Sharper Bounds for Chebyshev Moment Matching with Applications to Differential Privacy and Beyond

Cameron Musco
UMass Amherst
cmusco@cs.umass.edu

Christopher Musco
New York University
cmusco@nyu.edu

Lucas Rosenblatt
New York University
lucas.rosenblatt@nyu.edu

Apoorv Vikram Singh
New York University
apoorv.singh@nyu.edu

Abstract

We study the problem of approximately recovering a probability distribution given noisy measurements of its Chebyshev polynomial moments. We sharpen prior work, proving that accurate recovery in the Wasserstein distance is possible with more noise than previously known.

As a main application, our result yields a simple "linear query" algorithm for constructing a differentially private synthetic data distribution with Wasserstein-1 error $\tilde{O}(1/n)$ based on a dataset of n points in $[-1, 1]$. This bound is optimal up to log factors and matches a recent breakthrough of Boedihardjo, Strohmer, and Vershynin [Probab. Theory. Rel., 2024], which uses a more complex "superregular random walk" method to beat an $O(1/\sqrt{n})$ accuracy barrier inherent to earlier approaches.

We illustrate a second application of our new moment-based recovery bound in numerical linear algebra: by improving an approach of Braverman, Krishnan, and Musco [STOC 2022], our result yields a faster algorithm for estimating the spectral density of a symmetric matrix up to small error in the Wasserstein distance.

1 Introduction

The problem of recovering a probability distribution (or its parameters) by "matching" noisy estimates of the distribution's moments goes back over 100 years to the work of Chebyshev and Pearson [Pea94; Pea36; Fis11]. Moment matching continues to find a wide variety of applications, both in traditional statistical problems [KMV10; MV10; RSS14; WY19; WY20; FL23] and beyond. For example, moment matching is now widely used for solving eigenvalue estimation problems in numerical linear algebra and computational chemistry [WWAF06; CKSV18; CTU21; Che22].

One powerful and general result on moment matching for distributions with *bounded support* is that the method directly leads to approximations with small error in the Wasserstein-1 distance (a.k.a. earthmover's distance). Concretely, given a distribution p supported on $[-1, 1]$,[1] any distribution q for which $\mathbb{E}_{x\sim p}[x^i] = \mathbb{E}_{x\sim q}[x^i]$ for $i = 1, \ldots, k$ satisfies $W_1(p, q) = O(1/k)$, where W_1 denotes the Wasserstein-1 distance [KV17; CTU21]. I.e., to compute an ϵ-accurate approximation to p, it suffices to compute p's first $O(1/\epsilon)$ moments and to return any distribution q with the same moments.

Unfortunately, the above result is highly sensitive to noise, so is difficult to apply in the typical setting where, instead of p's exact moments, we only have access to *estimates* of the moments (e.g., computed from a sample). In particular, it can be shown that the accuracy of these estimates needs to be proportional to $1/2^k$ if we want to approximate p up to Wasserstein error $O(1/k)$ [JMSS23]. In other words, distribution approximation is *poorly conditioned* with respect to the standard moments.

1.1 Chebyshev moment matching

One way of avoiding the poor conditioning of moment matching is to move from the standard moments, $\mathbb{E}_{x\sim p}[x^i]$, to a better conditioned set of "generalized" moments. Specifically, significant prior work [WWAF06; WJF+16; BKM22] leverages *Chebyshev moments* of the form $\mathbb{E}_{x\sim p}[T_i(x)]$, where T_i is the i^{th} Cheybshev polynomial of the first kind, defined as:

$$T_0(x) = 1 \qquad T_1(x) = x \qquad T_i(x) = 2xT_{i-1}(x) - T_{i-2}(x), \text{ for } i \geq 2.$$

The Cheybshev moments are known to be less noise sensitive than the standard moments: instead of exponentially small error, $\tilde{O}(1/k)$ additive error[2] in computing p's first k Chebyshev moments suffices to find a distribution that is $O(1/k)$ close to p in Wasserstein distance (see, e.g., Lemma 3.1 in [BKM22]). This fact has been leveraged to obtain efficient algorithms for distribution estimation in a variety of settings. For example, Chebyshev moment matching leads to $O(n^2/\mathsf{poly}(\epsilon))$ time algorithms for estimating the eigenvalue distribution (i.e., the spectral density) of an $n \times n$ symmetric matrix A to error $\epsilon\|A\|_2$ in the Wasserstein distance [BKM22].

Cheybshev moment matching has also been leveraged for *differentially private synthetic data generation*. In this setting, p is the uniform distribution over a dataset $x_1, \ldots, x_n$. The goal is to find some q that approximates p, but in a differentially private way, which informally means that q cannot reveal too much information about any one data point, x_j [DNRRV09; RLP+20; MMSM22]. Such a q can be used to generate private synthetic data that is representative of the original data. One approach to solving this problem is to compute p's Chebyshev moments, and then add noise, which is known to ensure privacy [DR14]. Then, one can find a distribution q that matches the noised moments. It has been proven that, for a dataset of size n, this approach yields a differentially private distribution q that is $\tilde{O}(1/n^{1/3})$ close to p in Wasserstein distance [WJF+16].

[1] The result easily extends to p supported on any finite interval by shifting and scaling the distribution to $[-1, 1]$. For a general interval $[a, b]$, matching k moments yields error $O(|a - b|/k)$ in Wasserstein-1 distance.

[2] Throughout, we let $\tilde{O}(z)$ denote $O(z \log^c(z))$ for constant c.

1.2 Our contributions

Despite the success of Chebyshev moment matching, including for the applications discussed above, there is room for improvement. For example, for private distribution estimation, alternative methods can achieve nearly-optimal error $\tilde{O}(1/n)$ in Wasserstein distance for a dataset of size n [BSV24], improving on the $\tilde{O}(1/n^{1/3})$ bound known for moment matching. For eigenvalue estimation, existing moment matching methods obtain an optimal quadratic dependence on the matrix dimension n, but a suboptimal polynomial dependence on the accuracy parameter, ϵ [BKM22].

The main contribution of this work is to resolve these gaps by proving a sharper bound on the accuracy with which the Chebyshev moments need to be approximated to recover a distribution to high accuracy in the Wasserstein distance. Formally, we prove the following:

Theorem 1. *Let p, q be distributions supported on $[-1, 1]$. For any positive integer k, if the distributions' first k Chebyshev moments satisfy*

$$\sum_{j=1}^{k} \frac{1}{j^2} \left(\mathop{\mathbb{E}}_{x\sim p} T_j(x) - \mathop{\mathbb{E}}_{x\sim q} T_j(x) \right)^2 \leq \Gamma^2, \tag{1}$$

then, for an absolute constant c,[3]

$$W_1(p, q) \leq \frac{c}{k} + \Gamma. \tag{2}$$

As a special case, (1) holds if for all $j \in \{1, \ldots, k\}$,

$$\left| \mathop{\mathbb{E}}_{x\sim p} T_j(x) - \mathop{\mathbb{E}}_{x\sim q} T_j(x) \right| \leq \Gamma \cdot \sqrt{\frac{j}{1 + \log k}}.$$[4] \tag{3}

Theorem 1 characterizes the Chebyshev moment error required for a distribution q to approximate p in Wasserstein distance. The main requirement, (1), involves a weighted ℓ_2 norm with weights $1/j^2$, which reflects the diminishing importance of higher moments on the Wasserstein distance. Referring to (3), we obtain a bound of $W_1(p, q) \leq O(1/k)$ as long as q's j^{th} moment differs from p's by $\tilde{O}(\sqrt{j}/k)$. In contrast, prior work requires error $\tilde{O}(1/k)$ *for all of the first k moments* to ensure the same Wasserstein distance bound (Lemma 3.1, [BKM22]).

As a corollary of Theorem 1, we obtain the following algorithmic result:

Corollary 2. *Let p be a distribution supported on $[-1, 1]$. Given estimates $\hat{m}_1, \ldots, \hat{m}_k$ satisfying $\sum_{j=1}^{k} \frac{1}{j^2} \left(\mathbb{E}_{x\sim p} T_j(x) - \hat{m}_j\right)^2 \leq \Gamma^2$, Algorithm 1 returns a distribution q with $W_1(p, q) \leq c' \cdot \left(\frac{1}{k} + \Gamma\right)$ for a fixed constant c' in $\mathsf{poly}(k)$ time.*

Algorithm 1 simply solves a linearly-constrained least-squares regression problem to find a distribution q supported on a sufficiently fine grid whose moments are nearly as close to those of p as $\hat{m}_1, \ldots, \hat{m}_k$. We then obtain Corollary 2 by applying Theorem 1 to bound $W_1(p, q)$. The linear constraints ensure that q is positive and sums to one (i.e, that it is a valid distribution). This problem is easily solved using off-the-shelf software: in our experiments, we use a solver from MOSEK [DB16; MOS19].

Like prior work, our proof of Theorem 1 (given in Section 3) relies on tools from polynomial approximation theory. In particular, we leverage a constructive version of Jackson's theorem on

[3] Concretely, we prove a bound of $\frac{36}{k} + \Gamma$, although we believe the constants can be improved, at least to $\frac{2\pi}{k} + \Gamma$, and possibly further. See Section 3 for more discussion.

[4] Throughout, we let $\log k$ denote the natural logarithm of k, i.e., the logarithm with base e.

polynomial approximation of Lipschitz functions via "damped Chebyshev expansions" [Jac12]. Lipschitz functions are closely related to approximation in Wasserstein distance through the Kantorovich-Rubinstein duality: $W_1(p,q) = \max_{1-\text{Lip } f} \int_{-1}^{1} f(x)(p(x)-q(x))dx$. In contrast to prior work, we couple Jackson's theorem with a tight "global" characterization of the coefficient decay in the Chebyshev expansion of a Lipschitz function. In particular, we prove that any 1-Lipschitz function f with Chebyshev expansion $f = \sum_{j=0}^{\infty} c_j T_j$ has coefficients that satisfy $\sum_{j=1}^{\infty} j^2 c_j^2 = O(1)$. Prior work only leveraged the well-known "local" decay property, that the j^{th} coefficient has magnitude bounded by $O(1/j)$ [Tre19]. This property is implied by our bound, but much weaker.

1.3 Applications

We highlight two concrete applications of Theorem 1.

Differentially Private Synthetic Data. Privacy-enhancing technologies seek to protect individuals' data without preventing learning from the data. For theoretical guarantees of privacy, *differential privacy* [DR14] has become the industry standard, having been used in massive data products like the US Census, and included as a core tenet of the recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [Bid23; Abo18; AAS+19].

Concretely, we are interested in the ubiquitous notion of *approximate differential privacy*:

Definition 3 (Approximate Differential Privacy)**.** A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if, for all pairs of neighboring datasets X, X', and all subsets $\mathcal{B}$ of possible outputs:

$$\mathbb{P}[\mathcal{A}(X) \in \mathcal{B}] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{A}(X') \in \mathcal{B}] + \delta\,.$$

In our setting, a dataset X is a collection of n points in a bounded interval (without loss of generality, $[-1,1]$). Two datasets of size n are considered "neighboring" if all of their data points are equal except for one. Intuitively, Definition 3 ensures that the output of $\mathcal{A}$ is statistically indistinguishable from what the output would be if any one individual's data was replaced with something arbitrary.

There exist differentially private algorithms for a wide variety of statistical tasks [JL14; LLSY17; MTV+20]. One task of primary importance is *differentially private data synthesis*. Here, the goal is to generate *synthetic data* in a differentially private way that matches the original dataset along a set of relevant statistics or distributional properties. The appeal of private data synthesis is that, once generated, the synthetic data can be used for a wide variety of downstream tasks: a separate differentially private algorithm is not required for each potential use case.

