Toeplitz Low-Rank Approximation with Sublinear Query Complexity

Michael Kapralov[1], Hannah Lawrence[2], Mikhail Makarov[1],
Cameron Musco[3], and Kshiteej Sheth[1]

[1]EPFL
[2]MIT
[3]University of Massachusetts Amherst

Abstract

We present a sublinear query algorithm for outputting a near-optimal low-rank approximation to any positive semidefinite Toeplitz matrix $T \in \mathbb{R}^{d\times d}$. In particular, for any integer rank $k \leq d$ and $\epsilon, \delta > 0$, our algorithm makes $\widetilde{O}\left(k^2 \cdot \log(1/\delta) \cdot \text{poly}(1/\epsilon)\right)$ queries to the entries of T and outputs a rank $\widetilde{O}\left(k \cdot \log(1/\delta)/\epsilon\right)$ matrix $\widetilde{T} \in \mathbb{R}^{d\times d}$ such that $\|T - \widetilde{T}\|_F \leq (1+\epsilon) \cdot \|T - T_k\|_F + \delta\|T\|_F$. Here, $\|\cdot\|_F$ is the Frobenius norm and T_k is the optimal rank-k approximation to T, given by projection onto its top k eigenvectors. $\widetilde{O}(\cdot)$ hides polylog(d) factors. Our algorithm is *structure-preserving*, in that the approximation $\widetilde{T}$ is also Toeplitz. A key technical contribution is a proof that any positive semidefinite Toeplitz matrix in fact has a near-optimal low-rank approximation which is itself Toeplitz. Surprisingly, this basic existence result was not previously known. Building on this result, along with the well-established off-grid Fourier structure of Toeplitz matrices [Cybenko'82], we show that Toeplitz $\tilde{T}$ with near optimal error can be recovered with a small number of random queries via a leverage-score-based off-grid sparse Fourier sampling scheme.

Contents

1 Introduction.

In scientific computing, engineering, and signal processing, highly structured matrices – such as Toeplitz, circulant, hierarchical, and graph-structured matrices – arise in many problems, often due to the discretization of an underlying physical system. Such matrices are intensely studied and in many cases admit fast, even near-linear time algorithms for solving core linear algebraic problems.

We investigate the possibility of *sublinear* algorithms for such highly structured matrix classes, focusing in particular on Toeplitz matrices. A matrix $T \in \mathbb{R}^{d\times d}$ is Toeplitz if it is constant along its diagonals, i.e., for any $i, j, k, l \in [d]$, $i - j = k - l$ implies that $T_{i,j} = T_{k,l}$. Toeplitz matrices are ubiquitous in image and signal processing, where they arise as covariance matrices of stationary processes – i.e., when the covariance between two measurements depends only on their distance in space or time [17]. They also arise in queuing theory, the solution of differential and integral equations, control theory, approximation theory, and beyond – see [6] for a review of their many applications. Reversing the rows of a Toeplitz matrix yields a Hankel matrix. These matrices also find wide applications in signal processing, system identification, and numerical computing [16, 25].

A $d \times d$ Toeplitz matrix is specified by just $O(d)$ parameters, and Toeplitz matrices are a classic example of *low displacement rank matrices*. They admit fast algorithms for many problems. A Toeplitz matrix can be multiplied by a vector in $O(d \log d)$ time via a fast Fourier transform. Toeplitz linear systems can be solved exactly in $O(d^2)$ time via the Levinson algorithm, and to high precision in $O(d \log^2 d)$ time using randomized methods [39, 40]. A full Toeplitz eigendecomposition can be performed in $O(d^2 \log d)$ time [29]. However, little is known about algorithms for Toeplitz matrices with provable correctness on general input instances and running time or query complexity scaling *sublinearly* in the input size d.

1.1 Our contributions.

In this work, we study sublinear query algorithms for symmetric positive semidefinite (PSD) Toeplitz matrix low-rank approximation, which is a widely-studied problem [19,30] with applications to signal and image recovery [7, 23, 28], signal direction of arrival estimation [1, 21], and beyond. We show that a low-rank approximation to any PSD Toeplitz matrix with near-optimal error in the Frobenius norm can be computed using a sublinear number of queries to the input matrix. In particular, letting $\|M\|_F = \left(\sum_{i=1}^d \sum_{j=1}^d M_{ij}^2\right)^{1/2}$ denote the Frobenius norm of a matrix M, and letting $\widetilde{O}(\cdot)$ hide poly-logarithmic factors in the argument and in d, we have the following theorem:

Theorem 1 (Sublinear query Toeplitz low-rank approximation)**.** *There is a randomized algorithm that, given any PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$, $\epsilon, \delta \in (0, 1)$, and integer $k \leq d$, reads $\widetilde{O}\left(\frac{k^2 \log(1/\delta)}{\epsilon^6}\right)$ entries of T and returns a symmetric Toeplitz matrix $\widetilde{T}$ with rank $\widetilde{O}\left(\frac{k \log(1/\delta)}{\epsilon}\right)$ satisfying with probability at least $97/100$,*

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

where $T_k = \underset{B:\mathrm{rank}(B)\leq k}{\arg\min} \|T - B\|_F$ is the best rank-k approximation to T in the Frobenius norm.

Theorem 1 gives a near-relative error approximation, up to an additive $\delta\|T\|_F$. However, the dependence on δ in the sample complexity and output rank is just logarithmic. Thus, this additive term is comparable to the error that would be introduced in any practical algorithm due to round-off error. To the best of our knowledge, ours is the first sublinear query algorithm for Toeplitz low-rank approximation achieving near-relative error. See Section 1.2 for a detailed comparison to prior work.

Structure-preservation and bicriteria approximation. Observe that Theorem 1 outputs a low-rank Toeplitz matrix $\widetilde{T}$, even though the optimal low-rank approximation T_k will in general not be Toeplitz. This *structure-preserving low-rank approximation* is desirable in many applications [7, 10]. It allows fast methods for Toeplitz matrices to be applied to $\widetilde{T}$ itself, and is useful when the Toeplitz structure of the approximation has a physical meaning, in applications such as direction of arrival estimation [21]. As we will discuss, $\widetilde{T}$'s Toeplitz structure is also key to achieving sublinear query complexity – it allows us to recover $\widetilde{T}$ via query efficient sparse Fourier transform techniques.

Since $\widetilde{T}$ is Toeplitz, the approximation bound of Theorem 1 is *bicriteria* – i.e., $\widetilde{T}$ has rank larger than the input rank k. This is necessary since as ϵ, δ go to zero, $\|T - \widetilde{T}\|_F$ becomes arbitrarily close to $\|T - T_k\|_F$. However, one can find simple examples of T and k where there is a fixed gap between $\|T - T_k\|_F$ and the error of the best rank-k Toeplitz approximation to T. See Figure 1. It is an interesting open question if a sublinear query algorithm for Toeplitz low-rank approximation with the same near-relative error bound as Theorem 1 can be achieved without any bicriteria approximation, by abandoning the structure-preserving guarantee and allowing $\widetilde{T}$ to be non-Toeplitz.

$$T = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \qquad T_1 = \frac{2+\sqrt{2}}{4} \cdot \begin{bmatrix} 1 & \sqrt{2} & 1 \\ \sqrt{2} & 2 & \sqrt{2} \\ 1 & \sqrt{2} & 1 \end{bmatrix} \qquad T_{1,toep} = \frac{10}{9} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Figure 1: Example of a 3×3 positive semidefinite Toeplitz matrix T whose best rank-1 approximation $T_1 = \arg\min_{B:\text{rank}(B)\leq k} \|T - B\|_F$ differs from its best rank-1 Toeplitz approximation $T_{1,toep} = \arg\min_{B:\text{rank}(B)\leq k,\, \text{B is Toeplitz}} \|T - B\|_F$. We can check that $\|T - T_{1,toep}\|_F - \|T - T_1\|_F \approx 0.1271$. Thus, any Toeplitz low-rank approximation to T achieving error $(1+\epsilon)\|T - T_1\|_F$ for small enough ϵ must have rank > 1. I.e., it must be a bicriteria approximation, as in Theorem 1. T_1 is computed via projection onto T's top eigenvector. $T_{1,toep}$ is computed by observing that any rank-1 Toeplitz matrix must have all entries of the same magnitude. Thus, since T has all positive entries, the optimal approximation is the all ones matrix scaled by the mean entry in T.

Runtime. In Theorem 1 we focus on query complexity rather than runtime, and our algorithm requires time exponential in $\widetilde{O}(k/\epsilon)$ to identify $\widetilde{T}$ via a brute-force-search-based off-grid sparse Fourier transform algorithm. This is not sublinear time unless k is sublogarithmic in d. However, we conjecture that sublinear runtime is possible by adapting efficient off-grid sparse Fourier transform techniques [8] to find $\widetilde{T}$.

Existence proof. A key challenge in proving Theorem 1 is to demonstrate that a Toeplitz $\widetilde{T}$ achieving a near-optimal Frobenius norm low-rank approximation even *exists*. Surprisingly, this was not previously known. We show

Theorem 2 (Existence of near-optimal Toeplitz low-rank approximation)**.** *For any PSD Toeplitz $T \in \mathbb{R}^{d\times d}$, $\epsilon, \delta \in (0,1)$ and integer $k \leq d$ there exists a symmetric Toeplitz matrix $\widetilde{T} \in \mathbb{R}^{d\times d}$ of rank $\widetilde{O}\left(\frac{k\log(1/\delta)}{\epsilon}\right)$ such that*

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

where $T_k = \underset{B:\text{rank}(B)\leq k}{\arg\min} \|T - B\|_F$ is the best rank-k approximation to T in the Frobenius norm.

We prove a similar result for spectral norm low-rank approximation – see Theorem 5. One natural approach to proving Theorem 2 (and Theorem 5) is to note that, intuitively, if T is close to

low-rank, its optimal low-rank approximation T_k – given by projection onto its top k eigenvectors – should be nearly Toeplitz. Of course, rounding this matrix to be Toeplitz (by replacing the entries on each diagonal by their average) can increase its rank, but one could plausibly try to bound the increase in rank. This, however, would not naturally lead to a sublinear query algorithm for recovering a low rank approximation. We take a different approach, exploiting the classical Fourier structure of Toeplitz matrices, namely the Vandermonde decomposition [12]. In essence, instead of rounding the best rank-k approximation to Toeplitz, we round the matrix T itself to a rank-(almost)k matrix *in Fourier domain*, thereby preserving the Toeplitz property throughout. This lets us use a number of powerful ideas from the literature on recovering Fourier sparse functions from few measurements, and ultimately leads to a sublinear query algorithm – see Section 1.3 for a more detailed overview.

We believe that Theorem 2 is of interest beyond its application to proving Theorem 1. It remains an open question if the $\delta\|T\|_F$ term can be removed. Although we stated our guarantees as applying to exactly Toeplitz matrices, they extend fairly directly to near-Toeplitz input matrices. For example, our main existence result in Theorem 2 easily extends to arbitrary matrices which are close in Frobenius norm to a PSD Toeplitz matrix, by the triangle inequality. Moreover, the sublinear query complexity algorithm of Theorem 1 will extend to any matrix whose first column is close to the first column of a PSD Toeplitz matrix, according to the weighted ℓ_2 norm, as defined in Claim 4.3. It would be interesting to extend this further, e.g. to any matrix that is close to a PSD Toeplitz matrix in the Frobenius norm. Additionally, while $\widetilde{T}$ must have a rank larger than k when ϵ, δ are sufficiently small (see Figure 1), identifying the minimum rank required to achieve the given error bound is also a very interesting problem.

1.2 Related work.

Significant prior work in numerical linear algebra and signal processing has studied the low-rank approximation of Toeplitz and, relatedly, Hankel matrices [7, 19, 21, 23, 28, 30]. In signal processing applications, the input Toeplitz matrix T is often a covariance matrix and thus positive semidefinite.

Due to the fact that a $d \times d$ Toeplitz matrix can be multiplied by a vector in $O(d \log d)$ time via fast Fourier transform, a near-optimal low-rank approximation can be computed in near-linear time. In particular, Shi and Woodruff present an algorithm that outputs rank-k $\widetilde{T}$ with $\|T-\widetilde{T}\|_F \leq (1+\epsilon)\|T-T_k\|_F$ in $\widetilde{O}(d+\mathsf{poly}(k/\epsilon))$ time [33]. Significant other work focuses on computing a near-optimal low-rank approximation that preserves Toeplitz structure, as we do in Theorem 1. Unlike unconstrained low-rank approximation, where the optimal solution can be computed directly via eigendecomposition, no simple characterization of the optimal structure-preserving Toeplitz low-rank approximation is known [10]. Computing such an optimal approximation in polynomial time remains open outside the special cases of $k = 1$ and $k = d-1$ [10,20]. Practical heuristics apply a range of techniques, based on convex relaxation [7, 16, 28], alternating minimization [10, 37], and sparse Fourier transform [21]. Observe that our main result, Theorem 1 directly gives a near-relative error bicriteria approximation algorithm for the optimal structure-preserving Toeplitz low-rank approximation problem, since $\|T-T_k\|_F = \min_{B:\text{rank}(B)\leq k} \|T-B\|_F \leq \min_{B:\text{rank}(B)\leq k, B \text{ is Toeplitz}} \|T-B\|_F$.

Several works also investigate sublinear query algorithms for Toeplitz matrices [1,9,15,22,31]. In the signal processing community, these algorithms are often framed in terms of *sparse array methods*, which can be thought of as reading a small principal submatrix of T from which an approximation to the full matrix can be recovered. Most closely related to our work is that of Eldar, Li, Musco, and Musco [15], which focuses on approximating a PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$ given samples from a d-dimensional Gaussian distribution with covariance T. They focus on minimizing both the number of samples taken from the distribution, as well as the number of entries read from

each d-dimensional sample. For nearly low-rank T, they present an algorithm that estimates T via a low-rank approximation of the sample covariance matrix, computed from a small number of randomly selected entries of that matrix. One can check that this algorithm can be directly applied to T itself. It gives $\widetilde{O}(k^2 \cdot \log(1/\delta))$ query complexity, and outputs $\widetilde{T}$ with rank $O(k)$ achieving error

$$\|T - \widetilde{T}\|_F \leq C \cdot \sqrt{\mathrm{tr}(T) \cdot \left(\|T - T_k\|_2 + \frac{\mathrm{tr}(T - T_k)}{k} \right)} + \delta \|T\|_F,$$

for some constant $C \gg 1$. The above guarantee is non-standard, and due to the $\mathrm{tr}(T)$ term, it can be much weaker than a relative error low-rank approximation. For example, $\mathrm{tr}(T) \gg \|T - T_k\|_F$ if the eigenvalues of T decay quickly, which is typically the case in settings where low-rank approximation is applied. In fact, we can observe that the above guarantee is strictly weaker than that of Theorem 1: letting the eigenvalues of T be denoted $\lambda_1(T) \geq \ldots \geq \lambda_n(T) \geq 0$,

$$\mathrm{tr}(T) \cdot \|T - T_k\|_2 = \sum_{i=1}^{n} \lambda_i(T) \cdot \lambda_{k+1}(T) \geq \sum_{i=k+1}^{n} \lambda_i(T)^2 = \|T - T_k\|_F^2.$$

Finally, the algorithm of [15] outputs $\widetilde{T}$ which is not structure-preserving as in Theorem 1. Note that in this sampling model of [15], it is not possible to achieve a 'near-relative' error guarantee like we do. For example, even if the matrix is exactly rank-k, the sampling will lead to error approximately $\delta\|T\|_2$, with a polynomial in δ number of samples. In contrast, our algorithm (with exact access to the entries of T) gives error $\delta\|T\|_F$ with just a logarithmic dependence on δ. Consider e.g., the setting when T is just a matrix with every entry equal to α for $\alpha = \Theta(1)$. Approximating α to error $\pm\delta$ will require $1/\delta^2$ 'vector samples' in the model of [15], for any algorithm.

However, our algorithm is indeed robust to noise, and so a similar guarantee with some additional additive error is achievable in the setting of [15]. One can take $\mathsf{poly}(d, 1/\epsilon)$ vector samples and then using e.g., Claim 2.2 of [15] one can argue that the sample covariance matrix will approximate the true covariance matrix, in that its first column will be close to the first column of the true matrix in the weighted ℓ_2 norm, as defined in Claim 4.3. We can then directly apply our algorithm to this sample covariance matrix. An interesting open problem here is to improve the vector sample complexity to just depend on k, and in general, to explore vector/entrywise sample complexity in more depth, as is done in [15].

Beyond Toeplitz matrices, significant recent work has focused on sublinear time low-rank approximation algorithms for other structured matrix classes. This includes positive semidefinite matrices [4, 26], distance matrices [5, 18], and kernel matrices [2, 26, 41].

1.3 Technical overview.

The main results of the paper are the proof of the existence of a near optimal low-rank Toeplitz approximation and the algorithm to recover it, presented in Sections 3 and 4 respectively. In this section we give an overview of the techniques used to achieve both results.

We start by introducing some notation. From classical works on Toeplitz matrices [12] it is known that any PSD Toeplitz matrix admits a *Vandermonde decomposition* $T = F_S D F_S^*$, where D is a diagonal matrix with positive entries and F_S is a Fourier matrix. The Vandermonde decomposition is central to our work. We define Fourier matrices formally now.

Definition 1.1 (Frequency vector)**.** *For any frequency $f \in \mathbb{C}$, we define the* frequency vector *$v(f) \in \mathbb{C}^d$ as the column vector $[1, e^{2\pi i f}, \ldots, e^{2\pi i f(d-1)}]$.*

Definition 1.2 (Symmetric Fourier matrix)**.** *For any set $S = \{f_1, f_2, \ldots, f_s\} \subset [0, 1/2]$, let $F_S \in \mathbb{C}^{d \times 2s}$ be the Fourier matrix defined by the frequencies in S as $F_S = [F_{+S}; F_{-S}]$. For every $j \in \{1, 2, \ldots, s\}$ the j-th column of F_{+S} is $v(f_j)$. Similarly, for every $j \in \{1, 2, \ldots, s\}$ the j-th column of F_{-S} is $v(-f_j)$.*

The formal statement of the Vandermonde decomposition is as follows.

