \theta)\|_\circ > (3s/4)2\pi$. By triangle inequality, this further implies that

$$\|2\pi\beta(\theta' - \theta_q)\|_\circ > \|2\pi\beta(\theta - \theta_q)\|_\circ - \|2\pi\beta(\theta' - \theta)\|_\circ > (3s/4)2\pi - s\pi/2 = s\pi.$$

This implies that v_q will not increase as per Line 8 of Algorithm LOCATE1INNER. □

The next lemma essentially gives the final guarantee of the LOCATE1SIGNAL algorithm.

LEMMA 4.6. (VARIANT OF LEMMAS 7.15 AND 7.16 IN [11]) *Consider the parameter setting as described in Algorithm* LOCATE1SIGNAL. *The procedure* LOCATE1INNER *uses* R_{loc} *legal samples and then procedure* LOCATE1SIGNAL *runs* LOCATE1INNER D_{max} *times to output a frequency* $\widetilde{f}_0$ *such that* $|\widetilde{f}_0 - f_0|_\circ \lesssim \Delta'\sqrt{\Delta' d}$ *with probability at least* 0.9. *Moreover,* LOCATE1SIGNAL *has time and sample complexity* $\mathsf{poly}(k, \log(d/\delta))$.

Proof. Equipped with Lemma 4.4 the proof is identical to the proofs of Lemmas 7.15 and 7.16 in [11]. □

Finally, the success probability can be boosted by repeating the procedure $O(k)$ times and using the median trick.

LEMMA 4.1. *Let* z *be a* (ϵ, Δ')*-clustered signal around* f_0 *as per Definition 4.1 for any* ϵ *smaller than an absolute constant. Suppose* $f_0 \in \mathcal{I}$ *for an interval* $\mathcal{I} \subseteq [0,1]$ *of size* $|\mathcal{I}|$ *smaller than an absolute constant, with* $\mathcal{I}$ *known to the algorithm apriori. Then procedure* FREQUENCYRECOVERY1CLUSTER *returns an* $\widetilde{f}_0$ *such that with probability at least* $1 - 2^{-\Omega(k)}$,

$$|\widetilde{f}_0 - f_0|_\circ \lesssim \Delta'\sqrt{\Delta' d}.$$

Moreover, FREQUENCYRECOVERY1CLUSTER *has time and sample complexity of* $\mathsf{poly}(k, \log(d/\delta))$.

Proof. Follows the proof of original lemma, Lemma 7.17 in [11]. □

4.2 General case. Consider the setup and parameters of Lemma 2.1, where we have a general signal $x^*(t) = \sum_{f\in S} a_f e^{2\pi i f t}$ where $S \subseteq [0,1]$, $|S| \leq k$ and $g(t)$ is noise satisfying $\|g\|_d^2 \leq c\|x^*\|_d^2$ for a small enough constant $c > 0$. Our framework to recover frequencies from such an instance will be to first reduce a general instance to a *bounded* instance, these are instances where $supp(\widehat{x^*})$ is only contained in some interval of length $1/B$. We will use $B = \Theta(k^2)$. Then using hashing techniques, we will show how to recover frequencies from bounded instances by reducing them to one-cluster instances and then running one-cluster recovery primitive of Lemma 4.1.

First we introduce the basic hashing primitives that will be needed multiple times throughout this section. The following definition introduces the hash function which maps frequencies in the Fourier domain to B buckets, and the filter function which when convolved with gives access to the function containing frequencies restricted to any desired bucket. This is the standard notation identical to Definitions 6.3 and 6.8 in [11].

Definition 4.2. (Hashing and filtering notation) *Let $h_{\sigma,b} : [0,1] \to \{0,1,\ldots,B-1\}$ defined as follows for any $\sigma \in \mathbb{R}, b \in [0,1]$,*

$$h_{\sigma,b}(f^*) = round(\frac{B}{2\pi} \cdot (2\pi\sigma(f^* - b) \mod 2\pi)) \quad \forall f^* \in [0,1].$$

Let $G^{(j)}_{\sigma,b}$ be the function which when convolved with allows us to access the function restricted to bin j for all $j \in [B]$ (via G as per Lemma 3.4):

$$\widehat{G}^{(j)}_{\sigma,b}(f) = \widehat{G}^{dis}(\frac{j}{B} - \sigma f - \sigma b) := \sum_{i\in\mathbb{Z}} \widehat{G}(i + \frac{j}{B} - \sigma f - \sigma b) \quad \forall f \in [0,1],$$
$$G^{(j)}_{\sigma,b}(t) = DTFT(\widehat{G}^{(j)}_{\sigma,b}(f)).$$

When necessary, we will make explicit the parameters B, δ, w, k used in the construction of G as per Lemma 3.4 and therefore $G^{(j)}_{\sigma,b}$.

Next we present the notation for the functions obtained by convolving the input with the filter function $G^{(j)}_{\sigma,b}(t)$.

Definition 4.3. *Let $H(t)$ be as per Lemma 3.2 for parameters k, δ and $G^{(j)}_{\sigma,b}$ as per Def. 4.2. Let $z^{(j)} = (x^* \cdot H) * G^{(j)}_{\sigma,b}$, $g^{(j)} = (g \cdot H) * G^{(j)}_{\sigma,b}$ and $x^{(j)} = (x \cdot H) * G^{(j)}_{\sigma,b} = z^{(j)} + g^{(j)}$, for all $j \in [B]$.*

Finally we present the algorithm HashToBins and its guarantees from [11] which computes $z^{(j)}(t)$ for any integer t of one's choice for all $j \in [B]$. The proof is slightly modified to work with the DTFT (see Defn. 3.1).

```
Algorithm 4.6 HashToBins($x, H, G, B, \sigma, \alpha, b, \delta, w = 0.001$(default)))
Let $(G(t), \widehat{G}(f))$ be the filter functions as per Definition 3.4 with parameters $B, \delta, w$.
Let $W(t) = x \cdot H(t), D = O(\log(k/\delta))$ such that $|supp(G(t))| = 2BD$ and $V$ as $V[j] = G[j] \cdot W(\sigma(j-\alpha))e^{2\pi i \sigma b}$
    for $j \in [-BD, BD]$ (here $G[j] = G(j) \forall j \in [-BD, BD]$ is the discretization of $G(t)$).
Let $v \in \mathbb{R}^B$ as $u[j] = \sum_{i\in[-D,D)} V[j + iB] \; \forall j \in [B]$.
Return: FFT($v$);
```

Lemma 4.7. (Variant of Lemma 6.9 in [11]) *Let $u \in \mathbb{C}^B$ be the output of* HashToBins*$(x, H, G, B, \sigma, \alpha, b, w)$, and assume σ and $\sigma\alpha \in \mathbb{Z}$. Then $u[j] = x^{(j)}(\sigma\alpha) \; \forall j \in [B]$ where $x^{(j)}$ is as per Definition 4.3. Let $D = O(\log(k/\delta)/w)$ such that $|supp(G(t))| = 2BD$ as per Lemma 3.4. Then* HashToBins *takes the following samples from x - $\{x(\sigma(i - \alpha))\}_{i=-BD}^{BD}$, and runs in time $O(B \log(k/\delta)/w)$.*

Proof. The proof is identical to the proof of Lemma 6.9 in [11], but we just need to observe that in the last line of the proof where they conclude that

$$\widehat{u}[j] = \int_{-\infty}^{\infty} \widehat{W}(s) \cdot \widehat{G^{dis}}(j/B - \sigma s - \sigma b)e^{-2\pi i \sigma a s} ds,$$

for all $j \in [B]$, $\widehat{W}(f) = \widehat{x \cdot H}(f)$ is the continuous Fourier transform of $(x \cdot H)(t)$ which is considered to be a continuous signal from $\mathbb{R} \to \mathbb{C}$ (Here the continuous version of H as per Lemma 6.6 of [11], i.e. before discretization in time domain to obtain the H function as per Lemma 3.2). We however need to consider the discretized version of this signal. Now observe that since σ and σa are integers we have that,

$$\widehat{G}^{dis}(j/B - \sigma(s+i) - \sigma b)e^{-2\pi i \sigma a(s+i)} = \widehat{G}^{dis}(j/B - \sigma s - \sigma b)e^{-2\pi i \sigma a s},$$

for all $i \in \mathbb{Z}$, where $\widehat{G}^{dis}$ is as per Definition 4.2. Thus we can rewrite the definition of $\widehat{u}[j]$ as

$$\widehat{u}[j] = \int_0^1 \left(\sum_{i\in\mathbb{Z}} \widehat{W}(s+i) \right) \cdot \widehat{G^{dis}}(j/B - \sigma s - \sigma b)e^{-2\pi i \sigma a s} ds,$$

where $\sum_{i\in\mathbb{Z}} \widehat{W}(s+i)$ is the exactly the DTFT of $(x\cdot H)(t)$ when t is restricted to $\mathbb{Z}$. Thus $\widehat{u}(j) = (x\cdot H * G^{(j)}_{\sigma,b})(\sigma a) = x^{(j)}(\sigma a)$ where $G^{(j)}_{\sigma,b}(t)$ is as per definition 4.2 and $(x\cdot H)(t), G^{(j)}_{\sigma,b}(t)$ are both discrete time signals as is in our case. □

Equipped with these tools we now describe a clustering based pre-processing step, also described in [11], applied to general instances before explaining what are bounded instances and how to reduce to them.

Definition 4.4. (Clustering) *Consider H as per Lemma 3.2 for parameters k, δ. For any two frequencies f_1, f_2 in the support of $\widehat{x^*}(f)$ we say that $f_1 \sim f_2$ if their supports overlap after convolving with $\widehat{H}$, i.e. $supp(\widehat{H\cdot e^{2\pi i f_1 t}}) \cap supp(\widehat{H\cdot e^{2\pi i f_2 t}}) \neq \emptyset$. We cluster frequencies by taking a transitive closure under this relation $\sim$ and define $C_1, C_2, \ldots, C_l$, where $0 \le l \le k$, as the clusters. Thus, each $C_i \subseteq [0,1]$ is an interval, $C_1 \cup \ldots \cup C_l = supp(\widehat{x^*\cdot H}(f))$ and $C_i \cap C_j = \emptyset$ for any $i \neq j$.*

Remark 4.8. *Let $\Delta_h = |supp(\widehat{H}(f))|$ as per Lemma 3.2, and let the Δ parameter of Lemma 2.1 be set such that it satisfies $\Delta \ge k\Delta_h$. Note that thus width of any cluster is at most $k\Delta_h \le \Delta = \mathsf{poly}(k, \log(d/\delta))/d$.*

Δ will be fixed throughout this section as per Remark 4.8. This is the same Δ as set in the statement of Lemma 2.1. Equipped with these tools and notations, we now define a bounded instance as follows.

Definition 4.5. ($(\mathcal{I},\delta')$-bounded instance) *Let $\mathcal{I} \subseteq [0,1]$ be an interval satisfying $|\mathcal{I}| \le 1/B$, and $\epsilon > 0$ be a small enough constant. Let $x^*(t) = \sum_{f\in S} a_f e^{2\pi i f t}$ where $S \subseteq [0,1]$, $|S| \le k$. Let $H(t)$ be as per Lemma 3.2 for parameters k, δ, and $g(t)$ be noise. Then the instance $x(t) = (x^*(t)+g(t))\cdot H(t)$ is an $(\mathcal{I},\delta')$-bounded instance if the following is satisfied - All clusters C as per Defn. 4.4 applied to $x^*\cdot H$ satisfy $C \subseteq \mathcal{I}$ and $\int_{[0,1]\setminus\mathcal{I}} |\widehat{g\cdot H}(f)|^2 df \le \delta'$. The noise threshold $\mathcal{N}^2$ is defined as*

$$\mathcal{N}^2 = \frac{\delta}{d}\int_{[0,1]} |\widehat{x^*\cdot H}(f)|^2 df + \frac{1}{d}\int_{\mathcal{I}} |\widehat{g\cdot H}(f)|^2 df + k\delta'/d\epsilon.$$

We will only be able to recover "heavy" frequencies from a bounded instance, and thus next we present their definition.

Definition 4.6. (Heavy frequency) *Consider the setup of Definition 4.5. Call a frequency $f^* \in [0,1]$ heavy frequency if it satisfies*

$$\int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^*\cdot H}(f)|^2 df \gtrsim \frac{d\mathcal{N}^2}{k},$$

where the interval $[f^*-\Delta, f^*+\Delta]$ is modulo 1.

