Sharper Bounds for Chebyshev Moment Matching, with Applications

Cameron Musco CMUSCO@CS.UMASS.EDU
UMass Amherst

Christopher Musco CMUSCO@NYU.EDU
New York University

Lucas Rosenblatt LUCAS.ROSENBLATT@NYU.EDU
New York University

Apoorv Vikram Singh APOORV.SINGH@NYU.EDU
New York University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the problem of approximately recovering a probability distribution given noisy measurements of its Chebyshev polynomial moments. This problem arises broadly across algorithms, statistics, and machine learning. By leveraging a *global decay bound* on the coefficients in the Chebyshev expansion of any Lipschitz function, we sharpen prior work, proving that accurate recovery in the Wasserstein distance is possible with more noise than previously known. Our result immediately yields a number of applications:

1. We give 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 result of Boedihardjo, Strohmer, and Vershynin [Probab. Theory. Rel., 2024], which uses a more complex "superregular random walk" method.

2. We give an $\tilde{O}(n^2/\epsilon)$ time algorithm for the linear algebraic problem of estimating the spectral density of an $n \times n$ symmetric matrix up to ϵ error in the Wasserstein distance. Our result accelerates prior methods from Chen et al. [ICML 2021] and Braverman et al. [STOC 2022].

3. We tighten an analysis of Vinayak, Kong, Valiant, and Kakade [ICML 2019] on the maximum likelihood estimator for the statistical problem of "Learning Populations of Parameters", extending the parameter regime in which sample optimal results can be obtained.

Beyond these main results, we provide an extension of our bound to estimating distributions in $d > 1$ dimensions. We hope that these bounds will find applications more broadly to problems involving distribution recovery from noisy moment information.

Keywords: Moment Matching, Chebyshev Polynomials, Differential Privacy, Estimating Population of Parameters, Eigenvalue Estimation, Synthetic Data.

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 (Pearson, 1894, 1936; Fischer, 2011). Moment matching continues to find a wide variety of applications, both in traditional statistical problems (Kalai et al., 2010; Moitra and Valiant, 2010; Rabani et al., 2014; Wu and Yang, 2019, 2020; Fan and Li, 2023) and beyond. For example, moment matching is widely used for estimating eigenvalues in numerical linear algebra and computational chemistry (Weiße et al., 2006; Cohen-Steiner et al., 2018; Chen et al., 2021; Chen, 2022).

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. earth mover'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 (Kong and Valiant, 2017; Chen et al., 2021). In other words, to compute an ε-accurate approximation to p, it suffices to compute p's first $O(1/\varepsilon)$ moments and to return any distribution q with those moments.

Unfortunately, the above result is extremely sensitive to noise, and thus 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 moments must be estimated to accuracy $O(1/2^k)$ if we want to approximate p up to W_1 error of $O(1/k)$ (Jin et al., 2023). In other words, distribution approximation is *poorly conditioned* with respect to the standard moments.

1.1. Chebyshev moment matching

One way to avoid 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 (Weiße et al., 2006; Wang et al., 2016; Braverman et al., 2022) leverages *Chebyshev moments* of the form $\mathbb{E}_{x\sim p}[T_i(x)]$, where T_i is the i^{th} Chebyshev polynomial of the first kind, defined as: $T_0(x) = 1$, $T_1(x) = x$, and for $i \geq 2$, $T_i(x) = 2xT_{i-1}(x) - T_{i-2}(x)$.

The Chebyshev moments are known to be less noise sensitive than the standard moments: instead of exponentially small error, it has been shown that $\tilde{O}(1/k)$ 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., Braverman et al. (2022, Lemma 3.1)). This has enabled efficient algorithms for distribution estimation in various settings. For example, Chebyshev moment matching leads to $O(n^2/\mathsf{poly}(\varepsilon))$ time algorithms for estimating the eigenvalue distribution (i.e., the spectral density) of an $n \times n$ symmetric matrix A to error $\varepsilon\|A\|_2$ in the Wasserstein distance (Braverman et al., 2022).

Chebyshev moment matching has also been used for *differentially private synthetic data generation*. In this setting, p is uniform 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 (see Section 1.3 for more details). A differentially private 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

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 the Wasserstein-1 distance.
2. Throughout, we let $\tilde{O}(z)$ denote $O(z \log^c(z))$ for constant c.

ensures privacy) (Dwork and Roth, 2014). 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 W_1 distance (Wang et al., 2016).

1.2. Our contributions

Despite the success of Chebyshev moment matching, including for the applications discussed above, there is room for improvement. For private distribution estimation, alternative methods can achieve nearly-optimal error $\tilde{O}(1/n)$ in W_1 distance for a dataset of size n (Boedihardjo et al., 2024), 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 (Braverman et al., 2022).

The main contribution of this work is to resolve these gaps by proving a sharper bound on the accuracy to 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 and 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, (Braverman et al., 2022)).

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.*

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.

Algorithm 1 simply solves a linearly-constrained least-squares regression problem to find a distribution q supported on a sufficiently fine grid whose moments match those of p nearly as well as $\hat{m}_1, \ldots, \hat{m}_k$. Corollary 2 follows 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 distribution). This problem is easily solved using off-the-shelf software: in Section A.1 we implement our method using a solver from MOSEK (MOSEK ApS, 2019) and report some initial experimental results.

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 polynomial approximation of Lipschitz functions via "damped Chebyshev expansions" (Jackson, 1912). Lipschitz functions are closely related to approximation in Wasserstein distance through the Kantorovich-Rubinstein duality: $W_1(p, q) = \max_{1\text{-Lipschitz } f} \int_{-1}^{1} f(x)(p(x) - q(x))dx$.

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, in Lemma 13, 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 $O(1/j)$ (Trefethen, 2019). This property is implied by our bound, but is much weaker. We believe that our new decay bound may be of independent interest given the ubiquitous use of Chebyshev expansions across computational science, statistics, and beyond.

1.3. Applications

We highlight three concrete applications of our main bounds, Theorem 1 and Lemma 13, to algorithms for private synthetic data generation, spectral density estimation, and estimating populations of parameters. We suspect further applications exist.

Application 1: Differentially Private Synthetic Data. Privacy-enhancing technologies seek to protect individuals' data without preventing learning from the data. For theoretical guarantees of privacy, the industry standard is *differential privacy* (Dwork and Roth, 2014), which is used in massive data products like the US Census, and is a core tenet of the recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Biden, 2023; Abowd, 2018; Abowd et al., 2019).

Concretely, we are interested in the predominant 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:*

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

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 the would-be output had any one individual's data been replaced with something arbitrary.

There exist differentially private algorithms for many statistical tasks (Ji and Lipton, 2014; Li et al., 2017; Mireshghallah et al., 2020). One task of primary importance is *differentially private data synthesis*. Here, the goal is to generate *synthetic data* 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 (Hardt et al., 2012; Zhang et al., 2017; Rosenblatt et al., 2020; Liu et al., 2021; Abowd et al., 2019; Aydore et al., 2021; Rosenblatt et al., 2023; Domingo-Ferrer et al., 2021). Such methods offer strong empirical performance and a variety of theoretical guarantees, e.g., that the synthetic data can effectively answer a fixed set of data analysis queries with high accuracy (Hardt et al., 2012; McKenna et al., 2022). Recently, there has been interest in algorithms with more general distributional guarantees – e.g., statistical distance guarantees between the synthetic data and the original data (Wang et al., 2016; Boedihardjo et al., 2024; He et al., 2023). 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 that, in $O(n) + \mathsf{poly}(\varepsilon n)$ time, returns a distribution q satisfying, for a fixed constant c_1,*

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

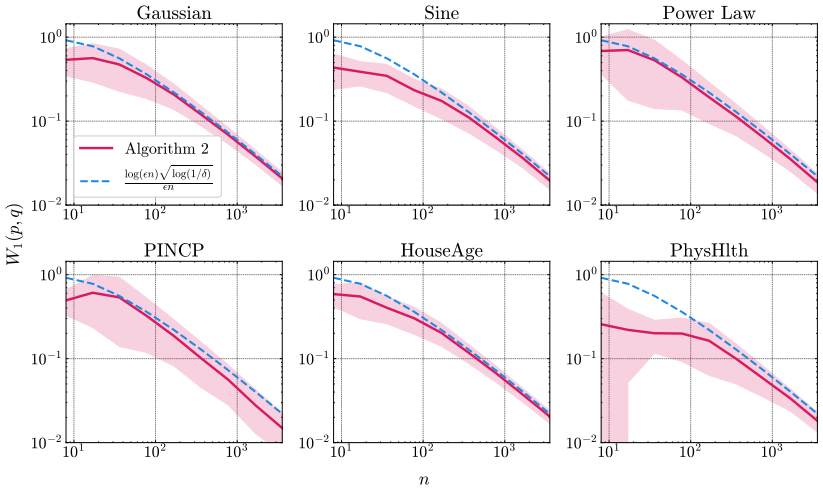
Moreover, for any $\beta \in (0, 1/2)$, $W_1(p, q) \leq \frac{c_1\sqrt{\log(1/\beta)+\log(\varepsilon n)}\sqrt{\log(\varepsilon n)\log(1/\delta)}}{\varepsilon n}$ w.p. $\geq 1 - \beta$.

Theorem 4 is proven in Section A. The returned distribution q is represented as a discrete distribution on $O(\varepsilon n)$ 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[5] a recent result of Boedihardjo, Strohmer, and Vershynin (Boedihardjo et al., 2024), who give an $(\varepsilon, 0)$-differentially private method with expected Wasserstein-1 error $O(\log^{3/2}(n)/(\varepsilon n))$, which is optimal up to logarithmic factors.[6] Like that method, we improve on a natural barrier of $\tilde{O}(1/(\varepsilon\sqrt{n}))$ error that is inherent to naive "private histogram" methods for approximation in the Wasserstein-1 distance (Xiao et al., 2010; Qardaji et al., 2013; Xu et al., 2013; Dwork and Roth, 2014; Zhang et al., 2016; Li et al., 2017). "Private *hierarchical* histogram" methods can also be shown to match the Wasserstein-1 error of $\tilde{O}(1/(\varepsilon n))$, albeit with worse polylog factors in n (Hay et al., 2010; Ghazi et al., 2023; Feldman et al., 2024).

The result of Boedihardjo et al. (2024) introduces a "superregular random walk" that directly adds 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 into the empirically popular *Select, Measure, Project* framework for differentially private synthetic data generation Vietri et al. (2022); Liu et al. (2021); McKenna et al. (2022). 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 Dwork et al. (2006); McSherry and Mironov (2009), and then recover q matching these noisy moments. We verify the strong empirical performance of the method in Section A.1. A similar method was

5. Our result is for *approximate* (ϵ, δ)-DP instead of exact $(\epsilon, 0)$-DP. However, we obtain a very good $\sqrt{\log(1/\delta)}$ dependence on the approximation parameter δ. Thus, we can set $\delta = 1/\mathsf{poly}(n)$ and match the accuracy of Boedihardjo et al. (2024) up to constant factors. In our experience, approximate DP results where δ can be chosen to be a vanishingly small polynomial in n are considered alongside exact DP results.
6. 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 of a dataset supported on $[-1, 1]$ requires error $\Omega(1/(\epsilon n))$. See e.g., (Kamath, 2020), Theorem 3.

Figure 1: Experimental validation of Algorithm 2 for private synthetic data. For each dataset, we collect subsamples of size n for different 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-1 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 (the empirical variance across trials). The blue dotted line plots the theoretical bound of Theorem 4, without any leading constant. See Section A.1 for further details.



analyzed in Wang et al. (2016), although that work obtains a much weaker Wasserstein error bound of $\tilde{O}(1/(\varepsilon n^{1/3}))$. Theorem 1's tighter connection between Chebyshev moment estimation and distribution approximation allows us to obtain a significantly better dependence on n.

He et al. (2023) also claims a fast, simple alternative to Boedihardjo et al. (2024). While their simplest method achieves error $\tilde{O}(1/\sqrt{n})$, they describe a more complex method that matches our $\tilde{O}(1/n)$ result up to a $\log(n)$ factor. While we could not find their implementation, future work could empirically compare synthetic data generators with Wasserstein distance guarantees. We note that Feldman et al. (2024) recently studied *instance optimal* private distribution estimation in Wasserstein distance; it would be interesting to explore if moment matching had applications there.

Application 2: Matrix Spectral Density Estimation. Spectral density estimation (SDE) is a central problem in numerical linear algebra. In the standard version of the problem, we are given a symmetric $n \times n$ matrix A with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n \in \mathbb{R}$. The goal is to output a distribution q that is close in Wasserstein distance to the uniform distribution over these eigenvalues, p. An approximate spectral density can be useful in determining properties of A's 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, where they are used to compute the "density of states" of quantum systems (Skilling, 1989; Silver and Röder, 1994; Moldovan et al., 2020). More recently, the problem has found applications in network science (Dong et al., 2019; Cohen-Steiner et al., 2018; Jin et al., 2024), deep learning (Cun et al., 1991; Pennington et al., 2018; Mahoney and Martin, 2019; Yao et al., 2020), optimization (Ghorbani et al., 2019), and beyond (Li et al., 2019; Chen et al., 2022).

Many popular SDE algorithms are based on Chebyshev moment matching (Weiße et al., 2006; Bhattacharjee et al., 2025; Chen, 2024). The i^{th} Chebyshev moment of the spectral density is equal to $\mathbb{E}_{x\sim p} T_i(x) = \frac{1}{n}\sum_{j=1}^{n} T_i(\lambda_j) = \text{Tr}(\frac{1}{n}T_i(A))$. Stochastic trace estimation methods such as Hutchinson's method can estimate this trace using a small number of matrix-vector products with $T_i(A)$ (Hutchinson, 1990; Meyer et al., 2021). 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 access to the entries of A, but can efficiently multiply the matrix by vectors (Avron and Toledo, 2011).