Theorem 3 (Vandermonde decomposition, Corollary 1 of [12])**.** *Any real-valued PSD Toeplitz matrix $T \in \mathbb{R}^{d \times d}$ of rank r can be expressed as $F_S D F_S^*$ where $F_S \in \mathbb{C}^{d \times r}$ is a symmetric Fourier matrix with the set of frequencies $S \subseteq [0, 1/2]$ satisfying $|S| = r/2$, and $D \in \mathbb{R}^{r \times r}$ is a diagonal matrix with r positive entries. Moreover, for any $f \in S$ the values in D corresponding to columns $v(f), v(-f)$, which we refer to as the weights of $f, -f$ respectively, are identical. Thus D is uniquely defined by values $\{a_f\}_{f \in S}$.*

A basic Fourier-based approach and why it fails. One approach to show the existence of a near optimal low-rank approximation, which itself is Toeplitz, is to show that retaining only the top (nearly) k entries in the diagonal matrix D in the Vandermonde decomposition $T = F_S D F_S^*$ and zeroing out the rest would give a near optimal low-rank approximation. This approach is natural, as this operation trivially preserves the Toeplitz structure. However, to formally argue that such an approach works would require us to relate the Vandermonde decomposition to the eigendecomposition of T. This is challenging in general because the Fourier matrix F_S could potentially be highly ill-conditioned [24], whereas the eigenvector matrix of T has condition number 1. Therefore, these decompositions could be very far from each other.

The special case of circulant matrices. For the special but important case of circulant matrices, as defined below, relating these two decompositions turns out to be much easier.

Definition 1.3. *A Toeplitz matrix $C \in \mathbb{R}^{d \times d}$ is called* circulant *if there exists a vector $c \in \mathbb{R}^d$ such that $C_{i,j} = c_{i+j-1 \mod d}$.*

Essentially, a circulant matrix $C \in \mathbb{R}^{d \times d}$ is a matrix composed of all cyclic permutations of some vector $c \in \mathbb{R}^d$. Any circulant matrix is a Toeplitz matrix. A Vandermonde decomposition of a symmetric circulant matrix has the property that all of its frequencies are multiples of $1/d$ [12]. Thus the F_S matrix is actually the discrete Fourier transform matrix, and its frequency vectors are orthogonal. In particular, the Vandermonde and eigendecompositions are identical in this special case, and we can easily argue that the best rank-k approximation to a symmetric circulant matrix itself is circulant! This is because the best rank-k approximation is given by retaining the top k elements of D in $F_S D F_S^*$, which has the circulant property by definition.

Toeplitz matrices. In a similar spirit to the easy case of circulant matrices mentioned above, for a general Toeplitz matrix T, it is natural to ask whether it is true that if the Vandermonde decomposition of T contains k frequencies with large corresponding values, then the number of large eigenvalues of T is also $\Omega(k)$. This is in general not the case. For example, if $S = \{f, -f, f + \epsilon, -f - \epsilon \ldots, f + (k/2)\epsilon, -f - (k/2)\epsilon\}$ for some $f \in [0, 1]$, D is the identity matrix and $\epsilon \to 0$, then both F_S and T tend to a rank 2 matrix, with the third eigenvalue tending towards 0. This happens because when the frequencies are close, their frequency vectors are highly correlated. Thus, for the purposes of our analysis it makes sense to consider a group of close frequencies as a single entity, and this observation motivates our proof plan.

We first analyze the case when all frequencies in S are 'close' (this analysis is presented in Section 3.1), then analyze the interaction of 'clusters' of frequencies in Section 3.2 and Section 3.3,

and finally derive the main results of Section 3, namely Theorem 5 and Theorem 2, in Section 3.4. Theorem 5 and Theorem 2 state that for any k, there exists a symmetric Toeplitz matrix $\widetilde{T}$ of rank *almost* k which is *almost* as good a low-rank approximation to T as T_k, the best rank-k approximation to T, in the spectral or Frobenius norm, respectively. We now give a high-level overview of the organization of Section 3 and constituent proofs.

Clustered case. In Subsection 3.1, we consider the case where the distance between each pair of frequencies in S is at most Δ for some $\Delta \in [0, 1/2]$ (we ultimately choose $\Delta = O(1/d)$). This case is referred to as the *clustered* case, formalized in the definition below.

Definition 1.4. *For $f^* \in [0, 1/2]$ and $\Delta \in [0, 1/2]$ we say that a Toeplitz matrix T with Vandermonde decomposition $T = F_S D F_S^*$ is (f^*, Δ)-clustered if $S \subset [f^* - \Delta, f^* + \Delta]$ for some $f^* \in [0, 1/2]$ and $\Delta \geq 0$. Thus we can write any $f \in S$ as $f = f^* + r_f$ for some $r_f \in [-\Delta, +\Delta]$.*

Intuitively, in this case all the columns (i.e. frequency vectors) of F_S are nearly identical, so we would expect F_S (and thus T) to be close to an almost constant rank matrix. Furthermore, we can even show that this matrix $\widetilde{T}$ is Toeplitz itself, as specified by the following lemma.

Lemma 1.5. *There exists a universal constant $C_1 > 0$ such that given any symmetric PSD Toeplitz $T = F_S D F_S^*$ that is (f^*, Δ)-clustered for some $f^* \in [0, 1/2]$ and $\Delta \leq 1/d$ and $0 < \varepsilon, \delta, \gamma < 1$ satisfying $\gamma \leq \varepsilon / (tr(T) 2^{C_1 \log^7 d})$ the following conditions hold. There exists a symmetric Toeplitz matrix $\widetilde{T} = F_{\widetilde{S}} \widetilde{D} F_{\widetilde{S}}^*$ of rank at most $O(\ell)$ for $\ell = O(\log d + \log(1/\delta))$ such that*

1. *$\widetilde{S} = \{f^* + j\gamma\}_{j=1}^{\ell+1} \cup \{f^* - j\gamma\}_{j=1}^{\ell+1}$ and $\widetilde{D}$ is a diagonal matrix such that for any $f \in \widetilde{S}$ the weights corresponding to f and $-f$ are identical.*
2. *$\|T - \widetilde{T}\|_F \leq \delta(\sum_{f \in S} a_f) + \varepsilon d$.*

Remark 1.6. *The value of the diagonal elements in $\widetilde{D}$ depends on ε, and in particular, the value of the diagonal elements would go to infinity as ε goes to zero. However, we do not state this tradeoff explicitly in the lemma, because we apply it later in Section 4 to obtain leverage score upper bounds on Fourier sparse functions based on the work of [8]. These bounds do not depend on the magnitude of the coefficients in the function, but only on its sparsity (as long as the coefficients are finite). Thus ε can be set to any strictly positive number of one's choice.*

Note that the error term in point 2 of the above theorem is negligible, as the dependence of the rank of $\widetilde{T}$ on $1/\delta$ is logarithmic. The proof of this result uses tools from polynomial approximation and is achieved in three steps. First, we observe that since the t-th entry in the first column of T is the linear combination of complex exponentials of form $e^{2\pi i (f^* + r_f)t}$ where $r_f \leq \Delta$, this column can be well approximated by a sum of polynomials of the form $p_j(t) = e^{(2\pi i f^* t)} \sum_{m=0}^{l} a_f \frac{(2\pi i r_f t)^m}{m!}$ instead through the Taylor series. We then show that this sum of polynomials can be approximated by another polynomial on complex exponentials of the form $\widetilde{p}(t) = e^{(2\pi i f^* t)} \sum_{m=-l-1}^{l+1} \alpha_m (e^{2\pi i m \gamma t})$, from which we can then finally construct the matrix $\widetilde{T}$. The details are presented in Section 3.1.

Relating Vandermonde and eigenvalue decompositions. We now discuss the high-level strategy of how we relate the Vandermonde and eigenvalue decompositions for general Toeplitz matrices. Central to our proof is the notion of a *bucket*. We divide the interval $[0, 1/2]$ into $d/2$ equal-sized sub-intervals, and refer to the frequencies of F_S (recall $T = F_S D F_S^*$ is the Vandermonde decomposition of T) in each sub-interval as a *bucket*. Each bucket corresponds to a group of close frequencies, corresponding to the **clustered case** discussed above. The *weight* of a bucket is the

sum of the coefficients in D with corresponding frequencies in that bucket. Using this notion, our first structural result that relates the two decompositions is as follows[1]:

Lemma 1.7. *For every $\lambda, k > 0$, if T has at least k buckets with weight at least λ, then T has at least $k/\log^3 d$ eigenvalues that have value $\Omega(d\lambda/\log d)$.*

At a high level, this result follows from first showing that the interaction between frequencies, measured as an inner product $v(f_1)^* v(f_2)$ of their frequency vectors, has a harmonic decay as a function of the distance between f_1 and f_2. Then, we subsample the set of all heavy buckets to ensure that interaction between any remaining pair of buckets is small. Finally, we use a strengthening of Gershgorin's circle theorem for block matrices to guarantee that the contribution to each eigenvalue is dominated by only one bucket.

In Subsection 3.3, we apply the same tools to show a complementary result[2]:

Lemma 1.8. *For every $\lambda \geq 0$, if all buckets of T have weight at most λ, then $\|T\|_2 \leq O(d\lambda \log d)$.*

In other words, a uniform bound on bucket weight implies a uniform bound over eigenvalues. Combined with the Lemma 1.7, this implies that the bucket structure characterizes the eigenvalue structure in some sense, up to polylogarithmic precision. These structural statements can also be seen as providing fine grained insights into the eigenvalue structure of arbitrary off-grid Fourier matrices beyond just condition number bounds [24]. This completes the overview of our approach of relating the potentially ill-conditioned Vandermonde decompostion and the eigenvalue decomposition for general Toeplitz matrices.

Putting it together: Toeplitz low-rank approximation in Frobenius and spectral norm. Finally, Subsection 3.4 uses the structural statements of Subsections 3.2 and 3.3 to prove Theorems 5 and 2.

We first prove Theorem 5, the high level idea of which is as follows. Consider the Toeplitz matrix $\widetilde{T}$ obtained by taking only those buckets of weight at least $\widetilde{\Omega}(\lambda_{k+1}(T)/d)$, which we will call *heavy*. By Lemma 1.8, $\|T - \widetilde{T}\|_2 \leq \widetilde{O}(\lambda_{k+1}(T)) \leq \widetilde{O}(1)\|T - T_k\|_2$. Moreover, by Lemma 1.7 the matrix $\widetilde{T}$ cannot contain more than $\widetilde{O}(k)$ heavy buckets. Each bucket corresponds to a Toeplitz matrix with clustered frequencies, so replacing each bucket with its $\widetilde{O}(1)$-rank approximation by Lemma 1.5 incurs the additive error of $\delta\|T\|_F$ and finishes the proof.

In general, however, a spectral norm low-rank approximation guarantee does not imply a Frobenius norm low-rank approximation guarantee. The main idea behind Theorem 2 is to consider the $\widetilde{T}$ obtained by taking a few more buckets than in the proof of Theorem 5. Doing so ensures that T is even closer to $\widetilde{T}$, such that even the top k eigenvalues of $T - \widetilde{T}$ can be bounded in terms of the $d-k$ smallest eigenvalues of T. In particular, let $\widetilde{T}$ be the Toeplitz matrix obtained by taking only those buckets of weight at least $\widetilde{\Omega}(\lambda_{\widetilde{O}(k/\epsilon)}(T)/d)$. Again by Lemma 1.8, $\widetilde{T}$ cannot contain more than $\widetilde{O}(k/\epsilon)$ heavy buckets. Lemma 1.7 bounds each of the top $k+1$ eigenvalues of $T-\widetilde{T}$ by $\widetilde{O}(\lambda_{\widetilde{\Omega}(k/\epsilon)}(T))$, so their contribution to $\|T - \widetilde{T}\|_F^2$ is at most $k\widetilde{O}(\lambda_{\widetilde{\Omega}(k/\epsilon)}(T)^2) \leq \epsilon \sum_{i=k+1}^{\widetilde{O}(k/\epsilon)} \lambda_i(T)^2 \leq \epsilon\|T - T_k\|_F^2$. On the other hand, since $T - \widetilde{T} \preceq T$, we can bound any eigenvalue of $T - \widetilde{T}$ except the top $k+1$ by the corresponding eigenvalue of T. Thus the remaining eigenvalues' contribution to $\|T - \widetilde{T}\|_F^2$ can be bounded by $\sum_{i=k+2}^{d} \lambda_i^2(T) = \|T - T_k\|_F^2$. Overall this results in $\|T - \widetilde{T}\|_F \leq (1+\epsilon)\|T - T_k\|_F$.

[1] Note that the theorem below is a natural relaxation of the claim that k frequencies with coefficients at least λ in the Vandermonde decomposition of a circulant matrix imply at least k eigenvalues of value at least λ.

[2] Again, note that the theorem below is a natural relaxation of the claim that if all frequencies in the Vandermonde decomposition of a circulant matrix have coefficients bounded by λ, then the spectral norm of T is bounded by $d\lambda$.

Replacing each bucket in $\widetilde{T}$ with the $\widetilde{O}(1)$-rank matrix from Lemma 1.5 incurs the additive error and completes the proof.

Sublinear query algorithm. Now we present the main ideas behind our recovery algorithm, Algorithm 1, and its theoretical guarantees as stated in Theorem 1.

We treat the $\widetilde{T} = F_{\widetilde{S}}\widetilde{D}F_{\widetilde{S}}^*$ that is guaranteed to exist by Theorem 2 as the true underlying matrix and T as a noisy version of it that we have access to. Note that T is completely determined by its first column $T_1 \in \mathbb{R}^d$ and the first column of $\widetilde{T}$ is $F_{\widetilde{S}}\widetilde{d}$ where $\widetilde{d} \in \mathbb{R}^{\widetilde{O}(k/\epsilon)}$ is the vector of values on the diagonal of $\widetilde{D}$.

This suggests the following strategy - if the algorithm knew $\widetilde{S}$ exactly, it could try to find $\widetilde{d}$ by solving the regression problem $\arg\min_{a\in\mathbb{R}^d} \|F_{\widetilde{S}}a - T_1\|_2$. We can solve this regression problem approximately without reading too many entries of T_1 using the technique of leverage score sampling [13], and using universal leverage score upper bounds for off-grid Fourier matrices that are independent of the set of frequencies $\widetilde{S}$ defining $F_{\widetilde{S}}$ [8, 15].

The first issue with this strategy is the fact that the error bounds in fitting the first column T_1 would not translate to error bounds on fitting the entire matrix T in the Frobenius norm. This is because each entry of T_1 appears a different number of times in T. For example $T_{1,1}$ appears d times whereas $T_{d,1}$ only appears twice. To circumvent this, we need to solve an alternate regression problem where each row of $[F_{\widetilde{S}}; T_1]$ is weighted differently to account for the asymmetry in fitting the first column T_1 versus fitting the entire matrix T. Using a geometric grouping technique we are able to obtain leverage score upper bounds for this modified regression problem that only suffer a logarithmic overhead compared to those known for the unweighted case [15].

The second issue is the assumption on the knowledge of $\widetilde{S}$ which can be circumvented by brute-force searching over the set of all possible $\widetilde{S}$ and choose the one with the smallest error.

This suggests the following algorithm. It first obtains a sample set containing a few entries of T_1 using this universal leverage score distribution that is valid for any off-grid Fourier matrix, then searches for all possible sets of $\widetilde{O}(k)$ frequencies that could be the set $\widetilde{S}$. For each such guess of $\widetilde{S}$ it finds an approximately optimal $\widetilde{d}$ by approximately solving the weighted regression problem described previously, using the same sample set that works for any set $\widetilde{S}$. Finally, it returns the best $\widetilde{d}$ among all the guesses.

To prove that this algorithm works with good probability is still challenging. This is because standard sample-efficient regression results based on leverage score sampling [32, 38] do not suffice in our setting for two reasons. First, we search over many possible $\widetilde{S}$ and thus solve many regression problems – we must take a union bound to argue that our sample set gives a good approximation for all these problems. This presents an issue for standard results, which typically require sample complexity depending linearly on $1/\eta$, where $1-\eta$ is the probability of success. Second, we require identifying a frequency set $\widetilde{S}$ with near minimal error – i.e., we must compare the errors of the many regression problems that we solve. Standard leverage score based sampling results however, typically do not output an estimate of the actual regression error, making it impossible to chose a near optimal $\widetilde{S}$. To overcome these issues, we use a two stage algorithm, as in [27]. Following techniques of [15], we first find $\tilde{S}_1$ achieving a constant factor of the optimal. We then find $\tilde{S}_2$ which gives a $(1+\epsilon)$ relative error fit to the residual remaining after regressing onto $\tilde{S}_1$. Our final frequency subset is $\tilde{S}_1 \cup \tilde{S}_2$. This approach allows us to use a modified analysis of leverage score sampling for fitting $\tilde{S}_2$, which both gives high probability bounds (with $\log(1/\eta)$ dependence for failure probability η) and regression error estimates, as required. A detailed description of the algorithm together with its analysis are presented in Algorithm 4.

2 Notation and preliminaries.

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

2.1 Notation.

For any functions $f, g : \mathbb{R} \to \mathbb{R}$, $f(n) \lesssim g(n)$ means that there exists a constant $C > 0$ such that $f(n) \leq Cg(n)$. For any positive integer n, let $[n] = \{1, 2, \ldots, n\}$. For any set $S \subset \mathbb{R}$, let $-S$ denote the set obtained by negating each element in S and let $+S$ denote S itself. 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. For any vector $x \in \mathbb{C}^d$, let $\|x\|_2 = \sqrt{x^*x}$ denote its ℓ_2 norm. For a matrix A with d columns, let $\|A\|_2 = \sup_{x \in \mathbb{C}^d} \|Ax\|_2 / \|x\|_2$ denote its spectral norm and $\|A\|_F = \sqrt{\sum_{i \in [d]} \sum_{j \in [d]} A_{i,j}^2}$ denote its Frobenius norm. For a square matrix A, let $tr(A)$ denote its trace.