Many methods for private data synthesis have been proposed [HLM12; ZCPSX17; LVW21; AAS+19; ABK+21; RHR+23; DSB21]. Such methods offer strong empirical performance and a variety of theoretical guarantees, e.g., that the generated synthetic data can effectively answer a fixed set of data analysis queries with high accuracy [HLM12; MMSM22]. Recently, there has been interest in algorithms with more general statistical guarantees – e.g., guarantees that the synthetic data comes from a distribution close in statistical distance to the original data [WJF+16; BSV24; HVZ23]. By leveraging Theorem 1, we contribute the following result to this line of work:

Theorem 4. *Let $X = \{x_1, \ldots, x_n\}$ be a dataset with each $x_j \in [-1,1]$. Let p be the uniform distribution on X. For any $\epsilon, \delta \in (0,1)$, there is an (ϵ, δ)-differentially private algorithm based on Chebyshev moment matching that, in $O(n) + \mathsf{poly}(\epsilon n)$ time, returns a distribution q satisfying for a*

fixed constant c_1,

$$\mathbb{E}[W_1(p,q)] \leq c_1 \frac{\log(\epsilon n)\sqrt{\log(1/\delta)}}{\epsilon n}.$$

Moreover, for any $\beta \in (0, 1/2)$, $W_1(p,q) \leq c_1 \frac{\sqrt{\log(1/\beta)+\log(\epsilon n)}\sqrt{\log(\epsilon n)\log(1/\delta)}}{\epsilon n}$ *with probability* $\geq 1-\beta$.

The distribution q returned by the algorithm behind Theorem 4 is represented as a discrete distribution on $O(\epsilon n)$ points in $[-1, 1]$, so can be sampled from efficiently to produce a synthetic dataset of arbitrary size. Typically, δ is chosen to be $1/\mathsf{poly}(n)$, in which case Theorem 4 essentially matches a recent break-through result of Boedihardjo, Strohmer, and Vershynin [BSV24], who give an $(\epsilon, 0)$-differentially private method with expected Wasserstein-1 error $O(\log^{3/2}(n)/(\epsilon n))$, which is optimal up to logarithmic factors.[5] Like that method, we improve on a natural barrier of $\tilde{O}(1/(\epsilon\sqrt{n}))$ error that is inherent to "private histogram" methods for approximation in the Wasserstein-1 distance [HRMS10; XWG10; QYL13; XZX+13; DR14; ZXX16; LLSY17].

The result of [BSV24] introduces a "superregular random walk" to directly add noise to $x_1, \ldots, x_n$ using a correlated distribution based on a Haar basis. Our method is simpler, more computationally efficient, and falls directly into the empirically popular *Select, Measure, Project* framework for differentially private synthetic data synthetis [VAA+22; LVW21]. In particular, as detailed in Algorithm 2, we compute the Chebyshev moments of p, add independent noise to each moment using the standard Gaussian mechanism [DKMMN06; MM09], and then recover q matching these noisy moments. We verify the strong empirical performance of the method in Section 6. A method similar to ours was analyzed in prior work [WJF+16], although that work obtains a Wasserstein error bound of $\tilde{O}(1/\epsilon n^{1/3})$. Our tighter connection between Chebyshev moment estimation and distribution approximation proven in Theorem 1 allows us to obtain a significantly better dependence on n.

We note that [HVZ23] also claims a faster and simpler alternative to [BSV24]. While the simplest method in that paper has error scaling with $\tilde{O}(1/\sqrt{n})$, they describe a more complex method that matches our $\tilde{O}(1/n)$ result up to a $\log(n)$ factor. While we are not aware of an implementation of that algorithm, empirically comparing alternative methods for generating synthetic data with Wasserstein distance guarantees would be a productive line of future work. Additionally, we note that, in concurrent work to ours, Feldman et al. study a stronger notion of *instance optimal* private distribution estimation in the Wasserstein distance [FMST24]. It would be interesting to explore if Cheybshev moment matching has any applications in this setting.

Matrix Spectral Density Estimation. Spectral density estimation (SDE) is a problem of central importance in numerical linear algebra. In the standard version of the problem, we are given a symmetric $n \times n$ matrix A, which has real-valued eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. Letting p denote the uniform distribution over these n eigenvalues, the goal is to output q which is close to p in the Wasserstein distance. An approximate spectral density can be useful in determining a variety of properties of A's eigenvalue spectrum – e.g., if its eigenvalues are decaying rapidly or if they follow a distribution characteristic of random matrices. Efficient SDE algorithms were originally studied in computational physics and chemistry, and are widely used to compute the "density of states" of quantum systems [Ski89; SR94; MAP20]. More recently, the problem has found applications in network science [DBB19; CKSV18; JKMSS24], deep learning [CKS91; PSG18; MM19; YGKM20], optimization [GKX19], and beyond [LXES19; CTU22].

[5] An $\Omega(1/(\epsilon n))$ lower bound on the expected Wasserstein error holds via standard 'packing lower bounds' which imply that even the easier problem of privately reporting the mean value of a dataset supported on $[-1, 1]$ requires error $\Omega(1/(\epsilon n))$. See e.g., [Kam20], Theorem 3.

Many popular SDE algorithms are based on Chebyshev moment matching [WWAF06]. The i^{th} Chebyshev moment of the spectral density is equal to $\mathbb{E}_{x\sim p} T_j(x) = \frac{1}{n}\sum_{j=1}^{n} T_i(\lambda_j) = \text{tr}(\frac{1}{n}T_i(A))$. This trace can be estimated using a small number of matrix-vector products with $T_i(A)$, using stochastic trace estimation techniques like Hutchinson's estimator [Hut90; MMMW21]. Since T_i is a degree-i polynomial, each matrix-vector product with $T_i(A)$ requires just i products with A. Thus, with a small number of products with A, we can obtain approximate moments for use in estimating p. Importantly, this approach can be applied even in the common *implicit* setting, where we do not have direct accress to the entries of A, but can efficiently multiply the matrix by vectors [AT11].

Recently, [BKM22] gave a theoretical analysis of Chebyshev moment-matching for SDE, along with the related Kernel Polynomial Method [WWAF06]. They show that when n is sufficiently large, specifically, $n = \tilde{\Omega}(1/\epsilon^2)$, then $\tilde{O}(1/\epsilon)$ matrix-vector products with A (and $\mathsf{poly}(1/\epsilon)$ additional runtime) suffice to output q with $W_1(p,q) \leq \epsilon \|A\|_2$, where $\|A\|_2 = \max_i |\lambda_i|$ is A's spectral norm.

While the result of [BKM22] also holds for smaller values of n, it suffers from a polynomially worse $1/\epsilon$ dependence in the number of matrix-vector products required. By leveraging Theorem 1, we resolve this issue, showing that $\tilde{O}(1/\epsilon)$ matrix-vector products suffice for *any* n. Roughly, by weakening the requirements on how well we approximate A's spectral moments, Theorem 1 allows us to decrease the accuracy with which moments are estimated, and thus the number of matrix-vector products used by Hutchinson's method. Formally, we prove:

Theorem 5. *There is an algorithm that, given $\epsilon \in (0,1)$, symmetric $A \in \mathbb{R}^{n\times n}$ with spectral density p, and upper bound[6] $S \geq \|A\|_2$, uses $\tilde{O}\left(\frac{1}{\epsilon}\right)$ matrix-vector products[7] with A and $\tilde{O}(n/\epsilon + 1/\epsilon^3)$ additional time to output a distribution q such that, with high probability, $W_1(p,q) \leq \epsilon S$.*

In the case when A is dense, Theorem 5 yields an algorithm that runs in $\tilde{O}(n^2/\epsilon + 1/\epsilon^3)$ time, which can be much faster than the $O(n^\omega)$ time required to compute p directly via a full eigendecomposition. In terms of matrix-vector products, the result cannot be improved by more than logarithmic factors. In particular, a recent lower bound on estimating the trace of a positive definite matrix [WZZ22] implies that $\Omega(1/\epsilon)$ matrix-vector products with A are necessary to approximate the spectral density p up to error $\epsilon\|A\|_2$ (see Appendix C for details). Thus, Theorem 5 resolves, up to logarithmic factors, the complexity of the SDE problem in the "matrix-vector query model" of computation, where cost is measured via matrix-vector products with A. Understanding this model has become a core topic in theoretical work on numerical linear algebraic, as it generalizes other important models like the matrix sketching and Krylov subspace models [SWYZ21]. Our work contributes to recent progress on establishing tight upper and lower bounds for central problems like linear system solving [BHSW20], eigenvector approximation [MM15; SER18], trace estimation [JPWZ24], and more [CDLLN23; BN23; ACK+24; CKHMM24].

2 Preliminaries

Before our main analysis, we introduce notation and technical preliminaries.

[6] The power method can compute S satisfying $\|A\|_2 \leq S \leq 2\|A\|_2$ using $O(\log n)$ matrix-vector products with A and $O(n)$ additional runtime [KW92]. In some settings, an upper bound on $\|A\|_2$ may be known a priori [JKMSS24].

[7] Formally, we prove a bound of $\min\left\{n, O\left(\frac{1}{\epsilon}\right)\cdot\left(1+\frac{\log^2(1/\epsilon)\log^2(1/(\epsilon\delta))}{n\epsilon}\right)\right\}$ matrix-vector products to succeed with probability $1-\delta$. For constant δ, this is at worst $O\left(\log^4(1/\epsilon)/\epsilon\right)$, but actually $O(1/\epsilon)$ for all $\epsilon = \Omega(n/\log^4 n)$.

Notation. We let $\mathbb{Z}_{\geq 0}$ denote the natural numbers and $\mathbb{Z}_{>0}$ denote the positive integers. For a vector $x \in \mathbb{R}^k$, we let $\|x\|_2 = \sqrt{\sum_{i=1}^k x_i^2}$ denote the Euclidean norm. We often work with functions from $[-1,1] \to \mathbb{R}$. For two such functions, f, g, we use the convenient inner product notation:

$$\langle f, g \rangle \stackrel{\text{def}}{=} \int_{-1}^{1} f(x)g(x)\,dx.$$

We will often work with products, quotients, sums, and differences of two functions f, g, which are denoted by $f \cdot g$, f/g, $f+g$, and $f-g$, respectively. E.g., $[f \cdot g](x) = f(x)g(x)$. For a function $f : [-1,1] \to \mathbb{R}$, we let $\|f\|_\infty$ denote $\|f\|_\infty = \max_{x\in[-1,1]} |f(x)|$ and $\|f\|_1 = \int_{-1}^{1} |f(x)|\,dx$.

Wasserstein Distance. This paper concerns the approximation of probability distributions in the Wasserstein-1 distance, defined below.

Definition 6 (Wasserstein-1 Distance)**.** Let p and q be two distributions on $\mathbb{R}$. Let $Z(p,q)$ be the set of all couplings between p and q, i.e., the set of distributions on $\mathbb{R} \times \mathbb{R}$ whose marginals equal p and q. Then the Wasserstein-1 distance between p and q is:

$$W_1(p,q) = \inf_{z \in Z(p,q)} \left[\mathop{\mathbb{E}}_{(x,y)\sim z} |x-y| \right].$$

The Wasserstein-1 distance measures the total cost (in terms of distance per unit mass) required to "transport" the distribution p to q. Alternatively, it has a well-known dual formulation:

Fact 7 (Kantorovich-Rubinstein Duality)**.** *Let p, q be as in Definition 6. Then $W_1(p,q) = \sup_{1\text{-Lipschitz } f} \langle f, p-q \rangle$, where $f : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz if $|f(x) - f(y)| \leq |x-y|$ for all $x, y \in \mathbb{R}$.*

Above we slightly abuse notation and use p and q to denote (generalized) probability density functions[8] instead of the distributions themselves. We will do so throughout the paper.

In our analysis, it will be convenient to assume when applying Fact 7 that, in addition to being 1-Lipschitz, f is smooth, i.e. that it is infinitely differentiable. Since any Lipschitz function can be arbitrarily well approximated by a smooth function, we can do so without changing the distance. In particular, for distributions on $[-1,1]$ we have:

$$W_1(p,q) = \sup_{\text{1-Lipschitz, smooth } f} \langle f, p-q \rangle. \tag{4}$$

Chebyshev Polynomials and Chebyshev Series. Our main result analyzes the accuracy of (noisy) Chebyshev polynomial moment matching for distribution approximation. The Chebyshev polynomials are defined in Section 1.1, and can alternatively be defined on $[-1,1]$ via the trigonometric definition, $T_j(\cos\theta) = \cos(j\theta)$. We use a few basic properties about these polynomials.