Recently, Braverman, Krishnan and Musco (Braverman et al., 2022) gave a theoretical analysis of Chebyshev moment-matching for SDE, along with the related Kernel Polynomial Method (Weiße et al., 2006). They show that, when n is sufficiently large, specifically, $n = \tilde{\Omega}(1/\varepsilon^2)$, then $\tilde{O}(1/\varepsilon)$ matrix-vector products with A (and $\text{poly}(1/\varepsilon)$ additional runtime) suffice to output q with $W_1(p,q) \leq \varepsilon \|A\|_2$, where $\|A\|_2 = \max_i |\lambda_i|$ is A's spectral norm.

While the result of Braverman et al. (2022) also holds for smaller values of n, it suffers from a polynomially worse $1/\varepsilon$ dependence in the number of matrix-vector products required. By leveraging Theorem 1, we resolve this issue, showing that $\tilde{O}(1/\varepsilon)$ matrix-vector products suffice for *any* n. Roughly, by weakening the requirements on how well we approximate A's spectral moments, Theorem 1 lets us decrease the accuracy with which moments are estimated, and thus reduces the number of matrix-vector products used by Hutchinson's estimator. Formally, in Section B, we prove:

Theorem 5 *There is an algorithm that, given $\varepsilon \in (0,1)$, symmetric $A \in \mathbb{R}^{n\times n}$ with spectral density p, and upper bound [7] $S \geq \|A\|_2$, uses $\tilde{O}\left(\frac{1}{\varepsilon}\right)$ matrix-vector products[8] with A and $\tilde{O}(n/\epsilon + 1/\varepsilon^3)$ additional runtime to output distribution q such that, with high probability, $W_1(p,q) \leq \varepsilon S$.*

When A is dense, Theorem 5 yields an algorithm that runs in $\tilde{O}(n^2/\varepsilon + 1/\varepsilon^3)$ time, much faster than the $O(n^\omega)$ time required to compute p directly via eigendecomposition. In terms of matrix-vector products, the result cannot be improved by more than logarithmic factors. In particular, existing lower bounds for estimating the trace of a positive definite matrix (Meyer et al., 2021; Woodruff et al., 2022) imply that $\Omega(1/\varepsilon)$ matrix-vector products with A are necessary to approximate the spectral density p up to error $\epsilon\|A\|_2$ (see Section H). Thus, Theorem 5 essentially resolves the complexity of the SDE problem in the "matrix-vector query model" of computation, where cost is measured via matrix-vector products with A. This model has become central to theoretical work on numerical linear algebra, as it generalizes other important models like the matrix sketching and Krylov subspace models (Meyer et al., 2021; Sun et al., 2021; Woodruff et al., 2022). Our work contributes to recent progress on establishing tight upper and lower bounds for problems such as linear system solving (Braverman et al., 2020), eigenvector approximation (Musco and Musco, 2015; Simchowitz et al., 2018), trace estimation (Jiang et al., 2024), and more (Chewi et al., 2023; Bakshi and Narayanan, 2023; Amsel et al., 2024; Chen et al., 2025).

7. 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 (Kuczyński and Woźniakowski, 1992). In some settings, an upper bound on $\|A\|_2$ may be known apriori (Jin et al., 2024).

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

Application 3: Estimating Populations of Parameters. Our final application is to a classical statistical problem that has been studied since at least the 1960s (Lord, 1965, 1969; Wood, 1999):

Problem 6 (Population of Parameters Estimation) *Let p be an unknown distribution over $[0, 1]$. Consider a set of N independent coins, each with unknown bias p_i drawn from the distribution p. For each coin i, we observe the outcome of t independent coin tosses $X_i \sim$ Binomial(t, p_i). The goal is find a distribution q that is close to p in Wasserstein-1 distance.*

Problem 6 is motivated by settings (medicine, sports, etc.) where we want to estimate the distribution of a parameter over a large population of N individuals, but we only have noisy measurements of that parameter through a potentially much smaller number of observations, t, per individual. A simple approach is to compute empirical estimates for $p_1, \ldots, p_N$ based on $X_1, \ldots, X_N$ and to return the resulting distribution of biases. Doing so achieves error $O(1/\sqrt{t} + 1/\sqrt{N})$ in Wasserstein distance. Interestingly, Tian, Kong, and Valiant (Tian et al., 2017) show that in the "small sample" regime when N is large compared to t, it is possible to do much better. In particular, when $t = O(\log N)$, they introduce a moment-matching method with error $O(1/t)$.

More recently, Vinayak et al. (2019) analyze the maximum likelihood estimator (MLE) for p. The MLE, which we denote by $\hat{p}_{\text{mle}}$, has a relatively simple form and can be computed efficiently. They prove that it matches the error of (Tian et al., 2017) in the small sample regime. Moreover, in the *medium sample regime*, where $t = O(N^{2/9-\varepsilon})$ for any $\varepsilon > 0$, the MLE achieves error $O(1/\sqrt{t \log N})$, which is still an improvement on the empirical estimator. Formally, they prove the following in Theorem 3.2 of Vinayak et al. (2019):

Theorem 7 (Vinayak et al. (2019, Theorem 3.2)) *For any fixed constant $\varepsilon > 0$ and for any $t \in \left[\Omega(\log N), O\left(N^{2/9-\varepsilon}\right)\right]$, with probability $99/100$,*

$$W_1(p, \hat{p}_{mle}) \leq O\left(\frac{1}{\sqrt{t \log N}}\right). \tag{4}$$

We are able to tighten this result by directly applying our new global bound on the Chebyshev coefficients of Lipschitz functions (Lemma 13). In particular, in Section C, we show how to increase the range of t in Theorem 7 to $t \in \left[\Omega(\log N), O\left(N^{1/4-\varepsilon}\right)\right]$. Moreover, Vinayak et al. (2019) propose a simple conjecture that would improve their bound to $t \in \left[\Omega(\log N), O\left(N^{2/3-\varepsilon}\right)\right]$. Combining our improvement with their conjecture would allow for $t \in \left[\Omega(\log N), O(N^{1-\varepsilon})\right]$, which is essentially optimal, as even if $t = \infty$ (i.e., we have access to the true parameter p_i), one cannot achieve an error better than $O\left(1/\sqrt{N}\right)$ in Wasserstein distance (see Section C for more details).

1.4. Extension to higher dimensions

Finally, we note that we extend our main theorem (Theorem 1), to arbitrary dimension $d > 1$ in Section D. Doing so requires two ingredients: 1) a high-dimensional generalization of our global Chebyshev coefficient decay bound, and 2) a constructive proof of Jackson's theorem in $d > 1$ dimensions, which shows that a *damped* truncated multivariate Chebyshev series well-approximates any Lipschitz function. As an application, we give an algorithm for differentially private synthetic data generation in $d > 1$ dimensions in Section E, proving that we can obtain expected Wasserstein error $\tilde{\mathcal{O}}\left(1/(\varepsilon n)^{1/d}\right)$, which matches prior work up to logarithmic factors (Boedihardjo et al., 2024).

2. Preliminaries

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

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 functions, f, g, we use the notation $\langle f, g\rangle := \int_{-1}^1 f(x)g(x)\,dx$. 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, which is defined below. Note that we only consider distributions supported on $[-1,1]$, but the definition generalizes to any distribution on $\mathbb{R}$ or $\mathbb{R}^d$.

Definition 8 (Wasserstein-1 Distance) *Let p and q be two distributions on $[-1,1]$. Let $Z(p,q)$ be the set of all couplings between p and q, i.e., the set of distributions on $[-1,1]\times[-1,1]$ 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 known as the Kantorovich-Rubinstein Dual:

Fact 9 *Let p, q be as in Definition 8. 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[9] instead of the distributions themselves. We will do so throughout the paper.

In our analysis, it will be convenient to work with functions that are smooth, i.e. that are infinitely differentiable. Since any Lipschitz function can be arbitrarily well approximated by a smooth function, we have from Fact 9 that for distributions on $[-1,1]$[10]:

$$W_1(p,q) = \sup_{\text{1-Lipschitz, smooth } f} \langle f, p-q\rangle. \tag{5}$$

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 also 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 (see e.g. (Hale, 2015))

Fact 10 (Boundedness and Orthogonality) *The Chebyshev polynomials satisfy:*

9. p and q might correspond to discrete distributions, in which case they will be sums of Dirac delta functions.
10. Since $\|p-q\|_1 \leq 2$, if $\|f-\tilde{f}\|_\infty \leq \epsilon$ for some approximation $\tilde{f}$, then $\left|\langle f, p-q\rangle - \langle \tilde{f}, p-q\rangle\right| \leq 2\epsilon$. Since any Lipschitz function can be arbitrarily well-approximated by a smooth function in the ℓ_∞ norm, taking a sup over Lipschitz functions or smooth Lipschitz functions is therefore equivalent.

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:

Definition 11 (Normalized Chebyshev Polynomials) *The j^{th}* normalized *Chebyshev polynomial, $\bar{T}_j := T_j / \sqrt{\langle T_j \cdot w, T_j \rangle}$. 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 (Trefethen, 2019, 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 (Lighthill, 1958). Formally, any density p can be replaced with a Lipschitz continuous density (which has a convergent Chebyshev series) that is arbitrarily close in W_1 distance and the same analysis goes through.

3. Main Analysis

In this section, we prove our main result, Theorem 1, along with Corollary 2. We require two main ingredients. The first is a constructive version of Jackson's theorem on polynomial approximation of Lipschitz functions (Jackson, 1930), with a modern proof provided in Braverman et al. (2022).

Fact 12 (Jackson's Theorem (Jackson, 1930)) *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 $f_k = \sum_{j=0}^{k} b_k^j \cdot \left\langle f \cdot w, \bar{T}_j \right\rangle \cdot \bar{T}_j$ 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 (Trefethen, 2019). 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^j has a simple closed form (Braverman et al., 2022, Eq. 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 12 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 12. In particular, we prove the following result:

Lemma 13 (Global Chebyshev Coefficient Decay) *Let $f : [-1, 1] \to \mathbb{R}$ be an ℓ-Lipschitz, smooth function, and let $c_j := \left\langle f \cdot w, \bar{T}_j \right\rangle$ for $j \in \mathbb{Z}_{\geq 0}$. Then, $\sum_{j=1}^{\infty} (jc_j)^2 \leq \frac{\pi}{2}\ell^2$.*

Lemma 13 implies the well known fact that $c_j = O(\ell/j)$ for $j \geq 1$ (Trefethen, 2008). However, it is a much stronger bound: if all we knew was that the Chebyshev coefficients are bounded by $O(\ell/j)$, then $\sum_{j=1}^{\infty} (jc_j)^2$ could be unbounded, whereas we give a bound of $O(\ell^2)$. 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 13 in Section 3.3. Before doing so, we show how it implies Theorem 1.

Proof [Proof of Theorem 1] By (5), 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 12. We have:

$$\begin{aligned}\langle f, p-q\rangle = \langle f_k, p-q\rangle + \langle f-f_k, p-q\rangle &\le \langle f_k, p-q\rangle + \|f-f_k\|_\infty \|p-q\|_1 \\ &\le \langle f_k, p-q\rangle + \frac{36}{k}. \end{aligned} \tag{6}$$

In the last step, we use that $\|f-f_k\|_\infty \le 18/k$ by Fact 12, 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 $(p-q)/w$:

$$\frac{p-q}{w} = \sum_{j=0}^{\infty} \left\langle \frac{p-q}{w}\cdot w, \bar{T}_j \right\rangle \bar{T}_j = \sum_{j=0}^{\infty} \langle p-q, \bar{T}_j\rangle \cdot \bar{T}_j = \sum_{j=1}^{\infty} \langle p-q, \bar{T}_j\rangle \cdot \bar{T}_j. \tag{7}$$

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 12 that $f_k = \sum_{j=0}^{k} c_j' \bar{T}_j$, where each c_j' satisfies $|c_j'| \le |c_j|$ for $c_j := \langle f\cdot w, \bar{T}_j\rangle$. Using (7), the fact $\langle \bar{T}_i \cdot w, \bar{T}_j\rangle = 0$ whenever $i \ne j$, and $\langle \bar{T}_j \cdot w, \bar{T}_j\rangle = 1$ for all j, gives:

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

Via Cauchy-Schwarz inequality and our global decay bound from Lemma 13, we then have:

$$\begin{aligned}\langle f_k, p-q\rangle = \sum_{j=1}^{k} j c_j' \cdot \frac{\langle p-q, \bar{T}_j\rangle}{j} &\le \left(\sum_{j=1}^{k} (jc_j')^2\right)^{1/2} \cdot \left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2} \\ &\le \left(\sum_{j=1}^{k} (jc_j)^2\right)^{1/2} \cdot \left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2} \\ &\le \sqrt{\pi/2}\left(\sum_{j=1}^{k} \frac{1}{j^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2}. \end{aligned} \tag{8}$$

Observing from Definition 11 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, Equation (1), to upper bound Equation (8) by Γ.

Plugging this bound into Equation (6), we deduce the main bound of Theorem 1:

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

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 Chebyshev 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.

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 12 by 2. As discussed in Braverman et al. (2022, 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 Equation (3). If $|\mathbb{E}_{x\sim p} T_j(x) - \mathbb{E}_{x\sim q} T_j(x)| = |\langle p - q, \bar{T}_j\rangle|/\sqrt{\pi/2} \leq \Gamma \cdot \sqrt{\frac{j}{1+\log k}}$ for all j then we have that $\sum_{j=1}^{k} \frac{1}{j^2}\langle p - q, T_j\rangle^2 \leq \frac{\Gamma^2}{1+\log k}\sum_{j=1}^{k}\frac{1}{j} \leq \Gamma^2$. ■

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 given in Section F.