A Hermitian matrix $A \in \mathbb{C}^{d \times d}$ is positive semidefinite (PSD) if for all $x \in \mathbb{C}^d$, $x^*Ax \geq 0$. Let $\lambda_1(A) \geq \ldots \geq \lambda_d(A) \geq 0$ 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 and 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 $A_k = \arg\min_{\text{rank } k \ \widetilde{A}} \|A - \widetilde{A}\|_2$ and $A_k = \arg\min_{\text{rank } k \ \widetilde{A}} \|A - \widetilde{A}\|_F$. Finally, for any vector $y \in \mathbb{R}^d$, let $T(y) \in \mathbb{R}^{d \times d}$ denote the symmetric Toeplitz matrix whose first column is y.

2.2 Fourier analytic and linear algebra tools.

Let T denote a $d \times d$ symmetric PSD Toeplitz matrix. We heavily rely on the Fourier structure of Toeplitz matrices.

While circulant matrices can be diagonalized by the discrete Fourier transform, this does not hold in general for Toeplitz matrices. However, a classical result called the Caratheodory-Fejer-Pisarenko decomposition (also called the Vandermonde Decomposition), which we also stated in Section 1.3, says that they still can be decomposed into a product of off-grid Fourier and diagonal matrices; this is formalized in the following result below.

Theorem 3. *Any real-valued PSD Toeplitz matrix $T \in \mathbb{R}^{d \times d}$ of rank r can be expressed as $F_S D F_S^*$ where $F_S \in \mathbb{C}^{d \times r}$ is a symmetric Fourier matrix with frequencies $S \subseteq [0, 1/2]$, $|S| = r/2$ and $D \in \mathbb{R}^{r \times r}$ is a diagonal matrix with r positive entries. Moreover, for any $f \in S$ the values in D corresponding to columns $v(f), v(-f)$, which we refer to as the weights of $f, -f$ respectively, are identical. Thus D is uniquely defined by values $\{a_f\}_{f \in S}$.*

We also define the wrap around distance between two freqencies as follows.

Definition 2.1 (Wrap around distance)**.** *The wrap around distance between any two frequencies $f, g \in [-1/2, 1/2]$ is defined as $|f - g|_\circ = \min\{|f_i - f_j|, 1 - |f_i - f_j|\}$.*

Using the cyclic property of the trace, we can show the following Lemma, which relates the entries in D to the eigenvalues of T.

Lemma 2.2. $2d\sum_{f\in S} a_f = tr(T) = \sum_{j=1}^{d} \lambda_j(T)$.

Proof. Note that $tr(FDF^*) = tr(T) = tr(U\Sigma U^T) = \sum_{j=1}^{d} \lambda_j(T)$. Here $T = U\Sigma U^T$ is the eigen-decomposition of T. Further, $tr(FDF^*) = tr(F^*FD)$. Since each diagonal entry of F^*F is d and D is just a diagonal matrix with entries $\{a_f\}_{f\in S}$, we get that $tr(F^*FD) = 2d\sum_{f\in S} a_f$. Thus $2d\sum_{f\in S} a_f = \sum_{j=1}^{d} \lambda_j(T)$. □

We will need the notion of *statistical leverage scores* [13,34], which are used to define non-uniform row sampling schemes. These schemes then enable randomized matrix compressions, which provide spectral approximation guarantees and preserve significant information [11, 26]. Their precise definition is as follows.

Definition 2.3 (Leverage score)**.** *For any $A \in \mathbb{C}^{d\times r}$, let $\tau_j(A)$ denote the leverage score of the j^{th} row of A:*

$$\tau_j(A) = \max_{y\in\mathbb{C}^s} \frac{|(Ay)_i|^2}{\sum_{j=1}^{d} |(Ay)_j|^2}. \tag{2}$$

Finally, we will repeatedly use Weyl's eigenvalue perturbation bound for Hermitian matrices.

Theorem 4 (Weyl's inequality)**.** *For any $n > 0$ and Hermitian matrices $B, C \in \mathbb{R}^{n\times n}$ and $A = B + C$, the following holds for all $i \in [n]$*

$$\lambda_i(B) - \|C\|_2 \le \lambda_i(A) \le \lambda_i(B) + \|C\|_2.$$

For a proof of Weyl's inequality, we refer the reader to Section 1.3 in [35].

3 Existence of a near optimal low-rank approximation which itself is Toeplitz.

The goal of this section is to prove Theorems 5 and 2. Recall that we are given PSD Toeplitz matrix T and $T = F_S D F_S^*$, where F_S is a symmetric Fourier matrix (recall Definition 1.2), with frequencies $S \subset [0, 1/2]$ and $D = diag(\{a_f\}_{f\in S})$ in its Vandermonde decomposition. We briefly discuss how this section is organized. In subsection 3.1, we consider the case when all the frequencies in S are very close to each other; this is the *clustered* case. Then in subsection 3.2, we tackle the general case when the frequencies in S are not necessarily clustered. We partition the frequency domain into buckets, where the weight of each bucket is the sum of the weights of all frequencies in S landing in that bucket. The formal definition of buckets is as follows.

Definition 3.1. *For any PSD Toeplitz matrix $T = F_S D F_S^*$, where F_S is a Fourier matrix with frequencies $S \subset [0, 1/2]$, $D = diag(\{a_f\}_{f\in S})$, and j is a positive integer, we define the j-th bucket B_j by*

$$B_j := \left[\frac{j-1}{d}, \frac{j}{d}\right) \cap S,$$

and let

$$w(B_j) = \sum_{f\in B_j} a_f,$$

denote the weight of the j-th bucket.

The main claim in this subsection is to show that having heavy buckets implies that T has many large eigenvalues. This claim is shown in two steps. First, we show the claim for Toeplitz matrices that have a well-separated spectrum as defined formally below.

Definition 3.2 ((λ, w)-well separated)**.** *A PSD Toeplitz matrix $T = F_S D F_S^*$ is said to be (λ, w)-well separated if all non-empty buckets of T have weight exactly λ, and for any two non-empty buckets, the minimum wrap-around distance between any two frequencies in these two buckets is at least w.*

Second, we show how to reduce from proving the claim for general Toeplitz matrices, to proving it for well separated Toeplitz matrices. In subsection 3.3, we use the same tools to complete the characterization relating buckets and eigenvalues. In particular, we show that if all buckets have weight smaller than λ, then all eigenvalues of T are smaller than $O(d\lambda \log d)$. Finally, in subsection 3.4, we show how to use the structural statements proven in subsections 3.2 and 3.3 to prove Theorems 2 and 5.

3.1 Case of clustered frequencies.

We begin with the case when all the frequencies in S are clustered, as formalized by Definition 1.4. We first state some preliminaries. Define the t^{th} entry of the k^{th} column denoted by $T_k(t)$ for $t \in [0, d-1]$ as follows

$$T_k(t) = \sum_{f \in S} a_f e^{2\pi i (f^* + r_f)(t-(k-1))} + \sum_{f \in S} a_f e^{-2\pi i (f^* + r_f)(t-(k-1))} \quad \forall k \in [d].$$

Then the following lemma, which follows trivially from expanding T's Vandermonde decomposition $F_S D F_S^*$, says that the k^{th} column of T is the evaluation of $T_k(t)$ at $t = 0, \ldots, d-1$.

Lemma 3.3. *For all $(m, k) \in [d] \times [d]$, $T_{m,k} = T_k(m-1)$.*

Proof. Denote by $F_{S,i}$ the i-th row of F_S. Then

$$T_{m,k} = (F_S D F_S^*)_{m,k} = F_{S,m} D F_{S,k}^* = \sum_{f \in S} a_f e^{2\pi i (f^* + r_f) m} e^{-2\pi i (f^* + r_f) k} + \sum_{f \in S} a_f e^{-2\pi i (f^* + r_f) m} e^{2\pi i (f^* + r_f) k} = \sum_{f \in S} a_f e^{2\pi i (f^* + r_f)(m-k)} + \sum_{f \in S} a_f e^{-2\pi i (f^* + r_f)(m-k)} = T_k(m-1).$$

This concludes the proof of the lemma. ∎

Our main result in this section is stated in the following lemma.

Lemma 1.5. *There exists a universal constant $C_1 > 0$ such that, given any symmetric PSD Toeplitz T that is (f^*, Δ)-clustered for some $f^* \in [0, 1/2]$ and $\Delta \leq 1/d$, and for $0 < \varepsilon, \delta, \gamma < 1$ satisfying $\gamma \leq \varepsilon / (\operatorname{tr}(T) 2^{C_1 \log^7 d})$, the following conditions hold. There exists a symmetric Toeplitz matrix $\widetilde{T} = F_{\widetilde{S}} \widetilde{D} F_{\widetilde{S}}^*$ of rank at most $O(\ell)$ for $\ell = O(\log d + \log(1/\delta))$ such that*

1. *$\widetilde{S} = \{f^* + j\gamma\}_{j=1}^{\ell+1} \cup \{f^* - j\gamma\}_{j=1}^{\ell+1}$, and $\widetilde{D}$ is a diagonal matrix such that for any $f \in \widetilde{S}$, the weights corresponding to f and $-f$ are identical.*
2. *$\|T - \widetilde{T}\|_F \leq \delta(\sum_{f \in S} a_f) + \varepsilon d$.*

Remark 3.4. *The value of the diagonal elements in $\widetilde{D}$ depends on ε, and in particular the value of the diagonal elements would go to infinity as ε goes to zero. However, we do not state this tradeoff explicitly in the lemma, since we use this lemma later in Section 4 to obtain leverage score upper bounds on Fourier sparse functions (based on the work of [8]). These bounds do not depend on the magnitude of the coefficients in the function; instead, they depend only on its sparsity, as long as the coefficients are finite. Thus ε can be set to any strictly positive number of one's choice.*

Proof. Define for all $t \in [-d, d]$

$$T(t) = \sum_{f \in S} a_f e^{2\pi i (f^* + r_f)t} + \sum_{f \in S} a_f e^{-2\pi i (f^* + r_f)t}. \tag{3}$$

Note that $T_k(t)$ is just the restriction of $T(t)$ to $t \in [-(k-1), d-k]$ for all $k \in [d]$. Thus we focus on uniformly approximating $T(t)$ over $t \in [-d, d]$, as this will yield uniform approximations for each $T_k(t)$ simultaneously.

The next Lemma shows how to approximate $T(t)$ with a modulated low degree polynomial. The proof of this Lemma is deferred to the end of this subsection.

Lemma 3.5. *For polynomials p_1, p_2 of degree $\ell = O(\log d + \log(1/\delta))$ defined as $p_1(t) = \sum_{m=0}^{\ell} \left[\sum_{f \in S} a_f \frac{(2\pi i r_f)^m}{m!}\right] t^m$ and $p_2(t) = \sum_{m=0}^{\ell} \left[\sum_{f \in S} a_f \frac{(-2\pi i r_f)^m}{m!}\right] t^m$, the following equality holds:*

$$|T(t) - e^{2\pi i f^* t} p_1(t) - e^{-2\pi i f^* t} p_2(t)| \leq \delta(\sum_{f \in S} a_f) \quad \forall t \in [-d, d].$$

Observe that we can write the $p_1(t)$ and $p_2(t)$ obtained from Lemma 3.5 as $p_1(t) = p_{even}(t) + i p_{odd}(t)$ and $p_2(t) = p_{even}(t) - i p_{odd}(t)$, where p_{even} and p_{odd} contain the even and odd powered terms in $\sum_{m=0}^{\ell} \left[\sum_{f \in S} a_f \frac{(2\pi i r_f)^m}{m!}\right] t^m$ respectively. Now we can use the following lemma, which is a minor variation of Lemma 8.8 in [8], to express $p_{even}(t)$ and $p_{odd}(t)$ as $O(\log d + \log(1/\delta))$-Fourier sparse functions. For completeness, we present its proof at the end of this section.

Lemma 3.6 (Lemma 8.8 in [8])**.** *Let $p(t)$ be a degree ℓ polynomial with coefficients $c_1, \ldots, c_\ell$, defined over $t \in [-d, d]$ and containing only even powers of t. For every $\varepsilon \in (0, 1)$ and $\gamma \in (0, 1)$ satisfying*

$$\gamma \leq \varepsilon / (d 2^{\Theta(\ell^3 \log(\ell))} \max_{1 \leq i \leq p} |c_i|),$$

there exists $\widetilde{p}(t) = \sum_{j=1}^{\ell+1} \alpha_j (e^{2\pi i (j\gamma)t} + e^{-2\pi i (j\gamma)t})$ such that

$$|p(t) - \widetilde{p}(t)| \leq \varepsilon \quad \forall t \in [-d, d].$$

If $p(t)$ has only odd powers of t, then there exists $\widetilde{p}(t) = -i \sum_{j=1}^{\ell+1} \beta_j (e^{2\pi i (j\gamma)t} - e^{-2\pi i (j\gamma)t})$ that satisfies the above guarantee.

We apply Lemma 3.6 to approximate $p_{even}(t)$ by $\widetilde{p}_{even}(t) = \sum_{j=1}^{\ell+1} \alpha_j (e^{2\pi i (j\gamma)t} + e^{-2\pi i (j\gamma)t})$ and $p_{odd}(t)$ by $\widetilde{p}_{odd}(t) = -i \sum_{j=1}^{\ell+1} \beta_j (e^{2\pi i (j\gamma)t} - e^{-2\pi i (j\gamma)t})$. Note that since $r_f \in [0, 1]$ for all $f \in S$, we can trivially upper bound the absolute value of the j^{th} coefficient of p_{even} and p_{odd} by $d \sum_{f \in S} a_f$. Thus we can choose any $\gamma \leq \varepsilon / (d (\sum_{f \in S} a_f) 2^{C_1 \log^7 d})$ while applying Lemma 3.6, since $\ell^3 \log(\ell) \leq$

$O(\log^7 d)$. Now define $\widetilde{T}(t)$ as follows:

$$\begin{aligned}\widetilde{T}(t) &= e^{2\pi i f^* t}(\widetilde{p}_{even}(t) + i\widetilde{p}_{odd}(t)) + e^{-2\pi i f^* t}(\widetilde{p}_{even}(t) - i\widetilde{p}_{odd}(t)) \\ &= \sum_{j=1}^{\ell+1} \alpha_j \left[(e^{2\pi i(f^*+j\gamma)t} + e^{-2\pi i(f^*+j\gamma)t}) + (e^{2\pi i(f^*-j\gamma)t} + e^{-2\pi i(f^*-j\gamma)t})\right] \\ &+ \sum_{j=1}^{\ell+1} \beta_j \left[(e^{2\pi i(f^*+j\gamma)t} + e^{-2\pi i(f^*+j\gamma)t}) - (e^{2\pi i(f^*-j\gamma)t} + e^{-2\pi i(f^*-j\gamma)t})\right] \\ &= \sum_{j=1}^{\ell+1} (\alpha_j + \beta_j) \left[(e^{2\pi i(f^*+j\gamma)t} + e^{-2\pi i(f^*+j\gamma)t})\right] + (\alpha_j - \beta_j) \left[(e^{2\pi i(f^*-j\gamma)t} + e^{-2\pi i(f^*-j\gamma)t})\right].\end{aligned}$$

Let $\widetilde{T}_k(t)$ be the restriction of $\widetilde{T}(t)$ to $t \in [-(k-1), d-k]$. Thus applying the error guarantees of Lemma 3.5 and 3.6 and applying triangle inequality, we get that for all $k \in [d]$

$$|T_k(t) - \widetilde{T}_k(t)| \leq \delta(\sum_{f\in S} a_f) + \varepsilon \quad \forall t \in [-(k-1), d-k].$$

Now $\widetilde{T}_k(t)$ for $k \in \{1, \ldots, d\}$ naturally defines the rank $4(\ell+1)$ symmetric Toeplitz matrix $\widetilde{T} = F_{\widetilde{S}} \widetilde{D} F_{\widetilde{S}}^*$, where $\widetilde{S} = \{f^* + j\gamma\}_{j=1}^{\ell+1} \cup \{f^* - j\gamma\}_{j=1}^{\ell+1}$ and $\widetilde{D} = diag(\{\alpha_j + \beta_j\}_{j=1}^{\ell+1} \cup \{\alpha_j - \beta_j\}_{j=1}^{\ell+1})$. The previous equation and this observation immediately that for $\widetilde{T}$,

$$\|T - \widetilde{T}\|_F \leq \sqrt{\sum_{(i,j)\in[d]\times[d]} (\delta(\sum_{f\in S} a_f) + \varepsilon)^2} = \delta d(\sum_{f\in S} a_f) + \varepsilon d.$$

We redefine δ as δ/d. This concludes the proof of statement 2 in Lemma 1.5. Lemma 2.2 implies that choosing $\gamma \leq \varepsilon/(d(\sum_{f\in S} a_f) 2^{C_1 \log^7 d}) = \varepsilon/(\mathrm{tr}(T) 2^{C_1 \log^7 d})$ suffices for the first claim of the lemma to hold. □

We now state the proof of Lemma 3.5.