Fact 8 (Boundedness and Orthogonality, see e.g. [Hal15])**.** *The Chebyshev polynomials satisfy:*

1. **Boundedness:** *$\forall x \in [-1,1]$ and $j \in \mathbb{Z}_{\geq 0}$, $|T_j(x)| \leq 1$.*
2. **Orthogonality:** *The Chebyshev polynomials are orthogonal with respect to the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$. In particular, for $i, j \in \mathbb{Z}_{\geq 0}$, $i \neq j$, $\langle T_i \cdot w, T_j \rangle = 0$.*

To obtain an orthonormal basis we also define the *normalized* Chebyshev polynomials as follows:

[8] p and q might correspond to discrete distributions, in which case they will be sums of Dirac delta functions.

Definition 9 (Normalized Chebyshev Polynomials)**.** The j^{th} *normalized* Chebyshev polynomial, $\bar{T}_j$, is defined as $\bar{T}_j \stackrel{\text{def}}{=} T_j / \sqrt{\langle T_j \cdot w, T_j \rangle}$. Note that $\langle T_j \cdot w, T_j \rangle$ equals π for $j = 0$ and $\pi/2$ for $j \geq 1$.

We define the *Chebyshev series* of a function $f : [-1, 1] \to \mathbb{R}$ as $\sum_{j=0}^{\infty} \left\langle f \cdot w, \bar{T}_j \right\rangle \bar{T}_j$. If f is Lipschitz continuous then the Chebyshev series of f converges absolutely and uniformly to f [Tre19, Theorem 3.1]. Throughout this paper, we will also write the Chebyshev series of generalized probability density functions, which could involve Dirac delta functions. This is standard in Fourier analysis, even though the Chebyshev series does not converge pointwise [Lig58]. Formally, any density p can be replaced with a Lipschitz continuous density (which has a convergent Cheybshev series) that is arbitrarily close in Wasserstein distance and the same analysis goes through.

3 Main Analysis

In this section, we prove our main result, Theorem 1, as well as Corollary 2. To do so, we require two main ingredients. The first is a constructive version of Jackson's theorem on polynomial approximation of Lipschitz functions [Jac30]. A modern proof can be found in [BKM22, Fact 3.2].

Fact 10 (Jackson's Theorem [Jac30])**.** *Let $f : [-1, 1] \to \mathbb{R}$ be an ℓ-Lipschitz function. Then, for any $k \in \mathbb{Z}_{>0}$, there are $k + 1$ constants $1 = b_k^0 > \ldots > b_k^k \geq 0$ such that the polynomial $f_k = \sum_{j=0}^{k} b_k^0 \cdot \left\langle f \cdot w, \bar{T}_j \right\rangle \cdot \bar{T}_j$ satisfies $\|f - f_k\|_\infty \leq 18\ell/k$.*

It is well-known that truncating the Chebyshev series of an ℓ-Lipschitz function f to k terms leads to error $O(\log k \cdot \frac{\ell}{k})$ in the ℓ_∞ distance [Tre19]. The above version of Jackson's theorem improves this bound by a $\log k$ factor by instead using a *damped* truncated Chebyshev series: each term in the series is multiplied by a positive scaling factor between 0 and 1. We will not need to compute these factors explicitly, but b_k^i has a simple closed form (see [BKM22, Equation 12]).

To bound the Wasserstein distance between distributions p, q, we need to upper bound $\langle f, p - q \rangle$ for every 1-Lipschitz f. The value of Fact 10 is that this inner product is closely approximated by $\langle f_k, p - q \rangle$. Since f_k is a damped Chebyshev series, this inner product can be decomposed as a difference between p and q's Chebyshev moments. Details will be shown in the proof of Theorem 1.

The second ingredient we require is a stronger bound on the decay of the Chebyshev coefficients, $\left\langle f \cdot w, \bar{T}_j \right\rangle$, which appear in Fact 10. In particular, we prove the following result:

Lemma 11 (Global Chebyshev Coefficient Decay)**.** *Let $f : [-1, 1] \to \mathbb{R}$ be an ℓ-Lipschitz, smooth function, and let $c_j \stackrel{\text{def}}{=} \left\langle f \cdot w, \bar{T}_j \right\rangle$ for $j \in \mathbb{Z}_{\geq 0}$. Then, $\sum_{j=1}^{\infty} (j c_j)^2 \leq \frac{\pi}{2} \ell^2$.*

Lemma 11 implies the well known fact that $c_j = O(\ell/j)$ for $j \geq 1$ [Tre08]. However, it is a much stronger bound: if all we knew was that the Chebyshev coefficients are bounded by $O(\ell/j)$, then $\sum_{j=1}^{\infty} (j c_j)^2$ could be unbounded. We show that it can in fact be bounded by $O(\ell^2)$. Informally, the implication is that not all coefficients can saturate the "local" $O(\ell/j)$ constraint at the same time, but rather obey a stronger global constraint, captured by a weighted ℓ_2 norm of the coefficients.

3.1 Proof of Theorem 1

We prove Lemma 11 in Section 3.3. Before doing so, we show how it implies Theorem 1.

Proof of Theorem 1. By (4), to bound $W_1(p, q)$, it suffices to bound $\langle f, p - q \rangle$ for any 1-Lipschitz,

smooth f. Let f_k be the approximation to any such f guaranteed by Fact 10. We have:

$$\begin{aligned}\langle f, p-q\rangle = \langle f_k, p-q\rangle + \langle f-f_k, p-q\rangle &\le \langle f_k, p-q\rangle + \|f-f_k\|_\infty \|p-q\|_1 \\ &\le \langle f_k, p-q\rangle + \frac{36}{k}. \end{aligned} \tag{5}$$

In the last step, we use that $\|f-f_k\|_\infty \le 18/k$ by Fact 10, and that $\|p-q\|_1 \le \|p\|_1 + \|q\|_1 = 2$. So, to bound $\langle f, p-q\rangle$ we turn our attention to bounding $\langle f_k, p-q\rangle$.

For technical reasons, we will assume from here on that p and q are supported on the interval $[-1+\delta, 1-\delta]$ for arbitrarily small $\delta \to 0$. This is to avoid an issue with the Chebyshev weight function $w(x) = 1/\sqrt{1-x^2}$ going to infinity at $x = -1, 1$. The assumption is without loss of generality, since we can rescale the support of p and q by a $(1-\delta)$ factor, and the distributions' moments and Wasserstein distance change by an arbitrarily small factor as $\delta \to 0$.

We proceed by writing the Chebyshev series of the function $(p-q)/w$:

$$\frac{p-q}{w} = \sum_{j=0}^{\infty} \left\langle \frac{p-q}{w}\cdot w, \bar{T}_j \right\rangle \bar{T}_j = \sum_{j=0}^{\infty} \langle p-q, \bar{T}_j\rangle \cdot \bar{T}_j = \sum_{j=1}^{\infty} \langle p-q, \bar{T}_j\rangle \cdot \bar{T}_j. \tag{6}$$

In the last step we use that both p and q are distributions so $\left\langle p-q, \bar{T}_0 \right\rangle = 1/\pi - 1/\pi = 0$.

Next, recall from Fact 10 that $f_k = \sum_{j=0}^{k} c'_j \bar{T}_j$, where each c'_j satisfies $|c'_j| \le |c_j|$ for $c_j \stackrel{\text{def}}{=} \langle f\cdot w, \bar{T}_j\rangle$. Using (6), the fact that $\langle \bar{T}_i \cdot w, \bar{T}_j\rangle = 0$ whenever $i \ne j$, and that $\langle \bar{T}_j \cdot w, \bar{T}_j\rangle = 1$ for all j, we have:

$$\langle f_k, p-q\rangle = \left\langle f_k\cdot w, \frac{p-q}{w}\right\rangle = \left\langle \sum_{j=0}^{k} c'_j \bar{T}_j \cdot w, \sum_{j=1}^{\infty} \langle p-q, \bar{T}_j\rangle \bar{T}_j \right\rangle = \sum_{j=1}^{k} c'_j \cdot \langle p-q, \bar{T}_j\rangle.$$

Via Cauchy-Schwarz inequality and our global decay bound from Lemma 11, we then have:

$$\begin{aligned}\langle f_k, p-q\rangle = \sum_{j=1}^{k} j c'_j \cdot \frac{\langle p-q, \bar{T}_j\rangle}{j} &\le \left(\sum_{j=1}^{k} (jc'_j)^2\right)^{1/2} \cdot \left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2} \\ &\le \left(\sum_{j=1}^{k} (jc_j)^2\right)^{1/2} \cdot \left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2} \\ &\le \sqrt{\pi/2}\left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2}. \end{aligned} \tag{7}$$

Observing from Definition 9 that $\langle p-q, \bar{T}_j\rangle/\sqrt{\pi/2}$ is exactly the difference between the j^{th} Chebyshev moments of p and q, we can apply the assumption of the theorem, (1), to upper bound (7) by Γ.

Plugging this bound into Equation (5), we conclude the main bound of Theorem 1:

$$W_1(p,q) = \sup_{\text{1-Lipschitz, smooth } f} \langle f, p-q\rangle \le \Gamma + \frac{36}{k}.$$

We note that the constants in the above bound can likely be improved. Notably, the 36 comes from multiplying the factor of 18 in Fact 10 by 2. As discussed in [BKM22, Appendix C.2], strong numerical evidence suggests that this 18 can be improved to π, leading to a bound of $\Gamma + \frac{2\pi}{k}$.

Finally, we comment on the special case in (3). If $|\mathbb{E}_{x\sim p} T_j(x) - \mathbb{E}_{x\sim q} T_j(x)| = |\langle p-q, \bar{T}_j\rangle|/\sqrt{\pi/2} \le \Gamma \cdot \sqrt{\frac{j}{1+\log k}}$ for all j then we have that $\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, T_j\rangle^2 \le \frac{\Gamma^2}{1+\log k}\sum_{j=1}^{k}\frac{1}{j} \le \Gamma^2$. □

Algorithm 1 Chebyshev Moment Regression

Input: Estimates $\hat{m}_1, \ldots, \hat{m}_k$ for the first k Chebyshev polynomial moments of a distribution p.
Output: A probability distribution q approximating p.

1: For $g = \lceil k^{1.5} \rceil$, let $\mathcal{C} = \{x_1, \ldots, x_g\}$ be the degree g Cheybshev nodes. I.e., $x_i = \cos\left(\frac{2i-1}{2g}\pi\right)$.
2: Let $q_1, \ldots, q_g$ solve the following optimization problem:

$$\begin{aligned} \min_{z_1,\ldots,z_g} \quad & \sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \sum_{i=1}^{g} z_i T_j(x_i)\right)^2 \\ \text{subject to} \quad & \sum_{i=1}^{g} z_i = 1 \text{ and } z_i \geq 0, \ \forall i \in \{1, \ldots, g\}. \end{aligned}$$

3: Return $q = \sum_{i=1}^{m} q_i \delta(x - x_i)$, where δ is the Dirac delta function.

3.2 Efficient recovery

The primary value of Theorem 1 for our applications is that, given sufficiently accurate estimates, $\hat{m}_1, \ldots, \hat{m}_k$, of p's Chebyshev moments, we can recover a distribution q that is close in Wasserstein-1 distance to p, even if there is no distribution whose moments exactly equal $\hat{m}_1, \ldots, \hat{m}_k$.

This claim is formalized in Corollary 2, whose proof is straightforward. We outline the main idea here. Recall the condition of the corollary, that $\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle p, \bar{T}_j\rangle\right)^2 \leq \Gamma^2$. Now, suppose we could solve the optimization problem:

$$q^* = \underset{\text{distributions } q \text{ on } [-1,1]}{\operatorname{argmin}} \sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle q, \bar{T}_j\rangle\right)^2.$$

Then by triangle inequality we would have:

$$\begin{aligned} \left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \langle q^*, \bar{T}_j\rangle\right)^2\right)^{1/2} &\leq \left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle q^*, \bar{T}_j\rangle\right)^2\right)^{1/2} + \left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle p, \bar{T}_j\rangle\right)^2\right)^{1/2} \\ &\leq 2\left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle p, \bar{T}_j\rangle\right)^2\right)^{1/2} \leq 2\Gamma. \end{aligned} \tag{8}$$

It then follows immediately from Theorem 1 that $W_1(p, q^*) \leq O\left(\frac{1}{k} + \Gamma\right)$, as desired.