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 (Ye and Tse, 1989; Kapoor and Vaidya, 1986; Andersen et al., 2003). In practice, the problem can also be solved efficiently using first-order methods like projected gradient descent (Wright and Recht, 2022).

3.3. Proof of Lemma 13

We conclude this section by proving Lemma 13, 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., Rivlin (1969)).

Fact 14 (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. For $j \geq 1$, $T_j'(x) = jU_{j-1}(x)$.*

Fact 15 (Orthogonality of Chebyshev polynomials of the second kind) *The Chebyshev polynomials of the second kind are orthogonal with respect to the weight function* $u(x) = \sqrt{1-x^2}$*, i.e.,*

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

With the above facts we can now prove Lemma 13.

Proof [Proof of Lemma 13] Let f be a smooth, ℓ-Lipschitz function, with Chebyshev expansion $f(x) = \sum_{j=0}^{\infty} c_j \bar{T}_j = \frac{1}{\sqrt{\pi}} c_0 T_0 + \sum_{j=1}^{\infty} \sqrt{\frac{2}{\pi}} c_j T_j$. Using Fact 14, we can write f's derivative as:

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

By the orthogonality property of Fact 15, we then have that

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

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

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

This completes the proof of the lemma. We remark that this bound cannot be improved, as it holds with equality for the function $f(x) = x$. ■

Acknowledgments

We thank Gregory Valiant for suggesting us the work on populations of parameters. 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

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.

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.

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.

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.

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

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.

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.

Rajarshi Bhattacharjee, Rajesh Jayaram, Cameron Musco, Christopher Musco, and Archan Ray. mproved spectral density estimation via explicit and implicit deflation. In *Proceedings of the 36th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2025.

Joseph R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Private measures, random walks, and synthetic data. *Probability Theory and Related Fields*, 189(1):569–611, 2024.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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.

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.

Tyler Chen. *Lanczos-based methods for matrix functions*. PhD thesis, University of Washington, 2022.

Tyler Chen. A spectrum adaptive kernel polynomial method. *The Journal of Chemical Physics*, 159 (11), 2023.

Tyler Chen. The Lanczos algorithm for matrix functions: a handbook for scientists. *arXiv:2410.11090*, 2024.

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.

Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. Randomized matrix-free quadrature for spectrum and spectral sum approximation. *arXiv:2204.01941*, 2022.

Tyler Chen, Feyza Duman Keles, Diana Halikias, Cameron Musco, and Christopher Musco. Near-optimal hierarchical matrix approximation from matrix-vector products. *Proceedings of the 36th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2025.

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.

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.

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.

Yann Le Cun, Ido Kanter, and Sara A. Solla. Eigenvalues of covariance matrices: Application to neural-network learning. *Phys. Rev. Lett.*, 66:2396–2399, 5 1991.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

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.

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.

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.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

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.

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.

Vitaly Feldman, Audra McMillan, Satchit Sivakumar, and Kunal Talwar. Instance-optimal private density estimation in the wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 37, pages 90061–90131, 2024.

Hans Fischer. *Chebyshev's and Markov's Contributions*, pages 139–189. Springer New York, 2011.

Badih Ghazi, Junfeng He, Kai Kohlhoff, Ravi Kumar, Pasin Manurangsi, Vidhya Navalpakkam, and Nachiappan Valliappan. Differentially private heatmaps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7696–7704, 2023.

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.

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.

Nicholas Hale. *Chebyshev Polynomials*, pages 203–205. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-540-70529-1.

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.

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.

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.

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.

Dunham Jackson. On approximation by trigonometric sums and polynomials. *Transactions of the American Mathematical Society*, 13(4):491–515, 1912.

Dunham Jackson. *The Theory of Approximation*, volume 11 of *Colloquium Publications*. American Mathematical Society, 1930.

Zhanglong Ji and Charles Lipton, Zachary C.and Elkan. Differential privacy and machine learning: a survey and review. *arXiv:1412.7584*, 2014.

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.

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.

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.

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.

Gautam Kamath. Cs 860: Algorithms for private data analysis, lecture 11 – packing lower bounds, 2020. URL http://www.gautamkamath.com/CS860notes/lec11.pdf.

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.

Richard M. Karp and Robert Kleinberg. Noisy binary search and its applications. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 881–890, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

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].

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.

Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45 (5):2218–2247, 10 2017.

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.

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.

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.

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.

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.

Michael J. Lighthill. *An introduction to Fourier analysis and generalised functions*. Cambridge University Press, 1958.

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.

Frederic M. Lord. A strong true-score theory, with applications. *Psychometrika*, 30(3):239–270, 1965.

Frederic M. Lord. Estimating true-score distributions in psychological testing (an empirical bayes estimation problem). *Psychometrika*, 34(3):259–299, 1969.

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.

J.C. Mason. Near-best multivariate approximation by fourier series, chebyshev series and chebyshev interpolation. *Journal of Approximation Theory*, 28(4):349–358, 1980. ISSN 0021-9045.

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.

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.

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.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey. *arXiv:2004.12254*, 2020.

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.

Dean Moldovan, Misa Andelkovic, and Francois Peeters. pybinding v0.9.5: a Python package for tight-binding calculations, 2020.

MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019. URL http://docs.mosek.com/9.0/toolbox/index.html.

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.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, 185:71–110, 1894.

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

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.

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

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.

Theodore J. Rivlin. *An introduction to the approximation of functions*. Dover Publications, 1969.

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.

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.

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.

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.

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.

John Skilling. *The Eigenvalues of Mega-dimensional Matrices*, pages 455–466. Springer Netherlands, 1989.

Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.

Xiaoming Sun, David P. Woodruff, Guang Yang, and Jialin Zhang. Querying a matrix through matrix-vector products. *ACM Trans. Algorithms*, 17(4), 2021. ISSN 1549-6325.

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

Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Lloyd N. Trefethen. Is Gauss Quadrature Better than Clenshaw–Curtis? *SIAM Review*, 50(1): 67–87, 2008.

Lloyd N. Trefethen. *Approximation Theory and Approximation Practice, Extended Edition*. SIAM-Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2019. ISBN 161197593X.

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.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

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.

Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. *Rev. Mod. Phys.*, 78:275–306, 2006.

G. R. Wood. Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics*, 27(5):1706 – 1721, 1999.

David Woodruff, Fred Zhang, and Richard Zhang. Optimal query complexities for dynamic trace estimation. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Stephen J. Wright and Benjamin Recht. *Optimization for Data Analysis*. Cambridge University Press, 2022.

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

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

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

Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal*, 22:797–822, 2013.

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.

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

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*, page 155–170, 2016.

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.

Appendix A. Private Synthetic Data Generation

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 (Kifer and Machanavajjhala, 2011). 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.[11]

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 16 (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:*

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

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 17 (Gaussian Mechanism (Dwork et al., 2006; Dwork and Roth, 2014)) *Let $f : \mathcal{X} \to \mathbb{R}^k$ be a function with ℓ_2-sensitivity $\Delta_{2,f}$ and let $\sigma^2 = \Delta_{2,f}^2 \cdot 2\ln(1.25/\delta)/\varepsilon^2$, where $\varepsilon, \delta \in (0, 1)$ 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 [Proof of Theorem 4] We analyze both the privacy and accuracy of Algorithm 2.

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

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

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

4: Set $k = \lceil 2\varepsilon n \rceil$.[12] 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.

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

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

By Fact 10, $\max_{x_i \in [-1,1]} |\bar{T}_j(x_i)| \leq \sqrt{2/\pi}$ 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 17 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 $\sigma^2 = \frac{16}{\pi}(1 + \log k)\ln(1.25/\delta)/(n^2\varepsilon^2)$. 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.

12. While we choose $k = \lceil 2\varepsilon n \rceil$ by default, any choice of $k = \lceil c\varepsilon n \rceil$ 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 defaults so that any

Accuracy. Algorithm 2 begins by rounding the dataset X so that every data point is a multiple of $1/\lceil \varepsilon n \rceil$. Let $\tilde{p}$ be the uniform distribution over the rounded dataset $\tilde{X}$. Using the transportation definition of the Wasserstein-1 distance, we obtain the bound:

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

In particular, we can transport p to $\tilde{p}$ by moving every unit of $1/n$ probability mass a distance of at most $1/2\lceil \varepsilon n \rceil$. 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 $W_1(p, q)$.

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:

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

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:

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

By triangle inequality and (10), $W_1(p, q) \leq W_1(\tilde{p}, q) + W_1(\tilde{p}, p) \leq W_1(\tilde{p}, q) + \frac{1}{2\lceil \varepsilon n \rceil}$. 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 $k = \lceil 2\varepsilon n \rceil$ 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 Section G 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 $r = 2\lceil \varepsilon n \rceil + 1$ variables and $r + 1$ constraints, which can be done to high accuracy in $\mathsf{poly}(\varepsilon n)$ time. ■

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.

A.1. Empirical Evaluation for Private Synthetic Data

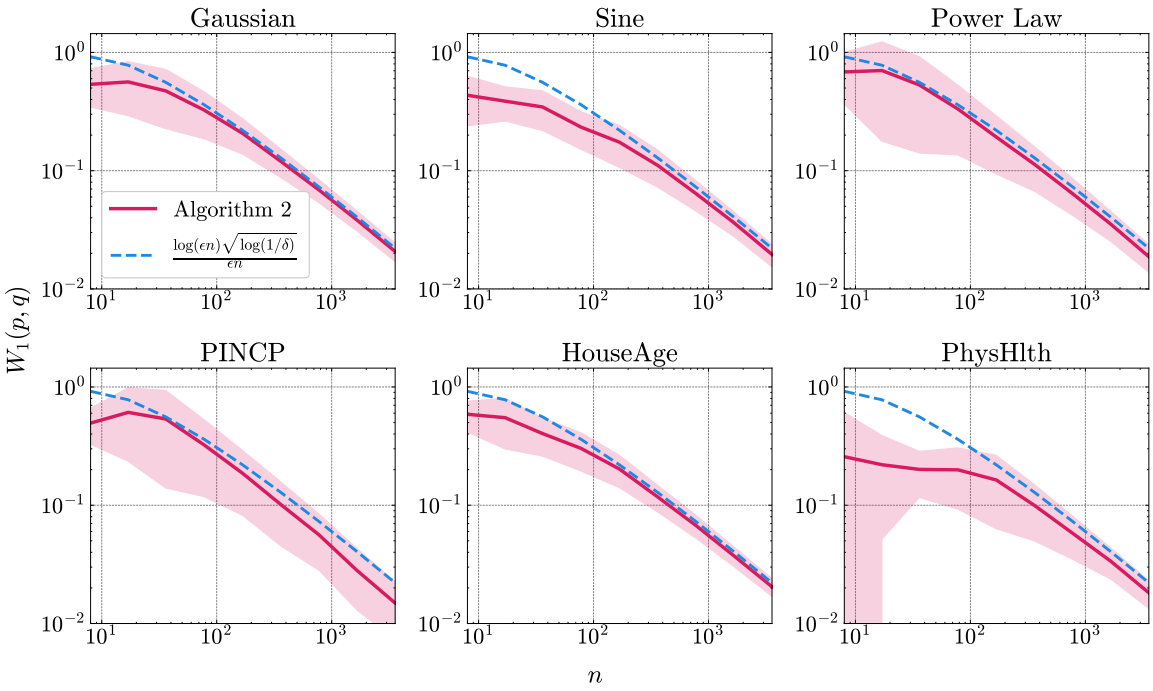


Figure 2: Experimental validation of Algorithm 2 for private synthetic data, as shown in Figure 1 in the main paper body. For each dataset, we collect subsamples of size n for different n. We plot the W_1 distance between the uniform distribution, p, over the subsample and a differentially private approximation, 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 (the empirical variance across trials). The blue dotted line plots the theoretical bound of Theorem 4, without any leading constant.

In this section, we give more details on our empirical evaluation of the application of our main result to differentially private synthetic data generation, as presented in Section A (and noted in Figure 1 in the main paper body). 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 (Diamond and Boyd, 2016; MOSEK ApS, 2019; Andersen et al., 2003). 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 (Ding et al., 2021). We use the 2018 ACS 1-Year data for the state of New York; we give results for the `PINCP` (personal income) column from this data. We also consider the California Housing dataset (Pace and Barry, 1997); we give results for the `HouseAge` (median

house age in district) column, from this data. Finally, we consider the CDC Diabetes Health Indicators dataset (Teboul, 2021; Kelly et al., 2024); 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 $\varepsilon = 0.5$ and $\delta = 1/n^2$; this is a standard setting for private synthetic data (McKenna et al., 2022; Rosenblatt et al., 2023). We use the default choice of $k = \lceil 2\varepsilon n \rceil$. In Figure 2 (Figure 1 in the main paper body), 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 choice 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 2.