Proof. First, using a Taylor expansion we can write $T(t)$ as follows

$$\begin{aligned}T(t) &= \sum_{f\in S} a_f e^{2\pi i(f^*+r_f)t} + \sum_{f\in S} a_f e^{-2\pi i(f^*+r_f)t} \\ &= e^{2\pi i f^* t} \left(\sum_{f\in S} a_f (\sum_{m=0}^{\infty} \frac{(2\pi i r_f t)^m}{m!})\right) + e^{-2\pi i f^* t} \left(\sum_{f\in S} a_f (\sum_{m=0}^{\infty} \frac{(-2\pi i r_f t)^m}{m!})\right) \\ &= e^{2\pi i f^* t} \sum_{m=0}^{\ell} \left[\sum_{f\in S} a_f \frac{(2\pi i r_f)^m}{m!}\right] t^m + e^{-2\pi i f^* t} \sum_{m=0}^{\ell} \left[\sum_{f\in S} a_f \frac{(-2\pi i r_f)^m}{m!}\right] t^m \\ &+ e^{2\pi i f^* t} \left(\sum_{f\in S} a_f (\sum_{m=\ell}^{\infty} \frac{(2\pi i r_f t)^m}{m!})\right) + e^{-2\pi i f^* t} \left(\sum_{f\in S} a_f (\sum_{m=\ell}^{\infty} \frac{(-2\pi i r_f t)^m}{m!})\right).\end{aligned}$$

Using the fact that $|r_f| \leq \Delta$ and $t \in [-d, d]$, we have that $|2\pi i r_f| < 10\Delta d$ for all $f \in S$. Also note that for any $m > 1, m! \geq (m/2)^{m/2}$. This implies that the following holds for all r_f and $t \in [-d, d]$

$$\begin{aligned}\left|\sum_{m=\ell}^{\infty} \frac{(2\pi i r_f t)^m}{m!}\right| &\leq \sum_{m=\ell}^{\infty} \left|\frac{(2\pi i r_f t)^m}{m!}\right| \leq \sum_{m=\ell}^{\infty} (10\Delta d)^m/(m/2)^{m/2} \\ &\leq \sum_{m=\ell}^{\infty} (10\Delta d/\sqrt{m/2})^m \leq \sum_{m=\ell}^{\infty} (10/100\sqrt{\log d})^m \\ &= \sum_{m=\ell}^{\infty} (0.1)^m \leq (0.1)^{O(\log(1/\delta))}/0.9 \leq \delta/2.\end{aligned}$$

Thus letting $p_1(t) = \sum_{m=0}^{\ell} \left[\sum_{f\in S} a_f \frac{(2\pi i r_f)^m}{m!}\right] t^m$ and $p_2(t) = \sum_{m=0}^{\ell} \left[\sum_{f\in S} a_j \frac{(-2\pi i r_f)^m}{m!}\right] t^m$, we get that

$$\begin{aligned}|T(t) - e^{2\pi i f^* t} p_1(t) - e^{-2\pi i f^* t} p_2(t)| &\leq \sum_{f\in S} a_j \left(\left|\sum_{m=\ell}^{\infty} \frac{(2\pi i r_f t)^m}{m!}\right| + \left|\sum_{m=\ell}^{\infty} \frac{(-2\pi i r_f t)^m}{m!}\right|\right) \\ &\leq \delta(\sum_{f\in S} a_j),\end{aligned}$$

where in the second inequality we used that all the a_j's are non-negative. □

We finish this subsection with the proof of Lemma 3.6.

Proof. The proof of this is almost identical to the proof of Lemma 8.8 in [8]. We first consider the case when the degree ℓ polynomial $p(t)$ has only even powers of t. Let $p(t) = \sum_{j=0}^{\ell/2} c_{2j} t^{2j}$. First, note we can write $\widetilde{p}(t)$ as

$$\begin{aligned}\widetilde{p}(t) &= \sum_{j=1}^{\ell+1} \alpha_j (e^{2\pi i(\gamma j)t} + e^{-2\pi i(\gamma j)t}) \\ &= \sum_{j=1}^{\ell+1} 2\alpha_j (\sum_{k=0:k \text{ even}}^{\infty} \frac{(2\pi i\gamma j t)^k}{k!}) \\ &= \sum_{k=0:k \text{ even}}^{\infty} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k \\ &= \sum_{k=0:k \text{ even}}^{\ell} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k + \sum_{k=\ell+1:k \text{ even}}^{\infty} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k \\ &= p(t) + (\sum_{k=0:k \text{ even}}^{\ell} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k - p(t)) + \sum_{k=\ell+1:k \text{ even}}^{\infty} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k.\end{aligned}$$

We will show that there exists some γ and $\alpha_1, \ldots, \alpha_{\ell+1}$ such that $\sum_{k=0:k \text{ even}}^{\ell} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k - p(t)$ is identically zero. The $\ell + 1 \times \ell + 1$ Vandermonde matrix A' is the same as in the proof of Lemma 8.8 in [8]; however, the $\ell + 1$ dimensional vector c' is slightly different: $c'_j = c_j j!/(2\pi i\gamma)^j$ for j even and $c'_j = 0$ for j odd. The rest of the argument is identical to the proof of Lemma 8.8 in [8].

Now, the argument to show that the absolute value of $C_2 = \sum_{k=\ell+1:k \text{ even}}^{\infty} 2\frac{(2\pi i\gamma t)^k}{j!} \sum_{j=1}^{\ell+1} \alpha_j j^k$ is less than ϵ for all $|t| \leq d$ (for the choice of γ and $\alpha_1, \ldots, \alpha_{\ell+1}$ shown to exist previously) is also nearly identical to the proof of Lemma 8.8 in [8]. The only difference is that, while upper bounding the value of $|C_2|$, we replace t^j with $|t|^j$ for all $j \in [\ell+1, \infty]$.

The proof for the case when $p(t)$ has only odd powered terms is identical to the proof discussed above, with the exception that the goal is to show the existence of $\widetilde{p}(t) = -i\sum_{j=1}^{\ell+1} \beta_j (e^{2\pi i(\gamma j)t} - e^{-2\pi i(\gamma j)t})$ satisfying the claim of the theorem. □

3.2 Many heavy buckets implies many large eigenvalues.

In this subsection, we present our first structural result relating the Vandermonde decomposition $T = F_S D F_S^*$ to the eigenvalue decomposition of $T = U\Sigma U^T$. We first bucket the frequencies in the Vandermonde decomposition, as described in Definition 3.1.

Equipped with this definition, we now state the main result of this section.

Lemma 1.7. *For every $\lambda, k > 0$, if T has at least k buckets with weight at least λ, then T has at least $k/\log^3 d$ eigenvalues that have value $\Omega(d\lambda/\log d)$.*

Before we prove Lemma 1.7, we state the following helper lemmas. We will then present the proof of Lemma 1.7 using these helper lemmas, and finally end this subsection with their proofs.

Lemma 3.7. *For any two frequencies $f, g \in [-1/2, 1/2]$,*

$$|v(f)^* v(g)| \leq O(1/|f-g|_\circ),$$

where $v(f), v(g)$ are the corresponding frequency vectors (recall Definition 1.1 of frequency vectors) and $|f-g|_\circ$ is the wrap-around distance between f and g.

The next lemma essentially is a strengthening of Gershgorin's circle theorem for block matrices.

Lemma 3.8 (Section 1.13 in [36], or Corollary 3.2 in [14])**.** *Let $A = (A_{ij}) \in \mathbb{C}^{dn \times dn}$ be a Hermitian matrix composed of blocks $A_{ij} \in \mathbb{C}^{d \times d}$. Then each A_{ii} is Hermitian, and the following is true:*

$$\|A\|_2 \leq \max_{i \in [n]} (\|A_{ii}\|_2 + \sum_{\substack{j \in [n] \\ j \neq i}} \|A_{ij}\|_2).$$

The final helper lemma will help us upper bound the norms of the off-diagonal blocks when applying Lemma 3.8.

Lemma 3.9. *For any $\lambda \geq 0$, $\sigma_1, \sigma_2 \in \{+, -\}$, sets of frequencies $S_1, S_2 \subseteq [0, 1/2]$ and corresponding diagonal weight matrices D_1, D_2 both of whose traces are at most λ, $A = D_1^{1/2} F_{\sigma_1 S_1}^* F_{\sigma_2 S_2} D_2^{1/2}$ satisfies $\|A\|_F \leq O(\lambda/d(\sigma_1 S_1, \sigma_2 S_2))$. Here recall that for any set of frequencies $S \subset [0, 1/2]$, $+S$ and $-S$ contain the frequencies in S and their negations, respectively, and $d(A, B)$ for any sets $A, B \subseteq [-1/2, 1/2]$ is the minimum wraparound distance between any two frequencies in A, B.*

We first consider a Toeplitz matrix with a well separated spectrum (as per Definition 3.2) and prove Lemma 1.7 for this case. Then, we reduce the general case to the well separated case, which proves Lemma 1.7 in full generality.

It can be seen from the following lemma that, essentially, we can reduce from a general Toeplitz matrix to the well-separated case.

Lemma 3.10. *For any $w, \lambda, k > 0$ and a PSD Toeplitz matrix T having at least k buckets with weight at least λ, there exists a (λ, w)-well separated PSD Toeplitz matrix $T_{separated}$ having at least k/dw non-empty buckets satisfying $T_{separated} \preceq T$.*

The following lemma essentially proves Lemma 1.7 for the special case of well-separated matrices.

Lemma 3.11. *For every $\lambda, w, k > 0$ and a (λ, w)-well separated T having at least k non-empty buckets, T has at least k eigenvalues that have value at least $\Omega(\lambda(d/\log d - \log d/w))$.*

Equipped with these helper lemmas, we are ready to present the proof of Lemma 1.7.

Proof. Refer to the k buckets with weight at least λ as *heavy buckets*. Without loss of generality, we can assume that each bucket has weight exactly λ. This can be seen by the following argument. We reduce the weight of each heavy bucket if needed to ensure that each heavy bucket has weight exactly λ. This change will only make T PSD smaller, and thus it suffices to lower bound eigenvalues of T after this operation. We can easily obtain a (λ, w)-well separated Toeplitz matrix from T as per Claim 3.10.

Consider the matrix $T_{separated}$ guaranteed to exist by Lemma 3.10. Thus, invoking Lemma 3.11 for $T_{separated}$ with parameters $\lambda, k/dw, w$ for $w = \log^3 d/d$, we get that $T_{separated}$ has at least $k/\log^3 d$ eigenvalues of value $\Omega(d\lambda(1/\log d - 1/\log^2 d)) = \Omega(d\lambda/\log d)$. Since $T_{separated} \preceq T$, T also has at least $k/\log^3 d$ eigenvalues of value $\Omega(d\lambda(1/\log d - 1/\log^2 d)) = \Omega(d\lambda/\log d)$. This finishes the proof of Lemma 1.7. □

Now we present the proof of the well separated case (Lemma 3.11).

Proof. Let $T = F_S D F_S^*$ denote the Vandermonde decomposition of T. Observe that $F_S D F_S^*$ has the same eigenvalues as $D^{1/2} F_S^* F_S D^{1/2}$. This is because, for any matrix A, A^*A and AA^* have the same eigenvalues; apply this fact to $A = F_S D^{1/2}$. It can be easily seen that $D^{1/2} F_S^* F_S D^{1/2}$ is a block matrix, where each of its blocks is of the form $D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}$ for some m, n, and D_m is the diagonal matrix containing weights of the frequencies in bucket B_m for every m. We can pad each block with enough zero rows and columns to ensure that the dimensions of each block are the same, without affecting its spectral norm. Thus we can express $D^{1/2} F_S^* F_S D^{1/2} = A + E$, where A contains the diagonal blocks with zero matrices on the off diagonal blocks, and E contains the off-diagonal blocks with zero matrices on the diagonal blocks. Our high level strategy is to prove the theorem statement for A by using the fact that it is block diagonal. Then, we show that E has small spectral norm, so the eigenvalues of $A + E$ are close to the eigenvalues of A.

Lower bounding largest eigenvalue of diagonal blocks. A is block diagonal with at least k blocks. Consider any of its blocks $D_m^{1/2} F_{B_m}^* F_{B_m} D_m^{1/2}$ for some m. We know that $F_{B_m} D_m F_{B_m}^*$ is $((m - 1/2)/d, 1/d)$-clustered. Thus by Lemma 1.5 with $\delta = 1/2d^{11}$ and $\varepsilon = \lambda/2d^{12}$, $F_{B_m} D_m F_{B_m}^*$ is $\delta(\sum_{f \in B_m} a_f) + \varepsilon d = \lambda/d^{11}$ close to a symmetric Toeplitz matrix $\widetilde{T}$ with rank at most $C \log d$ in the Frobenius norm for some universal constant $C > 0$. Thus by Weyl's inequality (Theorem 4), we get the following for any $k > C \log d$

$$\begin{aligned}
\lambda_k(F_{B_m} D_m F_{B_m}^*) &\leq \lambda_k(\widetilde{T}) + \|F_{B_m} D_m F_{B_m}^* - \widetilde{T}\|_2 \\
&\leq 0 + \|F_{B_m} D_m F_{B_m}^* - \widetilde{T}\|_F \\
&\leq \lambda/d^{11}.
\end{aligned}$$

Thus we get the following bound on the trace of $F_{B_m} D_m F_{B_m}^*$

$$\mathrm{tr}(F_{B_m} D_m F_{B_m}^*) \leq C \log d \lambda_1(F_{B_m} D_m F_{B_m}^*) + \lambda/d^{10}.$$

This implies that

$$\begin{aligned}\lambda_{max}(D_m^{1/2} F_{B_m}^* F_{B_m} D_m^{1/2}) &= \lambda_1(F_{B_m} D_m F_{B_m}^*) \\ &\geq \mathrm{tr}(F_{B_m} D_m F_{B_m}^*)/(C \log d) - \lambda/(Cd^{10} \log d) \\ &= d\lambda/(C \log d) - \lambda/(Cd^{10} \log d) = \Omega(d\lambda/\log d).\end{aligned}$$

Since A is a block diagonal matrix with at least k blocks, and each block has largest eigenvalue $\Omega(d\lambda/\log d)$, we get that A has at least k eigenvalues which are at least $\Omega(d\lambda/\log d)$ in value.

Upper bounding the contribution of off-diagonal blocks. In the rest of the proof, we bound the spectral norm of E. We do this by applying Lemma 3.8 to E where its diagonal blocks are 0 and any off diagonal block is of the form $D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}$ for some buckets B_m, B_n satisfying $m \neq n$. Consider an arbitrary but fixed bucket index m. In order to apply Lemma 3.8, we need to upper bound $\sum_{n:n\neq m} \|D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}\|_2 \leq \sum_{n:n\neq m} \|D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}\|_F$. We recall that for any n, $F_{B_n} = [F_{+B_n}; F_{-B_n}]$ as per the notation in Definition 1.2. We apply Lemma 3.9 to bound $\|D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}\|_F$ as follows

$$\begin{aligned}\|D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}\|_F &\leq \|D_m^{1/2} F_{+B_m}^* F_{+B_n} D_n^{1/2}\|_F + \|D_m^{1/2} F_{+B_m}^* F_{-B_n} D_n^{1/2}\|_F \\ &\quad + \|D_m^{1/2} F_{-B_m}^* F_{+B_n} D_n^{1/2}\|_F + \|D_m^{1/2} F_{-B_m}^* F_{-B_n} D_n^{1/2}\|_F \\ &\leq O(\lambda(1/d(+B_m, +B_n) + 1/d(+B_m, -B_n) + \\ &\quad 1/d(-B_m, +B_n) + 1/d(-B_m, -B_n))).\end{aligned}$$

Thus, we need to upper bound the following

$$\sum_{n:n\neq m} O(\lambda(1/d(+B_m, +B_n) + 1/d(+B_m, -B_n) + 1/d(-B_m, +B_n) + 1/d(-B_m, -B_n))).$$

It is easy to see that the maximum value the above expression can take will correspond to the case when, informally, there are as many buckets as possible in S, while still ensuring that any two buckets have separation at least w. More formally, this corresponds to the case when for every $n \in [m - d/(2w), m + d/(2w)]$, there is a bucket at distance $d_{m,n} = \Theta(|m - n|w)$ from B_m. Thus, the above expression is bounded by

$$\begin{aligned}&\sum_{n=m+1}^{m+d/(2w)} O(\lambda/((n-m)w)) + \sum_{n=m-d/(2w)}^{m-1} O(\lambda/((m-n)w)) \\ &\qquad \leq 2\sum_{l=1}^{d} O(\lambda/(lw)) \\ &\qquad = O(\lambda/w)\sum_{l=1}^{d} 1/l \\ &\qquad = O(\lambda \log d/w).\end{aligned}$$

Since any diagonal block has spectral norm 0 and the previous bound holds for all m, Lemma 3.8 applied to E implies that $\|E\|_2 \leq O(\lambda \log d/w)$. Now, applying Weyl's inequality for all i such that $\lambda_i(A) = \Omega(d\lambda/\log d)$, we have

$$\lambda_i(A+E) \geq \lambda_i(A) - \|E\|_2 = \Omega(d\lambda/\log d) - \|E\|_2 \geq \Omega(\lambda(d/\log d - \log d/w)).$$

Since there are at least k such indices i, this completes the proof of the Lemma. □

We now present the proof of Lemma 3.7.

Proof. Denote by $\delta = |f-g|_\circ$ the wrap around distance between f and g. Since $|v_f^* v_g| = |v_f^* v_g|$, we have the following.

$$|v(g)^* v(f)| = \left|\sum_{l=0}^{d-1} \exp(2\pi i \delta l)\right| = \left|\frac{\exp(2\pi i \delta d) - 1}{\exp(2\pi i \delta) - 1}\right| = \left|\frac{\exp(\pi i \delta d)}{\exp(\pi i \delta)} \cdot \frac{\sin(\pi \delta d)}{\sin(\pi \delta)}\right|.$$

Therefore, it holds that

$$|v(g)^* v(f)| = \left|\frac{\sin(\pi \delta d)}{\sin(\pi \delta)}\right| \leq \frac{1}{|\sin(\pi \delta)|} \leq O\left(\frac{1}{\delta}\right).$$

Here, the last inequality follows by the fact that $\sin(\pi x) \geq 2x$ for $x \in [0, 1/2]$, and the wrap around distance between any two frequencies lies between 0 and $1/2$. □

Now, we present the proof of Lemma 3.9.

Proof. We have the following for all i, j:

$$|(A)_{i,j}| = \sqrt{(D_1)_i (D_2)_j}|v(f_i)^* v(f_j)| \leq O(\sqrt{(D_1)_i (D_2)_j}/|f_i - f_j|) \leq O(\sqrt{(D_1)_i (D_2)_j}/d(\sigma_1 S_1, \sigma_2 S_2)),$$

where $(D_1)_i, (D_2)_j$ and f_i, f_j are the weights and frequencies of the i^{th} frequency in the $\sigma_1 S_1$ and j^{th} frequency in the $\sigma_2 S_2$ respectively. Also, note that the second to last inequality follows from Lemma 3.7. This implies

$$\begin{aligned}
\|A\|_F^2 &\leq O(\sum_{i,j} (D_1)_i (D_2)_j / d(\sigma_1 S_1, \sigma_2 S_2)^2) \\
&= O((\sum_i (D_1)_i)(\sum_j (D_2)_j)/d(\sigma_1 S_1, \sigma_2 S_2)^2) \\
&\leq O(\lambda^2 / d(\sigma_1 S_1, \sigma_2 S_2)^2).
\end{aligned}$$

Thus $\|A\|_F \leq O(\lambda / d(\sigma_1 S_1, \sigma_2 S_2))$. □

We finally end this sub-section with the proof of Claim 3.10.