The only catch with the argument above is that we cannot efficiently optimize over the entire set of distributions on $[-1, 1]$. Instead, we have to optimize over a sufficiently fine discretization. Specifically, we consider discrete distributions on a finite grid, choosing the Chebyshev nodes (of the first kind) instead of a uniform grid because doing so yields a better approximation, and thus allows for a coarser grid. Concretely, Corollary 2 is proven by analyzing Algorithm 1. The full analysis is given in Appendix A.

We note that the optimization problem solved by Algorithm 1 is a simple linearly constrained quadratic program with $g = O(k^{1.5})$ variables and $O(k^{1.5})$ constraints, so can be solved to high accuracy in $\mathsf{poly}(k)$ time using a variety of methods [YT89; KV86; ART03]. In practice, the problem can also be solved efficiently using first-order methods like projected gradient descent [WR22].

3.3 Proof of Lemma 11

We conclude this section by proving Lemma 11, our global decay bound on the Chebyshev coefficients of a smooth, Lipschitz function, which was key in the proof of Theorem 1. To do so we will leverage an expression for the derivatives of the Chebyshev polynomials of the first kind in terms of the Chebyshev polynomials *of the second kind*, which can be defined by the recurrence

$$U_0(x) = 1 \qquad U_1(x) = 2x \qquad U_i(x) = 2xU_{i-1}(x) - U_{i-2}(x), \text{ for } i \geq 2.$$

We have the following standard facts (see e.g., [Riv69]).

Fact 12 (Chebyshev Polynomial Derivatives)**.** *Let T_j be the j^{th} Chebyshev polynomial of the first kind, and U_j be the j^{th} Chebyshev polynomial of the second kind. Then, for $j \geq 1$, $T_j'(x) = jU_{j-1}(x)$.*

Fact 13 (Orthogonality of Chebyshev polynomials of the second kind)**.** *The Chebyshev polynomials of the second kind are orthogonal with respect to the weight function $u(x) = \sqrt{1-x^2}$. In particular,*

$$\int_{-1}^{1} U_i(x)U_j(x)u(x)\,dx = \begin{cases} 0, & \text{for } i \neq j \\ \frac{\pi}{2}, & \text{for } i = j\,. \end{cases}$$

With the above facts we can now prove Lemma 11.

Proof of Lemma 11. Let f be a smooth, ℓ-Lipschitz function, with Chebyshev expansion $f(x) = \sum_{j=0}^{\infty} c_j\bar{T}_j = \frac{1}{\sqrt{\pi}}c_0T_0 + \sum_{j=1}^{\infty}\sqrt{\frac{2}{\pi}}c_jT_j$. Using Fact 12, we can write f's derivative as:

$$f'(x) = \sum_{j=1}^{\infty}\sqrt{\frac{2}{\pi}}c_jT_j'(x) = \sqrt{\frac{2}{\pi}}\sum_{j=1}^{\infty} jc_jU_{j-1}(x)\,.$$

By the orthogonality property of Fact 13, we then have that

$$\int_{-1}^{1} f'(x)f'(x)u(x)\,dx = \frac{2}{\pi}\sum_{j=1}^{\infty} j^2c_j^2\frac{\pi}{2} = \sum_{j=1}^{\infty} j^2c_j^2.$$

Further, using that f is ℓ-Lipschitz and so $|f'(x)| \leq \ell$, and that the weight function $u(x) = \sqrt{1-x^2}$ is non-negative, we can upper bound this sum by

$$\sum_{j=1}^{\infty} j^2c_j^2 = \int_{-1}^{1} f'(x)f'(x)u(x)\,dx \leq \ell^2\int_{-1}^{1} u(x)\,dx = \frac{\pi\ell^2}{2}.$$

This completes the proof of the lemma. □

4 Private Synthetic Data

In this section, we present an application of our main result to differentially private synthetic data generation. We recall the setting from Section 1.3: we are given a dataset $X = \{x_1, \ldots, x_n\}$, where each $x_i \in [-1, 1]$, and consider the distribution p that is uniform on X. The goal is to design an (ϵ, δ)-differentially private algorithm that returns a distribution q that is close to p in Wasserstein distance. For the purpose of defining differential privacy (see Def. 3), we consider the "bounded"

notation of neighboring datasets, which applies to datasets of the same size [KM11]. Concretely, $X = \{x_1, \ldots, x_n\}$ and $X' = \{x'_1, \ldots, x'_n\}$ are *neighboring* if $x_i \neq x'_i$ for *exactly one* value of i.[9]

To solve this problem, we will compute the first n Chebyshev moments of p, then add noise to those moments using the standard *Gaussian mechanism*. Doing so ensures that the noised moments are (ϵ, δ)-differentially private. We then post-process the noised moments (which does not impact privacy) by finding a distribution q that matches the moments. The analysis of our approach follows directly from Theorem 1, although we use a slightly different method for recovering q than suggested in our general Algorithm 1: in the differential privacy setting, we are able to obtain a moderately faster algorithm that solves a regression problem involving $O(n)$ variables instead of $O(n^{1.5})$.

Before analyzing this approach, we introduce preliminaries necessary to apply the Gaussian mechanism. In particular, applying the mechanism requires bounding the ℓ_2 *sensitivity* of the function mapping a distribution p to its Chebyshev moments. This sensitivity is defined as follows:

Definition 14 (ℓ_2 Sensitivity)**.** Let $\mathcal{X}$ be some data domain (in our setting, $\mathcal{X} = [-1, 1]^n$) and let $f : \mathcal{X} \to \mathbb{R}^k$ be a vector valued function. The ℓ_2-sensitivity of f, $\Delta_{2,f}$, is defined as:

$$\Delta_{2,f} \stackrel{\text{def}}{=} \max_{\substack{\text{neighboring datasets} \\ X, X' \in \mathcal{X}}} \|f(X) - f(X')\|_2.$$

The Gaussian mechanism provides a way of privately evaluating any function f with bounded ℓ_2 sensitivity by adding a random Gaussian vector with appropriate variance. Let $\mathcal{N}(0, \sigma^2 I_k)$ denote a vector of k i.i.d. mean zero Gaussians with variance σ^2. We have the following well-known result:

Fact 15 (Gaussian Mechanism [DKMMN06; DR14])**.** *Let $f : \mathcal{X} \to \mathbb{R}^k$ be a function with ℓ_2-sensitivity $\Delta_{2,f}$ and let $\sigma^2 = \Delta_{2,f}^2 \cdot 2\ln(1.25/\delta)/\epsilon^2$, where $\epsilon, \delta \in (0, 1)$ are privacy parameters. Then the mechanism $\mathcal{M} = f(X) + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2 I_k)$ is (ϵ, δ)-differentially private.*

We are now ready to prove the main result of this section, Theorem 4, which follows by analyzing Algorithm 2. Note that Algorithm 2 is very similar to Algorithm 1, but we first round our distribution to be supported on a uniform grid, $\mathcal{G}$. Doing so will allow us to solve our moment regression problem over the same grid, which is smaller than the set of Chebyshev nodes used in Algorithm 1.

Proof of Theorem 4. We analyze both the privacy and accuracy of Algorithm 2.

Privacy. For a dataset $X = \{x_1, \ldots, x_n\} \in [-1, 1]^n$, let $f(X)$ be a vector-valued function mapping to the first $k = \lceil 2\epsilon n \rceil$ (as set in Algorithm 2) *scaled* Chebyshev moments of the uniform distribution over X. I.e.,

$$f(X) = \begin{bmatrix} 1 \cdot \frac{1}{n}\sum_{i=1}^n \bar{T}_1(x_i) \\ \frac{1}{\sqrt{2}} \cdot \frac{1}{n}\sum_{i=1}^n \bar{T}_2(x_i) \\ \vdots \\ \frac{1}{\sqrt{k}} \cdot \frac{1}{n}\sum_{i=1}^n \bar{T}_k(x_i) \end{bmatrix}$$

[9] Although a bit tedious, our results can be extended to the "unbounded" notation of neighboring datasets, where X and X' might differ in size by one, i.e., because X' is created by adding or removing a single data point from X.

Algorithm 2 Private Chebyshev Moment Matching

Input: Dataset $x_1, \dots, x_n \in [-1, 1]$, privacy parameters $\epsilon, \delta > 0$.

Output: A probability distribution q approximating the uniform distribution, p, on $x_1, \dots, x_n$.

1: Let $\mathcal{G} = \{-1, -1 + \frac{1}{\lceil \epsilon n \rceil}, -1 + \frac{2}{\lceil \epsilon n \rceil}, \dots, 1\}$. Let $r \stackrel{\text{def}}{=} |\mathcal{G}| = 2\lceil \epsilon n \rceil + 1$ and let $g_i = -1 + \frac{i-1}{\lceil \epsilon n \rceil}$ denote the i^{th} element of $\mathcal{G}$.

2: For $i = 1, \dots, n$, let $\tilde{x}_i = \text{argmin}_{y \in \mathcal{G}} |x_i - y|$. I.e., round x_i to the nearest multiple of $1/\lceil \epsilon n \rceil$.

3: Set $\sigma^2 = \frac{\frac{16}{\pi}(1 + \log k)\ln(1.25/\delta)}{\epsilon^2 n^2}$.

4: Set $k = \lceil 2\epsilon n \rceil$.[10] For $j = 1, \dots, k$, let $\hat{m}_j = \eta_j + \frac{1}{n}\sum_{i=1}^n \bar{T}_j(\tilde{x}_i)$, where $\eta_j \sim \mathcal{N}(0, j\sigma^2)$.

5: Let $q_0, \dots, q_r$ be the solution to the following optimization problem:

$$\begin{aligned} \min_{z_1, \dots, z_r} \quad & \sum_{j=1}^{k} \frac{1}{j^2} \left(\hat{m}_j - \sum_{i=1}^{r} z_i T_j(g_i) \right)^2 \\ \text{subject to} \quad & \sum_{i=1}^{r} z_i = 1 \text{ and } z_i \geq 0, \ \forall i \in \{1, \dots, r\}. \end{aligned}$$

6: Return $q = \sum_{i=1}^r q_i \delta(x - g_i)$, where δ is the Dirac delta function.

By Fact 8, $\max_{x_i \in [-1,1]} |\bar{T}_j(x_i)| \leq \sqrt{2/\pi}$ for $j \in \mathbb{Z}_{>0}$, so we have:

$$\Delta_{2,f}^2 = \max_{\substack{\text{neighboring datasets} \\ X, X' \in \mathcal{X}}} \|f(X) - f(X')\|_2^2 \leq \sum_{j=1}^{k} \frac{1}{jn^2} \cdot \frac{8}{\pi} \leq \frac{8}{\pi n^2}(1 + \log k). \tag{9}$$

For two neighboring datasets X, X', let $\tilde{X}$ and $\tilde{X}'$ be the rounded datasets computed in line 2 of Algorithm 2 – i.e., $\tilde{X} = \{\tilde{x}_1, \dots, \tilde{x}_n\}$. Observe that $\tilde{X}$ and $\tilde{X}'$ are also neighboring. Thus, it follows from Fact 15 and the sensitivity bound of eq. (9) that $\tilde{m} = f(\tilde{X}) + \eta$ is (ϵ, δ)-differentially private for $\eta \sim \mathcal{N}(0, \sigma^2 I_k)$ as long as $\sigma^2 = \frac{16}{\pi}(1 + \log k)\ln(1.25/\delta)/(n^2\epsilon^2)$. Finally, observe that $\hat{m}_j$ computed by Algorithm 2 is exactly equal to $\sqrt{j}$ times the j^{th} entry of such an $\tilde{m}$. So $\hat{m}_1, \dots, \hat{m}_k$ are (ϵ, δ)-differentially private. Since the remainder of Algorithm 2 simply post-processes $\hat{m}_1, \dots, \hat{m}_k$ without returning to the original data X, the output of the algorithm is also (ϵ, δ)-differentially private, as desired.