Appendix B. 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

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

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 (Braverman et al., 2022). 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 Hutchinson's method can approximate $\mathrm{Tr}(T_j(A))$ efficiently via multiplication of $T_j(A)$ with random vectors (Girard, 1987; Hutchinson, 1990). 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 (Chen, 2023). Using a standard analysis of Hutchinson's trace estimator (see, e.g., (Roosta-Khorasani and Ascher, 2015) or (Cortinovis and Kressner, 2022)) Braverman et al. (2022) prove the following:

Lemma 18 ((Braverman et al., 2022, Lemma 4.2)) *Let A be a matrix with $\|A\|_2 \leq 1$. 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 $g_1, \ldots, g_{\ell_j} \sim \text{Uniform}(\{-1,1\}^n)$ 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 $1-\alpha$, $|\hat{m}_j - m_j| \leq \sqrt{j}\gamma$.*

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

Theorem 19 *There is an algorithm that, given $\varepsilon \in (0,1)$, symmetric $A \in \mathbb{R}^{n \times n}$ with spectral density p, and upper bound $S \geq \|A\|_2$, uses $\min\left\{n, O\left(\frac{1}{\varepsilon} \cdot \left(1 + \frac{\log^2(1/\varepsilon)\log^2(1/(\varepsilon\delta))}{n\varepsilon}\right)\right)\right\}$ matrix-vector products with A and $\tilde{O}(n/\epsilon + 1/\varepsilon^3)$ additional time to output a distribution q such that, with probability at least $1-\delta$, $W_1(p,q) \leq \varepsilon S$.*

Proof First note that, if $\epsilon \leq 1/n$, the above result can be obtained by simply recovering A by multiplying by all $n \leq 1/\epsilon$ 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 $\|A\|_2 \leq 1$ and our goal is to prove that $W_1(p,q) \leq \epsilon$. 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 $\|A\|_2 \leq S \leq 2\|A\|_2$ can be computed using $O(\log n)$ matrix-multiplications with A via the power method (Kuczyński and Woźniakowski, 1992). Given such an S, Theorem 19 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 (Cohen-Steiner et al., 2018; Dong et al., 2019; Jin et al., 2024), $\|A\|_2$ is known apriori, so we can simply set $S = \|A\|_2$.

Choose $k = \hat{c}/\varepsilon$ for a sufficiently large constant $\hat{c}$ and apply Lemma 18 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,

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

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 $W_1(p,q) \leq \varepsilon$ 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 $k = O(1/\varepsilon)$.

In terms of computational complexity, Corollary 2 immediately yields a bound of $\mathsf{poly}(1/\varepsilon)$ time to solve the quadratic program in Algorithm 1. However, this runtime can actually be improved to $\tilde{\mathcal{O}}(1/\varepsilon^3)$ 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 $q_1^{\mathrm{LP}}, \ldots, q_g^{\mathrm{LP}}$ be any solution to the following linear program with variables $z_1, \ldots, z_g$:

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

We first verify that the linear program has a solution. To do so, note that, by Equation (37) in Section F, 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 $q^{\mathrm{LP}} = \sum_{i=1}^{g} q_i^{\mathrm{LP}} \delta(x - x_i)$ be the distribution formed by any solution to the linear program. We have that, for any j,

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

Setting $g = k^{1.5}\sqrt{1 + \log(k)}$ and plugging into Theorem 1, we conclude that $W_1(p, q^{\mathrm{LP}}) \leq O(1/k)$.

The linear program in Equation (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 (Lee and Sidford, 2014, 2015), which equals $\tilde{O}(1/\varepsilon^3)$ time since we chose $k = O(1/\epsilon)$. ■

Appendix C. Learning Populations of Parameters

In this section, we present the final application of our results to the "population of parameters problem" introduced as Problem 6 in Section 1.3. Unlike our prior two applications to differentially private synthetic data and spectral density estimation, we obtain an improvement on the prior work by applying the global Chebyshev coefficient decay bound from Lemma 13 directly, instead of applying the full moment matching bound from Theorem 1. We recall the problem statement below:

Problem 6 (Population of Parameters Estimation) *Let p be an unknown distribution over $[0,1]$. Consider a set of N independent coins, each with unknown bias p_i drawn from the distribution p. For each coin i, we observe the outcome of t independent coin tosses $X_i \sim$ Binomial(t, p_i). The goal is find a distribution q that is close to p in Wasserstein-1 distance.*

(Vinayak et al., 2019) shows that the maximum likelihood estimator (MLE) of p can be formulated as:

$$\hat{p}_{\text{mle}} \in \operatorname*{argmax}_{Q \in \mathcal{D}} \sum_{i=1}^{N} \log \int_0^1 \binom{t}{X_i} y^{X_i}(1-y)^{t-X_i} \mathsf{d}Q(y)\,, \tag{16}$$

where $\mathcal{D}$ denotes the set of all distributions on $[0,1]$. They prove that, in the *small sample regime*, when $t = O(\log N)$, the MLE obtains error $W_1(p, \hat{p}_{\text{mle}}) \le O(1/t)$. This improves on the naive estimator that simply returns a uniform distribution based on empirical estimates of $p_1, \ldots, p_N$, which gives Wasserstein error $O(1/\sqrt{t} + 1/\sqrt{N})$. Moreover, they prove that it is also possible to beat the naive estimator in the *medium sample regime*:

Theorem 7 (Vinayak et al. (2019, Theorem 3.2)) *For any fixed constant $\varepsilon > 0$ and for any $t \in \left[\Omega(\log N), O\left(N^{2/9-\varepsilon}\right)\right]$, with probability $99/100$,*

$$W_1(p, \hat{p}_{mle}) \le O\left(\frac{1}{\sqrt{t \log N}}\right). \tag{4}$$

We improve this result in the medium sample regime to hold for a wider range of t, showing:

Theorem 20 (Improvement in the Medium Sample Regime) *There exists an $\varepsilon > 0$, such that, for $t \in \left[\Omega(\log N), O\left(N^{1/4-\varepsilon}\right)\right]$, with probability at least $99/100$,*

$$W_1(p, \hat{p}_{mle}) \le O\left(\frac{1}{\sqrt{t \log N}}\right). \tag{17}$$

As will be discussed in Section C.2, under a natural conjecture from Vinayak et al. (2019), our approach can actually be used to extend the range for which (17) holds all the way to $t = O(N^{1-\epsilon})$ for any fixed constant ϵ, which is essentially optimal.

Notation. We begin by introducing notation used throughout this section. Unlike prior applications, Problem 6 involves distributions over $[0,1]$ instead of $[-1,1]$. For this reason, we use *shifted* Chebyshev polynomials, which we denote by $\tilde{T}_0(x), \tilde{T}_1(x), \ldots$, where the degree m shifted Chebyshev polynomial, $\tilde{T}_m$, is defined as $\tilde{T}_m(x) = T_m(2x-1)$. Note that the shifted Chebyshev polynomials are orthogonal on the range $[0,1]$ under weight function $w(2x-1)$, where $w(x) = \frac{1}{\sqrt{1-x^2}}$ is as defined in Fact 10. Also note that Jackson's theorem (Fact 12) and our global Chebyshev coefficient decay bound (Lemma 13) continue to hold up to small changes in constant factors when working with shifted Chebyshev polynomial expansions of Lipschitz functions on $[0,1]$.

C.1. Proof of Theorem 20

The approach from (Vinayak et al., 2019) centers on rewriting $\hat{p}_{\text{mle}}$ in terms of the *fingerprint* of the observed coin tosses, which can be shown to be a sufficient statistic for the estimation problem. Recall that the observations are $\{X_i\}_{i=1}^N$, where $X_i \sim \text{Binomial}(t, p_i)$. For $s \in \{0, 1, \ldots, t\}$, let n_s denote the number of coins that evaluate to 1 on s of the t tosses, i.e. $n_s = |\{i : X_i = s\}|$. Let h_s^{obs} denote the fraction of coins that evaluate to 1 on s tosses, i.e., $h_s^{\text{obs}} = n_s/N$. The fingerprint is defined as $\boldsymbol{h}^{\text{obs}} := (h_0^{\text{obs}}, \ldots, h_t^{\text{obs}})$. Similarly, for any distribution Q, let $\mathbb{E}_Q[h_j]$ denote the expected fraction of coins that evaluate to 1 on j out of t tosses when $X_1, \ldots, X_N$ are drawn from some distribution Q.

Vinayak et al. (2019) prove a result relating the Wasserstein error of $\hat{p}_{mle}$ to how closely the expected fingerprints under $\hat{p}_{mle}$ match the observed fingerprints. Like our Theorem 1 on matching moments, this result leverages the dual definition of Wasserstein distance involving Lipschitz functions (Fact 9) and proceeds by replacing f with a polynomial approximation, $\hat{f}$. Just as our proof depends on the Chebyshev coefficients of $\hat{f}$, their result depends on the coefficients of $\hat{f}$ when written in a *Bernstein polynomial basis*. In particular, let $B_j^t(x) = \binom{t}{j} x^j (1-x)^{t-j}$ denote the j^{th} Bernstein polynomial of degree t. Vinayak et al. (2019) works with a degree t approximation $\hat{f}$ of the form $\hat{f} = \sum_{j=0}^t b_j B_j^t(x)$. They prove the following:

$$W_1(p, \hat{p}_{\text{mle}}) \leq \sup_{\substack{\text{1-Lipschitz,} \\ \text{smooth } f}} \left[\inf_{\hat{f} = \sum_{j=0}^t b_j B_j^t(x)} \left[\underbrace{2 \left\| f - \hat{f} \right\|_\infty}_{\text{(a)}} + \underbrace{\sum_{j=0}^t b_j (\mathop{\mathbb{E}}_{p}[h_j] - h_j^{\text{obs}})}_{\text{(b)}} + \underbrace{\sum_{j=0}^t b_j (h_j^{\text{obs}} - \mathop{\mathbb{E}}_{\hat{p}_{\text{mle}}}[h_j])}_{\text{(c)}} \right] \right]. \tag{18}$$

Above and in the remainder of this section, $\left\| f - \hat{f} \right\|_\infty$ denotes $\max_{x \in [0,1]} |f(x) - \hat{f}(x)|$ (instead of our usual definition involving $x \in [-1, 1]$.) (Vinayak et al., 2019) bounds the terms (b) and (c) as follows:

Lemma 21 ((Vinayak et al., 2019, Lemmas 4.1 and 4.2)) *Term (b): With probability* $1 - \delta$,

$$\left| \sum_{j=0}^t b_j (\mathop{\mathbb{E}}_{p}[h_j] - h_j^{obs}) \right| \leq O\left(\max_j |b_j| \sqrt{\frac{\log 1/\delta}{N}} \right).$$

Term (c): For $3 \leq t \leq \sqrt{C_0 N} + 2$, *where* $C_0 > 0$ *is constant, with probability* $1 - \delta$,

$$\left| \sum_{j=0}^t b_j (h_j^{obs} - \mathop{\mathbb{E}}_{\hat{p}_{mle}}[h_j]) \right| \leq \max_j |b_j| \sqrt{2 \ln 2} \sqrt{\frac{t}{2N} \log \frac{4N}{t} + \frac{1}{N} \log \frac{3e}{\delta}}.$$

It remains to bound (a), i.e., $\left\| f - \hat{f} \right\|_\infty$, as well as $\max_j |b_j|$ which appears in both bounds above.

Doing so requires proving that there exist good uniform polynomial approximations to f that have bounded coefficients $b_0, \ldots, b_t$ in the Bernstein polynomial basis. Towards that end, Vinayak et al. (2019) prove the following key result:

Proposition 22 ((Vinayak et al., 2019, Proposition 4.2)) *Any 1-Lipschitz function on $[0,1]$ can be approximated by a degree t polynomial $\hat{f}(x) = \sum_{j=0}^{t} b_j B_j^t(x)$, such that, for any $k < t$,*

$$\left\|f - \hat{f}\right\|_\infty \leq O\left(\frac{1}{k}\right) \qquad \textit{and} \qquad \max_j |b_j| \leq \sqrt{k}(t+1)e^{k^2/t}.$$

Proposition 4.2 is proven by using Jackson's theorem (Fact 12) to approximate f by a degree k polynomial f_k. Recall that f_k is written as a linear combination of Chebyshev polynomials. (Vinayak et al., 2019) then obtain $\hat{f}$ by expressing these Chebyshev polynomials as linear combinations of Bernstein polynomials of degree t. Naturally, by using our Lemma 13 to give a better bound on the Chebyshev coefficients of f_k, we can improve their bound on the Bernstein polynomial coefficients, $b_0, \ldots, b_t$, of $\hat{f}$. Concretely, we show the following:

Proposition 23 (Improvement to (Vinayak et al., 2019, Proposition 4.2)) *Any 1-Lipschitz function on $[0,1]$ can be approximated by a degree t polynomial $\hat{f}(x) = \sum_{j=0}^{t} b_j B_j^t(x)$, such that, for any $k < t$,*

$$\left\|f - \hat{f}\right\|_\infty \leq O\left(\frac{1}{k}\right) \qquad \textit{and} \qquad \max_j |b_j| \leq (t+1)e^{k^2/t}.$$

Proof Let $f_k = \sum_{m=0}^{k} a_m \tilde{T}_m(x)$ be the damped truncated Chebyshev series of f as defined in Fact 12 (appropriately shifted and scaled to involve the shifted Chebyshev polynomials over $[0,1]$) [13]. Recall that, for all i, $a_i \leq 1$. From Fact 12, we have that $\|f - f_k\| \leq O(1/k)$. Any Chebyshev polynomials $\tilde{T}_m$ can be expressed as a linear combination of Bernstein polynomials of degree m:

$$\tilde{T}_m(x) = \sum_{i=0}^{m} (-1)^{m-i} \frac{\binom{2m}{2i}}{\binom{m}{i}} B_i^m(x) \qquad \text{((Vinayak et al., 2019, eq. 21))}\ .$$

Moreover, following (Vinayak et al., 2019), we can express any degree m Bernstein polynomial as an appropriate sum of Bernstein polynomial of higher degree t:

$$B_i^m(x) = \sum_{j=i}^{i+t-m} \frac{\binom{m}{i}\binom{t-m}{j-i}}{\binom{t}{j}} B_j^t(x) \qquad \text{((Vinayak et al., 2019, eq. 22))}\ .$$