Proof. Sort the k heavy buckets of T according to their central frequency, and index them by $[k]$. Now consider the Toeplitz matrix $T_{separated}$, obtained by taking the buckets indexed by the residue class 1 mod dw out of the k heavy buckets of T. There are at least k/dw such buckets, and the pairwise separation between any two such buckets is at least w. This implies $T_{separated}$ is (λ, w)-well separated and has k/dw non-empty buckets. Moreover, since $T_{separated}$ is obtained by taking a subset of the non-empty buckets of T, this implies that $T_{separated} \preceq T$. □

3.3 All buckets being light implies all eigenvalues are small.

In this subsection, we show that if all buckets of T have small weight, then all eigenvalues of T are small. This result, combined with the previous section, will suffice to prove Theorems 5 and 2. This is formalized in the lemma below.

Lemma 1.8. *For every $\lambda \geq 0$, if all buckets of T have weight at most λ, then $\|T\|_2 \leq O(d\lambda \log d)$.*

Proof. We need to upper bound $\|F_S D F_S^*\|_2 = \|D^{1/2} F_S^* F_S D^{1/2}\|_2$. Again, $A := D^{1/2} F_S^* F_S D^{1/2}$ is a block matrix, where its blocks are of the form $A_{m,n} := D_m^{1/2} F_{B_m}^* F_{B_n} D_n^{1/2}$ for some buckets B_m, B_n. Here F_{B_m} and D_m are the Fourier and diagonal weight matrices, respectively, of the m^{th} bucket. To prove the Lemma, we will apply Lemma 3.8 on A. To do so, we need to upper bound $\|A_{m,m}\|_2 + \sum_{n:n\neq m} \|A_{m,n}\|_2 \leq \|A_{m,m}\|_2 + \sum_{n:n\neq m} \|A_{m,n}\|_F$ for all m. Since $A_{m,m}$ is PSD, we upper bound $\|A_{m,m}\|_2$ as follows

$$\|A_{m,m}\|_2 \leq \mathrm{tr}(A_{m,m}) = d \sum_i (D_m)_i = dw(B_m) \leq d\lambda.$$

We can trivially upper bound $\|A_{m,m+1}\|_F$ by first applying Cauchy-Schwarz to each entry $(A_{m,m+1})_{i,j}$:

$$\begin{aligned} |(A_{m,m+1})_{i,j}| = \sqrt{(D_m)_i (D_{(m+1)})_j} |v(f_i)^* v(f_j)| &\leq \sqrt{(D_m)_i (D_{(m+1)})_j} \|v(f_i)\|_2 \|v(f_j)\|_2 \\ &= \sqrt{(D_m)_i (D_{(m+1)})_j}\, d, \end{aligned}$$

where $f_i, (D_m)_i$ are the i^{th} frequency and its corresponding weight in the m^{th} bucket respectively. Then we use the previous bound to get that the following holds for any m:

$$\|A_{m,m+1}\|_F \leq \sqrt{(\sum_i (D_m)_i)(\sum_j (D_m)_j) d^2} = d\lambda.$$

The same bound holds for $\|A_{m,m-1}\|_F$. Thus we obtain the bound

$$\|A_{m,m}\|_2 + \sum_{n:n\neq m} \|A_{m,n}\|_F \leq 3d\lambda + \sum_{n:n\notin\{m-1,m,m+1\}} \|A_{m,n}\|_F.$$

Since $d(+B_m, +B_{m+1})$ and $d(+B_m, +B_{m-1})$ could be arbitrarily close to 0 (adjacent buckets could have frequencies very close to each other), the bound $\|A_{m,m+1}\|_F \leq d\lambda$ is tight. For the remaining $n \notin \{m, m-1, m+1\}$, we use Lemma 3.9 to bound $\|A_{m,n}\|_F$ as follows:

$$\begin{aligned} \|A_{m,n}\|_F \leq O(\lambda(1/d(+B_m, +B_n) + 1/d(+B_m, -B_n) + \\ 1/d(-B_m, +B_n) + 1/d(-B_m, -B_n))). \end{aligned}$$

Similar to the proof in the previous subsection, it is easy to see that the maximum value $\sum_{n:n\notin\{m-1,m,m+1\}} \|A_{m,n}\|_F$ can take will correspond to the case when all buckets in $[0, 1/2]$ are non-empty. More formally, for every $n \in [m - d/2, m + d/2] \setminus \{m-1, m, m+1\}$, there is a bucket

at distance $\Theta(|m-n|/d)$ from bucket B_m. Thus we get that the above is bounded by

$$\begin{aligned}\sum_{n=m+2}^{m+d/2} O(\lambda/((n-m)/d)) + \sum_{n=m-d/2}^{m-2} O(\lambda/((m-n)/d)) &\\ \leq 2\sum_{l=1}^{d} O(\lambda/(l/d)) &\\ = O(d\lambda)\sum_{l=1}^{d} 1/l &\\ = O(d\lambda \log d). &\end{aligned}$$

Thus applying Lemma 3.8 to A, we get that $\|T\|_2 = \|A\|_2 \leq 3d\lambda + O(d\lambda \log d) = O(d\lambda \log d)$. This completes the proof of Lemma 1.8. □

3.4 Existence of Toeplitz low-rank approximation in the spectral and Frobenius norm.

In this section, we use Lemma 1.7 and Lemma 1.8 to show the existence of a near optimal low-rank approximation in the both the spectral and Frobenius norm, which itself is Toeplitz. The first claim is formalized in the following theorem.

Theorem 5. *For any PSD Toeplitz $T \in \mathbb{R}^{d\times d}$, $0 < \delta < 1$ and any integer $k \leq d$, there exists a symmetric Toeplitz matrix $\widetilde{T}$ of rank $\widetilde{O}(k\log(1/\delta))$ such that the following holds*

$$\|T - \widetilde{T}\|_2 \leq \widetilde{O}(1)\|T - T_k\|_2 + \delta\|T\|_F.$$

where $T_k = \underset{B:rank(B)\leq k}{\|T-B\|_2}$ is the best rank-k approximation to T in the spectral norm.

Proof. Let $T = F_S D F_S$ be T's Vandermonde decomposition and let $\lambda = \lambda_{k+1}(T)\log d/c'd$, where c' is the same constant appearing in the big-O notation in Lemma 1.7. Bucket the frequencies in S as per Definition 3.1. Let T_{heavy} be PSD Toeplitz matrix obtained by considering all buckets with weight strictly more than λ. Then $T - T_{heavy}$ is also a PSD Toeplitz matrix, which contains only buckets of T that have weight bounded by λ. Thus by Lemma 1.8 we have the following

$$\|T - T_{heavy}\|_2 \leq O(d\lambda \log d) = \widetilde{O}(1)\|T - T_k\|_2.$$

Now we claim that T has at most $(k+2)\log^3 d$ buckets with weight strictly more than $\lambda_{k+1}(T)\log d/c'd$. (If not, then by Lemma 1.7, T has at least $k+2$ eigenvalues with value strictly more than $\lambda_{k+1}(T)$; this contradicts the definition of $\lambda_{k+1}(T)$ being the $(k+1)^{th}$ largest eigenvalue of T.) Thus, T_{heavy} contains at most $(k+2)\log^3 d$ non-empty buckets. Let $r \leq (k+2)\log^3 d$ be the number of these heavy buckets, and define $F_{S_1}, \ldots, F_{S_r}$ and $D_1, D_2, \ldots, D_r$ to be the Fourier and diagonal weight matrices corresponding to these r heavy buckets which define T_{heavy}. We can therefore write T_{heavy} as

$$T_{heavy} = \sum_{m=1}^{r} F_{S_m} D_m F_{S_m}^*.$$

Since each $F_{S_m} D_m F_{S_m}^*$ is $(f, 1/d)$-clustered for some $f \in [0, 1/2]$, we use Theorem 1.5 to approximate each $F_{S_m} D_m F_{S_m}^*$ with a rank $O(\log d + \log(1/\delta))$ symmetric Toeplitz matrix $F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*$ that satisfies

$$\|F_{S_m} D_m F_{S_m}^* - F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*\|_F \leq \delta(\sum_i (D_m)_i) + \varepsilon d.$$

This implies that

$$\begin{aligned}
\left\| T - \sum_{m=1}^{r} F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m} \right\|_2 &\leq \left\| T - \sum_{m=1}^{r} F_{S_m} D_m F^*_{S_m} \right\|_2 + \left\| \sum_{m=1}^{r} (F_{S_m} D_m F^*_{S_m} - F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m}) \right\|_2 \\
&= \|T - T_{heavy}\|_2 + \left\| \sum_{m=1}^{r} (F_{S_m} D_m F^*_{S_m} - F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m}) \right\|_2 \\
&\leq \|T - T_{heavy}\|_2 + \sum_{m=1}^{r} \left\| F_{S_m} D_m F^*_{S_m} - F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m} \right\|_F \\
&\leq \widetilde{O}(1)\|T - T_k\|_2 + \delta(\sum_{m,i} (D_m)_i) + \varepsilon r d \\
&\leq \widetilde{O}(1)\|T - T_k\|_2 + \delta(\sum_{i} (D)_i) + \varepsilon r d.
\end{aligned}$$

Now using Lemma 2.2, we get the following:

$$\sum_i (D)_i = \sum_{i=1}^{d} \lambda_i(T)/d \leq \left(\sqrt{d \sum_{i=1}^{d} \lambda_i(T)^2} \right) /d = \|T\|_F/\sqrt{d}.$$

Setting $\varepsilon = \delta\|T\|_F/rd$, we finally get that

$$\left\| T - \sum_{m=1}^{r} F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m} \right\|_2 \leq \widetilde{O}(1)\|T - T_k\|_2 + \delta\|T\|_F.$$

Defining $\widetilde{T}$ as $\widetilde{T} = \sum_{m=1}^{r} F_{\widetilde{S}_m} \widetilde{D}_m F^*_{\widetilde{S}_m}$, which has rank at most $r(O(\log d + \log(1/\delta)) = \widetilde{O}(k \log(1/\delta))$, we find that $\widetilde{T}$ satisfies the claim of the theorem. □

Next, we state the our main result on the existence of a near optimal Toeplitz low-rank approximation in the Frobenius norm.

Theorem 2. *Given any PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$, $0 < \epsilon, \delta < 1$ and an integer $k \leq d$, let $r_1 = O(k \log^8 d/\epsilon)$ and $r_2 = O(\log d + \log(1/\delta))$. Then there exists a symmetric Toeplitz matrix $\widetilde{T} = F_{\widetilde{S}} \widetilde{D} F^*_{\widetilde{S}}$ of rank $r = 2r_1r_2 = \widetilde{O}((k/\epsilon)\log(1/\delta))$ such that,*

1. *$\|T - \widetilde{T}\|_F \leq (1+\epsilon)\|T - T_k\|_F + \delta\|T\|_F$, where $T_k = \underset{B:rank(B)\leq k}{\|T - B\|_F}$ is the best rank-k approximation to T in the Frobenius norm.*
2. *$F_{\widetilde{S}}, \widetilde{D}$ are Fourier and diagonal matrices respectively. The set of frequencies $\widetilde{S}$ can be partitioned into r_1 sets $\widetilde{S}_1, \ldots, \widetilde{S}_{r_1}$ where each $\widetilde{S}_i$ is as follows:*

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

3. *$f_i \in N$ for all $i \in [r_1]$, where $N := \{1/2d, 3/2d, \ldots, 1 - 1/2d\}$. Furthermore, $\gamma = \delta/(2^{C_2 \log^7 d})$, where $C_2 > 0$ is a fixed constant.*

Proof. Let $T = F_S D F_S$ be T's Vandermonde decomposition and let $\lambda = \lambda_{Ck\log^5 d/\epsilon}(T)\log d/c'd$, where c' is the same constant as in Lemma 1.7 and C is some large constant. Bucket the frequencies in S as per Definition 3.1. Let T_{heavy} be the matrix obtained by considering all buckets of T with weight strictly more than λ. Then $T - T_{heavy}$ is also a Toeplitz matrix defined by the remaining buckets, all of which have weight at most λ. Thus by Lemma 1.8, we have the following

$$\|T - T_{heavy}\|_2 \leq O(\lambda d \log d) = O(\log^2 d)\lambda_{Ck\log^5 d/\epsilon}(T).$$

Since $T - T_{heavy}$ just corresponds to retaining a subset of the frequencies in S and their corresponding weights, it is a PSD change. More formally, $0 \preceq T - T_{heavy} \preceq T$. Thus by Weyl's inequality, this implies that $\lambda_i(T - T_{heavy}) \leq \lambda_i(T)$ for all $i \in [d]$. Thus we have that

$$\begin{aligned}
\|T - T_{heavy}\|_F^2 &= \sum_{i=1}^{d} \lambda_i^2(T - T_{heavy}) = \sum_{i=1}^{k+1} \lambda_i^2(T - T_{heavy}) + \sum_{i=k+2}^{d} \lambda_i^2(T - T_{heavy}) \\
&\leq (k+1)\lambda_1^2(T - T_{heavy}) + \sum_{i=k+2}^{d} \lambda_i^2(T - T_{heavy}) \\
&\leq (k+1)\lambda_1^2(T - T_{heavy}) + \sum_{i=k+2}^{d} \lambda_i^2(T).
\end{aligned}$$

Here, we used the fact that $\lambda_i(T - T_{heavy}) \leq \lambda_1(T - T_{heavy})$ for all $1 \leq i \leq k+1$ and $\lambda_i(T - T_{heavy}) \leq \lambda_i(T)$ for $i > k+1$. Now, using that $\lambda_1(T - T_{heavy}) \leq O(\log^2 d)\lambda_{Ck\log^5 d/\epsilon}(T)$, we get the following:

$$\begin{aligned}
\|T - T_{heavy}\|_F^2 &\leq (k+1)\lambda_1^2(T - T_{heavy}) + \sum_{i=k+2}^{d} \lambda_i^2(T) \\
&\leq O(k\log^4 d)\lambda^2_{Ck\log^5 d/\epsilon}(T) + \sum_{i=k+2}^{d} \lambda_i^2(T) \\
&= \epsilon O(k\log^4 d/\epsilon)\lambda^2_{Ck\log^5 d/\epsilon}(T) + \sum_{i=k+2}^{d} \lambda_i^2(T) \\
&\leq \epsilon \sum_{i=k+1}^{O(k\log^4 d/\epsilon)} \lambda_i^2(T) + \sum_{i=k+2}^{d} \lambda_i^2(T) \\
&\leq \epsilon \sum_{i=k+1}^{d} \lambda_i^2(T) + \sum_{i=k+1}^{d} \lambda_i^2(T) \\
&= (1+\epsilon)\|T - T_k\|_F^2.
\end{aligned}$$

Therefore, we have that $\|T - T_{heavy}\|_F \leq (1+\epsilon)\|T - T_k\|_F$, where T_{heavy} contains all buckets with weight more than $\lambda_{Ck\log^5 d/\epsilon}\log d/c'd$. There are at most $r_1 = O(k\log^8 d/\epsilon)$ such buckets by Lemma 1.7. Let $F_{S_1}, \ldots, F_{S_{r_1}}$ and $D_1, D_2, \ldots, D_{r_1}$ to be the Fourier and diagonal weight matrices corresponding to these r_1 heavy buckets defining T_{heavy}. We can then write T_{heavy} as follows

$$T_{heavy} = \sum_{m=1}^{r_1} F_{S_m} D_m F_{S_m}^*.$$

Since each $F_{S_m} D_m F_{S_m}^*$ is $(f, 1/d)$-clustered for some $f \in N$, we use Theorem 1.5 to approximate each $F_{S_m} D_m F_{S_m}^*$ with a rank $O(\log d + \log(1/\delta))$ symmetric Toeplitz matrix $F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*$ that satisfies

$$\left\| F_{S_m} D_m F_{S_m}^* - F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^* \right\|_F \leq \delta(\sum_i (D_m)_i) + \varepsilon d.$$

This implies that

$$\begin{aligned}
\left\| T - \sum_{m=1}^{r_1} F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^* \right\|_F &\leq \left\| T - \sum_{m=1}^{r_1} F_{S_m} D_m F_{S_m}^* \right\|_F + \left\| \sum_{m=1}^{r_1} (F_{S_m} D_m F_{S_m}^* - F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*) \right\|_F \\
&= \|T - T_{heavy}\|_F + \left\| \sum_{m=1}^{r_1} (F_{S_m} D_m F_{S_m}^* - F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*) \right\|_F \\
&\leq \|T - T_{heavy}\|_F + \sum_{m=1}^{r_1} \|F_{S_m} D_m F_{S_m}^* - F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*\|_F \\
&\leq (1+\epsilon)\|T - T_k\|_F + \delta(\sum_{m,i} (D_m)_i) + \varepsilon r d \\
&\leq (1+\epsilon)\|T - T_k\|_F + \delta(\sum_i (D)_i) + \varepsilon r d.
\end{aligned}$$

Now using Lemma 2.2, we get the following:

$$\sum_i (D)_i = \sum_{i=1}^d \lambda_i(T)/d \leq \left(\sqrt{d \sum_{i=1}^d \lambda_i(T)^2} \right) /d = \|T\|_F/\sqrt{d}.$$

Setting $\varepsilon = \delta \|T\|_F / rd$, we finally have

$$\left\| T - \sum_{m=1}^{r_1} F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^* \right\|_F \leq (1+\epsilon) \left\| T - T_k \|_2 + \delta \|T\right\|_F .$$

Defining $\widetilde{T}$ as $\widetilde{T} = \sum_{m=1}^{r_1} F_{\widetilde{S}_m} \widetilde{D}_m F_{\widetilde{S}_m}^*$, which has rank at most $r_1(C(\log d + \log(1/\delta)) = \widetilde{O}((k/\epsilon)\log(1/\delta))$, we get that $\widetilde{T}$ satisfies the claim and point 1 of Theorem 2. Observe that the Toeplitz matrix corresponding to any bucket $F_{S_m} D_m F_{S_m}^*$ is $(f, 1/d)$-clustered for some $f \in N$. Thus, it follows by guarantee 1 of Lemma 1.5 that each $\widetilde{S}_m$ is of the form described in point 2 of Theorem 2. Since $T_{1,1} \leq \|T\|_F$, with this value of ε we can choose any $\gamma \leq \delta T_{1,1}/(\mathrm{tr}(T) r d^2 2^{C_1 \log^7 d}) = \delta/(2^{C_2 \log^7 d})$ for a large enough constant C_2. (Here, we used the fact that $\mathrm{tr}(T) = d T_{1,1}$.) This completes the proof of Theorem 2. □

4 Low-rank approximation with sublinear query complexity.

In this section, we present our main algorithm, Algorithm 1 and prove the corresponding Theorem 1, which shows that the algorithm outputs a near optimal low-rank approximation to T, while reading only sublinearly many entries. We treat $\widetilde{T}$, the low-rank Toeplitz approximation to T (guaranteed to exist from Theorem 2), as the true matrix, which we noisily access by reading entries of T.