Accuracy. Algorithm 2 begins by rounding the dataset X so that every data point is a multiple of $1/\lceil \epsilon n \rceil$. Let $\tilde{p}$ be the uniform distribution over the rounded dataset $\tilde{X}$. Then it is not hard to see from the transportation definition of the Wasserstein-1 distance that:

$$W_1(p, \tilde{p}) \leq \frac{1}{2\lceil \epsilon n \rceil}. \tag{10}$$

In particular, we can transport p to $\tilde{p}$ by moving every unit of $1/n$ probability mass a distance of at most $1/2\lceil \epsilon n \rceil$. Given (10), it will suffice to show that Algorithm 2 returns a distribution q that is close in Wasserstein distance to $\tilde{p}$. We will then apply triangle inequality to bound $W_1(p, q)$.

[10] While we choose $k = \lceil 2\epsilon n \rceil$ by default, any choice of $k = \lceil c\epsilon n \rceil$ for constant c suffices to obtain the bound of Theorem 4. Similarly, the grid spacing in $\mathcal{G}$ can made finer or coarse by a multiplicative constant. A larger k or a finer grid will lead to a slightly more accurate result at the cost of a slower algorithm. We chose defauts so that any error introduced from the grid and choice of k is swamped by error incurred from the noise added in Line 4. I.e., the error cannot be improved by more than a factor of two with difference choices. See the proof of Theorem 4 for more details.

To show that Algorithm 2 returns a distribution q that is close to $\tilde{p}$ in Wasserstein distance, we begin by bounding the moment estimation error:

$$E \stackrel{\text{def}}{=} \sum_{j=1}^{k} \frac{1}{j^2} \left(\hat{m}_j(p) - \langle \tilde{p}, T_j \rangle\right)^2,$$

where k is as chosen in Algorithm 2 and $\langle \tilde{p}, T_j \rangle = \frac{1}{n}\sum_{i=1}^{n} T_j(\tilde{x}_i)$. Let σ^2 and $\eta_1, \ldots, \eta_k$ be as in Algorithm 2. Applying linearity of expectation, we have that:

$$\mathbb{E}[E] = \mathbb{E}\left[\sum_{j=1}^{k} \frac{1}{j^2}\eta_j^2\right] = \sum_{j=1}^{k} \frac{1}{j^2}\mathbb{E}\left[\eta_j^2\right] = \sum_{j=1}^{k} \frac{1}{j^2} \cdot j\sigma^2 \leq (1+\log k)\sigma^2. \tag{11}$$

Now, let q be as in Algorithm 2. Using a triangle inequality argument as in Section 3.2, we have:

$$\Gamma^2 = \sum_{j=1}^{k} \frac{1}{j^2} \left(\langle q, T_j \rangle - \langle \tilde{p}, T_j \rangle\right)^2 \leq \sum_{j=1}^{k} \frac{1}{j^2} \left(\langle q, T_j \rangle - \hat{m}_j\right)^2 + \sum_{j=1}^{k} \frac{1}{j^2} \left(\langle \tilde{p}, T_j \rangle - \hat{m}_j\right)^2 \leq 2E.$$

Above we use that $\tilde{p}$ is a feasible solution to the optimization problem solved in Algorithm 2 and, since q is the optimum, $\sum_{j=1}^{k} \frac{1}{j^2} \left(\langle q, T_j \rangle - \hat{m}_j\right)^2 \leq \sum_{j=1}^{k} \frac{1}{j^2} \left(\langle \tilde{p}, T_j \rangle - \hat{m}_j\right)^2$. It follows that $\mathbb{E}[\Gamma^2] \leq 2\,\mathbb{E}[E]$, and, via Jensen's inequality, that $\mathbb{E}[\Gamma] \leq \sqrt{2\,\mathbb{E}[E]}$. Plugging into Theorem 1, we have for constant c:

$$\mathbb{E}[W_1(\tilde{p}, q)] \leq \mathbb{E}[\Gamma] + \frac{c}{k} \leq \sqrt{2(1+\log k)\sigma^2} + \frac{c}{k} = O\left(\frac{\log(\epsilon n)\sqrt{\log(1/\delta)}}{\epsilon n}\right). \tag{12}$$

By triangle inequality and (10), $W_1(p, q) \leq W_1(\tilde{p}, q) + W_1(\tilde{p}, p) \leq W_1(\tilde{p}, q) + \frac{1}{2\lceil \epsilon n \rceil}$. Combined with the bound above, this proves the accuracy claim of the theorem.

Recall from Section 3 that the constant c in Theorem 1 is bounded by 36, but can likely be replaced by 2π, in which case it can be checked that the $\frac{c}{k}$ term in (12) will be dominated by the $\sqrt{2(1+\log k)\sigma^2}$ term for our default of $k = \lceil 2\epsilon n \rceil$ in Algorithm 2. However, any choice $k = \Theta(\epsilon n)$ suffices to prove the theorem. We also remark that our bound on the expected value of $W_1(\tilde{p}, q)$ can also be shown to hold with high probability. See Appendix B for details.

We conclude by noting that, as in our analysis of Algorithm 1 (see Section 3.2), Algorithm 2 requires solving a linearly constrained quadratic program with $r = 2\lceil \epsilon n \rceil + 1$ variables and $r + 1$ constraints, which can be done to high accuracy in $\mathsf{poly}(\epsilon n)$ time. □

5 Spectral Density Estimation

In this section, we present a second application of our main result to the linear algebraic problem of Spectral Density Estimation (SDE). We recall the setting from Section 1.3: letting p be the uniform distribution over the eigenvalues given $\lambda_1 \geq \cdots \geq \lambda_n$ of a symmetric matrix $A \in \mathbb{R}^{n \times n}$, the goal is to find some distribution q that satisfies

$$W_1(p, q) \leq \epsilon \|A\|_2. \tag{13}$$

In many settings of interest, A is implicit and can only be accessed via matrix-vector multiplications. So, we want to understand 1) how many matrix-vector multiplications with A are required to achieve (13), and 2) how efficiently can we achieve (13) in terms of standard computational complexity.

We show how to obtain improved answers to these questions by using our main result, Theorem 1, to give a tighter analysis of an approach from [BKM22]. Like other SDE methods, that approach uses *stochastic trace estimation* to estimate the Chebyshev moments of p. In particular, let $m_1, \ldots, m_k$ denote the first k Chebyshev moments. I.e., $m_j = \frac{1}{n}\sum_{i=1}^{n} T_j(\lambda_i)$. Then we have for each j,

$$m_j = \frac{1}{n}\sum_{i=1}^{n} T_j(\lambda_i) = \frac{1}{n}\mathrm{tr}(T_j(A)),$$

where tr is the matrix trace. Stochastic trace estimation methods like Hutchinsons method can approximate $\mathrm{tr}(T_j(A))$ efficiently via multiplication of $T_j(A)$ with random vectors [Gir87; Hut90]. In particular, for any vector $g \in \mathbb{R}^n$ with mean 0, variance 1 entries, we have that:

$$\mathbb{E}[g^T T_j(A) g] = \mathrm{tr}(T_j(A)).$$

$T_j(A)g$, and thus $g^T T_j(A)g$, can be computed using j matrix-vector products with A. In fact, by using the Chebyshev polynomial recurrence, we can compute $g^T T_j(A)g$ for all $j = 1, \ldots, k$ using k total matrix-vector products:

$$T_0(A)g = g \qquad T_1(A)g = Ag \qquad \ldots \qquad T_j(A)g = 2AT_{j-1}(A)g - T_{j-2}(A)g.$$

Optimized methods can actually get away with $\lceil k/2 \rceil$ matrix-vector products [Che23]. Using a standard analysis of Hutchinson's trace estimator (see, e.g., [RA15] or [CK22]) Braverman et al. [BKM22] prove the following:

Lemma 16 ([BKM22, Lemma 4.2])**.** *Let A be a matrix with $\|A\|_2 \le 1$. Let C be a fixed constant, $j \in \mathbb{Z}_{>0}$, $\alpha, \gamma \in (0,1)$, and $\ell_j = \lceil 1 + \frac{C\log^2(1/\alpha)}{nj\gamma^2} \rceil$. Let $g_1, \ldots, g_{\ell_j} \sim \mathrm{Uniform}(\{-1,1\}^n)$ and let $\hat{m}_j = \frac{1}{\ell_j n}\sum_{i=1}^{\ell_j} g_i^\top T_j(A) g_i$. Then, with probability $1-\alpha$, $|\hat{m}_j - m_j| \le \sqrt{j}\gamma$.*

We combine this lemma with Theorem 1 to prove the following more precise version of Theorem 5:

Theorem 17. *There is an algorithm that, given $\epsilon \in (0,1)$, symmetric $A \in \mathbb{R}^{n\times n}$ with spectral density p, and upper bound $S \ge \|A\|_2$, uses $\min\left\{n, O\left(\frac{1}{\epsilon}\cdot\left(1 + \frac{\log^2(1/\epsilon)\log^2(1/(\epsilon\delta))}{n\epsilon}\right)\right)\right\}$ matrix-vector products with A and $\tilde{O}(n/\epsilon + 1/\epsilon^3)$ additional time to output a distribution q such that, with probability at least $1-\delta$, $W_1(p,q) \le \epsilon S$.*

Proof. First note that, if $\epsilon \le 1/n$, the above result can be obtained by simply recovering A by multiplying by all $n \le 1/\epsilon$ standard basis vectors. We can then compute a full eigendecomposition to extract A's spectral density, which takes $o(n^3)$ time. So we focus on the regime when $\epsilon > 1/n$.

Without loss of generality, we may assume from here forward that $\|A\|_2 \le 1$ and our goal is to prove that $W_1(p,q) \le \epsilon$. In particular, we can scale A by $1/S$, compute an approximate spectral density q with error ϵ, then rescale by S to achieve error ϵS. As mentioned in Section 1.3, an S satisfying $\|A\|_2 \le S \le 2\|A\|_2$ can be computed using $O(\log n)$ matrix-multiplications with A via the power method [KW92]. Given such an S, Theorem 17 implies an error bound of $2\epsilon\|A\|_2$. In some settings of interest for the SDE problem, for example when A is the normalized adjacency matrix of a graph [CKSV18; DBB19; JKMSS24], $\|A\|_2$ is known a priori, so we can simply set $S = \|A\|_2$.

Choose $k = \hat{c}/\epsilon$ for a sufficiently large constant $\hat{c}$ and apply Lemma 16 for all $j = 1, \ldots, k$ with $\gamma = \frac{1}{k\sqrt{1+\log k}}$, and $\alpha = \delta/k$. By a union bound, we obtain estimates $\hat{m}_1, \ldots, \hat{m}_k$ satisfying, for all j,

$$|\hat{m}_j - m_j| \le \sqrt{j}\gamma = \sqrt{j}\cdot\frac{1}{k\sqrt{1+\log k}}. \tag{14}$$

Applying Theorem 1 (specifically, (3)) and Corollary 2, we conclude that, using these moments, Algorithm 1 can recover a distribution q satisfying:

$$W_1(p,q) \leq \frac{2c'}{k}.$$

I.e., we have $W_1(p,q) \leq \epsilon$ as long as $\hat{c} \geq 2c'$. This proves the accuracy bound. We are left to analyze the complexity of the method. We first bound the total number of matrix-vector multiplications with A, which we denote by T. Since $\ell_j \leq \ell_{j-1}$ for all j, computing the necessary matrix-vector product to approximate m_j only costs ℓ_{j-1} additional products on top of those used to approximate m_{j-1}. So, recalling that $\ell_j = \lceil 1 + \frac{C\log^2(1/\alpha)}{nj\gamma^2} \rceil$, we have:

$$T = \left(1 + \frac{C\log^2(k/\delta)}{n\gamma^2}\right) + \left(1 + \frac{C\log^2(k/\delta)}{2n\gamma^2}\right) + \cdots + \left(1 + \frac{C\log^2(k/\delta)}{kn\gamma^2}\right).$$

Using the fact that $1 + 1/2 + \ldots + 1/k \leq 1 + \log(k)$ we can upper bound T by:

$$T = O\left(k + \frac{\log^2(k/\delta)\log(k)}{n\gamma^2}\right) = O\left(k + \frac{k^2\log^2(k/\delta)\log^2(k)}{n}\right),$$

which gives the desired matrix-vector product bound since $k = O(1/\epsilon)$.