For any heavy frequency f^* we would be interested in isolating the cluster of width $O(k\Delta)$ around it by making sure no other heavy frequency maps to the bucket f^* maps to. This is to ensure that the signal restricted to that bucket becomes a one-cluster signal, thus allowing the application of Lemma 4.1 to recover f^* approximately. Thus next we define what a "well-isolated" frequency is.

Definition 4.7. (Well-isolated frequency, from [11]) *f^* is well-isolated under the hashing (σ, b) if, for $j{=}h_{\sigma,b}(f^*)$, the signal $z^{(j)}$ (as per Defn. 4.3) satisfies*

$$\int_{[0,1]\setminus(f^*-200k\Delta, f^*+200k\Delta)} |\widehat{z}^{(j)}(f)|^2 df \lesssim \epsilon d\mathcal{N}^2/k.$$

4.2.1 Reducing from a bounded multi cluster instance to $\Theta(k^2)$ one cluster instances.

In this section, we show how to reduce a bounded instance to $B = \Theta(k^2)$ clustered instances.

This is presented as the main result of this section below, Lemma 4.9. It extends Lemma 7.6 of [11] for the case when we can only take samples on integer points in time domain. It essentially states that if we hash the at most k clusters in the Fourier spectrum of the input into $B = \Theta(k^2)$ buckets then all clusters are simultaneously isolated. Moreover the clusters which are heavy and the corresponding bins to which they hash have high SNR then they are essentially one-clustered as per Def. 4.1. We first state this main lemma, then proceed with sub lemmas and their proofs that will build to up to the proof of this main lemma.

Lemma 4.9. (Variant of Lemma 7.6 of [11] for on-grid samples) *Let $x(t) = (x^*(t) + g(t)) \cdot H(t)$ be a $(\mathcal{I}, \delta')$-bounded instance as per Definition 4.5. Let $G^{(j)}_{\sigma,b}(t)$ for all $j \in [B]$ be as per Def. 4.2 and Lemma 3.4 for parameters $B = \Theta(k^2), \delta = \epsilon\delta/k$, $w = 0.01$ and $k = k$. Apply the clustering procedure of Defn. 4.4 to $x^* \cdot H$ and let $k_C \leq k$ be the total number of clusters in $\widehat{x^* \cdot H}$. Let σ be a u.a.r* ***integer*** *in $[\frac{1}{200Bk\Delta}, \frac{1}{100Bk\Delta}]$ and b be a u.a.r. real number from $[0, \frac{1}{\sigma}]$. With prob. at least $1 - k_C^2/B - k_C/99k$ the following is true -*

Consider any cluster C with f^* the midpoint of C. Then f^* is well-isolated (Def. 4.7). Moreover for $j = h_{\sigma,b}(f^*)$ if,

1. *f^* is heavy as per Defn. 4.6,*
2. *$\int_{[0,1]} |\widehat{g}^{(j)}(f)|^2 df \lesssim \epsilon \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^* \cdot H}(f)|^2 df$ ($g^{(j)}$ as per Defn. 4.3),*

then $x^{(j)}$ (as per Definition 4.3) is an $(2\sqrt{\epsilon}, 200k\Delta)$-one clustered signal around f^* with interval $\mathcal{I}$.

The particular choice of σ is because we need to ensure that the time points at which HashToBins accesses the input as per Lemma 4.7 are integers. Thus σ and α must be chosen such that σ and $\sigma\alpha$ are both integers (see Lemma 4.7). This is done by choosing σ to be a random integer in a bounded interval and α such that $\sigma\alpha$ is an integer and it equals the time point at which we want to access the functions $x^{(j)}$. However this introduces complications because if σ is a random integer as opposed to a random real number in a bounded interval then the collision behaviour under the random hashing changes. Thus we first start with the main lemma that bounds the collision probability of two frequencies under the hash function of Definition 4.2. This proof is different from Claim 6.4 in [11] because σ is not being random real number in a bounded interval, we can only ensure small collision probability for frequencies not more than $1/B$ apart.

Lemma 4.10. (Hashing collision probabilities, analog of Claim 6.4 of [11]) *Let σ be a u.a.r. integer in $[\frac{1}{200Bk\Delta}, \frac{1}{100Bk\Delta}]$. Let $h_{\sigma,b}$ be as per Definition 4.2 for σ and arbitrary $b \in [0,1]$. Then,*

1. *For any $f_1, f_2 \in [0,1]$ s.t. $200k\Delta \leq |f_1 - f_2| < 200(B/2 - 0.5)k\Delta$, then $\mathbb{P}_\sigma[h_{\sigma,b}(f_1) = h_{\sigma,b}(f_2)] = 0$.*
2. *For any $f_1, f_2 \in [0,1]$ s.t. $200(B/2 - 0.5)k\Delta < |f_1 - f_2| < \frac{1}{B}$, then $\mathbb{P}_\sigma[h_{\sigma,b}(f_1) = h_{\sigma,b}(f_2)] \lesssim \frac{1}{B}$.*

Proof. For convenience, let $F = |f_1 - f_2|$. For a hash collision to occur, $h_{\sigma,b}(f_1) = h_{\sigma,b}(f_2)$, it must be that $2\pi\sigma F \in (s \cdot 2\pi - \frac{2\pi}{2B}, s \cdot 2\pi + \frac{2\pi}{2B})$ for some integer s as per the definition of $h_{\sigma,b}$ in Def. 4.2.

1. The proof of this case is unchanged when σ is restricted to integers. For completeness: if $200k\Delta \leq F < 200(B - 0.5)k\Delta$, then $\frac{2\pi}{B} \leq |2\pi\sigma F| < \frac{2\pi}{100Bk\Delta} 200(B/2 - 0.5)k\Delta = (1 - \frac{1}{2B})2\pi$. Thus collision is impossible.
2. The collision condition is equivalent to $\sigma F \in (s - \frac{1}{2B}, s + \frac{1}{2B})$ for some integer s. Note that the range of possible values of s is $\{\lfloor \frac{F}{200Bk\Delta} \rfloor, \ldots, \lceil \frac{F}{100Bk\Delta} \rceil\}$.The rest of the proof proceeds by a simple counting argument based on lengths of intervals. Assume $F < \frac{1}{B}$. For any integer s, there are at most $\frac{1/B}{F} + 1 = \frac{1}{BF} + 1$ integer values of σ such that $\sigma F \in (s - \frac{1}{2B}, s + \frac{1}{2B})$. Taking into account the total number of possible values of s, we get that the total number of such integer values σ is at most
$$\begin{aligned}(\frac{1}{BF} + 1) \cdot (\lceil \frac{F}{100Bk\Delta} \rceil - \lfloor \frac{F}{200Bk\Delta} \rfloor + 1) &\leq (\frac{1}{BF} + 1) \cdot (\frac{F}{200Bk\Delta} + 2) \\ &\lesssim \frac{1}{200B^2k\Delta}.\end{aligned}$$
Moreover, there are at least $\frac{(B-1)/B}{F} = \frac{B-1}{BF}$ integer values of σ such that $s + \frac{1}{2B} < \sigma F < (s+1) - \frac{1}{2B}$, i.e. such that σF lies in between this interval and the next one. Again, taking into account the total possible values of s we get that the total number of such integer σ is at least
$$\begin{aligned}(\frac{B-1}{BF}) \cdot (\max(\lceil \frac{F}{100Bk\Delta} \rceil - \lfloor \frac{F}{200Bk\Delta} \rfloor - 3, 1)) &\gtrsim (\frac{B-1}{BF}) \cdot (\frac{F}{200Bk\Delta}) \\ &\gtrsim \frac{1}{200Bk\Delta}.\end{aligned}$$
Note that using the notation $\lesssim$ and $\gtrsim$ is correct as $\frac{1}{BF}$ and $\frac{F}{200Bk\Delta}$ are at least $\Omega(1)$. Thus we can upper bound the collision probability upto constants by $\frac{1/200B^2k\Delta}{1/200Bk\Delta} = \frac{1}{B}$.

□

We need the following definition before proceeding.

Definition 4.8. (Nice cluster) *Consider the setup of Lemma 4.9. Let C be a cluster with midpoint f^* as per Defn. 4.4. Let $j = h_{\sigma,b}(f^*)$ and $\widehat{G}^{(j)}_{\sigma,b}(f)$ be the corresponding filter function as per Defn. 4.2. Then C is a nice cluster if $|\widehat{G}^{(j)}_{\sigma,b}(f)| \geq 1 - \epsilon\delta/k$ for every $f \in C$.*

Now we move on to state and prove the lemma that bounds the probability of isolating all frequencies (as per Definition 4.7) under the random hashing $h_{\sigma,b}$.

Lemma 4.11. (Variant of Lemma 7.19 of [11] for on-grid samples) *Consider setup of Lemma 4.9. Then with probability $1 - k_C^2/B - k_C/99k$ over the randomness in σ, b the following holds -*

For all clusters C with f^* being the midpoint of C, f^* is well isolated and C is nice (as per Defn. 4.8). Moreover, for $j = h_{\sigma,b}(f^*)$ and $z^{(j)}$ as per Defn. 4.3,

$$\int_{f^*-200k\Delta}^{f^*+200k\Delta} |\widehat{z}^{(j)}(f)|^2 df \in [1-\delta, 1] \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^* \cdot H}(f)|^2 df.$$

Proof. Consider any cluster C with f^* being its central frequency. We divide by cases and draw on Lemma 4.10.

- Due to the random shift b, the entire interval where $|\widehat{G}^{(j)}_{\sigma,b}(f)| \in [1 - \epsilon\delta/k, 1]$ is randomly shifted by σb (see Def. 4.2). Furthermore since the width of the interval containing f^* and where $|\widehat{G}^{(j)}_{\sigma,b}(f)| \in [1 - \epsilon\delta/k, 1]$ is at least $99k\Delta$ and the width of C is Δ, the entire cluster C lands in the same bucket as f^* and the region where $|\widehat{G}^{(j)}_{\sigma,b}(f)| \in [1 - \epsilon\delta/k, 1]$ with probability $1 - 1/99k$.
- If $200k\Delta \leq |f' - f^*| < \frac{1}{\sigma} - \frac{1}{\sigma B}$, then $\frac{2\pi}{B} \leq |2\pi\sigma(f^* - f')| \leq 2\pi - \frac{2\pi}{B}$ implies that the two frequencies are always mapped to different buckets as per Definition 4.2. As a result, they fall in the region where $|\widehat{G}^{(j)}_{\sigma,b}(f)| \leq \frac{\epsilon\delta}{k}$. Any such f' contributes at most $\frac{\epsilon\delta}{k}\int_{f'-\Delta}^{f'+\Delta} |\widehat{x^* \cdot H}(f)|^2 df$ to the energy in Fourier domain, and thus the total contribution of all such f' is at most $\frac{\epsilon\delta}{k}\int |\widehat{x^* \cdot H}(f)|^2 df \leq \epsilon d\mathcal{N}^2/k$ to the energy in Fourier domain.
- If $\frac{1}{\sigma} - \frac{1}{\sigma B} \leq |f' - f^*| \leq \frac{1}{B}$, then by Lemma 4.10, the probability of a hash collision is at most $\frac{1}{B}$, and by taking a union bound over the at most k_C such cluster midpoints f' in the spectrum of $\widehat{x^*}$, we have that no such f' lands in the same bucket as f^* with probability at least $1 - 1/B$. Thus, for any such f', the entire interval $f' \pm \Delta$ again falls in the $\frac{\epsilon\delta}{k}$-valued tail of $\widehat{G}$. This implies the total contribution from all such clusters with midpoint f' to the energy in the Fourier domain is at most $\frac{\epsilon\delta}{k}\int |\widehat{x^* \cdot H}(f)|^2 df \leq \epsilon d\mathcal{N}^2/k$

Points 2 and 3 imply that f^* is well-isolated. Point 1,2 and 3 implies that

$$\int_{f^*-200k\Delta}^{f^*+200k\Delta} |\widehat{z}^{(j)}(f)|^2 df \in [1-\delta, 1] \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^* \cdot H}(f)|^2 df.$$

with probability $1 - k_C/B - 1/99k$. Taking a union bound over at most k_C such clusters C, the proof of the Lemma is finished. Point 1 after the union bound over all k_C clusters implies all such clusters are nice (as per Defn. 4.8). □

Then the next lemma we need shows that even after multiplying with $\widehat{G}^{(j)}_{\sigma,b}$, the signal's energy in the time domain remains concentrated on $[0, d]$.