4.1 Reduction to weighted linear regression.

The near optimal Toeplitz low-rank approximation $\widetilde{T} \in \mathbb{R}^{d\times d}$ guaranteed to exist by Theorem 2 is of the form $\widetilde{T} = F_{\widetilde{S}}\widetilde{D}F_{\widetilde{S}}^*$, where $F_{\widetilde{S}} \in \mathbb{C}^{d\times r}$ and $\widetilde{D} \in \mathbb{R}^{r\times r}$ are Fourier and diagonal matrices, respectively, and $r = \widetilde{O}((k/\epsilon)\log(1/\delta))$. Algorithm 1 uses brute force search to find the frequencies in $\widetilde{S}$. In particular, it uses an approximate regression oracle to test the quality of any guess for the frequencies in $\widetilde{S}$, without reading many entries of T.

Recall from Theorem 2 that these frequencies lie in $r_1 = \tilde{O}(k/\epsilon)$ clusters of r_2 equispaced frequencies centered around points in $N = \{1/2d, 3/2d, \ldots, 1 - 1/2d\}$. Thus, our search space will be all subsets of r_1 elements of N, of which there at at most N^{r_1}. Formally, the frequencies of our Toeplitz low-rank approximation will lie in the following set:

Definition 4.1 (Frequency Search Space)**.** *Consider any positive integers d, r_1, r_2 with $r_1, r_2 < d$ and $\gamma \in (0, 1)$, Let $N = \{1/2d, 3/2d, \ldots, 1 - 1/2d\}$. For any set of frequencies B, let $S(B) = \bigcup_{b\in B}\bigcup_{1\le j\le r_2}\{b + \gamma j, b - \gamma j\}$. Let $\mathcal{N}_{d,r_1,r_2,\gamma} = \{S(B) : B \in N^{r_1}\}$.*

Since the first column of $\widetilde{T}$ defines the full matrix, the approximate regression oracle will simply attempt to fit the first column of $\widetilde{T}$ to be close to that of T. However, since different entries in the first column appear with different frequencies in the matrix, we require the following weighting function to translate error bounds in the first column into error bounds for approximating the entire matrix T in the Frobenius norm.

Definition 4.2. *Let $w \in \mathbb{R}^d$ be defined as follows:*

$$w_i = \begin{cases} \sqrt{d} & i = 1 \\ \sqrt{2(d-i+1)} & i > 1 \end{cases}$$

Let $W = \mathrm{diag}(w) \in \mathbb{R}^{d\times d}$.

We have the following immediate claim, which expresses the Frobenius norm difference between two symmetric Toeplitz matrices as the weighted ℓ_2 norm difference of their first columns.

Claim 4.3. *Let $T, \widetilde{T} \in \mathbb{R}^{d\times d}$ be symmetric Toeplitz matrices with first columns $T_1, \widetilde{T}_1 \in \mathbb{R}^d$ respectively. Then letting $W \in \mathbb{R}^{d\times d}$ be as in Definition 4.2, $\|T - \widetilde{T}\|_F = \|WT_1 - W\widetilde{T}_1\|_2$.*

Now, for a Toeplitz matrix $T = F_S D F_S^*$, the first column can be expressed as $T_1 = F_S a$ where $a \in \mathbb{R}^r$ contains the diagonal entries of D. Further, if T is real, a must place equal weight on the conjugate frequencies in F_S. Thus, we can in fact write $T_1 = F_S R_S a$ where $a \in \mathbb{R}^{r/2}$ and $R_S \in \mathbb{R}^{r\times r/2}$ collapses the $2|S| = r$ conjugate pair columns of F_S into $|S| = \frac{r}{2}$ real-valued columns, each corresponding to a degree of freedom of a. More formally:

Definition 4.4. *Let $S \subset [0, \frac{1}{2})$ with $|S| = \frac{r}{2}$. Then, define the matrix $R_S \in \mathbb{R}^{r\times\frac{r}{2}}$ by setting the j^{th} column $(R_S)_{:,j}$ equal to 0 everywhere except at j and $|S| + j$, corresponding to the j^{th} pair of conjugate frequencies, where it is equal to 1.*

From Claim 4.3 and the existence proof of Theorem 2 we have the following lemma, which shows that fitting $\widetilde{T}$ can be reduced to a real-valued linear regression problem:

Lemma 4.5 (Toeplitz Approximation via Weighted Regression)**.** *For any PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$ with first column T_1, $\epsilon, \delta \in (0, 1)$, and integer $k \le d$, let $\mathcal{N}_{d,r_1,r_2,\gamma}$ be as in Def. 4.1, where*

r_1, r_2, γ are as in Theorem 2. Let $W \in \mathbb{R}^{d\times d}$ be as in Def. 4.2. Then if $\widetilde{S}$ and $\widetilde{a} \in \mathbb{R}^{|\widetilde{S}|}$ satisfy for some $\alpha \geq 1$:

$$\|WF_{\widetilde{S}}R_{\widetilde{S}}\widetilde{a} - WT_1\|_2 \leq \alpha \cdot \min_{S\in\mathcal{N}_{d,r_1,r_2,\gamma}, a\in\mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2,$$

letting $\widetilde{T} \in \mathbb{R}^{d\times d}$ be the symmetric Toeplitz matrix $\widetilde{T} = F_{\widetilde{S}} \operatorname{diag}(\widetilde{a}) F_{\widetilde{S}}^*$,

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

Proof. Letting $\widetilde{T}_1$ be the first column of $\widetilde{T}$, we have $\widetilde{T}_1 = F_{\widetilde{S}}R_{\widetilde{S}}\widetilde{a}$ where $R_{\widetilde{S}}$ is defined as in Def. 4.4. Thus, by Claim 4.3,

$$\begin{aligned}\|T - \widetilde{T}\|_F = \|WT_1 - WF_{\widetilde{S}}R_{\widetilde{S}}\widetilde{a}\|_2 &\leq \alpha \cdot \min_{S\in\mathcal{N}_{d,r_1,r_2,\gamma}, a\in\mathbb{R}^{|S|}} \|WT_1 - WF_S R_S a\|_2 \\ &\leq (1+\epsilon)\alpha \cdot \|T - T_k\|_F + \alpha\delta\|T\|_F,\end{aligned}$$

where the last inequality follows from Theorem 2, which shows that there is some Toeplitz matrix $\widetilde{T}$ with frequency set in $\mathcal{N}_{d,r_1,r_2,\gamma}$ satisfying $\|T - \widetilde{T}\|_F \leq (1+\epsilon)\|T - T_k\|_F + \delta\|T\|_F$. □

4.2 Leverage score preliminaries.

Our goal is now to find $\widetilde{S} \in \mathcal{N}_{d,r_1,r_2,\gamma}$ and $\widetilde{a} \in \mathbb{R}^{|\widetilde{S}|}$ satisfying the approximate regression guarantee of Lemma 4.5 for $\alpha \leq 1+\epsilon$. We do this using leverage score sampling. It is well known that leverage score sampling can be used to approximately solve linear regression problems in a sample-efficient manner [32, 38]. In our setting, an additional challenge arises in that to find $\widetilde{S}, \widetilde{a}$ satisfying the bound of Lemma 4.5 we must solve many regression problems – corresponding to all possible subsets $\widetilde{S} \in \mathcal{N}_{d,r_1,r_2,\gamma}$ – and output one with near minimal error. This is challenging, since standard results on leverage score sampling for sample-efficient regression 1) typically do not succeed with very high probability, making it difficult to union bound over all possible $\widetilde{S}$, and 2) typically do not output an estimate of the actual regression error, making it impossible to chose a near optimal $\widetilde{S}, \widetilde{a}$ as needed in Lemma 4.5. We show how to overcome these issues in Sections 4.3 and 4.4.

Recall the notion of matrix leverage scores as defined in Definition 2.3. Our algorithm will sample rows of WF_SR_S and WT_1 via approximations to the leverage scores of WF_SR_S to solve the regression problem of Lemma 4.5. For technical reasons, we will sample with a mixture of a leverage score distribution and the uniform distribution, defined below:

Definition 4.6 (Leverage score sampling; Def. 2.7 of [27])**.** *For any number of samples m and a set of leverage score bounds $\{\widetilde{\tau}_j\}_{j\in[n]}$ with $T = \sum_{j=1}^d \widetilde{\tau}_j$, let $\mathcal{S} \in \mathbb{R}^{m\times d}$ be a sampling matrix with each row set independently to the i^{th} standard basis vector multiplied by $(m\cdot p_i)^{-1/2}$, with probability*

$$p_i = \frac{1}{2}\left(\frac{\widetilde{\tau}_i}{T} + \frac{1}{d}\right).$$

We also have the following unbiasedness property of the sampling matrix $\mathcal{S}$.

Claim 4.7. *Let $\mathcal{S}$ be defined as in 4.6. Then*

$$\mathsf{E}[\|\mathcal{S}x\|_2^2] = \|x\|_2^2.$$

Critically, we would like to take *a single set* of samples and use them to approximately minimize $\|WT_1 - WF_S R_S a\|_2$ for all $S \in \mathcal{N}_{d,r_1,r_2,\gamma}$ in order to solve the optimization problem of Lemma 4.5. We are able to do this due to the existence of *universal* leverage score bounds on Fourier matrices [3, 15], which are independent of the frequency subset S. We adapt these bounds to our weighted setting.

Lemma 4.8 (Fourier Matrix Leverage Score Bound)**.** *Let W be the weight matrix defined in 4.2. Then there exist non-negative numbers $\{\widetilde{\tau}_j\}_{j=1}^d$ such that the following hold for all frequency sets $S \subset [0, \frac{1}{2}), |S| = \frac{r}{2} \leq d$:*

1. *$\tau_j(WF_S M_S) \leq \tau_j(WF_S) \leq \widetilde{\tau}_j$ for all $j \in [d]$.*
2. *$\sum_{j=1}^d \widetilde{\tau}_j \leq O(r \log r \cdot \log d)$.*

Here, $F_S \in \mathbb{C}^{d \times r}$ is the symmetric Fourier matrix with frequency set S as defined in 1.2, and M_S is any real-valued matrix with r rows, which may depend on S. (For example, this includes the case $M_S = R_S$ as defined in Definition 4.4.) Finally, $\tau_j(\cdot)$ is as defined in Def. 2.3.

4.8 is similar to Corollary C.2 of [15], but applies to *weighted* Fourier matrices. In 4.6 we restate Corollary C.2 of [15] as 8, and use it to prove 4.8.

In our analysis, we will use the well known fact that sampling $\tilde{O}(r)$ rows of $A \in \mathbb{C}^{d \times r}$ according to their leverage scores yields a *subspace* embedding of A, which preserves the norms of all vectors in A's column span to small relative error. In particular:

Lemma 4.9 (Subspace Embedding [38])**.** *Given $A \in \mathbb{R}^{d \times r}$, failure probability $\eta \in (0, 1)$, and a set of leverage score upper bounds $\{\tilde{\tau}_j\}_{j \in [n]}$ satisfying $\tilde{\tau}_j \geq \tau_j(A)$ for all $j \in [d]$, let $\mathcal{S} \in \mathbb{R}^{m \times d}$ be a random sampling matrix drawn according to 4.6, using the distribution $\widetilde{\tau}_j$ and $m = O\big(\frac{T \log(1/\eta)}{\beta^2}\big)$ samples. Then, with probability $\geq 1 - \eta$, for all $x \in \mathbb{R}^r$,*

$$(1 - \beta)\|Ax\|_2 \leq \|\mathcal{S}Ax\|_2 \leq (1 + \beta)\|Ax\|_2.$$

4.3 Constant factor approximation.

We now use the leverage score bounds of Lemma 4.8 and the subspace embedding guarantee of Lemma 4.9 to show how to solve the optimization problem of Lemma 4.5 for constant α. We later show how to refine this to an $\alpha = 1 + \epsilon$ approximation, achieving our final bound.

Theorem 6 (Constant Factor Toeplitz Fitting)**.** *Consider the setting of Lemma 4.5. Let $\mathcal{S} \in \mathbb{R}^{m \times d}$ be a random sampling matrix drawn according to 4.6, using the distribution $\widetilde{\tau}_j$ defined in Def. 4.8 with $m = \tilde{O}\left(\frac{k^2 \log(1/\delta)}{\epsilon^2}\right)$. Let*

$$\tilde{S}, \tilde{a} = \underset{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}}{\arg\min} \|\mathcal{S}WF_S R_S a - \mathcal{S}WT_1\|_2.$$

Then with probability at least 98/100, $\tilde{S}$ and $\tilde{a}$ satisfy Lemma 4.5 with $\alpha = 41$. In other words,

$$\|WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - WT_1\|_2 \leq 41 \cdot \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2.$$

Proof. Define the optimal frequency set and coefficients by

$$S', a' = \text{argmin}_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2.$$

For any $S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}$ we can write

$$\|WF_S R_S a - WF_{S'} R_{S'} a'\|_2$$

as

$$\|W[F_S R_S, F_{S'} R_{S'}][a; -a']\|_2 = \|W[F_S, F_{S'}][R_S; R_{S'}][a; -a']\|_2.$$

Here, $[F_S, F_{S'}]$ is a Fourier matrix, and $[R_S; R_{S'}]$ fits the criteria of the matrix $M_{S,S'}$ in 4.8. Therefore, by 4.8 the leverage scores of $W[F_S, F_{S'}][R_S; R_{S'}]$ are upper bounded by $\tilde{\tau}_j$ and sum to $O(r \log(d) \log(r))$. Applying 4.9 with $A = W[F_S, F_{S'}][R_S; R_{S'}]$, $\beta = \frac{1}{2}$, $\eta = 1/(100|N|^{r_1})$, and the upper bounds $\tilde{\tau}_j$, we obtain a sampling matrix $\mathcal{S} \in \mathbb{R}^{m \times d}$ that takes $m = \widetilde{O}(\frac{k^2 \log(1/\delta)}{\epsilon^2})$ samples. By combining the probabilistic guarantee of 4.9 with a union bound over all $S \in \mathcal{N}_{d,r_1,r_2,\gamma}$, we get that the following holds with probability at least $1 - |N|^{r_1}\eta = \frac{99}{100}$ over the choice of $\mathcal{S}$:

$$\|\mathcal{S}W[F_S R_S, F_{S'} R_{S'}]x\|_2 \geq \frac{1}{2}\|W[F_S R_S, F_{S'} R_{S'}]x\|_2 \quad \forall S \in \mathcal{N}_{d,r_1,r_2,\gamma},\ x \in \mathbb{R}^r.$$

This implies for the particular case of $S = \tilde{S}$ and $a = \tilde{a}$ that the following holds with probability at least $\frac{99}{100}$:

$$\begin{aligned}
&\leq 2\|\mathcal{S}W[F_{\tilde{S}} R_{\tilde{S}}, F_{S'} R_{S'}][\tilde{a}; -a']\|_2 \\
&= 2\|\mathcal{S}WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - \mathcal{S}WF_{S'} R_{S'} a'\|_2 \\
&\leq 2(\|\mathcal{S}WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - \mathcal{S}WT_1\|_2 + \|\mathcal{S}WT_1 - \mathcal{S}WF_{S'} R_{S'} a'\|_2) \\
&\leq 2(\|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2 + \|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2) \\
&= 4\|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2.
\end{aligned}$$

The last inequality above followed by the definition of $\tilde{S}$ and $\tilde{a}$ as minimizing $\|WF_S R_S a - WT_1\|_2$ over all $S \in \mathcal{N}_{d,r_1,r_2,\gamma}$, $a \in \mathbb{R}^r$.

Moreover, by 4.7, $\mathsf{E}[\|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2^2] = \|WF_{S'} R_{S'} a' - WT_1\|_2^2$. Then by applying Markov's inequality, $\|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2^2 \leq 100\|WF_{S'} R_{S'} a' - WT_1\|_2^2$ with probability at least $\frac{99}{100}$. Finally, we return to the quantity of interest, $\|WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - WT_1\|_2$. By applying a union bound once more, the following then holds with probability at least $\frac{98}{100}$ over the choice of $\mathcal{S}$:

$$\begin{aligned}
\|WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - WT_1\|_2 &\leq \|WF_{S'} R_{S'} a' - WT_1\|_2 + \|WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - WF_{S'} R_{S'} a'\|_2 \\
&\leq \|WF_{S'} R_{S'} a' - WT_1\|_2 + 4\|\mathcal{S}WF_{S'} R_{S'} a' - \mathcal{S}WT_1\|_2 \\
&\leq \|WF_{S'} R_{S'} a' - WT_1\|_2 + 40\|WF_{S'} R_{S'} a' - WT_1\|_2 \\
&\leq 41\|WF_{S'} R_{S'} a' - WT_1\|_2 \\
&= 41 \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2.
\end{aligned}$$

This concludes the proof of the lemma. □

4.4 $(1+\epsilon)$-approximation.