In terms of computational complexity, Corollary 2 immediately yields a bound of $\mathsf{poly}(1/\epsilon)$ time to solve the quadratic program in Algorithm 1. However, this runtime can actually be improved to $\tilde{\mathcal{O}}(1/\epsilon^3)$ by taking advantage of the fact that $\hat{m}_1, \ldots, \hat{m}_k$ obey the stronger bound of (3) instead of just (1). This allows us to solve a linear program instead of a quadratic program. In particular, let $\mathcal{C}$ be a grid of Chebyshev nodes, as used in Algorithm 1. I.e., $\mathcal{C} = \{x_1, \ldots, x_g\}$ where $x_i = \cos\left(\frac{2i-1}{2g}\pi\right)$. Let $q_1^{\text{LP}}, \ldots, q_g^{\text{LP}}$ be any solution to the following linear program with variables $z_1, \ldots, z_g$:

$$\begin{aligned}
\text{minimize} \quad & 0 \\
\text{subject to} \quad & \sum_{i=1}^{g} z_i = 1 \\
& z_i \geq 0, \qquad \forall i \in \{1,\ldots,g\} \\
& \sum_{i=1}^{g} T_j(x_i) z_i \leq \hat{m}_j + \left(\sqrt{j}\gamma + \frac{j\sqrt{2\pi}}{g}\right), \ \forall j \in \{1,\ldots,k\} \\
& \sum_{i=1}^{g} T_j(x_i) z_i \geq \hat{m}_j - \left(\sqrt{j}\gamma + \frac{j\sqrt{2\pi}}{g}\right), \ \forall j \in \{1,\ldots,k\}.
\end{aligned} \tag{15}$$

We first verify that the linear program has a solution. To do so, note that, by Equation (16) in Appendix A, there exists a distribution $\tilde{p}$ supported on $\mathcal{C} = \{x_1, \ldots, x_g\}$, such that $|m_j(p) - m_j(\tilde{p})| \leq \frac{j\sqrt{2\pi}}{g}$. By (14) and triangle inequality, it follows that $\tilde{p}$ is a valid solution to the linear program.

Next, let $q^{\text{LP}} = \sum_{i=1}^{g} q_i^{\text{LP}} \delta(x - x_i)$ be the distribution formed by any solution to the linear program. We have that, for any j,

$$\left|m_j - \langle q^{\text{LP}}, T_j\rangle\right| \leq \left|\langle q^{\text{LP}}, T_j\rangle - \hat{m}_j\right| + |\hat{m}_j - m_j| \leq 2\sqrt{j}\gamma + \frac{j\sqrt{2\pi}}{g}.$$

Setting $g = k^{1.5}\sqrt{1+\log(k)}$ and plugging into Theorem 1, we conclude that

$$W_1(p, q^{\text{LP}}) \leq O(1/k).$$

The linear program in (15) has $g = \tilde{O}(k^{1.5})$ variables, boundary constraints for each variable, and $2k + 1$ other constraints. It follows that it can be solved in $\tilde{O}(gk \cdot \sqrt{k}) = \tilde{O}(k^3)$ time [LS14; LS15], which equals $\tilde{O}(1/\epsilon^3)$ time since we chose $k = O(1/\epsilon)$. □

6 Empirical Evaluation of Private Synthetic Data

In this section, we empirically evaluate the application of our main result to differentially private synthetic data generation, as presented in Section 4. Specifically, we implement the procedure given in Algorithm 2, which produces an (ϵ, δ)-differentially private distribution q that approximates the uniform distribution, p, over a given dataset $X = x_1, \ldots, x_n \in [-1, 1]$. We solve the linearly constrained least squares problem from Algorithm 2 using an interior-point method from MOSEK [DB16; MOS19; ART03]. We evaluate the error $W_1(p, q)$ achieved by the procedure on both real world data and data generated from known probability density functions (PDFs), with a focus on how the error scales with the number of data points, n.

For real world data, we first consider the American Community Survey (ACS) data from the Folktables repository [DHMS21]. We use the 2018 ACS 1-Year data for the state of New York; we

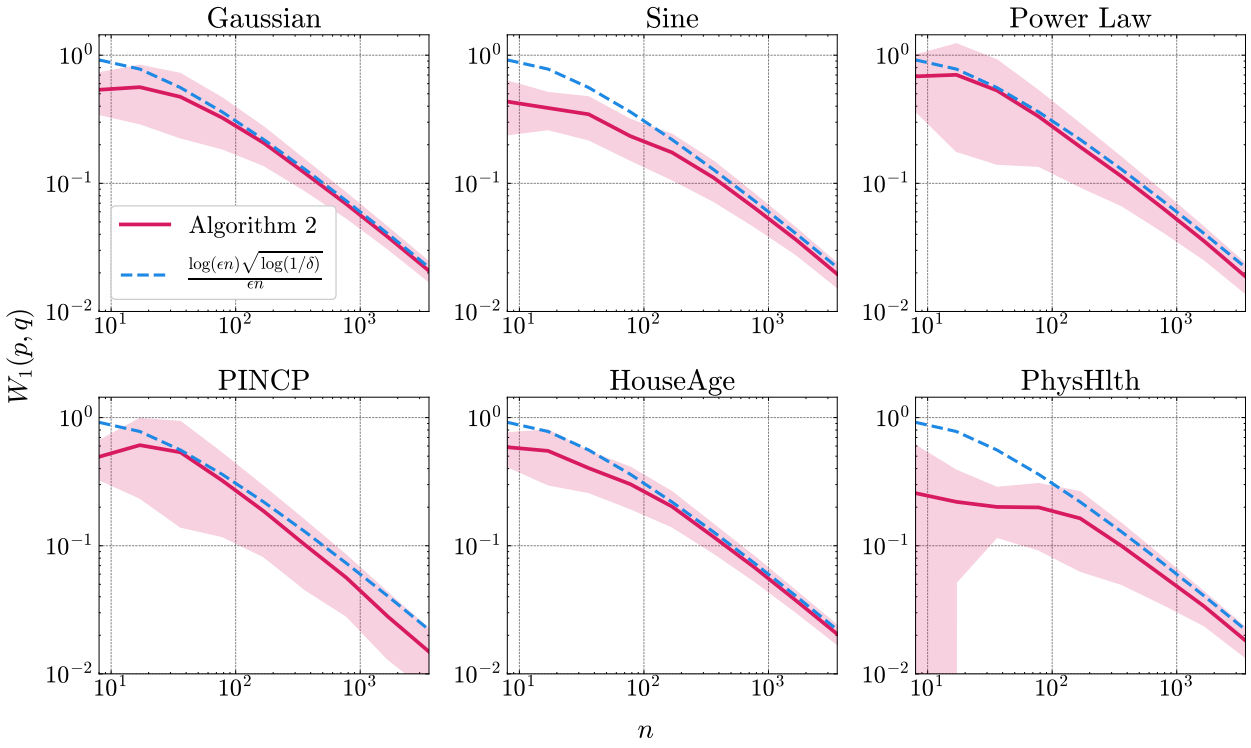

Figure 1: Experimental validation of Algorithm 2 for private synthetic data. For each dataset, we collect subsamples of size n for varying choices of n. We plot the W_1 distance between the uniform distribution, p, over the subsample and a differentially private approximation, q, constructed by Algorithm 2 with privacy parameters $\epsilon = 0.5$ and $\delta = 1/n^2$. As predicted by Theorem 4, the Wasserstein error scales as $\tilde{O}(1/n)$. The solid red line shows the mean of $W_1(p, q)$ over 10 trials, while the shaded region plots one standard deviation around the mean (based on the empirical variance across trials). The blue dotted line plots the theoretical bound of Theorem 4, without any leading constant.

give results for the `PINCP` (personal income) column from this data. We also consider the California Housing dataset [PB97]; we give results for the `HouseAge` (median house age in district) column, from this data. Finally, we consider the CDC Diabetes Health Indicators dataset [Teb21; KLN24]; we give results for the `PhysHlth` (number of physically unhealthy days) from this data. For each of these data sets, we collect uniform subsamples of size n for varying values of n.

In addition to the real world data, we generate datasets of varying size from three fixed probability distributions over $[-1, 1]$. We set the probability mass for $x \in [-1, 1]$ proportional to a chosen function $f(x)$, and equal to 0 for $x \notin [-1, 1]$. We consider the following choices for f: `Gaussian`, $f(x) = e^{-0.5x^2}$; `Sine`, $f(x) = \sin(\pi x) + 1$; and `Power Law`, $f(x) = (x + 1.1)^{-2}$.

For all datasets, we run Algorithm 2 with privacy parameters $\epsilon = 0.5$ and $\delta = 1/n^2$; this is a standard setting for private synthetic data [MMSM22; RHR+23]. We use the default choice of $k = \lceil 2\epsilon n \rceil$. In Figure 1, we plot the average Wasserstein error achieved across 10 trials of the method as a function of n. Error varies across trials due to the randomness in Algorithm 2 (given its use of the Gaussian mechanism) and due to the random choise of a subsample of size n.

As we can see, our experimental results strongly confirm our theoretical guarantees: the average W_1 error closely tracks our theoretical accuracy bound of $O\left(\log(\epsilon n)\sqrt{\log(1/\delta)}/\epsilon n\right)$ from Theorem 4, which is shown as a blue dotted line in Figure 1.

Acknowledgements

We thank Raphael Meyer for suggesting the lower bound on the number of matrix-vector multiplications required for spectral density estimation. We thank Tyler Chen for close proofreading and Gautam Kamath for helpful pointers to the literature. This work was partially supported by NSF Grants 2046235 and 2045590.

References

[Abo18] John Abowd. The US census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2867–2867, 2018 (cited on page 3).

[AAS+19] John Abowd, Robert Ashmead, Garfinkel Simson, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, and William Sexton. Census topdown: differentially private data, incremental schemas, and consistency with public knowledge. *US Census Bureau*, 2019 (cited on page 3).

[ACK+24] Noah Amsel, Tyler Chen, Feyza Duman Keles, Diana Halikias, Cameron Musco, and Christopher Musco. Fixed-sparsity matrix approximation from matrix-vector products. *arXiv:2402.09379*, 2024 (cited on page 5).

[ART03] Erling D. Andersen, Cornelis Roos, and Tamás Terlaky. On implementing a primal-dual interior-point method for conic quadratic optimization. *Mathematical Programming*, 95(2):249–277, 2003 (cited on pages 9, 16).

[AT11] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58(2), 2011 (cited on page 5).

[ABK+21] Sergul Aydore, William Brown, Michael Kearns, Krishnaram Kenthapadi, Luca Melis, Aaron Roth, and Ankit A Siva. Differentially private query release through adaptive projection. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 457–467, 2021 (cited on page 3).

[BN23] Ainesh Bakshi and Shyam Narayanan. Krylov methods are (nearly) optimal for low-rank approximation. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2023 (cited on page 5).

[Bid23] Joseph R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023 (cited on page 3).

[BSV24] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private measures, random walks, and synthetic data. *Probability Theory and Related Fields*, 189(1):569–611, 2024 (cited on pages 2–4).

[BHSW20] Mark Braverman, Elad Hazan, Max Simchowitz, and Blake Woodworth. The gradient complexity of linear regression. In *Proceedings of the 33rd Annual Conference on Computational Learning Theory (COLT)*, pages 627–647, 2020 (cited on page 5).

[BKM22] Vladimir Braverman, Aditya Krishnan, and Christopher Musco. Sublinear time spectral density estimation. In *Proceedings of the 54th Annual ACM Symposium on Theory of Computing (STOC)*, 2022 (cited on pages 1, 2, 5, 7, 8, 14).

[Che22] Tyler Chen. *Lanczos-based methods for matrix functions*. PhD thesis, University of Washington, 2022 (cited on page 1).

[Che23] Tyler Chen. A spectrum adaptive kernel polynomial method. *The Journal of Chemical Physics*, 159(11), 2023 (cited on page 14).

[CKHMM24] Tyler Chen, Feyza Duman Keles, Diana Halikias, Cameron Musco, and Christopher Musco. Near-optimal hierarchical matrix approximation from matrix-vector products. *arXiv:2407.04686*, 2024 (cited on page 5).

[CTU21] Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. Analysis of stochastic lanczos quadrature for spectrum approximation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021 (cited on page 1).

[CTU22] Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. Randomized matrix-free quadrature for spectrum and spectral sum approximation. *arXiv:2204.01941*, 2022 (cited on page 4).