Lemma 4.12. *Condition on the guarantee of Lemma 4.11 holding. Let C be any cluster with midpoint f^* such that f^* is heavy as per Defn. 4.6. Then $z^{(j)}(t)$ for $j = h_{\sigma,b}(f^*)$ satisfies,*

$$\sum_{t\in[-\infty,0]\cup[d,\infty]} |z^{(j)}(t)|^2 \lesssim \epsilon \sum_{t=-\infty}^{\infty} |z^{(j)}(t)|^2.$$

Proof. From Lemma 4.11, we know that C is nice and f^* is well isolated. Let $x^*_C(t) = \sum_{f\in supp(\widehat{x^*}):f\in C} v_f e^{2\pi i f t}$ Thus we have the following,

$$\int_{[0,1]} |\widehat{z}^{(j)}(f) - \widehat{x^*_C\cdot H}(f)|^2 df \lesssim \epsilon\delta/k \int_{[0,1]} |\widehat{x^*\cdot H}(f)|^2 df \le \epsilon d\mathcal{N}^2/k.$$

This implies we have the following by Cauchy-Schwarz,

$$\int_{[0,1]} |\widehat{z}^{(j)}(f)|^2 \gtrsim \int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df - \epsilon d\mathcal{N}^2/k \tag{4.5}$$

$$\gtrsim \int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df. \tag{4.6}$$

where we used the fact that f^* is heavy and thus $\epsilon d\mathcal{N}^2/k \le \epsilon\int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df$.

Now we have that for $z^{(j)} = (x^*_C\cdot H)(t) + g_C(t)$, g_C satisfies,

$$\sum_{t=-\infty}^{\infty} |g_C(t)|^2 = \int_{[0,1]} |\widehat{g_C}(f)|^2 df \lesssim (\epsilon\delta/k)\int_{[0,1]} |\widehat{x^*\cdot H}(f)|^2 \le \epsilon d\mathcal{N}^2/k.$$

Thus we have that,

$$\sum_{t\in[-\infty,0]\cup[d,\infty]} |z^{(j)}(t)|^2 \lesssim \sum_{t\in[-\infty,0]\cup[d,\infty]} |(x^*_C\cdot H)(t)|^2 + \epsilon d\mathcal{N}^2/k.$$

Now since x^*_C is an at most k-Fourier sparse function, by the properties of the H function as per Lemma 3.2, we have the following,

$$\begin{aligned}\sum_{t\in[-\infty,0]\cup[d,\infty]} |(x^*_C\cdot H)(t)|^2 &\le \epsilon \sum_{t\in[-\infty,\infty]} |(x^*_C\cdot H)(t)|^2 \\ &= \epsilon\int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df.\end{aligned}$$

and since f^* is heavy, $\epsilon\int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df + \epsilon d\mathcal{N}^2/k \lesssim \epsilon\int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df$. Thus we finally get the following,

$$\sum_{t\in[-\infty,0]\cup[d,\infty]} |z^{(j)}(t)|^2 \lesssim \epsilon\int_{[0,1]} |\widehat{x^*_C\cdot H}(f)|^2 df.$$

Combining this with equation 4.5 and applying Parseval's theorem completes the proof of the lemma. ∎

With these tools in place, we are ready to prove Lemma 4.9.

Proof. [Proof of Lemma 4.9] Condition on Lemma 4.11 being true. Consider any cluster C with midpoint f^* as per the Lemma satisfying the assumptions of the Lemma 4.9. Let $j = h_{\sigma,b}(f^*)$ and recall $z^{(j)} = (x^*\cdot H) * G^{(j)}_{\sigma,b}$ and $x^{(j)} = ((x^* + g)\cdot H) * G^{(j)}_{\sigma,b}$. Let $I_{f^*} = [f^* - 200k\Delta, f^* + 200k\Delta]$ modulo 1. Then combining the fact that f^* is heavy and it is well-isolated, we know that the following holds from guarantees of Lemma 4.11,

$$\int_{I_{f^*}} |\widehat{z}^{(j)}(f)|^2 df \gtrsim \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^*\cdot H}(f)|^2 df \ge d\mathcal{N}^2/k, \tag{4.7}$$

and the following as well

$$\int_{[0,1]\setminus I_{f^*}} |\widehat{z}^{(j)}(f)|^2 df \lesssim \epsilon d\mathcal{N}^2/k. \tag{4.8}$$

Now recall $\widehat{g}^{(j)} = \widehat{g \cdot H} \cdot \widehat{G}^{(j)}_{\sigma,b}$. Then the previous discussion implies the following,

$$
\begin{aligned}
(4.9)\qquad \int_{I_{f^*}} |\widehat{x}^{(j)}(f)|^2 df &\gtrsim \int_{I_{f^*}} |\widehat{z}^{(j)}(f)|^2 df - \int_{I_{f^*}} |\widehat{g}^{(j)}(f)|^2 df \quad \text{(Cauchy-Schwarz)}\\
(4.10)\qquad &\gtrsim \int_{I_{f^*}} |\widehat{z}^{(j)}(f)|^2 df \quad \text{(assumption 2 of Lemma 4.9 plus eqn. 4.7)}\\
(4.11)\qquad &\gtrsim \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^* \cdot H}(f)|^2 df \geq d\mathcal{N}^2/k \quad \text{(equation 4.7).}
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\int_{[0,1]\setminus I_{f^*}} |\widehat{x}^{(j)}(f)|^2 df &\leq 2\left(\int_{[0,1]\setminus I_{f^*}} |\widehat{z}^{(j)}(f)|^2 + |\widehat{g}^{(j)}(f)|^2 df\right) \quad \text{(Cauchy-Schwarz)}\\
&\lesssim \epsilon d\mathcal{N}^2/k + \epsilon \int_{I_{f^*}} |\widehat{z}^{(j)}(f)|^2 df \quad \text{(assumption 2 of Lemma 4.9 and eqns. 4.7 and 4.8)}\\
&\lesssim \epsilon \int_{I_{f^*}} |\widehat{x}^{(j)}(f)|^2 df \quad \text{(from eqn. 4.10 and } f^* \text{ being heavy).}
\end{aligned}
$$

Overall, we thus get,

$$
(4.12)\qquad \int_{I_{f^*}} |\widehat{x}^{(j)}(f)|^2 df \geq (1-\epsilon) \int_{[0,1]} |\widehat{x}^{(j)}(f)|^2 df.
$$

Next, our goal is to show that the $x^{(j)}$'s energy in time domain is concentrated in $[0, d]$. By Plancherel's theorem, we get the following

$$
(4.13)\qquad \sum_{t=0}^{d} |g^{(j)}(t)|^2 \leq \sum_{t=-\infty}^{\infty} |g^{(j)}(t)|^2 = \int_{[0,1]} |\widehat{g}^{(j)}(f)|^2 df \lesssim \epsilon \int_{[0,1]} |\widehat{z}^{(j)}(f)|^2 df = \epsilon \sum_{t=-\infty}^{\infty} |z^{(j)}(t)|^2,
$$

where the second last inequality used assumption 2 of Lemma 4.9. Combining this with Lemma 4.12, we have that

$$
\sum_{t=0}^{d} |g^{(j)}(t)|^2 \leq \sum_{t=-\infty}^{\infty} |g^{(j)}(t)|^2 \leq \epsilon \sum_{t=0}^{d} |z^{(j)}(t)|^2.
$$

Equipped with this inequality, we can bound the energy of $x^{(j)}$ in time domain over $[d]$ as follows

$$
\begin{aligned}
\sum_{t=0}^{d} |x^{(j)}(t)|^2 &\geq \sum_{t=0}^{d} |z^{(j)}(t)|^2 - \sum_{t=0}^{d} |g^{(j)}(t)|^2 - 2\sqrt{\sum_{t=0}^{d} |z^{(j)}(t)||g^{(j)}(t)|}\\
&\geq \sum_{t=0}^{d} |z^{(j)}(t)|^2 - \sum_{t=0}^{d} |g^{(j)}(t)|^2 - 2\sqrt{(\sum_{t=0}^{d} |z^{(j)}(t)|^2)(\sum_{t=0}^{d} |g^{(j)}(t)|^2)}\\
&\geq (1-2\sqrt{\epsilon}) \sum_{t=0}^{d} |z^{(j)}(t)|^2.
\end{aligned}
$$

Furthermore combining equation 4.13 and using Lemma 4.12, we can bound the energy of $x^{(j)}$ outside $[d]$ as follows,

$$
\begin{aligned}
\sum_{t\in[-\infty,0]\cup[d,\infty]} |x^{(j)}(t)|^2 &\lesssim \sum_{t=-[\infty,\infty]} |g^{(j)}(t)|^2 + \sum_{t\in[-\infty,0]\cup[d,\infty]} |z^{(j)}(t)|^2\\
&\lesssim \epsilon \sum_{t=0}^{d} |z^{(j)}(t)|^2.
\end{aligned}
$$

The above two inequalities imply the following

$$\sum_{t=0}^{d} |x^{(j)}(t)|^2 \geq (1-2\sqrt{\epsilon}) \sum_{t=-\infty}^{\infty} |x^{(j)}(t)|^2. \tag{4.14}$$

Finally we know that $f^* \in \mathcal{I}$. Thus, equations 4.14 and 4.12 imply $x^{(j)}$ is an $(2\sqrt{\epsilon}, 200k\Delta)$-clustered signal around f^* and thus completing the proof of Lemma 4.9. Since we conditioned on Lemma 4.11, this holds with probability $1 - k_C^2/B - k_C/99k$ simultaneously for all clusters C of $\widehat{x^* \cdot H}$. ∎

4.2.2 Frequency recovery of bounded instances.

Lemmas 4.9 and 4.1 imply the following lemma which shows how to recover all "recoverable" frequencies of a bounded instance with high probability.

LEMMA 4.13. *Let $x(t) = (x^*(t) + g(t)) \cdot H(t)$ be an $(\mathcal{I}, \delta')$-bounded instance with noise threshold $\mathcal{N}^2$ as per the setup of Definition 4.5. Let $G(t)$ for all $j \in [B]$ be as per Def. 4.2 and Lemma 3.4 for parameters $B = \Theta(k^2), \epsilon\delta/k$ and $w = 0.01, k$. Let Δ be as per Lemma 2.1. Apply the clustering procedure of Defn. 4.4 to $x^* \cdot H$ and let $k_C \leq k$ be the total number of clusters. Let σ be a u.a.r. integer in $[\frac{1}{200Bk\Delta}, \frac{1}{100Bk\Delta}]$ and b be a u.a.r. real number in $[0, \frac{1}{\sigma}]$. Then one can find a list L of B frequencies in $[0,1]$ in time $\mathsf{poly}(k, \log(d/\delta))$ such that with probability at least $1 - k_C^2/B - k_C/99k - o(1/k)$ the following holds - For all clusters C with f^* being the midpoint of C and $j = h_{\sigma,b}(f)$, f^* is well-isolated (Def. 4.7). Moreover, if*

1. *f^* is heavy as per Defn. 4.6 and,*
2. *$\int_{[0,1]} |\widehat{g}^{(j)}(f)|^2 df \lesssim \epsilon \int_{f^*-\Delta}^{f^*+\Delta} |\widehat{x^* \cdot H}(f)|^2 df$ ($g^{(j)}$ as per Def. 4.3),*

then there exists a $f \in L$ such that $|f^* - f|_\circ \lesssim k\Delta\sqrt{k\Delta d}$.

Proof. Condition on Lemma 4.9 holding true. Our procedure to find L is just to run FREQUENCYRECOVERY1CLUSTER on each of the B instances corresponding to each hash bucket and thus we get a list of B frequencies, one frequency per each hash bucket. Consider any cluster C of $x^* \cdot H$ with midpoint f^* satisfying the conditions of the Lemma. Let $j = h_{\sigma,b}(f^*)$, then Lemma 4.9 implies that $x^{(j)} = (x^* + g) \cdot H * G_{\sigma,b}^{(j)}$ is $(\sqrt{\epsilon}, 200k\Delta)$-one clustered around f^*. We then apply Lemma 4.1 to the j^{th} bucket, that is on $x^{(j)}$. We do this by returning the j^{th} element of the output of HASHTOBINS$(x, H, B, \sigma, \alpha/\sigma, b, \delta, w)$ to implement time domain access $x^{(j)}(\alpha)$ for any $\alpha \in [d]$ as needed by the algorithm of Lemma 4.1. This implies that we can recover a f such that $|f - f^*|_\circ \lesssim k\Delta\sqrt{k\Delta d}$ with probability $1 - 2^{-\Omega(k)}$. We take a union bound for Lemma 4.1 to succeed for all $j \in [B]$. This happens with probability $1 - B \cdot 2^{-\Omega(k)} = 1 - o(1/k)$. Then we union bound this with Lemma 4.9 succeeding. This completes the proof of the lemma. ∎

4.2.3 Reducing a general instance to a bounded instance.

Consider the setup and parameters of Lemma 2.1, that is a general signal $x(t) = x^*(t) + g(t)$, where $x^*(t) = \sum_{f \in S} a_f e^{2\pi i f t}$ where $S \subseteq [0,1]$, $|S| \leq k$ and $g(t)$ is arbitrary. In this section we explain the reduction from such a general instance to $B = \Theta(k^2)$ bounded instances. To achieve this, we convolve with the filter function $G_{\sigma,b}^{(j)}(t)$ as per Definition 4.2 with $\sigma = 1$ and b is real number chosen uniformly at random between $[0,1]$. The parameters used in the G filter function as per Definition 3.4 to construct $G_{1,b}^{(j)}(t)$ is B and $w = 1/\mathsf{poly}(k)$ and δ.