Combining the two equations above, we have that, for $m < t$,

$$\tilde{T}_m(x) = \sum_{i=0}^{m} (-1)^{m-i} \frac{\binom{2m}{2i}}{\binom{m}{i}} \sum_{j=i}^{i+t-m} \frac{\binom{m}{i}\binom{t-m}{j-i}}{\binom{t}{j}} B_j^t(x) =: \sum_{j=0}^{t} C(t,m,j) B_j^t(x)\,,$$

see Vinayak et al. (2019, eq. 23). The Lemma 4.4 of Vinayak et al. (2019) then gives us that

$$|C(t,m,j)| \leq (t+1)e^{m^2/t}\,. \tag{19}$$

Next, we choose $\hat{f}$ to be:

$$\hat{f} = \sum_{j=0}^{t} b_j B_j^t(x) := \sum_{m=0}^{k} a_m \left(\sum_{j=0}^{t} C(t,m,j) B_j^t(x) \right) = \sum_{m=0}^{k} a_m \tilde{T}_m(x) = f_k.$$

13. We use $a_0, \ldots, a_k$ to denote the damped coefficients to avoid confusion with the coefficients $b_0, \ldots, b_t$ above

Above, $b_j = \sum_{m=0}^{k} a_m C(t,m,j)$. Using the fact that $|C(t,0,j)| \leq (t+1)$ alongside our global Chebyshev coefficient decay bound from Lemma 13 we can bound each coefficients b_j as follows:

$$
\begin{aligned}
|b_j| &\leq |a_0 C(t,0,j)| + \left|\sum_{m=1}^{k} a_m C(t,m,j)\right| \\
&= |a_0 C(t,0,j)| + \left|\sum_{m=1}^{k} m a_m \frac{C(t,m,j)}{m}\right| \\
&\leq (t+1) + \left(\sum_{m=1}^{k} m^2 a_m^2\right)^{1/2} \cdot \left(\sum_{m=1}^{k} \frac{C(t,m,j)^2}{m^2}\right)^{1/2} && (20) \\
&\leq (t+1) + \left(\sum_{m=1}^{k} m^2 a_m^2\right)^{1/2} \cdot \left((t+1)^2 \sum_{m=1}^{k} \frac{e^{2m^2/t}}{m^2}\right)^{1/2} && (21) \\
&\leq (t+1) + \left(\sum_{m=1}^{k} m^2 a_m^2\right)^{1/2} \cdot \left((t+1)^2 e^{2k^2/t} \sum_{m=1}^{k} \frac{1}{m^2}\right)^{1/2} \\
&\leq (t+1) + \left(\sum_{m=1}^{k} m^2 a_m^2\right)^{1/2} \cdot (t+1)\left(e^{k^2/t}\sqrt{\frac{\pi^2}{6}}\right) && (22) \\
&\leq (t+1) + C_1' \cdot (t+1)\left(e^{k^2/t}\sqrt{\frac{\pi^2}{6}}\right) \leq C_1 (t+1) e^{k^2/t}, && (23)
\end{aligned}
$$

where $C_1, C_1' > 0$ are some absolute constants. Since f is a 1-Lipschitz function on $[0,1]$, we can let $|f(x)| \leq 1/2$, as we can shift f can such that its range is bounded between $[-1/2, 1/2]$. It can be checked that this implies that $|a_0| \leq 1$. Equation (20) follows by combining the fact that $|a_0| \leq 1$ with the bound on $|C(t,0,j)|$, and Cauchy Schwarz inequality. Equation (21) follows from the bound on $|C(t,m,j)|$ in Equation (19). Equation (22) follows from the fact that $\sum_{m=1}^{\infty} 1/m^2 \leq \pi^2/6$. Let $\sum_{m=0}^{\infty} c_m \tilde{T}_m(x)$ be the shifted Chebyshev series of f. Then, we know from Fact 12 that $|a_m| \leq |c_m|$, and from the global Chebyshev coefficient decay Lemma 13 that $\sum_{m=1}^{\infty} m^2 c_m^2 \leq C_1'$, for some constant C_1'. This proves the first inequality of Equation (23).

Equation (23) gives us the bound on the coefficients $|b_j|$. Using the fact that $\left\|f - \hat{f}_k\right\|_\infty \leq O(1/k)$, the proposition follows. ■

We are now ready to prove our main theorem from this section.

Proof [Proof of Theorem 20] For a 1-Lipschitz function f, let $\hat{f}(x) = \sum_{j=0}^{t} b_j B_j^t(x)$ denote a degree t Bernstein polynomial approximation to f. We use Equation (18) to bound $W_1(p, \hat{p}_{\text{mle}})$. Specifically, by Proposition 23, there is a choice of $\hat{f}(x)$ which ensures that the (a) term can be bounded by $2\left\|f - \hat{f}\right\|_\infty \leq O(1/k)$. Moreover, we will have that $\max_j |b_j| \leq (t+1)e^{k^2/t}$, for $k < t$. We will set $k = \sqrt{t \log(N^c)}$ for a small constant $c > 0$ to be chosen later. Note that since we require $k < t$ for Proposition 23 to hold, doing so requires $t = \Omega(\log N)$. With this choice of k, we have that $\max_j |b_j| \leq (t+1)N^c$. We can then plug this coefficient bound into Lemma 21 to

show that, with probability $99/100$, and $3 \le t \le \sqrt{C_0 N} + 2$,

$$\begin{aligned} W_1(p, \hat{p}_{\text{mle}}) &\le O\left(\frac{1}{k}\right) + O\left(\max_j |b_j| \sqrt{\frac{1}{N}}\right) + O\left(\max_j |b_j| \sqrt{\frac{t}{2N} \log \frac{4N}{t} + \frac{1}{N}}\right) \\ &\le O\left(\frac{1}{\sqrt{t \log N}}\right) + O\left((t+1) N^c \sqrt{\frac{t}{N} \log N}\right) . \\ &= O\left(\frac{1}{\sqrt{t \log N}}\right) + O\left(\sqrt{\frac{t^3}{N^{1-2c}} \log N}\right) . \end{aligned}$$

For any target constant ϵ, we can choose our constant c so that $N^\epsilon = N^c \log N$. We can then check that, as long as $t = O\left(N^{1/4-\epsilon}\right)$, $O\left(\sqrt{\frac{t^3}{N^{1-2c}} \log N}\right) = O\left(\sqrt{\frac{1}{t \log N}}\right)$, which proves the theorem. ■

C.2. Conjectured Improvement

Vinayak et al. (2019) conjecture that the range of t for which Theorem 7 holds can be improved. In particular, they conjecture that the bound on the coefficients in the proof of Proposition 23 can be improved to $|C(t, m, j)| \le e^{m^2/t}$ for $j = 0, \ldots, t$. Moreover, they conjecture that the bound on (c) in Equation (18) can be improved to $O\left(\max_j |b_j| \sqrt{\log(1/\delta)/N}\right)$. If these conjectures hold, the range of t can be improved to $t \in \left[\Omega(\log N), O\left(N^{2/3-\varepsilon}\right)\right]$. If we additionally include our improvement from Proposition 23, we would obtain a further improvement to $t \in \left[\Omega(\log N), O\left(N^{1-\varepsilon}\right)\right]$.

We note that improving the upper limit on t to $O\left(N^{1-\varepsilon}\right)$ is essentially the best that we can hope for. In particular, there exist distributions that are $1/\sqrt{N}$ far away in W_1 distance that would need N independent coins to distinguish between them, even if $t = \infty$. Consider two distributions with q_1 and q_2 such that q_1 has probability mass of $1/2 + 1/\sqrt{N}$ on 0 and $1/2 - 1/\sqrt{N}$ on 1, and q_2 has probability mass of $1/2 - 1/\sqrt{N}$ on 0 and $1/2 + 1/\sqrt{N}$ on 1. It is easy to see that $W_1(q_1, q_2) = 2/\sqrt{N}$. The coins drawn from q_1 or q_2 have biases of either 0 or 1. So, in this case, a single coin toss does not provide any less information than infinite coin tosses. By standard information-theoretic arguments (Karp and Kleinberg, 2007), $\Omega(1/N)$ independent samples are required to distinguish between q_1 and q_2 with probability greater than $1/2$. Accordingly, when $t = \Omega(N)$, we can no longer achieve error better than the $O(1/\sqrt{t} + 1/\sqrt{N})$ bound given by the naive estimator.

Appendix D. Multivariate Generalization of Theorem 1

In this section, we generalize our Theorem 1 to d-dimensions. To prove this, we look at the Chebyshev series of multivariate functions. The Wasserstein-1 distance and its dual is analogously defined in d-dimensions.

Definition 24 (Wasserstein-1 Distance, Euclidean Metric) *Let p and q be two distributions on $[-1, 1]^d$. Let $Z(p, q)$ be the set of all couplings between p and q, i.e., the set of distributions on*

$[-1,1]^d \times [-1,1]^d$ *whose marginals equal* p *and* q*. The Wasserstein-*1 *distance between* p *and* q *is:*

$$W_1(p,q) = \inf_{z \in Z(p,q)} \left[\mathop{\mathbb{E}}_{(x,y)\sim z} \|x-y\|_2 \right],$$

where $\|x-y\|_2$ *denotes the Euclidean distance.*

Like W_1 in 1-dimension, the Wasserstein distance in d-dimension also measures the total cost (in terms of distance per unit mass) required to "transport" the distribution p to q. Its dual form is as follows.

Fact 25 (Kantorovich-Rubinstein Duality in d-Dimensions) *Let* p, q *be as in Definition 24. Then,*

$$W_1(p,q) = \sup_{1\text{-Lipschitz, smooth } f} \int_{[-1,1]^d} f(x) \cdot (p(x) - q(x)) \mathsf{d}x \,,$$

where $f : [-1,1]^d \to \mathbb{R}$ *is a smooth,* 1*-Lipschitz function under the Euclidean metric.*

D.1. Multivariate Chebyshev Series

We use the fact that if $f : [-1,1]^d \to \mathbb{R}$ is smooth, it has a uniformly and absolutely convergent multivariate Chebyshev series (Mason, 1980)

$$f(x) = \sum_{K \in \mathbb{Z}^d_{\geq 0}} C_K T_K(x) \,,$$

where for $x = (x_1, \ldots, x_d) \in [-1,1]^d, K = (k_1, \ldots, k_d) \in \mathbb{Z}^d_{\geq 0}$, $T_K(x) = \prod_{i=1}^d T_{k_i}(x_i)$, and C_K is the K-th Chebyshev coefficients of f, and $T_{k_i}(x_i)$ is the k_i-th Chebyshev polynomial of the first kind. First, we will note a few facts about the multivariate Chebyshev polynomials, which are easily derived using properties of the univariate Chebyshev polynomials.

Definition 26 (Chebyshev Polynomials in d Dimensions) *Let* $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$*, and let* $K = (k_1, \ldots, k_d) \in \mathbb{Z}^d_{\geq 0}$*. The* K*-th Chebyshev polynomial of the first kind is denoted by* $T_K(x)$*, and is defined as:*

$$T_K(x) := \prod_{i=1}^d T_{k_i}(x_i) \,,$$

where $T_{k_i}(x_i)$ *is the* k_i*-th Chebyshev polynomial of the first kind in one dimension. Let* $\mathsf{W}(x)$ *denote the* weight function *defined as*

$$\mathsf{W}(x) := \prod_{i=1}^d \frac{1}{\sqrt{1-x_i^2}} \,.$$

Definition 27 (Inner Product in d-Dimensions) *The inner product of two functions* $f, g : [-1,1]^d \to \mathbb{R}$ *is defined as:*

$$\langle f, g \rangle := \int_{[-1,1]^d} f(x) g(x) \mathsf{d}x \,.$$

Fact 28 (Orthogonality of Chebyshev Polynomials in d-Dimensions) *Let $K_1, K_2 \in \mathbb{Z}^d_{\geq 0}$. Let $\mathrm{nnz}(K)$ denote the number of non-zero entries in $K \in \mathbb{Z}^d_{\geq 0}$. The higher dimensional Chebyshev polynomials satisfy the following orthogonality property:*

$$\langle T_{K_1}, W \cdot T_{K_2} \rangle = \int_{[-1,1]^d} T_{K_1}(x) T_{K_2}(x) \mathsf{W}(x) \mathsf{d}x = \begin{cases} 0 & \text{if } K_1 \neq K_2 \\ \frac{\pi^d}{2^{\mathrm{nnz}(K_1)}} & \text{if } K_1 = K_2 \end{cases}.$$

Definition 29 (Normalized d-Dimensional Chebyshev Polynomials) *The normalized d-dimensional Chebyshev polynomial $T^d_K(x)$, for $K \in \mathbb{Z}^d_{\geq 0}$ is defined as:*

$$\bar{T}_K(x) := \frac{T_K(x)}{\sqrt{T_K, W \cdot T_K}} = \sqrt{\frac{2^{\mathrm{nnz}(K)}}{\pi^d}} T_K(x)\,.$$

With the notations and definitions in place, we can now state the multivariate Jackson's theorem. The following theorem shows that the damped, truncated Chebyshev series of a smooth function is a good uniform multivariate polynomial approximation to the function.

Theorem 30 (Multivariate Jackson's Theorem) *Let $f : [-1,1]^d \to \mathbb{R}$ be an ℓ-Lipschitz smooth function, and for $K \in \mathbb{Z}^d_{>0}$, let $c_K = \left\langle f, W \cdot \bar{T}_K \right\rangle$. Then the polynomial*

$$\tilde{f}(x) = \sum_{K \in \{0,\ldots,2m-2\}^d} \tilde{c}_K T_K(x)$$

satisfies that

$$\left\| \tilde{f} - f \right\|_\infty \leq \frac{9\ell d}{m}, \text{ and } |\tilde{c}_K| \leq |c_K|\,, \text{ for } K \in \mathbb{Z}^d_{\geq 0}\,.$$

We now prove the theorem in Section D.3. With the high dimensional Jackson's theorem in place, we now prove the multivariate global Chebyshev coefficient decay lemma.