[CDLLN23] Sinho Chewi, Jaume De Dios Pont, Jerry Li, Chen Lu, and Shyam Narayanan. Query lower bounds for log-concave sampling. In *Proceedings of the 64th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2023 (cited on page 5).

[CKSV18] David Cohen-Steiner, Weihao Kong, Christian Sohler, and Gregory Valiant. Approximating the spectrum of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1263–1271, 2018 (cited on pages 1, 4, 14).

[CK22] Alice Cortinovis and Daniel Kressner. On randomized trace estimates for indefinite matrices with an application to determinants. *Foundations of Computational Mathematics*, 22(3):875–903, 2022 (cited on page 14).

[CKS91] Yann Le Cun, Ido Kanter, and Sara A. Solla. Eigenvalues of covariance matrices: application to neural-network learning. *Phys. Rev. Lett.*, 66:2396–2399, May 1991 (cited on page 4).

[DB16] Steven Diamond and Stephen Boyd. CVXPY: a Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016 (cited on pages 2, 16).

[DHMS21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: new datasets for fair machine learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021 (cited on page 16).

[DSB21] Josep Domingo-Ferrer, David Sánchez, and Alberto Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021 (cited on page 3).

[DBB19] Kun Dong, Austin R. Benson, and David Bindel. Network density of states. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1152–1161, 2019 (cited on pages 4, 14).

[DKMMN06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT*, pages 486–503, 2006 (cited on pages 4, 11).

[DNRRV09] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 381–390, 2009 (cited on page 1).

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014 (cited on pages 1, 3, 4, 11).

[FL23] Zhiyuan Fan and Jian Li. Efficient algorithms for sparse moment problems without separation. In *Proceedings of the 36th Annual Conference on Computational Learning Theory (COLT)*, pages 3510–3565, 2023 (cited on page 1).

[FMST24] Vitaly Feldman, Audra McMillan, Satchit Sivakumar, and Kunal Talwar. Instance-optimal private density estimation in the wasserstein distance. *arXiv:2406.19566*, 2024 (cited on page 4).

[Fis11] Hans Fischer. *Chebyshev's and Markov's contributions.* In *A History of the Central Limit Theorem: From Classical to Modern Probability Theory.* Springer New York, 2011, pages 139–189 (cited on page 1).

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2232–2241, 2019 (cited on page 4).

[Gir87] Didier Girard. Un algorithme simple et rapide pour la validation croisee géénéralisée sur des problémes de grande taille. Technical report, École nationale supérieure d'informatique et de mathématiques appliquées de Grenoble, 1987 (cited on page 14).

[Hal15] Nicholas Hale. *Chebyshev polynomials*. In *Encyclopedia of Applied and Computational Mathematics*. Springer Berlin Heidelberg, 2015, pages 203–205 (cited on page 6).

[HLM12] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, 2012 (cited on page 3).

[HRMS10] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1–2):1021–1032, 2010 (cited on page 4).

[HVZ23] Yiyun He, Roman Vershynin, and Yizhe Zhu. Algorithmically effective differentially private synthetic data. In *Proceedings of the 36th Annual Conference on Computational Learning Theory (COLT)*, pages 3941–3968, 2023 (cited on pages 3, 4).

[Hut90] Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990 (cited on pages 5, 14).

[Jac12] Dunham Jackson. On approximation by trigonometric sums and polynomials. *Transactions of the American Mathematical Society*, 13(4):491–515, 1912 (cited on page 3).

[Jac30] Dunham Jackson. *The Theory of Approximation*, volume 11 of *Colloquium Publications*. American Mathematical Society, 1930 (cited on page 7).

[JL14] Zhanglong Ji and Charles Lipton Zachary C.and Elkan. Differential privacy and machine learning: a survey and review. *arXiv:1412.7584*, 2014 (cited on page 3).

[JPWZ24] Shuli Jiang, Hai Pham, David P. Woodruff, and Qiuyi Zhang. Optimal sketching for trace estimation. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024 (cited on page 5).

[JKMSS24] Yujia Jin, Ishani Karmarkar, Christopher Musco, Apoorv Singh, and Aaron Sidford. Faster spectral density estimation and sparsification in the nuclear norm. In *Proceedings of the 37th Annual Conference on Computational Learning Theory (COLT)*, 2024 (cited on pages 4, 5, 14).

[JMSS23] Yujia Jin, Christopher Musco, Aaron Sidford, and Apoorv Vikram Singh. Moments, random walks, and limits for spectrum approximation. In *Proceedings of the 36th Annual Conference on Computational Learning Theory (COLT)*, 2023 (cited on page 1).

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC)*, pages 553–562, 2010 (cited on page 1).

[Kam20] Gautam Kamath. Cs 860: algorithms for private data analysis, lecture 11 – packing lower bounds, 2020 (cited on page 4).

[KV86] Sanjeev Kapoor and Pravin M. Vaidya. Fast algorithms for convex quadratic programming and multicommodity flows. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing (STOC)*, pages 147–159, 1986 (cited on page 9).

[KLN24] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. UCI Machine Learning Repository, Diabetes Health Indicators Dataset. https://www.archive.ics.uci.

edu/dataset/891/cdc+diabetes+health+indicators, 2024. [Accessed 11-07-2024] (cited on page 17).

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011 (cited on page 11).

[KV17] Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45(5):2218–2247, October 2017 (cited on page 1).

[KW92] J. Kuczyński and H. Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992 (cited on pages 5, 14).

[LS14] Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: solving linear programs in $\tilde{O}(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 424–433, 2014 (cited on page 16).

[LS15] Yin Tat Lee and Aaron Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 230–249, 2015 (cited on page 16).

[LLSY17] Ninghui Li, Min Lyu, Dong Su, and Weining Yang. *Differential Privacy: From Theory to Practice*. Synthesis Lectures on Information Security, Privacy, and Trust. Springer, 2017 (cited on pages 3, 4).

[LXES19] Ruipeng Li, Yuanzhe Xi, Lucas Erlandson, and Yousef Saad. The eigenvalues slicing library (EVSL): algorithms, implementation, and software. *SIAM Journal on Scientific Computing*, 41(4):C393–C415, 2019 (cited on page 4).

[Lig58] Michael J. Lighthill. *An introduction to Fourier analysis and generalised functions*. Cambridge University Press, 1958 (cited on page 7).

[LVW21] Terrance Liu, Giuseppe Vietri, and Steven Z. Wu. Iterative methods for private synthetic data: unifying framework and new methods. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021 (cited on pages 3, 4).

[MM19] Michael Mahoney and Charles Martin. Traditional and heavy tailed self regularization in neural network models. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4284–4293, 2019 (cited on page 4).

[MMSM22] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11):2599–2612, 2022 (cited on pages 1, 3, 17).

[MM09] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the Netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 627–636, 2009 (cited on page 4).

[MMMW21] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David Woodruff. Hutch++: optimal stochastic trace estimation. *Proceedings of the 4th Symposium on Simplicity in Algorithms (SOSA)*, 2021 (cited on page 5).

[MTV+20] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: a survey. *arXiv:2004.12254*, 2020 (cited on page 3).

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 93–102, 2010 (cited on page 1).

[MAP20] Dean Moldovan, Misa Andelkovic, and Francois Peeters. pybinding v0.9.5: a Python package for tight-binding calculations, 2020 (cited on page 4).

[MOS19] MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.* 2019 (cited on pages 2, 16).

[MM15] Cameron Musco and Christopher Musco. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 1396–1404, 2015 (cited on page 5).

[PB97] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997 (cited on page 17).

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, 185:71–110, 1894 (cited on page 1).

[Pea36] Karl Pearson. Method of moments and method of maximum likelihood. *Biometrika*, 28(1/2):34–59, 1936 (cited on page 1).

[PSG18] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1924–1932, 2018 (cited on page 4).

[QYL13] Wahbeh Qardaji, Weining Yang, and Ninghui Li. Understanding hierarchical methods for differentially private histograms. *Proceedings of the VLDB Endowment*, 6(14):1954–1965, 2013 (cited on page 4).

[RSS14] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 207–224, 2014 (cited on page 1).

[Riv69] Theodore J. Rivlin. *An introduction to the approximation of functions*. Dover Publications, 1969 (cited on page 10).

[RA15] Farbod Roosta-Khorasani and Uri M. Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015 (cited on page 14).

[RHR+23] Lucas Rosenblatt, Anastasia Holovenko, Taras Rumezhak, Andrii Stadnik, Bernease Herman, Julia Stoyanovich, and Bill Howe. Epistemic parity: reproducibility as an evaluation metric for differential privacy. *Proceedings of the VLDB Endowment*, 2023 (cited on pages 3, 17).

[RLP+20] Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. Differentially private synthetic data: applied evaluations and enhancements. *arXiv:2011.05537*, 2020 (cited on page 1).

[SR94] Richard N. Silver and H. Röder. Densities of states of mega-dimensional hamiltonian matrices. *International Journal of Modern Physics C*, 5(4):735–753, 1994 (cited on page 4).

[SER18] Max Simchowitz, Ahmed El Alaoui, and Benjamin Recht. Tight query complexity lower bounds for PCA via finite sample deformed wigner law. In *Proceedings of the 50th Annual ACM Symposium on Theory of Computing (STOC)*, 2018 (cited on page 5).

[Ski89] John Skilling. *The eigenvalues of mega-dimensional matrices*. In *Maximum Entropy and Bayesian Methods*. Springer Netherlands, 1989 (cited on page 4).

[SWYZ21] Xiaoming Sun, David P. Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *ACM Trans. Algorithms*, 17(4), 2021 (cited on page 5).

[Teb21] Alex Teboul. Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset, 2021. [Accessed 11-07-2024] (cited on page 17).

[Tre08] Lloyd N. Trefethen. Is Gauss Quadrature Better than Clenshaw–Curtis? *SIAM Review*, 50(1):67–87, 2008 (cited on page 7).

[Tre19] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM-Society for Industrial and Applied Mathematics, 2019 (cited on pages 3, 7).

[VAA+22] Giuseppe Vietri, Cedric Archambeau, Sergul Aydore, William Brown, Michael Kearns, Aaron Roth, Ankit Siva, Shuai Tang, and Steven Z Wu. Private synthetic data for multitask learning and marginal queries. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022 (cited on page 4).

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019 (cited on page 26).

[WJF+16] Ziteng Wang, Chi Jin, Kai Fan, Jiaqi Zhang, Junliang Huang, Yiqiao Zhong, and Liwei Wang. Differentially private data releasing for smooth queries. *Journal of Machine Learning Research*, 17(51):1–42, 2016 (cited on pages 1, 3, 4).

[WWAF06] Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. *Rev. Mod. Phys.*, 78:275–306, 2006 (cited on pages 1, 5).

[WZZ22] David Woodruff, Fred Zhang, and Richard Zhang. Optimal query complexities for dynamic trace estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022 (cited on pages 5, 27).

[WR22] Stephen J. Wright and Benjamin Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022 (cited on page 9).

[WY19] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019 (cited on page 1).

[WY20] Yihong Wu and Pengkun Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981–2007, 2020 (cited on page 1).

[XWG10] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on knowledge and data engineering*, 23(8):1200–1214, 2010 (cited on page 4).

[XZX+13] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal*, 22:797–822, 2013 (cited on page 4).

[YGKM20] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: neural networks through the lens of the Hessian. In *2020 IEEE International Conference on Big Data*, pages 581–590, 2020 (cited on page 4).

[YT89] Yinyu Ye and Edison Tse. An extension of Karmarkar's projective algorithm for convex quadratic programming. *Mathematical Programming*, 44(1):157–179, 1989 (cited on page 9).

[ZCPSX17] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017 (cited on page 3).

[ZXX16] Jun Zhang, Xiaokui Xiao, and Xing Xie. PrivTree: a differentially private algorithm for hierarchical decompositions. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*, pages 155–170, 2016 (cited on page 4).