Theorem 6 combined with Lemma 4.5 yields a $\tilde{O}\left(\frac{k^2 \log(1/\delta)}{\epsilon^2}\right)$ query algorithm for outputting $\tilde{T}$ with rank $\tilde{O}(k \log(1/\delta)/\epsilon)$ and $\|T - \tilde{T}\|_F = O(1)\|T - T_k\|_F + \delta\|T\|_F$. To prove Theorem 1 we need to improve this constant factor approximation to $(1+\epsilon)$. We do this using recently developed guarantees for high probability relative error active regression via leverage score sampling [27]. Importantly, we first compute a constant error solution via Theorem 6. We then show that we can fit the *residual* of this approximation to high accuracy via leverage score sampling. Formally,

Theorem 7. *Consider the setting of Lemma 4.5 and let $\tilde{S} \in \mathcal{N}_{d,r_1,r_2,\gamma}$ and $\tilde{a} \in \mathbb{R}^{|\tilde{S}|}$ satisfy the lemma with $\alpha = O(1)$. Let $T_R = T_1 - F_{\tilde{S}} R_{\tilde{S}} \tilde{a}$ be their residual in fitting the first column of T.*

Let $\mathcal{S} \in \mathbb{R}^{m \times d}$ be a random sampling matrix drawn according to 4.6, using the distribution $\widetilde{\tau}_j$ defined in Def. 4.8 with $m = \tilde{O}\left(\frac{k^2 \log(\frac{1}{\delta})}{\epsilon^6}\right)$. Let

$$S', a' = \underset{S \in \mathcal{N}_{d,2r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}}{\arg\min} \|\mathcal{S}WF_S R_S a - \mathcal{S}WT_R\|_2.$$

Then with probability at least 99/100, letting $\bar{S} = \tilde{S} \cup S'$ and $\bar{a} = [\tilde{a}, a']$, we have that $\bar{S}, \bar{a}$ satisfy Lemma 4.5 with $\alpha = (1+\epsilon)$. I.e.,

$$\|WF_{\bar{S}} R_{\bar{S}} \bar{a} - WT_1\|_2 \leq (1+\epsilon) \cdot \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2,$$

and thus, letting $\bar{T} = F_{\bar{S}} \operatorname{diag}(\bar{a}) F_{\bar{S}}^*$,

$$\|T - \bar{T}\|_F \leq (1+3\epsilon)\|T - T_k\|_F + 2\delta\|T\|_F.$$

Note that $\bar{S}$ output by Theorem 7 does not lie in $\mathcal{N}_{d,r_1,r_2,\gamma}$. Since $S' \in \mathcal{N}_{d,2r_1,r_2,\gamma}$ and $\tilde{S} \in \mathcal{N}_{d,r_1,r_2,\gamma}$, we have $\bar{S} \in \mathcal{N}_{d,3r_1,r_2,\gamma}$. Lemma 4.5 allows this – it simply means that the rank of the corresponding Toeplitz matrix $\bar{T}$ may be three times as large as if $\bar{S}$ were in $\mathcal{N}_{d,r_1,r_2,\gamma}$.

Proof. Let $OPT = \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_S R_S a - WT_1\|_2$. Observe that

$$\|WF_{\bar{S}} R_{\bar{S}} \bar{a} - WT_1\|_2 = \|WF_{S'} R_{S'} a' + WF_{\tilde{S}} R_{\tilde{S}} \tilde{a} - WT_1\|_2 = \|WF_{S'} R_{S'} a' - WT_R\|_2.$$

Thus, to prove the theorem, it suffices to show that

$$\|WF_{S'} R_{S'} a' - WT_R\|_2 \leq (1+\epsilon) \cdot OPT. \tag{4}$$

Further note that by the assumption that $\tilde{S}, \tilde{a}$ satisfy Lemma 4.5 with $\alpha = O(1)$ we have:

$$\|WT_R\|_2 \leq \alpha \cdot OPT = O(OPT).$$

By Markov's inequality, since by construction $\mathsf{E}[\|\mathcal{S}WT_R\|_2^2] = \|WT_R\|_2^2$, with probability at least 999/1000, this also gives that $\|\mathcal{S}WT_R\|_2 \leq \sqrt{1000} \cdot \alpha \cdot OPT = O(OPT)$.

For a given $S \in \mathcal{N}_{d,2r_1,r_2,\gamma}$ define:

$$a_S = \underset{a \in \mathbb{R}^{|S|}}{\arg\min} \|WF_S R_S a - WT_R\|_2$$

and

$$a'_S = \underset{a \in \mathbb{R}^{|S|}}{\arg\min} \|\mathcal{S}WF_S R_S a - \mathcal{S}WT_R\|_2.$$

Using that $\|WT_R\|_2 = O(OPT)$ and applying triangle inequality, we have

$$\begin{aligned} \|WF_S R_S a_S - WT_R\|_2 &\leq \|WT_R\|_2 \\ \|WF_S R_S a_S\|_2 - \|WT_R\|_2 &\leq \|WT_R\|_2 \\ \|WF_S R_S a_S\|_2 &= O(OPT). \end{aligned} \tag{5}$$

Similarly, using that $\|SWT_R\|_2 = O(OPT)$ with good probability, for all S we have $\|\mathcal{S}WF_S R_S a'_S\|_2 = O(OPT)$. Further, as in Theorem 6, with probability at least $1 - |N|^{2r_1}\eta = 99/100$ for $\eta =$

$1/(100|N|^{2r_1})$, the subspace embedding guarantee of Lemma 4.9 holds for all WF_SR_S simultaneously for $\beta = O(1)$, giving that for all $S \in \mathcal{N}_{d,2r_1,r_2,\gamma}$

$$\|WF_SR_Sa'_S\|_2 = O(OPT). \tag{6}$$

Given (5) and (6), to prove the theorem it suffices to show the following claim:

Claim 4.10. *With probability at least* 99/100, *for any* $S \in \mathcal{N}_{d,2r_1,r_2,\gamma}$ *and any* $a \in \mathbb{R}^{|S|}$ *with* $\|WF_SR_Sa\|_2 = O(OPT)$,

$$\left|\|WF_SR_Sa - WT_R\|_2^2 - \|\mathcal{S}WF_SR_Sa - \mathcal{S}WT_R\|_2^2 - C\right| \leq \epsilon \cdot OPT^2,$$

where C *is a fixed constant that may depend on* $\mathcal{S}$ *and* WT_R, *but does not depend on* S *or* a.

I.e., for any frequency set S and coefficient vector $a \in \mathbb{R}^{|S|}$ where $\|WF_SR_Sa\|_2 = O(OPT)$, the sampled regression cost, after shifting by a fixed constant, approximates the true regression cost up to additive error $\epsilon \cdot OPT^2$. This ensures that:

$$\begin{aligned}\|WF_{S'}R_{S'}a' - WT_R\|_2^2 &= \|WF_{S'}R_{S'}a'_{S'} - WT_R\|_2^2 \\ &\leq \|\mathcal{S}WF_{S'}R_{S'}a'_{S'} - \mathcal{S}WT_R\|_2^2 + C + \epsilon \cdot OPT^2 \\ &\leq \|\mathcal{S}WF_{S^*}R_{S^*}a^* - \mathcal{S}WT_R\|_2^2 + C + \epsilon \cdot OPT^2,\end{aligned}$$

where $S^*, a^* = \arg\min_{S \in \mathcal{N}_{d,2r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_SR_Sa - WT_R\|_2$. The first inequality follows from Claim 4.10, which can be applied since $\|WF_{S'}R_{S'}a'_{S'}\|_2 = O(OPT)$ by (6). The second inequality follows since $S', a'_{S'} = \arg\min_{S \in \mathcal{N}_{d,2r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|\mathcal{S}WF_SR_Sa - \mathcal{S}WT_R\|_2$. Applying Claim 4.10 again to S^*, a^*, which is valid by (5), we continue to bound:

$$\begin{aligned}\|WF_{S'}R_{S'}a' - WT_R\|_2^2 &\leq \|\mathcal{S}WF_{S^*}R_{S^*}a^* - \mathcal{S}WT_R\|_2^2 + C + \epsilon \cdot OPT^2 \\ &\leq \|WF_{S^*}R_{S^*}a^* - WT_R\|_2^2 + 2\epsilon \cdot OPT^2. \end{aligned}\tag{7}$$

Finally, note that since we allow $S^* \in \mathcal{N}_{d,2r_1,r_2,\gamma}$, we have:

$$\begin{aligned}\|WF_{S^*}R_{S^*}a^* - WT_R\|_2^2 &\leq \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_SR_Sa + WF_{\tilde{S}}R_{\tilde{S}}\tilde{a} - WT_R\|_2^2 \\ &= \min_{S \in \mathcal{N}_{d,r_1,r_2,\gamma}, a \in \mathbb{R}^{|S|}} \|WF_SR_Sa - WT_1\|_2^2 = OPT^2.\end{aligned}$$

Combined with (7), this gives that $\|WF_{S'}R_{S'}a' - WT_R\|_2^2 \leq (1+2\epsilon) \cdot OPT^2$, which, after taking a square root and adjusting ϵ by a constant yields (4) and in turn the theorem. ☐

We finally present the proof of Claim 4.10 below.

Proof. Claim 4.10 can be proven following the same approach as Theorem 3.4 of [27]. For simplicity of notation, let $z := WT_R$. We define a set of 'bad indices' where the relative size of z_j is significantly larger than the leverage score $\tilde{\tau}_j$. The regression error on these bad indices will not be well approximated via leverage score sampling. However, this is ok, since no WF_SR_Sa can do a good job fitting these indices, given the leverage score bounds on WF_SR_S. Formally, let

$$\mathcal{B} = \left\{j \in [d] : \frac{z_j^2}{OPT^2} \geq \frac{\tilde{\tau}_j}{\epsilon^2}\right\}.$$

Further, we let $\bar{z} \in \mathbb{R}^d$ be equal to z, except with $\bar{z}_j = 0$ for all $j \in \mathcal{B}$. Importantly $\mathcal{B}$ and $\bar{z}$ are defined independently of any specific S. For any $S, a \in \mathbb{R}^{|S|}$ with $\|WF_S R_S a\|_2 = O(OPT)$ and any $j \in \mathcal{B}$, we have by the definition of the leverage score (Def. 2.3),

$$\begin{aligned}\left|(WF_S R_S a)_j\right|^2 &\leq \|WF_S R_S A\|_2^2 \cdot \tilde{\tau}_j \\ &\leq O(OPT^2) \cdot \frac{\epsilon^2 \cdot |z_j|^2}{OPT^2} \\ &= O(\epsilon^2 \cdot |z_j|^2).\end{aligned}$$

This gives that

$$\left|(WF_S R_S a)_j - z_j\right|^2 - \left|(WF_S R_S a)_j - \bar{z}_j\right|^2 = \left|(WF_S R_S a)_j - z_j\right|^2 - \left|(WF_S R_S a)_j\right|^2 = (1 \pm O(\epsilon)) \cdot z_j^2.$$

Thus, for $C_1 = \sum_{j \in \mathcal{B}} |z_j|^2 = \|z - \bar{z}\|_2^2$,

$$\left| \|WF_S R_S a - z\|_2^2 - \|WF_S R_S a - \bar{z}\|_2^2 - C_1 \right| = O(\epsilon) \cdot OPT^2, \tag{8}$$

where we use that $\|z\|_2^2 = O(OPT^2)$. Using the same proof, and the fact that with probability at least 999/1000 by Markov's inequality, $\|\mathcal{S}z\|_2^2 = O(OPT^2)$, we have for $C_2 = \|\mathcal{S}(z - \bar{z})\|_2^2$,

$$\left| \|\mathcal{S}WF_S R_S a - \mathcal{S}z\|_2^2 - \|\mathcal{S}WF_S R_S a - \mathcal{S}\bar{z}\|_2^2 - C_2 \right| = O(\epsilon) \cdot OPT^2. \tag{9}$$

Observe that C_1 and C_2 only depend on the sampling matrix $\mathcal{S}$ and truncated vector $\bar{z}$, whose definition is independent of any specific frequency set S or coefficient vector a. Thus, with (8) and (9) in place, to prove Claim 4.10 it suffices to show that, for all $S \in \mathcal{N}_{d,r_1,r_2,\gamma}$ and $a \in \mathbb{R}^{|S|}$ with $\|WF_S R_S a\|_2 = O(OPT)$,

$$\left| \|WF_S R_S a - \bar{z}\|_2^2 - \|\mathcal{S}WF_S R_S a - \mathcal{S}\bar{z}\|_2^2 \right| = O(\epsilon) \cdot OPT^2. \tag{10}$$

Observe that by definition of $\mathcal{B}$, the entries of $\bar{z}$ are bounded by $\bar{z}_j^2 \leq OPT^2 \cdot \frac{\tilde{\tau}_j}{\epsilon^2}$. Similarly, since by assumption $\|WF_S R_S a\|_2^2 = O(OPT^2)$ and by the definition of the leverage scores (Def. 2.3) $(WF_S R_S a)_j^2 \leq OPT^2 \cdot \tilde{\tau}_j$. Thus, sampling entries with probabilities proportional to their leverage scores as in Def. 4.6 ensures that for any fixed S and $a \in \mathbb{R}^{|s|}$, by a standard Bernstein bound, (10) holds with high probability. This bound can then be extended to hold to all S and a via an ϵ-net analysis as follows. Fix an S. For simplicity we assume by scaling that $\|z\|_2 = 1$ and $OPT = \Theta(1)$. By Claim 3.8 in the version 1 of [27], to prove (10) it suffices to show that the following holds with high probability, for all $y \in \mathcal{N}_\epsilon$,

$$\left| \|\mathcal{S}y - \mathcal{S}\bar{z}\|_2^2 - \|y - \bar{z}\|_2^2 \right| \leq \epsilon$$

where $\mathcal{N}_\epsilon$ is an ϵ-net of the set $\{WF_S R_S a : \|WF_S R_S a\|_2 \leq 1\}$. By a standard volume argument, it is known that one can construct such a net with $\log |\mathcal{N}_\epsilon| = \widetilde{O}(r_1 r_2 \log(1/\epsilon))$. We will show equation (10) for a fixed S and all $a \in \mathbb{R}^{|S|}$ using Bernstein's inequality and a union bound. We have that $\mathbb{E}[\|\mathcal{S}y - \mathcal{S}\bar{z}\|_2^2] = \|y - \bar{z}\|_2^2 = O(1)$. Additionally by definition $|\bar{z}_i|^2 \leq \frac{\widetilde{\tau}_i}{\epsilon^2}$ for all i. Similarly by the definition of leverage scores $|y_i|^2 \leq \widetilde{\tau}_j$. This implies the following,

$$|y_i - \bar{z}_i|^2 = O\left(\frac{\widetilde{\tau}_i}{\epsilon^2}\right)$$

for all i. By the construction of $\mathcal{S}$, we have the following,

$$|[\mathcal{S}y - \mathcal{S}\bar{z}]_i|^2 \leq \frac{\sum_i \widetilde{\tau}_i}{m\widetilde{\tau}_i} \cdot \frac{\widetilde{\tau}_i}{\epsilon^2} \leq \widetilde{O}\left(\frac{r_1 r_2}{m\epsilon^2}\right)$$

for all i. Thus, by applying a Bernstein bound, we get the following,

$$\begin{aligned} Pr[|\|\mathcal{S}y - \mathcal{S}\bar{z}\|_2^2 - \|y - \bar{z}\|_2^2| > \epsilon] &\leq 2\exp\left(-\widetilde{\Omega}\left(\frac{\epsilon^4 m}{r_1 r_2}\right)\right) \\ &\leq \eta/|\mathcal{N}_\epsilon| \end{aligned}$$

for $m = \widetilde{O}(r_1 r_2 \log(|\mathcal{N}_\epsilon|/\eta)/\epsilon^4)$. Taking a union bound over all y and by Claim 3.8 of [27], we get that equation (10) holds for a fixed S and all $a \in \mathbb{R}^{|S|}$ with probability at least $1 - \eta$. Setting $\eta = \frac{1}{100 \cdot |\mathcal{N}_{d,2r_1,r_2,\gamma}|} = \frac{1}{100 \cdot |N|^{2r_1}}$ in the previous corollary, we have that for

$$\begin{aligned} m &= \tilde{O}\left(\frac{r_1 r_2}{\epsilon^4} \log(|\mathcal{N}_\epsilon| \cdot 100|N|^{2r_1})\right) \\ &= \tilde{O}\left(\frac{r_1 r_2}{\epsilon^4} \left(\log(|\mathcal{N}_\epsilon|) + \log(|N|^{2r_1})\right)\right) \\ &= \tilde{O}\left(\frac{r_1 r_2}{\epsilon^4} (r_1 r_2 + r_1)\right) \\ &= \tilde{O}\left(\frac{k^2 \log(1/\delta)}{\epsilon^6}\right), \end{aligned}$$

equation (10) holds for all $S \in \mathcal{N}_{d,2r_1,r_2,\gamma}$ simultaneously with probability at least 99/100. This completes the proof of the claim. □

4.5 Full algorithm.

Equipped with these tools, we now describe the recovery algorithm and its guarantees.

Algorithm 1: ToeplitzRecovery

1: **Input:** Query access to $T \in \mathbb{R}^{d\times d}$, k, ϵ, δ.
2: **Init:** Set $r_1 = O(k \log^8 d/\epsilon)$, $r_2 = O(\log d + \log(1/\delta))$, $\eta = \frac{1}{100|N|^{r_1}}$, $\widetilde{\tau}_j$ as defined in 4.8, $\mathcal{N}_{d,r_1,r_2,\gamma}$ and $\mathcal{N}_{d,2r_1,r_2,\gamma}$ as defined in 4.1, and $\gamma = \delta/(2^{C_2 \log^7(d)})$ as in 2.
3: Draw $\mathcal{S}_1$ according to 4.6, using $\widetilde{\tau}_j$ with $m = \tilde{O}\left(\frac{k^2 \log(\frac{1}{\delta})}{\epsilon^2}\right)$ samples.
4: Set $\tilde{S}, \tilde{a} = \arg\min_{S\in\mathcal{N}_{d,r_1,r_2,\gamma}, a\in\mathbb{R}^{|S|}} \|\mathcal{S}_1 W F_S R_S a - \mathcal{S}_1 W T_1\|_2$.
5: Set $T_R = T_1 - F_{\tilde{S}} R_{\tilde{S}} \tilde{a}$.
6: Draw $\mathcal{S}_2$ according to 4.6, using $\widetilde{\tau}_j$ with $m = \tilde{O}\left(\frac{k^2 \log(\frac{1}{\delta})}{\epsilon^6}\right)$ samples.
7: Set $S', a' = \arg\min_{S\in\mathcal{N}_{d,2r_1,r_2,\gamma}, a\in\mathbb{R}^{|S|}} \|\mathcal{S}_2 W F_S R_S a - \mathcal{S}_2 W T_R\|_2$.
8: Set $\bar{S} := \tilde{S} \cup S'$, $d := \bar{a} = [\tilde{a}; a']$, and $F = F_{\bar{S}}$.
9: **Return:** F, d.