Before stating and proving the main result of this section, We first make a few important observations about this filtering operation. The claim below formalizes the behavior of the hash function $h_{1,b}$, and its proof trivially follows from Definition 4.2.

CLAIM 4.14. *The hash function $h_{\sigma,b}$ as per Definition 4.2 for $\sigma = 1$ and $b \in [0,1]$ partitions $[0,1]$ into B intervals/buckets $[b-1/2B, b+1/2B], [b+1/2B, b+3/2B], \ldots, [b-3/2B, b-1/2B]$ where each interval is modulo 1. Convolving the input with $G_{1,b}^{(j)}$ gives access to the input only containing frequencies in the j^{th} such interval/bucket.*

We will use the HASHTOBINS primitive as per Lemma 4.7 to access the function containing frequencies restricted to a such a bucket/interval. We now define the notion of a badly cut cluster needed to argue that all clusters land in the region where the filter function G has value almost 1.

Definition 4.9. *Let C be any cluster as per Definition 4.4. Suppose there exists a j such that $C \subseteq I_j := [b + j/2B, b + (j+2)/2B]$. Let $I_j^{inner} := [b + j/2B + w/2B, b + (j+2)/2B - w/2B]$. We say that C is* “badly cut” *if $C \not\subseteq I_j^{inner}$.*

Since the width of any cluster is $\mathsf{poly}(k, \log(d/\delta))/d = o(1/k^2) = o(1/B)$, we can show the following Lemma.

Lemma 4.15. *With probability at least* 0.99 *over the randomness in b, no cluster in $x^* \cdot H$ as per Definition 4.4 is badly cut.*

Proof. Since the width of any cluster is at most $\mathsf{poly}(k, \log(d/\delta))/d$ and the width of the good region I_j^{inner} in any interval I_j is at least $(1-w)/B = \Theta(1/k^2)$, the probability that a fixed cluster is badly cut is at most $\mathsf{poly}(k, \log(d/\delta))/d$. Since there are at most k clusters, taking a union bound finishes the proof since $k^3 \cdot \mathsf{poly}(k, \log(d/\delta))/d << 0.01$. □

We choose a b uniformly at random and condition on the event guaranteed to hold with probability 0.99 as per the previous Lemma 4.15. We now state the main claim of this section below which shows that this event is enough to guarantee that the instance corresponding to each bin $j \in [B]$ is bounded.

Lemma 4.16. *Choose b u.a.r. in $[0,1]$. Let $z^{(j)}(t) = ((x^* + g)\cdot H) * G_{1,b}^{(j)}(t)$ for $G_{1,b}^{(j)}$ as per Def. 4.2 and Lemma 3.4 for parameters $B, w = 1/\mathsf{poly}(k), \delta$. Let C_j be the union of all clusters of $x^* \cdot H$ (see Def. 4.4 for clusters) in interval I_j (see Def. 4.9 for I_j) and $S_j = supp(\widehat{x^*}) \cap C_j$. Then with probability at least 0.9 over choice of b, for all $j \in [B]$ $z^{(j)}(t) = (x^{(j)} \cdot H)(t) + (g^{(j)} \cdot H)(t)$, where $x^{(j)}(t) = \sum_{f \in S_j} a_f e^{2\pi i f t}$, is a $(I_j, \delta\|x^*\|_d^2/k)$-bounded instance (as per Def. 4.5). Furthermore for any interval $I \subseteq I_j$,*

$$\int_I |\widehat{g^{(j)} \cdot H}(f)|^2 df \lesssim \int_I |\widehat{g \cdot H}(f)|^2 df + (\delta/k)\|x^*\|_d^2.$$

Now we state the proof this Lemma.

Proof. [Proof of Lemma 4.16] Lemma 4.15, implying $C_j \subseteq I_j^{inner}$, combined with the fact that $|\widehat{G}_{1,b}^{(j)}(f)| \in [1, 1-\delta/k]$ for all $f \in I_j^{inner}$, $|\widehat{G}_{1,b}^{(j)}(f)| \leq 1$ for all $f \in I_j$ from Lemma 3.4 implies the following,

$$\begin{aligned}
\int_I |\widehat{g^{(j)} \cdot H}|^2 df &= \int_I |\widehat{x^{(j)} \cdot H}(f) - \widehat{x^* \cdot H} \cdot \widehat{G}_{1,b}^{(j)}(f) - \widehat{g \cdot H} \cdot \widehat{G}_{1,b}^{(j)}(f)|^2 df \\
&\leq 2\int_I |\widehat{g \cdot H}(f)|^2 df + \frac{2\delta}{k}\int_I |\widehat{x^* \cdot H}(f)|^2 df \\
&\lesssim \int_I |\widehat{g \cdot H}(f)|^2 df + \frac{\delta}{k}\int_{[0,1]} |\widehat{x^* \cdot H}(f)|^2 df \lesssim \int_I |\widehat{g \cdot H}(f)|^2 df + \frac{\delta}{k}\|x^*\|_d^2.
\end{aligned}$$

Furthermore $|\widehat{G}_{1,b}^{(j)}(f)| \leq \delta/k$ for all $f \in [0,1] \setminus I_j$ from Lemma 3.4 implies the following,

$$\begin{aligned}
\int_{[0,1]\setminus I_j} |\widehat{g^{(j)} \cdot H}(f)|^2 df &= \int_{[0,1]\setminus I_j} |(\widehat{x^* \cdot H} + \widehat{g}) \cdot \widehat{G}_{1,b}^{(j)}(f) - 0|^2 df \quad (\widehat{x^{(j)} \cdot H}(f) = 0 \forall f \in [0,1] \setminus I_j) \\
&\leq \frac{\delta}{k}\left(\int_0^1 |\widehat{x^* \cdot H}(f)|^2 df + \int_0^1 |\widehat{g \cdot H}(f)|^2 df\right) \\
&\lesssim \frac{\delta}{k}\|x^*\|_d^2,
\end{aligned}$$

where in the last line we used the assumption of Lemma 2.1 that $\|g\|_d^2 \leq c\|x^*\|_d^2$ for some small enough constant $c > 0$. This implies that $z^{(j)}(t) = (x^{(j)} \cdot H)(t) + g^{(j)}(t)$ is a $(I_j, \delta\|x^*\|_d^2/k)$-bounded instance (recall Def. 4.5) because the previous equation implies that $\int_{[0,1]\setminus I_j} |\widehat{g^{(j)} \cdot H}(f)|^2 df \lesssim \frac{\delta}{k}\|x^*\|_d^2$ and $x^{(j)}$, an at most k-Fourier sparse function, has all its clusters in I_j whose width is $1/B$. □

Equipped with these reductions, we are now ready to finish the proof of Lemma 2.1. The proof essentially reduces a general instance to B bounded instances and then applies the algorithm of Lemma 4.13 to recover frequencies from each of these bounded instances.

Proof. [Proof of Lemma 2.1] Consider the setup as described in Lemma 2.1. We apply Lemma 4.16 to x. Then we know that with probability at least 0.9, $z^{(j)}(t) = (x^{(j)} \cdot H)(t) + g^{(j)}(t)$, where $x^{(j)}(t) = \sum_{f \in S_j} a_f e^{2\pi i f t}$ is an $(I_j, \delta\|x^*\|_d^2/k)$ bounded instance for all $j \in [B]$. Now consider the j^{th} such bounded instance. Then the noise threshold in the j^{th} bounded instance $x^{(j)}$ $\mathcal{N}_j^2$ as per Def. 4.5 satisfies,

$$\mathcal{N}_j^2 = \frac{1}{d}\int_{I_j} |\widehat{g}^{(j)}(f)|^2 df + \frac{\delta}{d\epsilon}\|x^*\|_d^2 + \frac{\delta}{d}\int_{[0,1]} |\widehat{x^{(j)} \cdot H}(f)|^2 df \tag{4.15}$$

$$\lesssim \frac{1}{d}\int_{I_j} |\widehat{g \cdot H}(f)|^2 df + \frac{\delta}{d}\|x^*\|_d^2 + \frac{\delta}{d}\int_{C_j} |\widehat{x^* \cdot H}(f)|^2 df \tag{4.16}$$

$$\lesssim \frac{1}{d}(\int_{[0,1]} |\widehat{g}(f)|^2 df + \delta\|x^*\|_d^2) = \frac{1}{d}(\|g\|_d^2 + \delta\|x^*\|_d^2) = \mathcal{N}^2. \tag{4.17}$$

where in the second inequality we used the guarantee of Lemma 4.16 to upper bound $\int_{I_j} |\widehat{g}^{(j)}(f)|^2 df$ and in third inequality we used the fact that $\int_{I_j} |\widehat{g \cdot H}(f)|^2 df \leq \|g \cdot H\|_d^2 \leq \|g\|_d^2 = \int_{[0,1]} |\widehat{g}(f)|^2 df$ (noise $g(t)$ outside $[d]$ is 0). Now we apply Lemma 4.13 to $z^{(j)}$. Then we know that all clusters in C_j (set of all clusters of $x^{(j)}$) with midpoint f^* that do not satisfy both the conditions of Lemma 4.13 have no guarantee of being recovered. Call C_{unrec} the union of all such clusters. First the amount of energy corresponding to clusters C_j with midpoints that are not heavy (as per Def. 4.6) is at most

$$|C_j| d\mathcal{N}_j^2/k \leq d\mathcal{N}^2, \tag{4.18}$$

since all clusters are disjoint as per 4.4. Now consider the random hashing (σ, b) and corresponding filters $G_{\sigma,b}^{(\cdot)}$as per Lemma 4.13, then we know that all clusters $C \in C_j$ with midpoint f^* and $i = h_{\sigma,b}(f^*)$ satisfying

$$\int_{[f^*-\Delta, f^*+\Delta]} |\widehat{x^{(j)} \cdot H}(f)|^2 df \lesssim \int_{[0,1]} |\widehat{g^{(j)} \cdot H} \cdot \widehat{G}_{\sigma,b}^{(i)}(f)|^2 df,$$

also have no guarantee of being recovered. Since the clusters are disjoint and also well-isolated simultaneously from Lemma 4.13, thus mapping to different bins, the total amount of energy lost due to such low SNR clusters is at most the following,

$$\begin{aligned}
\sum_{i \in [B]} \int_{[0,1]} |\widehat{g^{(j)} \cdot H} \cdot \widehat{G}_{\sigma,b}^{(i)}(f)|^2 df &\lesssim \int_{[0,1]} |\widehat{g^{(j)} \cdot H}(f)|^2 (\sum_{i \in [B]} |\widehat{G}_{\sigma,b}^{(i)}(f)|^2|) df \\
&\lesssim \int_{[0,1]} |\widehat{g^{(j)} \cdot H}(f)|^2 df = \int_{I_j} |\widehat{g^{(j)} \cdot H}(f)|^2 df + \int_{[0,1] \setminus I_j} |\widehat{g^{(j)} \cdot H}(f)|^2 df \\
&\lesssim \int_{I_j} |\widehat{g^{(j)} \cdot H}(f)|^2 df + \delta\|x^*\|_d^2/k \quad \text{(since } z^{(j)} \text{ is a bounded instance)} \\
&\lesssim \int_{I_j} |\widehat{g \cdot H}(f)|^2 + \delta\|x^*\|_d^2/k \quad \text{(from guarantee of Lemma 4.16)},
\end{aligned}$$

where the first two inequalities follow from Def. 4.4 and Properties I-III of Lemma 3.4. Summing this up over all $j \in [B]$, we get that the total energy of such low SNR clusters is at most

$$\int_{[0,1]} |\widehat{g \cdot H}(f)|^2 df + \delta\|x^*\|_d^2 = \|g \cdot H\|_d^2 + \delta\|x^*\|_d^2 \leq \|g\|_d^2 + \delta\|x^*\|_d^2 = d\mathcal{N}^2. \tag{4.19}$$