Lemma 31 (Multivariate Global Chebyshev Coefficient Decay) *Let $f : [-1,1]^d \to \mathbb{R}$ be a smooth, ℓ-Lipschitz function. For $K \in \mathbb{Z}^d_{\geq 0}$, let $c_K = \left\langle f, W \cdot \bar{T}_K \right\rangle$. Then, we have that*

$$\sum_{K \in \mathbb{Z}^d_{\geq 0}} \|K\|_2^2 \, c_K^2 \leq d\ell^2 \frac{\pi^d}{2}\,.$$

Proof Let $f : [-1,1]^d \to \mathbb{R}$ be a smooth, ℓ-Lipschitz function, with Chebyshev series

$$f(x) = \sum_{K \in \mathbb{Z}^d_{\geq 0}} \mathsf{C}_K T_K(x)\,.$$

Let $K = (k_1, \ldots, k_d) \in \mathbb{Z}^d_{\geq 0}$. Since f is ℓ-Lipschitz, it follows that $\|\nabla f\|_2^2 \leq \ell^2$. Consequently, for $i \in [d]$, we have:

$$\left(\frac{\partial f}{\partial x_i}\right)^2 (x') \leq \ell^2 \text{ at any } x' \in [-1,1]^d\,.$$

Therefore, we get that for $x = (x_1, \ldots, x_d) \in [-1,1]^d$,

$$\sum_{i=1}^{d} \int_{[-1,1]^d} \left(\frac{\partial f}{\partial x_i}\right)^2 \cdot \frac{\sqrt{1-x_i^2}}{\prod_{j\neq i\in[d]} \sqrt{1-x_j^2}} \mathrm{d}x \leq d\ell^2 \frac{\pi^d}{2} \,. \tag{24}$$

The upper bound follows from the fact that $0 \leq \left(\frac{\partial f}{\partial x_i}\right)^2 \leq \ell^2$, $\int_{-1}^{1} \sqrt{1-x_i^2} = \pi/2$, and that $\int_{-1}^{1} 1/\sqrt{1-x_i^2} = \pi$. We multiply $\left(\frac{\partial f}{\partial x_i}\right)^2$ by $\frac{\sqrt{1-x_i^2}}{\prod_{j\neq i\in[d]} \sqrt{1-x_j^2}}$ and integrate from $[-1,1]^d$ to exploit the orthogonality property of the Chebyshev polynomials of the first and second kind.

We now evaluate the LHS of Equation (24). Computing the gradient, we have from Fact 14 that for $i \in [d]$:

$$\frac{\partial f}{\partial x_i} = \sum_{k_i=0}^{\infty} \sum_{(k_1,\ldots,k_{i-1},k_{i+1}\ldots,k_d)\in\mathbb{Z}_{\geq 0}^{d}} k_i C_{(k_1,\ldots,k_i,\ldots,k_d)} \prod_{j\neq i\in[d]} T_{k_j}(x_j) U_{k_i-1}(x_i) \,.$$

We consider the square of the above expression. The orthogonality property of Chebyshev polynomials ensures that only the squared terms contribute non-zero values to the integral in Equation (24):

$$\sum_{k_i=0}^{\infty} \sum_{(k_1,\ldots,k_{i-1},k_{i+1}\ldots,k_d)\in\mathbb{Z}_{\geq 0}^{d-1}} k_i^2 C_{(k_1,\ldots,k_i,\ldots,k_d)}^2 \prod_{j\neq i\in[d]} (T_{k_j}(x_j))^2 (U_{k_i-1}(x_i))^2 \,. \tag{25}$$

Using the above equations, we evaluate the LHS of Equation (24) and get that

$$\sum_{i=1}^{d} \int_{[-1,1]^d} \left(\frac{\partial f}{\partial x_i}\right)^2 \cdot \frac{\sqrt{1-x_i^2}}{\prod_{j\neq i\in[d]} \sqrt{1-x_j^2}} \mathrm{d}x = \sum_{K=(k_1,\ldots,k_d)\in\mathbb{Z}_{\geq 0}^{d}} \frac{\pi^d}{2^{\mathrm{nnz}(K)}} \|K\|^2 \mathsf{C}_K^2 \,,$$

by using the orthogonality property of Chebyshev polynomials of the first and second kind and by inspecting the Equation (25). Therefore, combining above with Equation (24), we get that

$$\sum_{K=(k_1,\ldots,k_d)\in\mathbb{Z}_{\geq 0}^{d}} \frac{\pi^d}{2^{\mathrm{nnz}(K)}} \|K\|^2 \mathsf{C}_K^2 \leq d\ell^2 \frac{\pi^d}{2}$$

Note that these coefficients are not normalized. To get the normalized Chebyshev coefficients, we use the fact that $\bar{T}_K = \sqrt{\frac{2^{\mathrm{nnz}(K)}}{\pi^d}} T_K$. We let $f(x) = \sum_{K\in N^d} c_K \bar{T}_K(x)$, which yields

$$\sum_{K\in\mathbb{Z}_{\geq 0}^{d}} \|K\|^2 c_K^2 \leq d\ell^2 \frac{\pi^d}{2} \,.$$

■

D.2. Proof of Multivariate Generalization of Theorem 1

With the multivariate Jackson's theorem and the global Chebyshev coefficient decay lemma in place, we can now prove the multivariate version of our main theorem.

Theorem 32 *Let p, q be distributions supported on $[-1,1]^d$. For any $K \in \mathbb{Z}^d_{\geq 0}$, if the distributions' normalized Chebyshev moments satisfy*

$$\sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\mathop{\mathbb{E}}_{x \sim p} \bar{T}_K(x) - \mathop{\mathbb{E}}_{x \sim q} \bar{T}_K(x) \right)^2 \leq \Gamma^2, \tag{26}$$

where $\mathbf{0} = (0,\dots,0) \in \mathbb{R}^d$, then, for an absolute constant c,

$$W_1(p,q) \leq \frac{cd}{m} + \sqrt{\frac{d\pi^d}{2}}\Gamma. \tag{27}$$

Proof By Fact 25, to bound $W_1(p,q)$, it suffices to bound $\langle f, p-q\rangle$ for any 1-Lipschitz, smooth f. Let f_m be the approximation to any such f guaranteed by Theorem 30. We have:

$$\begin{aligned}\langle f, p-q\rangle = \langle f_m, p-q\rangle + \langle f - f_m, p-q\rangle &\leq \langle f_m, p-q\rangle + \|f-f_m\|_\infty \|p-q\|_1 \\ &\leq \langle f_m, p-q\rangle + \frac{36d}{m}.\end{aligned} \tag{28}$$

In the last step, we use that $\|f-f_m\|_\infty \leq 18d/m$ by Theorem 30, and that $\|p-q\|_1 \leq \|p\|_1 + \|q\|_1 = 2$. So, to bound $\langle f, p-q\rangle$, we turn our attention to bounding $\langle f_m, p-q\rangle$.

For technical reasons, we will assume from here on that p and q are supported on the interval $[-1+\delta, 1-\delta]^d$ for arbitrarily small $\delta \to 0$. This is to avoid an issue with the Chebyshev weight function $W(x) = \prod_{i=1}^d 1/\sqrt{1-x_i^2}$, for $x = (x_1,\dots,x_d)$ going to infinity at $x_i = -1, 1$. 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 $(p-q)/W$:

$$\frac{p-q}{W} = \sum_{K \in \mathbb{Z}^d_{\geq 0}}^{\infty} \left\langle \frac{p-q}{W} \cdot W, \bar{T}_K \right\rangle \bar{T}_K = \sum_{K \in \mathbb{Z}^d_{\geq 0}} \langle p-q, \bar{T}_K\rangle \cdot \bar{T}_K = \sum_{K \neq \mathbf{0}, K \in \mathbb{Z}^d_{\geq 0}}^{\infty} \langle p-q, \bar{T}_K\rangle \cdot \bar{T}_K, \tag{29}$$

where $\mathbf{0} = (0,\dots,0) \in \mathbb{R}^d$. In the last step we use that both p and q are distributions so $\left\langle p-q, \bar{T}_{\mathbf{0}} \right\rangle = 0$.

Next, recall from Theorem 30 that $f_m = \sum_{K \in \{0,\dots,m\}^d} \tilde{c}_K \bar{T}_K$, where each $\tilde{c}_K$ satisfies $|\tilde{c}_K| \leq |c_K|$ for $c_K := \langle f \cdot W, \bar{T}_K\rangle$. Using (29), the fact that $\langle \bar{T}_K \cdot w, \bar{T}_{K'}\rangle = 0$ whenever $K \neq K'$, and that $\langle \bar{T}_K \cdot W, \bar{T}_K\rangle = 1$ for all K, we have:

$$\begin{aligned}\langle f_m, p-q\rangle = \left\langle f_m \cdot W, \frac{p-q}{W} \right\rangle &= \left\langle \sum_{K \in \{0,\dots,m\}^d} \tilde{c}_K \bar{T}_K \cdot W, \sum_{K \neq \mathbf{0}, K \in \mathbb{Z}^d_{\geq 0}} \langle p-q, \bar{T}_j\rangle \bar{T}_j \right\rangle \\ &= \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \tilde{c}_K \cdot \langle p-q, \bar{T}_K\rangle.\end{aligned}$$

Via Cauchy-Schwarz inequality and our high-dimensional global decay bound from Lemma 31, we then have:

$$\begin{aligned}
\langle f_m, p-q\rangle &= \sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \|K\|_2 \tilde{c}_K \cdot \frac{\langle p-q, \bar{T}_K\rangle}{\|K\|_2} \\
&\leq \left(\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \|K\|_2^2 \tilde{c}_K^2\right)^{1/2} \cdot \left(\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2^2}\langle p-q, \bar{T}_K\rangle^2\right)^{1/2} \\
&\leq \left(\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \|K\|_2^2 c_K^2\right)^{1/2} \cdot \left(\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2^2}\langle p-q, \bar{T}_j\rangle^2\right)^{1/2} \\
&\leq \sqrt{\frac{d\pi^d}{2}} \left(\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2^2}\langle p-q, \bar{T}_K\rangle^2\right)^{1/2}. \qquad (30)
\end{aligned}$$

We can apply the assumption of the theorem, (26), to upper bound (30) by Γ.

Plugging this bound into Equation (28), we conclude the main bound of Theorem 32:

$$W_1(p,q) = \sup_{\text{1-Lipschitz, smooth } f} \langle f, p-q\rangle \leq \sqrt{\frac{d\pi^d}{2}}\Gamma + \frac{36d}{m}.$$

■

Remark 33 (Efficient Recovery in d Dimensions) *We note that given sufficiently accurate Chebyshev moments, we can back out a distribution close to the original distribution in Wasserstein-1 distance. The Algorithm 1 immediately generalizes to the d-dimensional setting; see Section 3.2 for the details in 1 dimension. We leave the details to the reader.*

We now give a constructive proof of the multivariate Jackson's theorem.

D.3. Proof of Multivariate Jackson's Theorem

We extend the 1-dimensional constructive proof of Jackson's theorem in Braverman et al. (2022) to d dimensions. To prove the multivariate Jackson's theorem, we will use Fourier analysis. We first define the Fourier series of a function in d dimensions. We start with a few standard preliminary definitions found in any standard text on Fourier analysis, such as Stein and Shakarchi (2011).

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an ℓ-Lipschitz function, i.e., $|f(x) - f(y)| \leq \ell \|x-y\|_2 \; \forall x, y \in \mathbb{R}^d$. We say that $f \in L^2([-\pi,\pi]^d)$ if $\int_{[-\pi,\pi]^d} |f(x)|^2 \, \mathrm{d}x < \infty$.

Definition 34 (Periodic Function) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is 2π periodic if $f(x) = f(x + 2\pi K)$ for all $x \in \mathbb{R}^d$ and $K \in \mathbb{Z}^d$. Formally, this is known as coordinate-wise periodic, but we will refer to it as periodic for simplicity.*

Definition 35 (Even Function) *Let $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. The function $f : \mathbb{R}^d \to \mathbb{R}$ is even if $f(x_1, \dots, x_d) = f(|x_1|, \dots, |x_d|)$ for all $x \in \mathbb{R}^d$.*

Definition 36 (Fourier Series) *Let $f \in L^2([-\pi,\pi]^d)$ be a 2π periodic function. The function f can be written via a Fourier series as:*

$$f(x) = \sum_{K\in\mathbb{Z}^d} \hat{f}(K) e^{\mathrm{i}\langle k,x\rangle}, \text{ where } \hat{f}(K) = \frac{1}{(2\pi)^d}\int_{[-\pi,\pi]^d} f(x)e^{-\mathrm{i}\langle K,x\rangle}\mathsf{d}x\,,$$

and $\mathrm{i} = \sqrt{-1}$. For $K \in \mathbb{Z}^d$, $\hat{f}(K)$ is called the Fourier coefficient of f.

Claim 37 (Convolution Theorem) *Let $f, g \in L^2([-\pi,\pi]^d)$ be 2π-periodic functions with Fourier coefficients $\left\{\hat{f}(K)\right\}_{K\in\mathbb{Z}^d}$ and $\{\hat{g}(K)\}_{K\in\mathbb{Z}^d}$ respectively. Let h be their convolution:*

$$h(x) := [f * g](x) = \int_{[-\pi,\pi]^d} f(u)g(x-u)\mathsf{d}u\,.$$

The Fourier coefficients of h, $\left\{\hat{h}(K)\right\}_{K\in\mathbb{Z}^d}$, are given by:

$$\hat{h}(K) = (2\pi)^d \cdot \hat{f}(K)\hat{g}(K)\,.$$

We now build a multivariate version of the Jackson kernel, a key ingredient in the proof of the multivariate Jackson's theorem. Braverman et al. (2022) define the Jackson kernel in one dimension, which we generalize to d-dimensions by just multiplying the one-dimensional Jackson kernel in each dimension.