A Proof of Corollary 2

In this section, we give the full proof of Corollary 2. We require the following basic property about the Chebyshev nodes:

Lemma 18 (Chebyshev Node Approximation)**.** *Let $\mathcal{C} = \{x_1, \ldots, x_g\}$ be the degree g Chebyshev nodes. I.e., $x_i = \cos\left(\frac{2i-1}{2g}\pi\right)$. Let $r_{\mathcal{C}} : [-1,1] \to \mathcal{C}$ be a function that maps a point $x \in [-1,1]$ to the point $y \in \mathcal{C}$ that minimizes $\left|\cos^{-1}(x) - \cos^{-1}(y)\right|$, breaking ties arbitrarily. For any $x \in [-1,1]$, $\left|\cos^{-1}(x) - \cos^{-1}(r_{\mathcal{C}}(x))\right| \leq \frac{\pi}{2g}$.*

Proof. For any two consecutive points x_i, x_{i+1} in the $\mathcal{C}$,

$$\left|\cos^{-1}(x_i) - \cos^{-1}(x_{i+1})\right| = \frac{\pi}{g}.$$

Since $\cos^{-1}(x)$ is non-increasing, for any $x \in [x_{i+1}, x_i]$, $\cos^{-1}(x) \in [\cos^{-1}(x_i), \cos^{-1}(x_{i+1})]$. So, $\cos^{-1}(x)$ has distance at most $\frac{\pi}{2g}$ from either $\cos^{-1}(x_i)$ or $\cos^{-1}(x_{i+1})$. Additionally, we can check that $\left|\cos^{-1}(x) - \cos^{-1}(x_1)\right| \leq \frac{\pi}{2g}$ for any $x < x_1$ and $\left|\cos^{-1}(x) - \cos^{-1}(x_g)\right| \leq \frac{\pi}{2g}$ for any $x > x_g$. □

With Lemma 18 in place, we are ready to prove Corollary 2.

Proof of Corollary 2. Let $\mathcal{C}$ and $r_{\mathcal{C}} : [-1,1] \to \mathcal{C}$ be as in Lemma 18. For $i \in \{1,\dots,g\}$, let Y_i be the set of points in $[-1,1]$ that are closest to $x_i \in \mathcal{C}$, i.e., $Y_i = \{x \in [-1,1] : r_{\mathcal{C}}(x) = x_i\}$. Let $\tilde{p}$ be a distribution supported on the set $\mathcal{C}$ with mass $\int_{Y_i} p(x)\,dx$ on $x_i \in \mathcal{C}$. For all $j \in 1,\dots,k$ we have:

$$\begin{aligned}
\left|\langle p, \bar{T}_j\rangle - \langle \tilde{p}, \bar{T}_j\rangle\right| &= \left|\sum_{i=1}^{g} \int_{Y_i} \bar{T}_j(x)p(x)\,dx - \left(\int_{Y_i} p(x)\,dx\right)\bar{T}_j(x_i)\right| \\
&= \left|\sum_{i=1}^{g} \left(\int_{Y_i} p(x)\,dx\right)\bar{T}_j(y_i) - \left(\int_{Y_i} p(x)\,dx\right)\bar{T}_j(x_i)\right| \qquad \text{(for some } y_i \in Y_i\text{)} \\
&\le \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\,dx\right)\left|\bar{T}_j(y_i) - \bar{T}_j(x_i)\right| \\
&= \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\,dx\right)\cdot\sqrt{\frac{2}{\pi}}\cdot\left|\cos(j\cos^{-1}(y_i)) - \cos(j\cos^{-1}(x_i))\right| \\
&\le \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\,dx\right)\cdot\sqrt{\frac{2}{\pi}}\cdot\frac{j\pi}{2g} = \frac{j\sqrt{\pi/2}}{g} \qquad (16)
\end{aligned}$$

The second equality follows from the intermediate value theorem. The first inequality follows by triangle inequality. The third equality follows by the trigonometric definition of the (normalized) Chebyshev polynomials. The second inequality follows from Lemma 18 and the fact that the derivative of $\cos(jx)$ is bounded by j. The bound in (16) then yields:

$$\left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \le \frac{\sqrt{\pi k/2}}{g}. \qquad (17)$$

Observe also that, since $\tilde{p}$ is supported on $\mathcal{C}$, it is a valid solution to the optimization problem solved by Algorithm 1. Accordingly, we have that:

$$\left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\hat{m}_j - \langle q, \bar{T}_j\rangle\right)^2\right)^{1/2} \le \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\hat{m}_j - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \qquad (18)$$

Applying triangle inequality, followed by (18), triangle inequality again, and finally (17), we have:

$$\begin{aligned}
\left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \langle q, \bar{T}_j\rangle\right)^2\right)^{1/2} &\le \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \hat{m}_j\right)^2\right)^{1/2} + \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\hat{m}_j - \langle q, \bar{T}_j\rangle\right)^2\right)^{1/2} \\
&\le \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \hat{m}_j\right)^2\right)^{1/2} + \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\hat{m}_j - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \\
&\le 2\left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \hat{m}_j\right)^2\right)^{1/2} + \left(\sum_{j=1}^{k}\frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \\
&\le 2\Gamma + \frac{\sqrt{2\pi k}}{g}.
\end{aligned}$$

Setting $g = \lceil k^{1.5}\rceil$, we can apply Theorem 1 to conclude that, for a fixed constant c',

$$W_1(p,q) \le \frac{c}{k} + 2\Gamma + \frac{\sqrt{\pi/2}}{k} \le c'\cdot\left(\frac{1}{k} + \Gamma\right). \qquad \square$$

B Theorem 4 High Probability Bound

In this section, we prove the high probability bound on Wasserstein distance stated in Theorem 4, which follows from a standard concentration bound for sub-exponential random variables [Wai19]. We recall that a random variable X is subexponential with parameters (ν, α) if:

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\nu^2\lambda^2/2} \qquad \text{for all} \qquad |\lambda| \leq \frac{1}{\alpha}.$$

We require the following well-known fact that a chi-square random variable with one degree of freedom is subexponential:

Fact 19 (Sub-Exponential Parameters [Wai19, Example 2.8])**.** *Let $\eta \sim \mathcal{N}(0, \sigma^2)$. Then, η^2 is sub-exponential random variable with parameters $(2\sigma^2, 4\sigma^2)$.*

We also require the following concentration inequality for a sum of sub-exponential random variable:

Fact 20 ([Wai19, Equation 2.18])**.** *Consider independent random variables $\gamma_1, \ldots, \gamma_k$, where, $\forall j \in 1, \ldots, k$, γ_j is sub-exponential with parameters (ν_j, α_j). Let $\nu_* = \sqrt{\sum_{j=1}^k \nu_j^2}$ and $\alpha_* = \max\{\alpha_1, \ldots, \alpha_k\}$. Then we have:*

$$\mathbb{P}\left[\sum_{j=1}^k (\gamma_j - \mathbb{E}[\gamma_j]) \geq t\right] \leq \begin{cases} \exp\left(\frac{-t^2}{2\nu_*^2}\right) & \text{for } 0 \leq t \leq \frac{\nu_*^2}{\alpha_*}, \\ \exp\left(\frac{-t}{2\alpha_*}\right) & \text{for } t > \frac{\nu_*^2}{\alpha_*}. \end{cases}$$

Proof of high-probability bound of Theorem 4. Recalling the proof of the expectation bound of Theorem 4 from Section 4, it suffices to bound $E = \sum_{j=1}^k \frac{1}{j^2} \left(\hat{m}_j(p) - \langle \tilde{p}, T_j \rangle\right)^2$ with high probability. Let $\gamma_j = \eta_j^2/j^2$, where $\eta_j \sim \mathcal{N}(0, j\sigma^2)$ is as in Algorithm 2. Then recall that $E = \sum_{j=1}^k \gamma_j$.

From Fact 19, γ_j is a sub-exponential random variable with parameter $(2\sigma^2/j, 4\sigma^2/j)$. We can then apply Fact 20, for which we have $\nu_* = \sqrt{\sum_{j=1}^k 4\sigma^4/j^2} \leq 2\pi\sigma^2/\sqrt{6}$ and $\alpha_* = 4\sigma^2$. For any failure probability $\beta \in (0, 1/2)$, setting $t = 8\log(1/\beta)\sigma^2$, we conclude that:

$$\mathbb{P}\left[E - \mathbb{E}[E] \geq 8\log(1/\beta)\,\sigma^2\right] \leq \beta\,.$$

Recalling from Equation (11) that $\mathbb{E}[E] \leq (1 + \log k)\sigma^2$, we conclude that $E \leq 8\log(1/\beta)\,\sigma^2 + (1 + \log k)\sigma^2$ with probability at least $1 - \beta$.

The rest of the details follow as before. In particular, as in Equation (12), we can bound:

$$W_1(p, q) \leq \sqrt{2}\Gamma + \frac{36}{k} + \frac{1}{2\lceil \epsilon n \rceil},$$

where $\Gamma \leq \sqrt{2E}$. Plugging in $k = \lceil 2\epsilon n \rceil$ (as chosen in Algorithm 2) and recalling that $\sigma^2 = \frac{16}{\pi}(1 + \log k)\ln(1.25/\delta)/(\epsilon^2 n^2)$, we conclude that with probability $\geq 1 - \beta$, for a fixed constant c,

$$W_1(p, q) \leq c\left(\frac{\sqrt{\log(\epsilon n) + \log(1/\beta)}\sqrt{\log(\epsilon n)\log(1/\delta)}}{\epsilon n}\right). \qquad \square$$

C Spectral Density Estimation Lower Bound

In this section, we provide a lower bound on the number of matrix-vector multiplications required for spectral density estimation. We first need the following theorem from Woodruff et al. [WZZ22], which shows that estimating the trace of a positive semi-definite matrix A to within a multiplicative error of $(1 \pm \epsilon)$ requires $\Omega(1/\epsilon)$ number of matrix-vector multiplications with A.

Theorem 21 (Restated [WZZ22, Theorem 4.2])**.** *For all $\delta > 0$ and $\epsilon = O\left(1/\sqrt{\log(1/\delta)}\right)$, any algorithm that is given matrix-vector multiplication access to a positive semi-definite (PSD) input matrix $A \in \mathbb{R}^{n \times n}$ with $\|A\|_2 \leq 1$, $n/4 \leq \mathrm{tr}(A) \leq n$ and succeeds with probability at least $1 - \delta$ in outputting an estimate $\tilde{t}$ such that $|\tilde{t} - \mathrm{tr}(A)| \leq \epsilon \cdot \mathrm{tr}(A)$ requires $\Omega\left(\frac{\log(1/\delta)}{\epsilon}\right)$ matrix-vector multiplications with A.*

As a corollary of this result, we obtain the following lower bound, which shows that Theorem 5 is tight up to $\log(1/\epsilon)$ factors:

Corollary 22. *Any algorithm that is given matrix-vector multiplication access to a symmetric matrix A with spectral density p and $\|A\|_2 \leq 1$ requires $\Omega\left(\frac{\log(1/\delta)}{\epsilon}\right)$ matrix-vector multiplications with A to output a distribution q such that $W_1(p, q) \leq \epsilon$.*

Proof. The proof is via a direct reduction. Consider a PSD matrix A with $\|A\|_2 \leq 1$, $n/4 \leq \mathrm{tr}(A) \leq n$, and spectral density p. Suppose we have a spectral density estimate q of p such that $W_1(p, q) \leq \epsilon/4$. We claim that $\tilde{t} = n \cdot \int_{-1}^{1} xq(x)\,dx$ yields a relative error approximate to A's trace, implying that computing such a q requires $\Omega(\log(1/\delta)/\epsilon)$ by Theorem 21.

In particular, applying Kantorovich-Rubinstein duality (Fact 7) with the 1-Lipschitz functions $f(x) = x$ and $f(x) = -x$, we have that:

$$\int_{-1}^{1} xp(x)\,dx - \int_{-1}^{1} xq(x)\,dx \leq \epsilon/4 \qquad \text{and} \qquad \int_{-1}^{1} xq(x)\,dx - \int_{-1}^{1} xp(x)\,dx \leq \epsilon/4. \tag{19}$$

We have that $\int_{-1}^{1} xp(x)\,dx = \frac{1}{n}\mathrm{tr}(A)$. So (19) implies that $\tilde{t} = n \cdot \int_{-1}^{1} xq(x)\,dx$ satisfies:

$$|\tilde{t} - \mathrm{tr}(A)| \leq n \cdot \epsilon/4 \leq \epsilon \cdot \mathrm{tr}(A). \qquad \square$$