Combining equations 4.19 and 4.18 we get that $\int_{C_{unrec}} |\widehat{x^* \cdot H}(f)|^2 df \lesssim d\mathcal{N}^2$. Morever taking a union bound over all $j \in [B]$ for Lemma 4.13 to succeed for all $z^{(j)}$, we get that this event happens with probability

$1 - (\sum_{j\in[B]} |C_j|^2/B + |C_j|/99k) \geq 1 - O(k^2)/B - k/99k \geq 0.9$ for $B = O(k^2)$ (since $\sum_{j\in[B]} |C_j|^2 \leq O((\sum_{j\in[B]} |C_j|)^2) = O(k^2)$). If we let L to be the union of the outputs of Lemma 4.13 on all $j \in [B]$, then from the claim of Lemma 4.13 we get that S_{heavy} as defined as per Lemma 2.1 satisfies the following,

$$\begin{aligned}\|x^* - x^*_{S_{heavy}}\|_d^2 &\lesssim \int_{[0,1]} |\widehat{x^* \cdot H}(f) - \widehat{x^*_{S_{heavy}} \cdot} H(f)|^2 df \\ &= \int_{C_{unrec}} |\widehat{x^* \cdot H}(f)|^2 df \\ &\lesssim d\mathcal{N}^2.\end{aligned}$$

We can implement sample access to $z^{(j)}(t)$ for any integer t that the algorithm of Lemma 4.13 demands by returning $u[j]$ for $u = \textsc{HashToBins}(x, H, G, 1, t, b, w)$. A remark here is that whenever HashToBins requires to access the input outside $[d]$ we output 0, from property 5 of Lemma 3.2 this only increases the noise by a additive $\delta\|x^*\|_d^2$ factor. The correctness of this follows from Lemma 4.7 and thus the total running time and sample complexity suffers a multiplicative overhead of $O(B\log(k/\delta)/w)$ on top of running the algorithm of Lemma 4.13. Thus the overall running time and sample complexity is still $\mathsf{poly}(k, \log(d/\delta))$. □

5 Sublinear Time Algorithms for Toeplitz Matrices

In this section, we prove our main sublinear time robust Toeplitz matrix approximation result (Theorem 1.1) and describe its applications to sublinear time Toeplitz low-rank approximation and covariance estimation. In Section 5.1 we present the proof of a heavy-light decomposition result using the off-grid frequency recovery algorithm of Lemma 2.1 and the existence result of Theorem 2.1, as briefly discussed in Section 2.2. Next, in Section 5.2 we use this heavy-light decomposition result and sublinear time approximate regression techniques for Fourier sparse functions to prove Theorem 1.1. Finally, in Section 5.3 we use Theorem 1.1 to prove our sublinear time low-rank approximation (Theorem 1.2) and covariance estimation (Theorem 1.3) results.

5.1 Heavy Light Decomposition. In this subsection, we present the proof of our heavy-light decomposition as discussed in Section 2.2. The formal statement is as follows.

Lemma 5.1. *Consider the input setup of Theorem 1.1. Let $\widetilde{T} = F_S D F_S^*$ be as guaranteed to exist by Theorem 2.1 for T, $\epsilon = 0.1$, δ and k. Let $E^k = E + T - \widetilde{T}$. Assume $\|E^k\|_F \leq c\|\widetilde{T}\|_F$ for some small enough constant $c > 0$. Then in $\mathsf{poly}(k, \log(d/\delta))$ time we can find a list of frequencies $L \subseteq [0, 1]$ of size $|L| = \mathsf{poly}(k, \log(d/\delta))$ satisfying the following with probability 0.9 - Let $S_{heavy} \subseteq \widetilde{S}$ defined as follows,*

$$S_{heavy} = \{f \in S : \exists f' \in L \text{ s.t. } |f - f'|_\circ \leq \mathsf{poly}(k, \log(d/\delta))/d\}.$$

Furthermore for every $f \in S_{heavy}$, if $(1 - f) \mod 1 \notin S_{heavy}$ then add it to S_{heavy}. Then there exists a diagonal D^{heavy} and $\widetilde{T}^{heavy} = F_{S_{heavy}} D^{heavy} F^*_{S_{heavy}}$ such that $\widetilde{T}^{light} = \widetilde{T} - \widetilde{T}^{heavy}$ satisfies

$$\|\widetilde{T}^{light}\|_F \lesssim \|E^k\|_F + \delta\|T\|_F.$$

To prove this, we first need the following helper claim for PSD matrices which will also be useful at a later stage in the paper.

Claim 5.2. (Equation (5) in [4]) *The following holds for any PSD matrix $A \in \mathbb{R}^{d\times d}$ -*

$$\|A\|_F \leq \|A_{[0:d/2,0:d/2]}\|_F + \|A_{[d/2:d,d/2:d]}\|_F.$$

Using the above claim, we now state our main helper lemma about the structural properties of symmetric Toeplitz matrices that are nearly PSD. The following lemma essentially says that if a nearly PSD Toeplitz matrix has large Frobenius norm, then the first half of the first column must have large ℓ_2 norm as well.

Lemma 5.3. *Let $\widetilde{T}$ be a $d \times d$ symmetric Toeplitz matrix, and suppose that there exists a PSD Toeplitz matrix T satisfying $\|\widetilde{T} - T\|_F \leq 0.001\|T\|_F$. Then the following holds,*

$$\|\widetilde{T}_{[0:d/2,0]}\|_2 \geq \frac{0.49}{\sqrt{d}}\|\widetilde{T}\|_F.$$

Proof. Since T is PSD, using Lemma 5.2 we know that

$$\|T\|_F \le \|T_{[0:d/2,0:d/2]}\|_F + \|T_{[d/2:d,d/2:d]}\|_F. \tag{5.20}$$

Since T is also Toeplitz, $T_{[0:d/2,0:d/2]} = T_{[d/2:d,d/2:d]}$. Thus we have $\|T_{[0:d/2,0:d/2]}\|_F \ge 0.5\|T\|_F$. Since $\|\widetilde{T}_{[0:d/2,0:d/2]} - T_{[0:d/2,0:d/2]}\|_F \le \|\widetilde{T} - T\|_F \le 0.001\|T\|_F \le 0.01\|\widetilde{T}\|_F$, we get that $\|\widetilde{T}_{[0:d/2,0:d/2]}\|_F \ge 0.49\|\widetilde{T}\|_F$. Finally we have the following.

$$\begin{aligned}\|\widetilde{T}_{[0:d/2,0:d/2]}\|_F^2 &= (d/2)T_{[0,0]}^2 + \sum_{i\in[1,d/2]} (d-2i)\widetilde{T}_{[i,0]}^2 \quad \text{(since } \widetilde{T}_{[0:d/2,0:d/2]} \text{ is symmetric Toeplitz)}\\ &\le \sum_{i\in[d/2]} d\widetilde{T}_{[i,0]}^2 = d\|\widetilde{T}_{[0:d/2,0]}\|_2^2.\end{aligned} \tag{5.21}$$

Thus we get that $\sqrt{d}\|\widetilde{T}_{[0:d/2,0]}\|_2 \ge \|\widetilde{T}_{[0:d/2,0:d/2]}\| \ge 0.49\|\widetilde{T}\|_F$. □

Equipped with this helper lemma and our sublinear time off-grid recovery result of Lemma 2.1, we are ready to present the proof of Lemma 5.1.

Proof. [Proof of Lemma 5.1] For $\widetilde{T} = F_S D F_S^*$ suppose that $S = \{f_1, f_2, \dots, f_{\widetilde{O}(k)}\}$ and $D = diag([v_1, \dots, v_{\widetilde{O}(k)}])$. Let $\widetilde{T}_1(t) = \sum_{j=1}^{\widetilde{O}(k)} v_j e^{2\pi i f_j t}$, a $\widetilde{O}(k)$ Fourier sparse function. Then it is easy to see from expanding out $\widetilde{T} = F_S D F_S^*$ that the first column of $\widetilde{T}$ is defined by $\widetilde{T}_1(t)$ for $t \in \{0, \dots, d-1\}$. A minor technicality compared to the description in the tech-overview is that rather than working with a random column $\widetilde{T}_{[0:d,j]}$ for $j \sim \{0, \dots, d/2\}$, we will work with the $d/2$ sized chunk of it $\widetilde{T}_{[j:j+d/2,j]}$ which by the virtue of $\widetilde{T}$ being Toeplitz is equal to the first half of the first column $\widetilde{T}_{[0:d/2,0]}$. Thus this $d/2$ sized chunk of the j^{th} column has identical Fourier spectrum compared to the first column.

First observe that $\mathbb{E}_{i\sim[d/2]}[\|E^k_{[i,i+d/2,i]}\|_2^2] \le \|E^k\|_F^2$, thus applying Markov's inequality the following holds with probability 0.99 for an $i \sim [d/2]$.

$$\|E^k_{[i:i+d/2,i]}\|_2^2 \le (100/d)\|E^k\|_F^2. \tag{5.22}$$

We apply Lemma 5.3 to $\widetilde{T}$, this is possible because point 3 of Theorem 2.1 in Section 3 implies that there exists some PSD Toeplitz matrix T' such that $\|\widetilde{T} - T'\|_F \le \delta\|\widetilde{T}\|_F$. We thus get the following by combining equation 5.22, $\|E^k\|_F \le c\|\widetilde{T}\|_F$ with Lemma 5.3,

$$\begin{aligned}\|E^k_{[i,i+d/2,i]}\|_2 &\le (10/\sqrt{d})\|E^k\|_F\\ &\le (10c/\sqrt{d})\|\widetilde{T}\|_F\\ &\lesssim c\|\widetilde{T}_{[0:d/2,0]}\|_2.\end{aligned} \tag{5.23}$$

Fix this i and condition on this event that equations 5.23 and 5.22 hold. Let $x^*(t) = \widetilde{T}_1(t)$, $g(t) = E^k_i(t+i) := E^k_{[i+t,i]}$ and $x(t) = x^*(t) + g(t)$ for $t \in [d/2]$. We can access $x(t)$ by querying the input $T + E$ at index $[i+t, i]$ for any $t \in [d/2]$. Let $H(t)$ be the function as per Lemma 3.2 for parameters $\widetilde{O}(k), \delta$. Apply Lemma 2.1 to $x = x^* + g$ to obtain a list of frequencies L of size $\mathsf{poly}(k)$ and let $S_{heavy} \subseteq S$ be the set of all $f \in S = supp(\widehat{x^*})$ such that there exists some $f' \in L$ satisfying $|f - f'| \lesssim \Delta k\sqrt{\Delta k d} = \mathsf{poly}(k, \log(d/\delta))/d$. Let $S^{light} = S \setminus S^{heavy}$. Observe that since the width of each set $\widetilde{S}_i$ as per point 2 in Theorem 2.1 in Section 3 is $\widetilde{O}(\gamma) = \widetilde{O}(1/2^{\mathsf{poly}\log d}) = o(\Delta)$ (Δ is as per Lemma 2.1), each $\widetilde{S}_i$ is either completely in S^{heavy} or completely in $\widetilde{S} \setminus S^{heavy}$. Let D^{heavy} contain the diagonal entries of D corresponding to S^{heavy}, $\widetilde{T}^{heavy} = F_{S^{heavy}} D^{heavy} F^*_{S^{heavy}}$ and $\widetilde{T}_1^{heavy}$ be its first column. Then we have the following,

$$\|\widetilde{T}_1(t) - \widetilde{T}_1^{heavy}(t)\|_{d/2}^2 \lesssim \|g\|_{d/2}^2 + \delta\|\widetilde{T}_1\|_{d/2}^2 \tag{5.24}$$

$$= \|E^k_{[i:i+d/2,i]}\|_2^2 + \delta\|\widetilde{T}_1\|_{d/2}^2 \quad \text{(Definition of } g = E^k_{[i:i+d/2,i]}) \tag{5.25}$$

$$\le \|E^k_{[i:i+d/2,i]}\|_2^2 + \delta\|T\|_F^2. \tag{5.26}$$

where the first inequality follows from the guarantee of Lemma 2.1. We now state an important caveat below.