Definition 38 (Jackson Kernel) *For $x_i \in \mathbb{R}$, $m \in \mathbb{Z}_{>0}$, let $b_1 : \mathbb{R} \to \mathbb{R}$ be the following function:*

$$b_1(x_i) := \left(\frac{\sin(mx_i/2)}{\sin(x_i/2)}\right)^4 = \sum_{k_1=-2m+2}^{2m-2} \hat{b_1}(k_1)e^{\mathrm{i}k_1x_i}\,,$$

where the Fourier coefficients $\hat{b_1}(-2m+2),\ldots,\hat{b_1}(0),\ldots,\hat{b_1}(2m-2)$ equals to:

$$\hat{b_1}(-k_1) = \hat{b_1}(k_1) = \sum_{t=-m}^{m-k_1}(m-|t|)\cdot(m-|t+k_1|) \quad \text{for } k_1 = 0,\ldots,2m-2\,. \tag{31}$$

Note that $\hat{b_1}(0) \geq \cdots \geq \hat{b_1}(2m-2)$. Let $\mathbf{0} = (0,\ldots,0) \in \mathbb{R}^d$, $x = (x_1,\ldots,x_d) \in \mathbb{R}^d$, and b be the following trigonometric polynomial:

$$b(x_1,\ldots,x_d) := \prod_{i=1}^d b_1(x_i) = \prod_{i=1}^d \left(\frac{\sin(mx_i/2)}{\sin(x_i/2)}\right)^4 = \sum_{K\in\{-2m+2,\ldots,2m-2\}^d} \hat{b}(K)e^{\mathrm{i}\langle K,x\rangle}\,.$$

From Equation (31), we have that $\hat{b}(\mathbf{0}) \geq \hat{b}(K)$, for $K \neq \mathbf{0}$.

We also need the following fact from Braverman et al. (2022) about Jackson's kernel.

Fact 39 ((Braverman et al., 2022, Theorem C.5)) *For $x_i \in \mathbb{R}$, $m \in \mathbb{Z}_{>0}$, the one-dimensional Jackson's Kernel b_1, defined in Definition 38, satisfies the following*

$$\frac{\int_0^\pi x_i b_1(x_i)\mathsf{d}x_i}{\int_0^\pi b_1(x_i)\mathsf{d}x_i} \leq \frac{8.06}{m}\,.$$

We are not ready to prove that a truncated and "damped" Fourier series of f is a *good* uniform approximation to f.

Theorem 40 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a ℓ-Lipschitz continuous, 2π-periodic function. For $m \in \mathbb{Z}_{>0}$, let $b : \mathbb{R}^d \to \mathbb{R}$ be the Kernel from Definition 38. The function $\tilde{f}(x) = \frac{1}{\hat{b}(\mathbf{0})(2\pi)^d} \int_{[\pi,\pi]^d} b(u) f(x-u) \mathsf{d}u$ satisfies:*

$$\left\| \tilde{f} - f \right\|_\infty \leq \frac{9\ell d}{m}.$$

Moreover, the Fourier coefficients of $\tilde{f}$, $\left\{ \hat{\tilde{f}}(K) \right\}_{K \in \mathbb{Z}^d}$, are given by:

$$\hat{\tilde{f}}(K) = \frac{\hat{b}(K)}{\hat{b}(\mathbf{0})} \hat{f}(K),$$

where $\hat{b}(\mathbf{0}) \geq \hat{b}(K)$ for all $K \in \mathbb{Z}^d$, and for $K \notin \{-2m+2, \ldots, 2m-2\}^d$, we have that $\hat{\tilde{f}}(K) = 0$.

Proof Let $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$. Note that $\hat{b}(\mathbf{0}) = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi]^d} b(x) \mathsf{d}x$, whence we get that $\frac{1}{\hat{b}(\mathbf{0})(2\pi)^d} \int_{[-\pi,\pi]^d} b(u) \mathsf{d}u = 1$. Therefore, by the definition of $\tilde{f}$, we get that:

$$\left| \tilde{f}(x) - f(x) \right| \leq \int_{[-\pi,\pi]^d} \frac{1}{\hat{b}(\mathbf{0})(2\pi)^d} b(u) \cdot |f(x) - f(x-u)| \, \mathsf{d}u.$$

Since f is ℓ-Lipschitz, we have that $|f(x) - f(x-u)| \leq \ell \|u\|_2$. Therefore, we get that:

$$\begin{aligned}
&\max_x \left| \tilde{f}(x) - f(x) \right| \\
&\quad \leq \int_{[-\pi,\pi]^d} \frac{1}{\hat{b}(\mathbf{0})(2\pi)^d} b(u) \cdot \ell \|u\|_2 \, \mathsf{d}u \\
&\quad \leq \int_{[-\pi,\pi]^d} \frac{1}{\hat{b}(\mathbf{0})(2\pi)^d} b(u) \cdot \ell \|u\|_1 \, \mathsf{d}u && (\because \|\cdot\|_2 \leq \|\cdot\|_1) \\
&\quad = \frac{\ell}{\hat{b}(\mathbf{0})(2\pi)^d} \cdot \sum_{i=1}^d \left(\int_{-\pi}^{\pi} |u_i| \, b_1(u_i) \mathsf{d}u_i \right) \cdot \left(\prod_{j \in [d], j \neq i} \int_{-\pi}^{\pi} b_1(u_j) \mathsf{d}u_j \right) && \left(\because \|u\|_1 = \sum_{i=1}^d |u_i| \right) \\
&\quad = \ell \sum_{i=1}^d \frac{\int_{-\pi}^{\pi} |u_i| \, b_1(u_i) \mathsf{d}u_i}{\int_{-\pi}^{\pi} b_1(u_i) \mathsf{d}u_i} \cdot \prod_{j \in [d], j \neq i} \frac{\int_{-\pi}^{\pi} b_1(u_j) \mathsf{d}u_j}{\int_{-\pi}^{\pi} b_1(u_j) \mathsf{d}u_j} && \text{(By Definition 38)} \\
&\quad = \ell \sum_{i=1}^d \frac{\int_0^{\pi} u_i b_1(u_i) \mathsf{d}u_i}{\int_0^{\pi} b_1(u_i) \mathsf{d}u_i} \leq (8.06) d \frac{\ell}{m},
\end{aligned}$$

where the last inequality follows from Fact 39. We now reason about the Fourier coefficients of $\tilde{f}$. Note that for $K \notin \{-2m+2, \ldots, 2m-2\}^d$, we have that $\hat{\tilde{f}}(K) = 0$. For $K \in \{-2m+2, \ldots, 2m-2\}^d$, we have by the convolution theorem (Claim 37) that:

$$\hat{\tilde{f}}(K) = \frac{\hat{b}(K)}{\hat{b}(\mathbf{0})} \cdot \hat{f}(K).$$

Using the fact from Definition 38 that $\hat{b}(\mathbf{0}) \geq \hat{b}(K)$, for $K \neq \mathbf{0}$, and the fact that $\hat{b}(K) = 0$ for $K \notin \{-2m+2, \ldots, 2m-2\}^d$, we get the desired result. ■

We now prove the multivariate Jackson's theorem for a smooth, ℓ-Lipschitz function $f : [-1,1] \to \mathbb{R}$. To do so, we construct a mapping to a periodic function h with period 2π and then apply the previous theorem.

Theorem 30 (Multivariate Jackson's Theorem) *Let $f : [-1,1]^d \to \mathbb{R}$ be an ℓ-Lipschitz smooth function, and for $K \in \mathbb{Z}^d_{\geq 0}$, let $c_K = \left\langle f, W \cdot \bar{T}_K \right\rangle$. Then the polynomial*

$$\tilde{f}(x) = \sum_{K \in \{0,\ldots,2m-2\}^d} \tilde{c}_K T_K(x)$$

satisfies that

$$\left\| \tilde{f} - f \right\|_\infty \leq \frac{9\ell d}{m}, \text{ and } |\tilde{c}_K| \leq |c_K|, \text{ for } K \in \mathbb{Z}^d_{\geq 0}.$$

Proof [Proof of Theorem 30] Let $(\cos\theta_1, \ldots, \cos\theta_d) \in [-1,1]^d$. We will use the identity that for $K = (k_1, \ldots, k_d) \in \mathbb{Z}^d_{\geq 0}$,

$$T_K(\cos\theta_1, \ldots, \cos\theta_d) = \prod_{i=1}^d \cos(k_i\theta_i).$$

Consider the Lipschitz continuous function $f : [-1,1]^d \to \mathbb{R}$ with Chebyshev expansion coefficients c_K for $K \in \mathbb{Z}^d_{\geq 0}$, where $c_K = \left\langle r, W \cdot \bar{T}_K \right\rangle$. We transform f into a periodic function as follows: For $\Theta = (\theta_1, \ldots, \theta_d) \in [-\pi, 0]^d$, let $g(\Theta) = f(\cos\theta_1, \ldots, \cos\theta_d)$ and let $h(\Theta) = g(-|\theta_1|, \ldots, -|\theta_d|)$ for $\Theta \in [-\pi, \pi]^d$. The function $h : [\pi, \pi]^d \to \mathbb{R}$ is a periodic and even function (Definition 35). Since the function is even, one can check that its Fourier series can be written as follows: For $\Theta = (\theta_1, \ldots, \theta_d) \in [-\pi, \pi]^d$,

$$h(\Theta) = \sum_{K=(k_1,\ldots,k_d) \in \mathbb{Z}^d_{\geq 0}} \alpha_K \prod_{i=1}^d \cos(k_i\theta_i),$$

where:

$$\alpha_K = \frac{2^{\mathrm{nnz}(K)}}{(2\pi)^d} \int_{[-\pi,\pi]^d} h(\Theta) \left(\prod_{i=1}^d \cos(k_i\theta_i) \right) \mathsf{d}\Theta = \frac{2^{\mathrm{nnz}(K)}}{\pi^d} \int_{[-\pi,0]^d} g(\Theta) \prod_{i=1}^d \cos(k_i\theta_i) \mathsf{d}\Theta.$$

Let $x = (x_1, \ldots, x_d) \in [-1,1]^d$. Using that fact that for $i \in [d]$, $\frac{\mathsf{d}}{\mathsf{d}x_i} \cos^{-1}(x_i) = \frac{1}{\sqrt{1-x_i^2}}$, we get

$$\int_{[-\pi,0]^d} g(\Theta) \prod_{i=1}^d \cos(k_i\theta_i) \mathsf{d}\Theta = \int_{[-1,1]^d} f(x) T_K(x) W(x) \mathsf{d}x.$$

Using the above equations, we conclude that for $K \in \mathbb{Z}^d_{\geq 0}$, the Fourier coefficients of h are just a scaling depending on K of the Chebyshev coefficients of f, and we get by Definition 29 that:

$$\sqrt{\frac{2^{\mathrm{nnz}(K)}}{\pi^d}} c_K = \alpha_K. \tag{32}$$

Algorithm 3 d-Dimension Private Chebyshev Moment Matching

Input: Dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n \in [-1,1]^d$, privacy parameters $\epsilon, \delta > 0$.

Output: A probability distribution q approximating the distribution, $p := \text{Unif}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

1: Let $\mathcal{G} = \left\{-1, -1 + \frac{1}{\lceil (\varepsilon n)^{1/d} \rceil}, -1 + \frac{2}{\lceil (\varepsilon n)^{1/d} \rceil}, \ldots, 1\right\}^d$. Let $r := (2\lceil (\varepsilon n)^{1/d} \rceil + 1)$ and for $J = (j_1, \ldots, j_d) \in [r]^d$ let $g_J = \left(-1 + \frac{j_1 - 1}{\lceil (\varepsilon n)^{1/d} \rceil}, \ldots, -1 + \frac{j_d - 1}{\lceil (\varepsilon n)^{1/d} \rceil}\right)$ denote the J^{th} element of $\mathcal{G}$.

2: For $i = 1, \ldots, n$, let $\tilde{\mathbf{x}}_i = \text{argmin}_{\boldsymbol{y} \in \mathcal{G}} |\mathbf{x}_i - \boldsymbol{y}|$. I.e., round $\mathbf{x}_i$ to the nearest point in the grid $\mathcal{G}$.

3: Set $\sigma^2 = \frac{4 \cdot 2^d}{\pi^d} \mathsf{S} \ln(1.25/\delta)/(n^2 \varepsilon^2)$, where $\mathsf{S} = \sum_{K \in \{0,\ldots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2}$. See Lemma 43 for the bound on S.

4: Set $m = \lceil 2(\varepsilon n)^{1/d} \rceil$. For $K \in \{0, \ldots, m\}^d \setminus \mathbf{0}$, let $\hat{m}_K = \eta_K + \frac{1}{n} \sum_{i=1}^n \bar{T}_K(\tilde{\mathbf{x}}_i)$, where $\eta_K \sim \mathcal{N}(0, \|K\|_2 \sigma^2)$.

5: Let $\{q_J\}_{J \in [r]^d}$ be the solution to the following optimization problem:

$$\begin{aligned} \min_{\{z_J\}_{J \in [r]^d}} \quad & \sum_{K \in \{0,\ldots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\hat{m}_K - \sum_{J \in [r]^d} z_J \bar{T}_K(g_J) \right)^2 \\ \text{subject to} \quad & \sum_{J \in [r]^d} z_J = 1 \text{ and } z_J \geq 0, \ \forall J \in [r]^d. \end{aligned}$$

6: Return $q = \sum_{J \in [r]^d} q_J \delta(x - g_J)$, where δ is the Dirac delta function.