Theorem 1. *Assume we are given query access to a PSD Toeplitz matrix $T \in \mathbb{R}^{d\times d}$ and parameters k, ϵ, δ. Let $r = \widetilde{O}(\frac{k}{\epsilon}\log(1/\delta))$. Then Algorithm 1 returns $F \in \mathbb{C}^{d\times\ell}, d \in \mathbb{R}^\ell$ such that $F\,\mathrm{diag}(d)F^*$ is a symmetric Toeplitz matrix of rank at most $\ell = O(r)$ and the following holds.*

1. *Algorithm 1 makes* $\tilde{O}\left(\frac{k^2 \log(\frac{1}{\delta})}{\epsilon^6}\right)$ *queries to* T.

2. $\|T - F\,\mathrm{diag}(d)F^*\|_F \le (1+3\epsilon)\|T - T_k\|_F + 2\delta\|T\|_F$ *with probability at least* $97/100$.

Proof. By 6, with probability at least $98/100$, $\tilde{S}$ and $\tilde{a}$ satisfy 4.5 with $\alpha = 41$. The conditions for 7 are then satisfied, so with conditional probability at least $99/100$, the Toeplitz matrix $\bar{T} := F\,\mathrm{diag}(d)F^*$ satisfies

$$\|T - \bar{T}\|_F \le (1+3\epsilon)\|T - T_k\|_F + 2\delta\|T\|_F.$$

By a union bound, the matrix $\bar{T} := F\,\mathrm{diag}(d)F^*$ based on the output of Algorithm 1 then satisfies $\|T - \bar{T}\|_F \le (1+3\epsilon)\|T - T_k\|_F + 2\delta\|T\|_F$ with probability at least $97/100$. The overall sample complexity is $\tilde{O}\left(\frac{k^2 \log(\frac{1}{\delta})}{\epsilon^6}\right)$, as this is the sample complexity of $\mathcal{S}_2$ and dominates that of $\mathcal{S}_1$. □

4.6 Leverage score bounds for weighted Fourier matrices.

In this subsection, we focus on proving the leverage score upper bounds for weighted Fourier matrices of Lemma 4.8. Recall the statement of Lemma 4.8 was as follows.

Lemma 4.8. *Let* $F_S \in \mathbb{C}^{d\times r}$ *be any Fourier matrix with symmetric (in the sense of 3) frequency set* $S \subset [0,1], |S| = r \le d, r \le d$, *and let* W *be as defined in 4.2 and and* M_S *be any real valued matrix with* r *rows. Then there exist non-negative numbers* $\widetilde{\tau}_j$ *for all* $j \in [d]$ *such that the following hold:*

1. $\tau_j(WF_SM_S) \le \tau_j(WF_S) \le \widetilde{\tau}_j$ *for all* $j \in [d]$.

2. $\sum_{j=1}^d \widetilde{\tau}_j \le O(r\log(d)\log(r))$.

To prove this lemma, we will use the following three helper lemmas.

Lemma 4.11. *Let* $A \in \mathbb{C}^{d\times r}$ *be a matrix, and let* D *be a diagonal matrix with positive entries satisfying*

$$\alpha \le D_{ii}^2 \le \beta.$$

Then,

$$\tau_i(DA) \le \frac{\beta}{\alpha}\tau_i(A).$$

Proof. By the definition of leverage scores, we have the following:

$$\tau_i(DA) = \max_{y\in\mathbb{C}^r} \frac{|DAy|_i^2}{\sum_{j=1}^d |DAy|_j^2}.$$

For any y, the numerator is given by $|DAy|_i^2 = D_{ii}^2|Ay|_i^2$. The denominator satisfies

$$\sum_{j=1}^d |DAy|_j^2 \ge \alpha \sum_{j=1}^d |Ay|_j^2.$$

As a result,

$$\max_{y\in\mathbb{C}^r} \frac{|DAy|_i^2}{\sum_{j=1}^d |DAy|_j^2} = \max_{y\in\mathbb{C}^r} \frac{D_{ii}^2|Ay|_i^2}{\sum_{j=1}^d |DAy|_j^2} \le \max_{y\in\mathbb{C}^r} \frac{\beta|Ay|_i^2}{\alpha\sum_{j=1}^d |Ay|_j^2}.$$

This proves that $\tau_i(DA) \le \frac{\beta}{\alpha}\tau_i(A)$ as desired. □

Lemma 4.12. *Let $A \in \mathbb{C}^{d\times r}$ be a matrix, and let $B \in \mathbb{C}^{d'\times r}$ be a matrix formed by taking any subset $R \subseteq [1,\dots,d]$ of the rows of A. Then the leverage score of any row b_i in B is at least the leverage score of row b_i in A.*

Proof. Let b_i correspond to the i^{th} row of A and the i'^{th} row of B. Then by the definition of leverage scores we have the following.

$$\begin{aligned}\tau_i(A) &= \max_y \frac{|Ay|_i^2}{\sum_{j=1}^d |Ay|_j^2} \\ &\le \max_y \frac{|Ay|_i^2}{\sum_{j\in R} |Ay|_j^2} \\ &= \max_y \frac{|By|_{i'}^2}{\sum_{j=1}^{d'} |By|_j^2} \\ &= \tau_{i'}(B).\end{aligned}$$

This completes the proof of the lemma. □

Lemma 4.13. *Let $A \in \mathbb{C}^{d\times r}$ be a matrix, and let $B \in \mathbb{C}^{d\times r'}$ be a matrix formed by taking r' linear combinations of the columns of A (i.e. there is some matrix $M \in \mathbb{C}^{r\times r'}$ such that $B = AM$). Then the i^{th} leverage score of B, $\tau_i(B)$, is at most the i^{th} leverage score of A, $\tau_i(A)$.*

Proof. We again rely on the maximization characterization of leverage scores:

$$\begin{aligned}\tau_i(B) &= \max_y \frac{|By|_i^2}{\sum_{j=1}^d |By|_j^2} \\ &= \max_y \frac{|AMy|_i^2}{\sum_{j=1}^d |AMy|_j^2} \\ &\le \max_z \frac{|Az|_i^2}{\sum_{j=1}^d |Az|_j^2} \\ &= \tau_i(A).\end{aligned}$$

where the second inequality follows from the fact that $\{My | y \in \mathbb{C}^{r'}\} \subseteq \{z | z \in \mathbb{C}^r\}$. □

Equipped with these lemmas, the high level strategy for bounding the leverage scores of WF_SM_S is as follows. By 4.13, it suffices to bound the leverage scores of WF_S. To do so, first we will bucket the rows of WF_S into submatrices, such that within each submatrix, the weights vary by at most a constant factor. We will then be able to bound the leverage scores of each submatrix using Lemma 4.11, and apply Lemma 4.12 to ensure that these upper bounds remain valid upper bounds for the entire matrix WF_S.

Formally, begin by dividing WF_S into submatrices as follows.

Definition 4.14. *For all $i \in [\log d]$, let W_i be the submatrix of W consisting of rows with indices in $[d(1-1/2^{i-1})+1, d(1-1/2^i)]$. For the edge case $i = (\log d)+1$, let W_i consist of only the last row of WF_S. For convenience, let $\mathcal{R}_i$ denote the index set of rows corresponding to W_i, and let $r_i = |\mathcal{R}_i|$ denote the number of rows in W_i. Note also that $r_i \le \frac{d}{2^i}$ for $i \in [\log d]$, and $r_{\log d+1} = 1$.*

Then we can write WF_S as $WF_S = [(W_1F_S)^T; \ldots; (W_{\log d+1}F_S)^T]^T$. Now consider any $i \in [\log d + 1]$ and W_iF_S. Since W_i is diagonal and $d/2^{i-1} \leq (W_i)^2_{j,j} \leq 2d/2^{i-1}$ for all $j \in [r_i]$, applying Lemma 4.11 with $\beta = 2\alpha = 2d/2^i$ we get the following for all $i \in [\log d + 1]$:

$$\tau_j(W_iF_S) \leq 2\tau_j(F_{S,i}) \quad \forall j \in [r_i].$$

Here, $F_{S,i} \in \mathbb{C}^{r_i \times s}$ is the matrix consisting of all rows of F_S in the index set $\mathcal{R}_i$. Note that the column span of $F_{S,i}$ and the first r_i rows of F_S is identical. This is because the j^{th} column of $F_{S,i}$ is a constant times the j^{th} column of the matrix formed by considering the first r_i rows of F_S. Since the first r_i rows of F_S forms a $r_i \times r$ Fourier matrix, their leverage scores are then identical as well. Finally, we appeal to Corollary C.2 of [15], restated below:

Corollary 8 (Corollary C.2 of [15])**.** *For any positive integers d and $s \leq d$, there is an explicit set of values $\widetilde{\tau}_1^{(s)}, \ldots, \widetilde{\tau}_d^{(s)} \in (0, 1])$ such that, for any Fourier matrix $F_S \in \mathbb{C}^{d \times s}$ with leverage scores $\tau_1, \ldots, \tau_d$,*

$$\forall j, \widetilde{\tau}_j^{(s)} \geq \tau_j.$$
$$\sum_{j=1}^{d} \widetilde{\tau}_j^{(s)} = O(s \log s).$$

Thus by Corollary 8 (C.2 of [15]), we easily obtain the following claim.

Claim 4.15. *For any $i \in [\log(d/r)]$ and any $j \in [r_i]$ define $\widetilde{\tau}_{j,i}$ as follows.*

$$\widetilde{\tau}_{j,i} = \min\left(1, \frac{r}{\min(j, d/2^i + 1 - j)}, \frac{O(r^6 \log^3(r+1))}{(d/2^i)}\right).$$

For any $\log(d/r) < i \leq \log(d)$ and $j \in [r_i]$ let $\widetilde{\tau}_{j,i} = 1$. Then we have the following.

1. *$\tau_j(W_iF_S) \leq \widetilde{\tau}_{j,i}$ for all $i \in [\log d]$ and $j \in [r_i]$.*
2. *$\sum_{j=1}^{d/2^i} \widetilde{\tau}_j \leq O(r \log r)$ for all $i \in [\log d]$.*

Equipped with these tools, we can now easily finish the proof of Lemma 4.8.

Proof. Define $\widetilde{\tau}_j$ values for $j \in [d]$ as follows.

$$\widetilde{\tau}_j = \widetilde{\tau}_{i,j'}. \tag{11}$$

where i, j' are chosen such that $d(1 - 1/2^{i-1}) + 1 \leq j \leq d(1 - 1/2^i)$ and $j = d(1 - 1/2^{i-1}) + 1 + j'$, and $\widetilde{\tau}_{i,j}$ are obtained from Claim 4.15. Finally, using Lemma 4.12 we can easily conclude that these upper bounds also serve as upper bounds on the leverage scores of WF_S, and by 4.13, upper bounds on the leverage scores of WF_SM_S as well. This completes the proof of Lemma 4.8. ∎

5 Conclusion.

In this paper, we study the design of sublinear algorithms for obtaining low-rank approximations of positive semidefinite Toeplitz matrices. Given query access to any such matrix $T \in \mathbb{R}^{d \times d}$, one can

trivially reconstruct it exactly by reading its first column, i.e. by reading d entries. Our main result is that for any k, ϵ, δ, there exists a symmetric Toeplitz $\widetilde{T}$ of rank $\widetilde{O}((k/\epsilon)\log(1/\delta))$ satisfying

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

where $T_k = \underset{B:rank(B)\leq k}{\arg\min} \|T - B\|_F$ is the best rank-k approximation T in the Frobenius norm. Surprisingly, such an existence result – that there exists a near optimal low-rank approximation to T which is itself Toeplitz – was not known before. We obtain this result by proving new results about the low rank structure of off-grid Fourier matrices, which we believe to be of independent interest. We also present an algorithm that reconstructs such a $\widetilde{T}$ by reading only $\widetilde{O}(k^2\log(1/\delta)\mathsf{poly}(1/\epsilon))$ entries of T, beating the trivial bound of d queries and thus achieving sublinear query complexity. We now present some of the main open problems raised by this work.

1. Is the additive error term in Theorem 2 necessary? Also, what is the minimum rank of required to achieve the guarantee of Theorem 2?

2. Is it possible to design a sublinear time low-rank approximation algorithm that recovers a $\widetilde{T}$ achieving the guarantee of Theorem 1?

3. Is it possible to design an algorithm with sublinear query complexity, or even sublinear runtime, that can recover a $\widetilde{T}$ satisfying the spectral norm low-rank approximation guarantee of Theorem 5?

4. Can the existence of a structure preserving low-rank approximation be proven for non-PSD Toeplitz matrices, similar to Theorem 2? Can a sublinear query and sublinear time algorithm be designed to recover near optimal low-rank approximations to non-PSD Toeplitz matrices?

6 Acknowledgements.

Michael Kapralov and Mikhail Makarov's work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 759471). Hannah Lawrence is supported by the Fannie and John Hertz Foundation and the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. Cameron Musco's work is supported by an Adobe Research grant, a Google Research Scholar Award, and NSF Grants No. 2046235 and No. 1763618.

7 References

[1] Yuri I Abramovich, Nicholas K Spencer, and Alexei Y Gorokhov. Positive-definite Toeplitz completion in DOA estimation for nonuniform linear antenna arrays. II. Partially augmentable arrays. *IEEE Transactions on Signal Processing*, 47(6):1502–1521, 1999.

[2] Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 141–160, 2020.

[3] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. A universal sampling method for reconstructing signals with simple Fourier

transforms. In *Proceedings of the 51st Annual ACM Symposium on Theory of Computing (STOC)*, pages 1051–1063, 2019.

[4] Ainesh Bakshi, Nadiia Chepurko, and David 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)*, pages 506–516, 2020.

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

[6] James R Bunch. Stability of methods for solving Toeplitz systems of equations. *SIAM Journal on Scientific and Statistical Computing*, 6(2):349–364, 1985.

[7] Jian-Feng Cai, Xiaobo Qu, Weiyu Xu, and Gui-Bo Ye. Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction. *Applied and Computational Harmonic Analysis*, 41(2):470–490, 2016.

[8] Xue Chen, Daniel M Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 741–750, 2016.

[9] Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[10] Moody T Chu, Robert E Funderlic, and Robert J Plemmons. Structured low rank approximation. *Linear Algebra and its Applications*, 366:157–172, 2003.

[11] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 163–172, 2015.

[12] George Cybenko. Moment problems and low rank Toeplitz approximations. *Circuits, Systems and Signal Processing*, 1(3):345–366, 1982.

[13] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[14] Carlos Echeverría, Jörg Liesen, and Reinhard Nabben. Block diagonal dominance of matrices revisited: bounds for the norms of inverses and eigenvalue inclusion sets. *Linear Algebra and its Applications*, 553:365–383, 2018.

[15] Yonina C. Eldar, Jerry Li, Cameron Musco, and Christopher Musco. Sample efficient Toeplitz covariance estimation. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 378–397, 2020.

[16] Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

[17] Robert M Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[18] Piotr Indyk, Ali Vakilian, Tal Wagner, and David P Woodruff. Sample-optimal low-rank approximation of distance matrices. In *Proceedings of the 32nd Annual Conference on Computational Learning Theory (COLT)*, pages 1723–1751. PMLR, 2019.

[19] Mariya Ishteva, Konstantin Usevich, and Ivan Markovsky. Factorization approach to structured low-rank approximation with applications. *SIAM Journal on Matrix Analysis and Applications*, 35(3):1180–1204, 2014.

[20] Hanna Knirsch, Markus Petz, and Gerlind Plonka. Optimal rank-1 Hankel approximation of matrices: Frobenius norm and spectral norm and Cadzow's algorithm. *Linear Algebra and its Applications*, 629:1–39, 2021.

[21] Hamid Krim and Mats Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996.

[22] Hannah Lawrence, Jerry Li, Cameron Musco, and Christopher 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)*, pages 4796–4800, 2020.

[23] Franklin T Luk and Sanzheng Qiao. A symmetric rank-revealing Toeplitz matrix decomposition. *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, 14(1):19–28, 1996.

[24] Ankur Moitra. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 821–830, 2015.

[25] Joakim Munkhammar, Lars Mattsson, and Jesper Rydén. Polynomial probability distribution estimation using the method of moments. *PloS One*, 12(4), 2017.

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

[27] Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond. In *Proceedings of the 63rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2022.

[28] Gregory Ongie and Mathews Jacob. A fast algorithm for convolutional structured low-rank matrix recovery. *IEEE Transactions on Computational Imaging*, 3(4):535–550, 2017.

[29] Victor Y Pan and Zhao Q Chen. The complexity of the matrix eigenproblem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing (STOC)*, pages 507–516, 1999.

[30] Haesun Park, Lei Zhang, and J Ben Rosen. Low rank approximation of a Hankel matrix by structured total least norm. *BIT Numerical Mathematics*, 39(4):757–779, 1999.

[31] Heng Qiao and Piya Pal. Gridless line spectrum estimation and low-rank Toeplitz matrix compression using structured samplers: A regularization-free approach. *IEEE Transactions on Signal Processing*, 65(9):2221–2236, 2017.

[32] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[33] Xiaofei Shi and David P Woodruff. Sublinear time numerical linear algebra for structured matrices. In *AAAI Conference on Artificial Intelligence*, 2019.

[34] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[35] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[36] Christiane Tretter. *Spectral theory of block operator matrices and applications*. World Scientific, 2008.

[37] Ruiping Wen and Yaru Fu. Toeplitz matrix completion via a low-rank approximation algorithm. *Journal of Inequalities and Applications*, 2020(1):1–13, 2020.

[38] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[39] Yuanzhe Xi, Jianlin Xia, Stephen Cauley, and Venkataramanan Balakrishnan. Superfast and stable structured solvers for Toeplitz least squares via randomized sampling. *SIAM Journal on Matrix Analysis and Applications*, 35(1):44–72, 2014.

[40] Jianlin Xia, Yuanzhe Xi, and Ming Gu. A superfast structured solver for Toeplitz linear systems via randomized sampling. *SIAM Journal on Matrix Analysis and Applications*, 33(3):837–858, 2012.

[41] Taisuke Yasuda, David Woodruff, and Manuel 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)*, pages 7055–7063, 2019.