REMARK 5.4. *It may happen that $\widetilde{T}^{heavy}$ has complex entries, this can happen when there is some $f \in S^{heavy}$ such that $1-f \notin S^{heavy}$. However, discarding the imaginary part of entries in $\widetilde{T}^{heavy}$ can only lead to reducing $\|\widetilde{T}_1(t) - \widetilde{T}_1^{heavy}(t)\|_{d/2}$, thus the bound of equation 5.26 still holds. The removal of imaginary parts can be achieved by adding $1-f$ to S^{heavy} for all such $f \in S^{heavy}$, and the coefficients of $f, 1-f$ in D^{heavy} will be equal to half of the corresponding coefficients in D, thus they will still be equal by point 3 of Theorem 2.1 in Section 3.*

Now define the symmetric Toeplitz matrix $\widetilde{T}^{light} = \widetilde{T} - \widetilde{T}^{heavy}$, and its first column $\widetilde{T}_1^{light}(t) = \widetilde{T}_1(t) - \widetilde{T}^{heavy}(t)$. Recall we know that every $\widetilde{S}_i$ (defined as per point 2 of Theorem 2.1 in Section 3) is either completely in S^{heavy} or completely out of it, and for every $\widetilde{S}_i \in S^{heavy}$, $-\widetilde{S}_i \in S^{heavy}$ ($-\widetilde{S}_i$ defined in point 3 of Thm. 2.1). Let S' be set of all $S_i = -\widetilde{S}_i \cup \widetilde{S}_i$ for $\widetilde{S}_i \in S^{heavy}$ such that $-\widetilde{S}_i$ was not in S^{heavy}, but we added it to make $\widetilde{T}^{heavy}$ real in Remark 5.4. This implies $\widetilde{T}^{light} = \sum_{S_i \in S \setminus S^{heavy}} F_{S_i} D_i F_{S_i}^* + (1/2) \sum_{S_i \in S'} F_{S_i} D_i F_{S_i}^*$ where each S_i, D_i as per point 3 of Theorem 2.1 in Section 3. Let $\widetilde{T}^a = \sum_{S_i \in S \setminus S^{heavy}} F_{S_i} D_i F_{S_i}^*$ and $\widetilde{T}^b = \sum_{S_i \in S'} F_{S_i} D_i F_{S_i}^*$. Then by point 3 of Theorem 2.1 in Section 3 we can say that there exists PSD Toeplitz matrices T^a, T^b such that,

$$\|\widetilde{T}^a - T^a\|_F \le \delta \|T^a\|_F$$
$$\|\widetilde{T}^b - T^b\|_F \le \delta \|T^b\|_F$$

Let $T^{light} = T^a + (1/2)T^b$, thus T^{light} is also PSD Toeplitz. Thus T^{light} satisfies

$$\|\widetilde{T}^{light} - T^{light}\|_F \le \delta(\|T^a\|_F + \|T^b\|_F) \le O(\delta)\|T^{light}\|_F \ll 0.001\|T^{light}\|_F. \tag{5.27}$$

On the other hand equation 5.26 implies the following for $\widetilde{T}^{light}$.

$$\|\widetilde{T}^{light}_{[0:d/2,0]}\|_2^2 = \|\widetilde{T}_1(t) - \widetilde{T}^{heavy}(t)\|_{d/2}^2 \lesssim \|E^k_{[i:i+d/2,i]}\|_2^2 + \delta\|T\|_F^2. \tag{5.28}$$

Thus equations (5.28) and (5.27) allow us to apply Lemma 5.3 to upper bound $\|\widetilde{T}^{light}\|_F \lesssim \sqrt{d}\|\widetilde{T}^{light}_{[0:d/2,0]}\|_2$ to get the following.

$$\|\widetilde{T}^{light}\|_F^2 \lesssim d\|E^k_{[i:i+d/2,i]}\|_F^2 + \delta d\|T\|_F^2$$
$$\implies \|\widetilde{T}^{light}\|_F \lesssim \|E^k\|_F + \delta d\|T\|_F.$$

where in the last line we used equation 5.22. Adjusting δ by $1/d$ factor (this is feasible by losing log factors as the dependence on δ is $\log(1/\delta)$), we finish the proof of point 2. of the Lemma. □

5.2 Noisy Toeplitz Recovery.

Equipped with Lemma 5.1 and Theorem 2.1, in this subsection, we present the proof of Theorem 1.1 which is our main sublinear time Toeplitz matrix approximation result.

Proof. [Proof of Theorem 1.1] Consider $\widetilde{T}^{heavy}$,L and $\widetilde{T}^{light}$ as per the statement of Lemma 5.1. Lemmas 5.1 and 2.1 imply the following for $\widetilde{T}^{heavy} = F_{S^{heavy}} D^{heavy} F^*_{S^{heavy}}$,

$$\begin{aligned}\|T + E - \widetilde{T}^{heavy}\|_F &\le \|E\|_F + \|T - \widetilde{T}\|_F + \|\widetilde{T} - \widetilde{T}^{heavy}\|_F \\ &\lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F + \|\widetilde{T}^{light}\|_F \\ &\lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F.\end{aligned}$$

Let $N = \{1/2d, 3/2d, \ldots, 1-1/2d\}$. Apply Lemma 5.1 and let L be the set of $\mathsf{poly}(k, \log d)$ frequencies returned by Lemma 5.1 and $L' = \{f \in N : \exists f' \in L \text{ s.t. } |f' - f| \le \mathsf{poly}(k, \log(d/\delta))/d\}$. This implies $|L'| \le \mathsf{poly}(k, \log(d/\delta))$. Let $S(L') = \bigcup_{f \in L'} \bigcup_{1 \le j \le r_2} \{f + \gamma j, f - \gamma j\}$ where γ, r_2 are as per Theorem 2.1. Thus $|S(L')| \le \mathsf{poly}(k, \log(d/\delta))$ This implies S^{heavy} as per Theorem 2.1 satisfies $S^{heavy} \subseteq S(L')$. Now we will solve the following regression problem approximately.

$$\min_{D: D \text{ is diagonal}} \|T + E - F_{S(L')} D F^*_{S(L')}\|_F. \tag{5.29}$$

We will first show that the optimal solution of the above optimization problem satisfies the guarantees of Theorem 1.1, then we will show how to obtain a constant factor approximate solution to the above problem in $\mathsf{poly}(k, \log(d/\delta))$ time.

Let $E^k = E + T - \widetilde{T}$ as defined in Lemma 5.1, and first consider the case when $\|E^k\|_F \geq c\|\widetilde{T}\|_F$ where c is the constant as per Lemma 5.1. This is the case when the noise is noticeably larger compared to the true input, and thus the guarantees of Lemma 5.1 are not guaranteed to hold. In this case, returning 0 as the solution of the regression problem won't be too bad. Formally we have the following.

$$\begin{aligned}\min_{D:D \text{ is diagonal}} \|T + E - F_{S(L')}DF^*_{S(L')}\|_F &\leq \|T + E\|_F \quad (\text{for } D = 0)\\ &= \|T - \widetilde{T} + \widetilde{T} + E\|_F\\ &\leq \|T - \widetilde{T}\|_F + \|\widetilde{T}\|_F + \|E\|_F\\ &\lesssim \|T - T_k\|_F + \delta\|T\|_F + \|\widetilde{T}\|_F + \|E\|_F\\ &\lesssim \|T - T_k\|_F + \delta\|T\|_F + \|E^k\|_F + \|E\|_F\\ &\lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F.\end{aligned}$$

where we used Theorem 2.1 in the third last inequality, $\|E^k\|_F \geq c\|\widetilde{T}\|_F$ in the second last inequality and the definition of E^k in the last inequality.

Now consider the other case when $\|E^k\|_F \leq c\|\widetilde{T}\|_F$. Then the requirements needed by Lemma 5.1 hold. Now, from the structure of S as per points 2 and 3 of Lemma 2.1 we know that $S^{heavy} \subseteq S(L')$. This implies the following,

$$\begin{aligned}\min_{D:D \text{ is diagonal}} \|T + E - F_{S(L')}DF^*_{S(L')}\|_F &\leq \|T + E - \widetilde{T}^{heavy}\|_F\\ &\leq \|E\|_F + \|T - \widetilde{T}\|_F + \|\widetilde{T} - \widetilde{T}^{heavy}\|_F\\ &= \|E\|_F + \|T - \widetilde{T}\|_F + \|\widetilde{T}^{light}\|_F\\ &\lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F,\end{aligned}$$

where in the last inequality we used point 2. of Lemma 5.1 to bound $\|\widetilde{T}^{light}\|_F$ and Theorem 2.1 to bound $\|T - \widetilde{T}\|_F$. Thus, in all cases, we can conclude the following.

$$\min_{D:D \text{ is diagonal}} \|T + E - F_{S(L')}DF^*_{S(L')}\|_F \lesssim \|T - T_k\|_F + \|E\|_F + \delta\|T\|_F. \tag{5.30}$$

We will now applying the following lemma that is a corollary of Lemma 5.7 of [19] that allows us to find a D that is a constant factor approximate solution to the regression problem of equation (5.29) in time only depending polynomially on $|S(L')| = \mathsf{poly}(k, \log(d/\delta))$.

Lemma 5.5. (Corollary of Lemma 5.7 of [19]) *There is an algorithm such that given any matrix $B \in \mathbb{R}^{d\times d}$ and set $M \subset [0, 1]$ of size $|M| = m$, it runs in time at most $\mathsf{poly}(m)$ and returns a diagonal $D' \in \mathbb{R}^{m\times m}$ that satisfies the following with probability* 0.99,

$$\|B - F_M D' F^*_M\|_F \lesssim \min_{\substack{D\in\mathbb{R}^{m\times m}:\\ D \text{ is diagonal}}} \|B - F_M D F^*_M\|_F,$$

where F_M is a Fourier matrix as per Def. 2.1.

Applying the previous lemma to $B = T + E$ and $M = S(L')$, we can find a D' in time $\mathsf{poly}(k, \log(1/\delta))$ such that the following holds with probability at least 0.99,

$$\begin{aligned}\|T - F_{S(L')}D'F^*_{S(L')}\|_F &\leq \|T + E - F_{S(L')}D'F^*_{S(L')}\|_F + \|E\|_F\\ &\lesssim \|E\|_F + \|T - T_k\|_F + \delta\|T\|_F\\ &\lesssim \max\{\|E\|_F, \|T - T_k\|_F\} + \delta\|T\|_F.\end{aligned}$$

where in the second last inequality we used equation (5.30). This finishes the proof of the main theorem. □

We end this section by presenting the proof of Lemma 5.5. This follows the proof of Lemma 5.7 in [19] with the modification that D is restricted to be a diagonal, however we restate the proof here for completeness.

Proof. Let $\widehat{D} = \arg\min_{\text{Diagonal } D} \|B - F_M D F_M^*\|_F$. Let $S_1, S_2^T \in \mathbb{R}^{s\times d}$ be independent sampling matrices as per Claim A.1 of [19] for $\epsilon = \delta = 0.1$ and the leverage score distribution of Corollary C.2 of [19]. Then $s = O(m\log^2(m))$. Let $D' = \arg\min_{\text{Diagonal } D} \|S_1 B S_2 - S_1 F_M D F_M^* S_2\|_F$. D' can be found in $\mathsf{poly}(s) = \mathsf{poly}(m)$ time. Our strategy to prove the lemma will have two steps. First we will state an inequality and show how that implies the lemma, and then we will prove the inequality. We will show that the following holds with probability at least 0.97, for all diagonal $D \in \mathbb{R}^{m\times m}$.

$$\|S_1 F_M D F_M^* S_2 - S_1 B S_2\|_F = (1 \pm 0.1)\|F_M D F_M^* - B\|_F \pm 100\|F_M \widehat{D} F_M^* - B\|_F. \tag{5.31}$$

Equipped with the previous inequality, we can use it to prove the lemma as follows.

$$\begin{aligned}\|F_M D' F_M^* - B\|_F &\le 1.1\|S_1 F_M D' F_M^* S_2 - S_1 B S_2\|_F + 100\|F_M \widehat{D} F_M^* - B\|_F\\ &\le 1.1\|S_1 F_M \widehat{D} F_M^* S_2 - S_1 B S_2\|_F + 100\|F_M \widehat{D} F_M^* - B\|_F\\ &\le (1.1)^2\|F_M \widehat{D} F_M^* - B\|_F + (100\cdot 1.1 + 100)\|F_M \widehat{D} F_M^* - B\|_F\\ &\le 212\|F_M \widehat{D} F_M^* - B\|_F.\end{aligned} \tag{5.32}$$

where the first and second last inequality follow by applying equation (5.31) and the second inequality follows from the optimality of D' for the subsampled regression problem.

Now we focus on proving equation (5.31). First using the triangle inequality we write $\|S_1 F_M D F_M^* S_2 - S_1 B S_2\|_F$ as follows.

$$\|S_1 F_M D F_M^* S_2 - S_1 B S_2\|_F = \|S_1 F_M D F_M^* S_2 - S_1 F_M \widehat{D} F_M^* S_2\|_F \pm \|S_1 F_M \widehat{D} F_M^* S_2 - S_1 B S_2\|_F. \tag{5.33}$$

Now observe that since S_1, S_2 are independent leverage score sampling matrices, they are unbiased estimators of the norm of any vector in $\mathbb{R}^d$. That is, for any $X \in \mathbb{R}^{d\times r}$ for any r, $\mathbb{E}[\|S_1 X\|_F^2] = \mathbb{E}[\|X^T S_2\|_F^2] = \|X\|_F^2$ for both $i = 1, 2$. Thus applying Markov's inequality we get that with probability at least 0.99,

$$\|S_1 F_M \widehat{D} F_M^* - S_1 B\|_F^2 \le 100\|F_M \widehat{D} F_M^* - B\|_F^2.$$

Applying Markov's inequality again over the randomness of S_2, we get the following with probability at least 0.99,

$$\begin{aligned}\|S_1 F_M \widehat{D} F_M^* S_2 - S_1 B S_2\|_F^2 &\le 100\|S_1 F_M \widehat{D} F_M^* - S_1 B\|_F^2\\ &\le 100^2\|F_M \widehat{D} F_M^* - B\|_F^2\end{aligned}$$

Thus taking a union bound over the randomness in S_1, S_2, we get that the following holds with probability at least 0.98.

$$\|S_1 F_M \widehat{D} F_M^* S_2 - S_1 B S_2\|_F \le 100\|F_M \widehat{D} F_M^* - B\|_F \tag{5.34}$$

Finally, due to Claim A.1 and Corollary C.2 of [19], since S_1, S_2 are independent leverage score sampling matrices taking $s = O(m\log^2(m))$ samples the following subspace embedding property holds with probability at least 0.99.

$$\begin{aligned}\|S_1[F_M; F_M]y\|_2^2 &= (1 \pm 0.01)\|[F_M; F_M]y\|_2^2 \quad \forall y \in \mathbb{C}^{2m} \text{ and}\\ \|y^*[F_M; F_M]^* S_2\|_2^2 &= (1 \pm 0.01)\|[F_M; F_M]y\|_2^2 \quad \forall y \in \mathbb{C}^{2m}\end{aligned}$$

This guarantee applied to S_1 implies the following.

$$\|S_1 F_M D F_M^* - S_1 F_M \widehat{D} F_M^*\|_F = (1 \pm 0.01)\|F_M D F_M^* - F_M \widehat{D} F_M^*\|_F$$

And then applying the subspace embedding guarantee for S_2 finally gives us the following.

$$\begin{aligned}\|S_1 F_M D F_M^* S_2 - S_1 F_M \widehat{D} F_M^* S_2\|_F &= (1 \pm 0.01)\|S_1 F_M D F_M^* - S_1 F_M \widehat{D} F_M^*\|_F\\ &= (1 \pm 0.01)^2\|F_M D F_M^* - F_M \widehat{D} F_M^*\|_F.\end{aligned}$$

Plugging the previous equation and 5.34 back into 5.33 we get 5.31. This completes the overall proof of Lemma 5.5. □

5.3 Sublinear Time Low-Rank Approximation and Covariance Estimation. Our goal in this section is to present the proofs of the sublinear time low-rank approximation and covariance estimation results of Theorems 1.2 and 1.3 respectively. The proof of Theorem 1.2 easily follows by applying Theorem 1.1 for $E = 0$. The main Lemma we will need to apply our framework to prove Theorem 1.3 is the following that allows us to bound the magnitude of the noise.

Lemma 5.6. *Consider $d \times d$ PSD Toeplitz T. Let k be an integer and $\epsilon > 0$. Given samples $x_1, \dots, x_s \sim \mathcal{N}(0, T)$, let $X \in \mathbb{R}^{d \times s}$ be a matrix whose i^{th} column is $x_i/\sqrt{s}$ for all $i \in [s]$. If $s = \widetilde{O}(k^4/\epsilon^2)$, then the following holds with probability at least* 0.98

$$\|XX^T - T\|_F \lesssim \sqrt{\|T - T_k\|_2 \operatorname{tr}(T) + \frac{\|T - T_k\|_F \operatorname{tr}(T)}{k}} + \epsilon \|T\|_2.$$

Assuming Lemma 5.6, we now present the proof of Theorem 1.3.

Proof. [Proof of Theorem 1.3] The proof follows easily from applying Theorem 1.1 with $E = XX^T - T$, $\delta = \epsilon/\mathsf{poly}(d)$ and bounding the Frobenius norm of E using Lemma 5.6 with $s = \widetilde{O}(k^4/\epsilon^2)$. Note that $\delta\|T\|_F \le \epsilon\|T\|_2$ since $\delta = \epsilon/\mathsf{poly}(d)$ thus $\log(1/\delta) = O(\log(d/\epsilon))$. This bounds the Vector Sample Complexity (VSC). Note that Theorem 1.1 only accesses $\mathsf{poly}(k, \log(d/\epsilon))$ entries of XX^T and any of its $(i,j)^{th}$ entry is equal to $\sum_{k=1}^s x_{k,i} x_{k,j}$. Thus each entry access to XX^T requires reading 2 entries, the i and j^{th} entries, from each sample. Thus we get that the Entry Sample Complexity (ESC) is $\mathsf{poly}(k, \log(d/\epsilon))$. □

Now we proceed by presenting the proof of Lemma 5.6.

Proof. [Proof of Lemma 5.6] Let $T = U\Sigma U^T$ be the eigenvalue decomposition of T where $\Sigma \succeq 0$ is diagonal. Let $P_k = U_k U_k^T$ be the projection matrix onto the subspace spanned by the top-k eigenvectors. Using the rotational invariance of the Gaussian distribution, we have that X is distributed as $X \sim U\Sigma^{1/2}G$, where $G \in \mathbb{R}^{d \times s}$ is a matrix with each entry distributed independently as $\frac{1}{\sqrt{s}}\mathcal{N}(0,1)$. Then we upper bound $\|XX^T - T\|_F$ by the following three terms using the triangle inequality,

$$\|XX^T - T\|_F \le \|XX^T - P_k XX^T P_k\|_F + \|P_k XX^T P_k - T_k\|_F + \|T_k - T\|_F.$$

We finish the proof by taking a union bound over the following two claims, which bound the first and second terms of the expression above.

Claim 5.7. *If $s = \widetilde{O}(k)$, then with probability at least* 0.99*,*

$$\|XX^T - P_k XX^T P_k\|_F \lesssim \sqrt{\|T - T_k\|_2 \operatorname{tr}(T) + \frac{\|T - T_k\|_F \operatorname{tr}(T)}{k}}.$$

Claim 5.8. *For any $\epsilon > 0$. If $s = \widetilde{O}(k^4/\epsilon^2)$, then with probability at least* 0.99*,*

$$\|P_k XX^T P_k - T_k\|_F \le \epsilon\|T\|_2.$$

Note that since T is PSD, $\|T - T_k\|_F = \sqrt{\sum_{j=k+1}^d \lambda_j^2(T)} \le \sqrt{\lambda_1(T)(\sum_{j=1}^d \lambda_j)} = \sqrt{\|T - T_k\|_2 tr(T)}$. Thus we ignore the third term $\|T - T_k\|_F$ in the upper bound on $\|XX^T - T\|_F$ in the big-Oh term of Claim 5.7. This completes the proof of Lemma 5.6. □

We now proceed by proving Claims 5.7 and 5.8.

Proof. [Proof of Claim 5.7] We know that $X = U\Sigma^{1/2}G$ and P_k is a rank-k projection matrix. Thus, we can apply the projection cost-preserving sketch property of the Gaussian distribution from Lemma 12 and Theorem 27 of [15].

As a result, if $s = \Omega(k/\gamma^2)$, then the following holds with probability at least 0.99:

$$\|P_k X - X\|_2^2 = (1 \pm \gamma)\|P_k U\Sigma^{1/2} - U\Sigma^{1/2}\|_2^2 \pm \frac{\gamma}{k}\|P_k U\Sigma^{1/2} - U\Sigma^{1/2}\|_F^2. \tag{5.35}$$

Fix $\gamma = O(1)$. Now we use the fact that for any matrices A, B, $\|AA^* - BB^*\|_F \leq \|AB^* - BB^*\|_F + \|AA^* - AB^*\|_F \leq \|A - B\|_2(\|A\|_F + \|B\|_F)$. Applying this fact to $A = P_kX$ and $B = X$, we get that

$$\begin{aligned}
\|XX^T - P_kXX^TP_k\|_F &\leq \|P_kX - X\|_2(\|X\|_F + \|P_kX\|_F) \\
&\leq O\left(\sqrt{\|P_kU\Sigma^{1/2} - U\Sigma^{1/2}\|_2^2 + \frac{\|P_kU\Sigma^{1/2} - U\Sigma^{1/2}\|_F^2}{k}}\|X\|_F\right) \text{(from equation 5.35)} \\
&= O\left(\sqrt{\|T^{1/2} - T_k^{1/2}\|_2^2 + \frac{\|T^{1/2} - T_k^{1/2}\|_F^2}{k}}\|X\|_F\right) \\
&= O\left(\sqrt{\|T - T_k\|_2\|X\|_F^2 + \frac{\|T - T_k\|_F\|X\|_F^2}{k}}\right).
\end{aligned}$$

In the second line, we used that $\|P_kX\|_F \leq \|X\|_F$, and the last line follows by observing that $U\Sigma^{1/2} = T^{1/2}, P_kU\Sigma^{1/2} = T_k^{1/2}$. Now observe that $\|X\|_F^2 = \|T^{1/2}G\|_F^2$. Thus by applying the Johnson-Lindenstrauss lemma to each row of $T^{1/2}$ and taking a union bound over the d rows of $T^{1/2}$, we have that as long as $s = \Omega((1/\epsilon^2)\log(d/\delta))$, $\|X\|_F^2 \leq (1+\epsilon)\|T^{1/2}\|_F^2 = (1+\epsilon)\operatorname{tr}(T)$ with probability $1-\delta$. Fixing $\epsilon = \delta = 0.001$, taking a union bound over this event and the projection cost-preserving sketch property, and plugging the bound $\|X\|_F^2 \leq O(\operatorname{tr}(T))$ into the last line of the derivation above, we finish the proof of the claim. □

Finally, we prove Claim 5.8. This proof follows the proof strategy of Theorem 7.1. in [48]but adapted for the case when the covariance matrix is exactly rank k.

Proof. [Proof of Claim 5.8] For any $d \times d$ rank-k matrix A, we have that

$$\|A\|_2 = \max_{x \in \mathcal{S}^{k-1}} |x^TU_k^TAU_kx| \leq \frac{1}{1-2\epsilon} \max_{z \in \mathcal{N}_\epsilon} |z^TU_k^TAU_kz|,$$

where U_k is the matrix containing the top-k eigenvectors of A, $\mathcal{S}^{k-1}$ is the unit sphere in k dimensions, and $\mathcal{N}_\epsilon$ is an ϵ-net of $\mathcal{S}^{k-1}$. Fix $\epsilon = 1/4$, then using Lemma 5 of [54] which uses volumetric arguments to bound the size of $\mathcal{N}_\epsilon$, we have that $|\mathcal{N}_\epsilon| \leq 17^k$. Now let $E = P_kXX^TP_k - T_k = T_k^{1/2}GG^TT_k^{1/2} - T_k$. Observe that E has rank at most k. Letting U_k be the matrix of the top-k eigenvectors of E, we have that

$$\|E\|_2 \leq 2\max_{x \in \mathcal{N}_\epsilon} |x^TU_k^TEU_kx|.$$

This implies the following:

$$\begin{aligned}
\mathbb{P}(\|E\|_2 > t) &\leq \mathbb{P}(\max_{x \in \mathcal{N}_\epsilon} |x^TU_k^TEU_kx| > t/2) \\
&\leq \sum_{x \in \mathcal{N}_\epsilon} \mathbb{P}(|x^TU_k^TEU_kx| > t/2).
\end{aligned}$$

Thus we need to focus on upper bounding $\mathbb{P}(|x^TU_k^TEU_kx| > t/2)$ for a fixed $x \in \mathcal{S}^{k-1}$. Let $y = U_kx$, which implies that $y \in \mathcal{S}^{d-1}$. Thus we need to bound $\mathbb{P}(|y^TEy| > t/2)$ for some $y \in \mathcal{S}^{d-1}$. We have that

$$\begin{aligned}
y^TEy &= \frac{1}{s}\sum_{i \in [s]} \left(y^TT_k^{1/2}g_ig_i^T(T_k^{1/2})^Ty - y^TT_ky\right) \\
&= \frac{1}{s}\sum_{i \in [s]} \left(Z_i^2 - \mathbb{E}[Z_i^2]\right),
\end{aligned}$$

where each $Z_i = y^TT_k^{1/2}g_i$ and $g_i \sim \mathcal{N}(0, I_{d\times d})$. Thus each $Z_i \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = \|y^TT_k^{1/2}\|_2^2$. As a result, $\frac{1}{s}\sum_{i\in[s]}\left(Z_i^2 - \mathbb{E}[Z_i^2]\right)$ is a chi-squared random variable. Using standard chi-squared concentration bound of Laurent-Massart [38], we have the following

$$\mathbb{P}(|y^T Ey| \geq t/2) \leq \exp\left\{-\Omega\left(s \min\left(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right)\right)\right\}.$$

Thus, setting $t = \epsilon\sigma^2$ and $s = O\left(\frac{1}{\epsilon^2}\log\left(\frac{|\mathcal{N}_{1/4}|}{\delta}\right)\right)$ yields that the above upper bound on the failure probability is at most $\delta/|\mathcal{N}_{1/4}|$. Therefore,

$$\mathbb{P}(\|E\|_2 > \epsilon\sigma^2) \leq \delta$$

Thus we get that if $s = \widetilde{O}(k/\epsilon^2)$, then with probability at least 0.99,

$$\|E\|_2 \leq \epsilon\|y^T U_k \Sigma^{1/2}\|_2^2 \leq \epsilon tr(T_k) \leq \epsilon k\|T\|_2.$$

This further implies that $\|E\|_F \leq \epsilon k^{1.5}\|T\|_2$. Setting $\epsilon = \epsilon/k^{1.5}$, we conclude the proof of the claim. □

6 Acknowledgements

Cameron Musco's work is supported by an Adobe Research grant and NSF Grants No. 2046235 and No. 1763618. The authors acknowledge useful discussions with Michael Kapralov, Mikhail Makarov, and Hannah Lawrence.

7 References

[1] Y. I. Abramovich, N. K. Spencer, and A. Y. Gorokhov, *Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays. II. Partially augmentable arrays*, IEEE Transactions on Signal Processing, 47 (1999), pp. 1502–1521.

[2] T. D. Ahle, M. Kapralov, J. B. Knudsen, R. Pagh, A. Velingker, D. P. Woodruff, and A. Zandieh, *Oblivious sketching of high-degree polynomial kernels*, in Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2020, pp. 141–160.

[3] B. M. Asl and A. Mahloojifar, *A low-complexity adaptive beamformer for ultrasound imaging using structured covariance matrix*, IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 59 (2012), pp. 660–667.

[4] K. M. Audenaert, *A norm compression inequality for block partitioned positive semidefinite matrices*, Linear algebra and its applications, 413 (2006), pp. 155–176.

[5] A. Bakshi, N. Chepurko, and D. P. Woodruff, *Robust and sample optimal algorithms for PSD low rank approximation*, in Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2020, pp. 506–516.

[6] A. Bakshi and D. Woodruff, *Sublinear time low-rank approximation of distance matrices*, Advances in Neural Information Processing Systems 31 (NeurIPS), 31 (2018).

[7] P. Boufounos, V. Cevher, A. C. Gilbert, Y. Li, and M. J. Strauss, *What's the frequency, Kenneth?: Sublinear Fourier sampling off the grid*, Algorithmica, 73 (2015), pp. 261–288.

[8] M. J. Brookes, J. Vrba, S. E. Robinson, C. M. Stevenson, A. M. Peters, G. R. Barnes, A. Hillebrand, and P. G. Morris, *Optimising experimental design for meg beamformer imaging*, Neuroimage, 39 (2008), pp. 1788–1802.

[9] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM Journal on Scientific and Statistical Computing, 6 (1985), pp. 349–364.

[10] J.-F. Cai, X. Qu, W. Xu, and G.-B. Ye, *Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction*, Applied and Computational Harmonic Analysis, 41 (2016), pp. 470–490.

[11] X. Chen, D. M. Kane, E. Price, and Z. Song, *Fourier-sparse interpolation without a frequency gap*, in Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 2016, pp. 741–750.

[12] Y. Chen, Y. Chi, and A. J. Goldsmith, *Exact and stable covariance estimation from quadratic sampling via convex programming*, IEEE Transactions on Information Theory, 61 (2015), pp. 4034–4059.

[13] M. T. Chu, R. E. Funderlic, and R. J. Plemmons, *Structured low rank approximation*, Linear Algebra and its Applications, 366 (2003), pp. 157–172.

[14] D. Cohen, S. Tsiper, and Y. C. Eldar, *Analog-to-digital cognitive radio: Sampling, detection, and hardware*, IEEE Signal Processing Magazine, 35 (2018), pp. 137–166.

[15] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, *Dimensionality reduction for k-means clustering and low rank approximation*, in Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC), 2015, pp. 163–172.

[16] R. Cohen and Y. C. Eldar, *Sparse Doppler sensing based on nested arrays*, IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 65 (2018), pp. 2349–2364.

[17] G. Cybenko, *Moment problems and low rank Toeplitz approximations*, Circuits, Systems and Signal Processing, 1 (1982), pp. 345–366.

[18] G. R. de Prony, *Essai experimental et analytique: sur les lois de la dilatabilite des fluides elastique et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, a differentes temperatures*, Journal Polytechnique ou Bulletin du Travail fait a l'Ecole Centrale des Travaux Publics, (1795).

[19] Y. C. Eldar, J. Li, C. Musco, and C. Musco, *Sample efficient Toeplitz covariance estimation*, in Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2020, pp. 378–397.

[20] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, *Hankel matrix rank minimization with applications to system identification and realization*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 946–977.

[21] D. R. Fuhrmann, *Application of Toeplitz covariance estimation to adaptive beamforming and detection*, IEEE Transactions on Signal Processing, 39 (1991), pp. 2194–2198.

[22] M. Ghadiri, *On symmetric factorizations of Hankel matrices*, Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS), (2023).

[23] A. C. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, *Recent developments in the sparse Fourier transform: A compressed Fourier transform for big data*, IEEE Signal Processing Magazine, 31 (2014), pp. 91–100.

[24] G. H. Golub and C. F. Van Loan, *Matrix computations*, JHU press, 2013.

[25] R. M. Gray, *Toeplitz and circulant matrices: A review*, Foundations and Trends in Communications and Information Theory, 2 (2006), pp. 155–239.

[26] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, *Nearly optimal sparse Fourier transform*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC), 2012, pp. 563–578.

[27] ——, *Simple and practical algorithm for sparse Fourier transform*, in Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2012.

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

[29] P. Indyk, M. Kapralov, and E. Price, *(Nearly) sample-optimal sparse Fourier transform*, in Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2014.

[30] P. Indyk, A. Vakilian, T. Wagner, and D. P. Woodruff, *Sample-optimal low-rank approximation of distance matrices*, in Proceedings of the 32nd Annual Conference on Computational Learning Theory (COLT), PMLR, 2019, pp. 1723–1751.

[31] M. Ishteva, K. Usevich, and I. Markovsky, *Factorization approach to structured low-rank approximation with applications*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 1180–1204.

[32] Y. Jin, D. Liu, and Z. Song, *Super-resolution and robust sparse continuous Fourier transform in any constant dimension: Nearly linear time and sample complexity*, in Proceedings of the 34th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2023.

[33] M. Kapralov, *Sparse Fourier transform in any constant dimension with nearly-optimal sample complexity in sublinear time*, in Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC), 2016.

[34] M. Kapralov, H. Lawrence, M. Makarov, C. Musco, and K. Sheth, *Toeplitz low-rank approximation with sublinear query complexity*, in Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2023, pp. 4127–4158.

[35] H. Knirsch, M. Petz, and G. Plonka, *Optimal rank-1 Hankel approximation of matrices: Frobenius norm and spectral norm and Cadzow's algorithm*, Linear Algebra and its Applications, 629 (2021), pp. 1–39.

[36] H. Krim and M. Viberg, *Two decades of array signal processing research: the parametric approach*, IEEE signal processing magazine, 13 (1996), pp. 67–94.

[37] ———, *Two decades of array signal processing research: the parametric approach*, IEEE Signal Processing Magazine, 13 (1996), pp. 67–94.

[38] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Annals of Statistics, (2000), pp. 1302–1338.

[39] H. Lawrence, J. Li, C. Musco, and C. Musco, *Low-rank Toeplitz matrix estimation via random ultra-sparse rulers*, in Proceedings of the 2020 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020, pp. 4796–4800.

[40] F. T. Luk and S. Qiao, *A symmetric rank-revealing Toeplitz matrix decomposition*, Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, 14 (1996), pp. 19–28.

[41] J. Ma, G. Y. Li, and B. H. Juang, *Signal processing in cognitive radio*, Proceedings of the IEEE, 97 (2009), pp. 805–823.

[42] J. Munkhammar, L. Mattsson, and J. Rydén, *Polynomial probability distribution estimation using the method of moments*, PloS One, 12 (2017).

[43] C. Musco and C. Musco, *Recursive sampling for the Nyström method*, in Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.

[44] G. Ongie and M. Jacob, *A fast algorithm for convolutional structured low-rank matrix recovery*, IEEE Transactions on Computational Imaging, 3 (2017), pp. 535–550.

[45] V. Y. Pan and Z. Q. Chen, *The complexity of the matrix eigenproblem*, in Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC), 1999, pp. 507–516.

[46] H. Park, L. Zhang, and J. B. Rosen, *Low rank approximation of a Hankel matrix by structured total least norm*, BIT Numerical Mathematics, 39 (1999), pp. 757–779.

[47] H. Qiao and P. Pal, *Gridless line spectrum estimation and low-rank Toeplitz matrix compression using structured samplers: A regularization-free approach*, IEEE Transactions on Signal Processing, 65 (2017), pp. 2221–2236.

[48] A. Rinaldo, *Lecture 7*, in Lecture notes on Advanced Statistical Theory, CMU [link], 2017.

[49] C. S. Ruf, C. T. Swift, A. B. Tanner, and D. M. Le Vine, *Interferometric synthetic aperture microwave radiometry for the remote sensing of the earth*, IEEE Transactions on geoscience and remote sensing, 26 (1988), pp. 597–611.

[50] T. Sarlos, *Improved approximation algorithms for large matrices via random projections*, in 2006 47th annual IEEE symposium on foundations of computer science (FOCS'06), IEEE, 2006, pp. 143–152.

[51] X. Shi and D. P. Woodruff, *Sublinear time numerical linear algebra for structured matrices*, in AAAI Conference on Artificial Intelligence, 2019.

[52] D. L. Snyder, J. A. O'Sullivan, and M. I. Miller, *The use of maximum likelihood estimation for forming images of diffuse radar targets from delay-doppler data*, IEEE Transactions on Information Theory, 35 (1989), pp. 536–548.

[53] R. Wen and Y. Fu, *Toeplitz matrix completion via a low-rank approximation algorithm*, Journal of Inequalities and Applications, 2020 (2020), pp. 1–13.

[54] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Foundations and Trends in Theoretical Computer Science, 10 (2014), pp. 1–157.

[55] X. Wu, W.-P. Zhu, and J. Yan, *A Toeplitz covariance matrix reconstruction approach for direction-of-arrival estimation*, IEEE Transactions on Vehicular Technology, (2017).

[56] Y. Xi, J. Xia, S. Cauley, and V. Balakrishnan, *Superfast and stable structured solvers for Toeplitz least squares via randomized sampling*, SIAM Journal on Matrix Analysis and Applications, 35 (2014), pp. 44–72.

[57] J. Xia, Y. Xi, and M. Gu, *A superfast structured solver for Toeplitz linear systems via randomized sampling*, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 837–858.

[58] T. Yasuda, D. Woodruff, and M. Fernandez, *Tight kernel query complexity of kernel ridge regression and kernel k-means clustering*, in Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 7055–7063.