We observe that the mapping from f to h preserves the ℓ-Lipschitz property. The function $h : [\pi, \pi]^d \to \mathbb{R}$ is periodic and an even function (Definition 35), and is ℓ-Lipschitz. Let $\tilde{h}$ be the function obtained by applying Jackson's theorem (Theorem 40) to h. We know that $\tilde{h}$ is an even function since h is even and Jackson's Kernel b, which h is convolved with, is also even. Recall that the Fourier series coefficients of $\tilde{h}$, denoted by $\hat{\tilde{h}}(K)$, are 0 for $K \notin \{0, \ldots, 2m-2\}^d$. Finally, let $\tilde{f} : [-1,1]^d \to \mathbb{R}$ be defined as

$$\tilde{f}(\cos\theta_1, \ldots, \cos\theta_d) := \tilde{h}(\theta_1, \ldots, \theta_d)\,.$$

By Equation (32), we get that the Chebyshev coefficients of $\tilde{f}$ are exactly $\frac{\hat{b}(K)}{\hat{b}(\mathbf{0})} c_K$. Note from Theorem 40 that $\left|\frac{\hat{b}(K)}{\hat{b}(\mathbf{0})}\right| \leq 1$, therefore, we get the $\tilde{f}$ is the damped Chebyshev truncated series of f.

Moreover, we have that $\left\|\tilde{f} - f\right\|_\infty = \left\|\tilde{h} - h\right\|_\infty \leq \frac{9\ell d}{m}$, where the inequality follows from Theorem 40. This completes the proof. ■

Appendix E. Differentially Private Synthetic Data Generation in Higher Dimensions

Theorem 41 *Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a dataset with each $\mathbf{x}_j \in [-1,1]^d$, for $d \geq 2$. 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 $\mathsf{poly}\left(n, \varepsilon, \delta, 2^d\right)$ time, returns a distribution q*

satisfying for a fixed constant c_4,

$$\mathbb{E}[W_1(p,q)] \le c_4 d \left(\frac{1+\ln^{0.5}(1.25/\delta)}{n\varepsilon}\right)^{1/d}.$$

Proof We analyze both the privacy and accuracy of Algorithm 3.

Privacy. For a dataset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in [-1,1]^{d\times n}$, where each data-point $\mathbf{x}_i \in [-1,1]^d$. Let $f(X)$ be a vector-valued function mapping to the $K \in \{0,\ldots,m\}^d \setminus \mathbf{0}$ *scaled* Chebyshev moments of the uniform distribution over X. I.e.,

$$f(X)_K = \frac{1}{\sqrt{\|K\|_2}} \cdot \frac{1}{n}\sum_{i=1}^{n} \bar{T}_K(\mathbf{x}_i)\,,$$

where $f(X)_K$ denotes the K-th entry of the vector $f(X)$, and $\bar{T}_K(\mathbf{x})$ is the K-th normalized multivariate Chebyshev polynomial.

We will show that Algorithm 3 is (ε,δ)-differentially private and that the output of the algorithm is close in Wasserstein distance to the true moments of the uniform distribution over X. By Definition 29, $\max_{\mathbf{x}_i\in[-1,1]^d} |\bar{T}_K(x_i)| \le \sqrt{2^{\mathrm{nnz}(K)}/\pi^d}$ for $K \in \mathbb{Z}^d_{\ge 0}$, so we have:

$$\begin{aligned}
\Delta^2_{2,f} = \max_{\substack{\text{neighboring datasets}\\ X,X'\in\mathcal{X}}} \|f(X)-f(X')\|_2^2 &\le \sum_{K\in\{0,\ldots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2}\cdot\frac{1}{n^2}\cdot\frac{4\cdot 2^{\mathrm{nnz}(K)}}{\pi^d}\\
&\le \frac{4\cdot 2^d}{\pi^d n^2}\sum_{K\in\{0,\ldots,m\}^d\setminus\mathbf{0}}\frac{1}{\|K\|_2}\\
&= \frac{4\cdot 2^d}{\pi^d n^2}\mathsf{S}\,, \qquad (33)
\end{aligned}$$

where $\mathsf{S} = \sum_{K\in\{0,\ldots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2}$. For two neighboring datasets X, X', let $\tilde{X}$ and $\tilde{X}'$ be the rounded datasets computed in line 2 of Algorithm 3 – i.e., $\tilde{X} = \{\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_n\}$. Observe that $\tilde{X}$ and $\tilde{X}'$ are also neighboring. Thus, it follows from Fact 17 and the sensitivity bound of Equation (33) that $\tilde{m}_K = f(\tilde{X})_K + \eta_K$ is (ε,δ)-differentially private for $\eta_K \sim \mathcal{N}(0,\sigma^2)$ as long as $\sigma^2 = \frac{4\cdot 2^d}{\pi^d}\mathsf{S}\ln(1.25/\delta)/(n^2\varepsilon^2)$. Finally, observe that $\hat{m}_K$ computed by Algorithm 3 is exactly equal to $\sqrt{\|K\|_2}$ times the K^{th} entry of such an $\tilde{m}$. So $\{\hat{m}_K\}_{K\in\{0,\ldots,m^d\}\setminus\mathbf{0}}$ are (ε,δ)-differentially private. Since the remainder of Algorithm 2 simply post-processes $\{\hat{m}_K\}_{K\in\{0,\ldots,m^d\}\setminus\mathbf{0}}$ without returning to the original data X, the output of the algorithm is also (ε,δ)-differentially private, as desired.

Accuracy. The Algorithm 3 begins by rounding the dataset X so that every coordinate of every data point is a multiple of $1/\lceil(\varepsilon n)^{1/d}\rceil$. 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:

$$W_1(p,\tilde{p}) \le \frac{d}{2\lceil(\varepsilon n)^{1/d}\rceil}\,. \qquad (34)$$

In particular, we can transport p to $\tilde{p}$ by moving every unit of $1/n$ probability mass a distance of at most $1/2\lceil(\varepsilon n)^{1/d}\rceil$, along each of the d coordinates. Given eq. (34), it will suffice to show that

Algorithm 3 returns a distribution q that is close in Wasserstein distance to $\tilde{p}$. We will apply triangle inequality to bound $W_1(p, q)$.

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:

$$E := \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\hat{m}_K(\tilde{p}) - \langle \tilde{p}, \bar{T}_K \rangle \right)^2,$$

where m is as chosen in Algorithm 3 and $\langle \tilde{p}, T_K \rangle = \frac{1}{n} \sum_{i=1}^n \bar{T}_K(\tilde{\mathbf{x}}_i)$. Let σ^2 and $\{\eta_K\}_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}}$ be as in Algorithm 3. Applying linearity of expectation, we have that:

$$\begin{aligned}
\mathbb{E}[E] = \mathbb{E}\left[\sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \eta_K^2 \right] &= \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \mathbb{E}\left[\eta_K^2\right] \\
&= \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \cdot \|K\|_2 \, \sigma^2 \\
&\leq \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2} \sigma^2 = \sigma^2 \mathsf{S}, \qquad (35)
\end{aligned}$$

where we recall that $\mathsf{S} = \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2}$.

Now, let q be as in Algorithm 3. Using the triangle inequality argument as in Section 3.2, we have:

$$\begin{aligned}
\Gamma^2 &= \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\langle q, \bar{T}_K \rangle - \langle \tilde{p}, \bar{T}_K \rangle \right)^2 \\
&\leq \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\langle q, \bar{T}_K \rangle - \hat{m}_j \right)^2 + \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\langle \tilde{p}, \bar{T}_K \rangle - \hat{m}_j \right)^2 \\
&\leq 2E.
\end{aligned}$$

Above we use that $\tilde{p}$ is a feasible solution to the optimization problem solved in Algorithm 3 and, since q is the optimum, $\sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \left(\langle q, \bar{T}_K \rangle - \hat{m}_j \right)^2 \leq \sum_{K \in \{0,\dots,m\}^d \setminus \mathbf{0}} \frac{1}{\|K\|_2^2} \left(\langle \tilde{p}, \bar{T}_K \rangle - \hat{m}_j \right)^2$. 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 32, we have for constant c:

$$\begin{aligned}
\mathbb{E}[W_1(\tilde{p}, q)] &\leq \sqrt{\frac{d\pi^d}{2}} \mathbb{E}[\Gamma] + \frac{cd}{m} \\
&\leq \sqrt{\frac{d\pi^d}{2}} \sqrt{2\mathsf{S}\sigma^2} + \frac{cd}{m}. \qquad (36)
\end{aligned}$$

From the bound on σ^2 computed above, and from the upper bound on S in Lemma 43, we get that

$$\begin{aligned}\sqrt{\frac{d\pi^d}{2}}\sqrt{2\mathsf{S}\sigma^2} &\le \sqrt{\frac{d\pi^d}{2}}\sqrt{\frac{8\ln(1.25/\delta)\cdot 2^d}{\pi^d}}\cdot\frac{S}{n\varepsilon}\\ &\le \sqrt{\frac{d\pi^d}{2}}\sqrt{\frac{8\ln(1.25/\delta)\cdot 2^d}{\pi^d}}\cdot\frac{4(\pi e)^{d/2}}{2^d}\cdot\frac{m^{d-1}}{dn\varepsilon}\\ &= 8\sqrt{\ln(1.25/\delta)}(\pi e/2)^{d/2}\cdot\frac{m^{d-1}}{\sqrt{d}n\varepsilon}\,.\end{aligned}$$

Therefore, for some absolute constant c_2, setting $m = c_2\left(\frac{dn\varepsilon}{\ln^{0.5}(1.25/\delta)}\right)^{\frac{1}{d}}$, we get from Equation (36),

$$\mathbb{E}[W_1(\tilde{p},q)] \le c_3\cdot d\cdot\left(\frac{\ln^{0.5}(1.25/\delta)}{n\varepsilon}\right)^{1/d},$$

for an absolute constant c_3. By triangle inequality $W_1(p,q)\le W_1(p,\tilde{p})+W_1(\tilde{p},q)$ and using the bound on $W_1(p,\tilde{p})$ in Equation (34), we get that for an absolute constant c_4,

$$\mathbb{E}[W_1(p,q)] \le c_4 d\left(\frac{1+\ln^{0.5}(1.25/\delta)}{n\varepsilon}\right)^{1/d}.$$

Runtime. The number of points in the grid in Algorithm 3 is upper bounded by $|\mathcal{G}| = (1+2\lceil\varepsilon n\rceil^{1/d})^d = O\left(2^d\lceil n\varepsilon\rceil^{1/d}\right)$. The number of Chebyshev moments we calculate is less than $(m+1)^d = O\left(2^d\lceil(dn\varepsilon)\rceil^{1/d}\right)$. Since the optimization problem runs in polynomial time in its variables and constraints, we get that the running time of the algorithm is bounded by $\mathsf{poly}\left(n,\varepsilon,\delta,2^d\right)$. ■

Remark 42 (Comparison to Boedihardjo et al. (2024)) *We remark that Boedihardjo et al. (2024) use the ℓ_∞ metric instead of the ℓ_2 metric for the Wasserstein-1 distance, and they achieve an error of $\mathbb{E}[W_1(p,q)] \le O\left(\frac{\log^{1.5}(\varepsilon n)}{\varepsilon n}\right)^{1/d}$. Since the Wasserstein-1 distance in the ℓ_∞ metric is bounded by 1, their bound is non-vacuous for $d = O(\log n)$. For $d = O(\log n)$, our bound matches their bound to $\log(n)$-factors.*

Finally, we show how to upper bound S.

Lemma 43 *Let $m\in\mathbb{Z}_{>0}$ and $d\ge 2$. Then, we have that*

$$\mathsf{S} := \sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}}\frac{1}{\|K\|_2} \le \frac{4(\pi e)^{d/2}}{2^d}\cdot\frac{m^{d-1}}{d}\,.$$

Proof Note that the function $\frac{1}{\|K\|_2}$ is decreasing in $\|K\|_2$. Moreover, for $K\in\{0,\dots,m\}^d\setminus\mathbf{0}$, $\|K\|_2\ge 1$. Thus, we can upper bound the sum by the integral as follows:

$$\sum_{K\in\{0,\dots,m\}^d\setminus\mathbf{0}}\frac{1}{\|K\|_2} \le d + \int_{K\in[0,m]^d,\|K\|_2\ge 1}\frac{1}{\|K\|_2}\mathsf{d}K\,.$$

Using the fact that for $K \in \{0,\ldots,m\}^d \setminus \mathbf{0}$, $\|K\|_2 \leq m\sqrt{d}$, and that the $m\sqrt{d}$ ball contains the hypercube $[0,m]^d$, we get that

$$\int_{K\in[0,m]^d, \|K\|_2\geq 1} \frac{1}{\|K\|_2} \mathsf{d}K \leq \frac{1}{2^d} \int_{1\leq\|K\|_2\leq m\sqrt{d}} \frac{1}{\|K\|_2} \mathsf{d}K\,,$$

where the factor of $1/2^d$ comes due to symmetry and the fact that the set $K \in [0,m]^d$, i.e., the vectors in K contains non-negative entries, which is only $1/2^d$ fraction of vectors in the set $\left\{K : 1 \leq \|K\|_2 \leq m\sqrt{d}\right\}$. To compute the integral, we can transition to polar coordinates. Let $\|K\|_2 = r$, then we get that

$$\int_{1\leq\|K\|_2\leq m\sqrt{d}} \frac{1}{\|K\|_2} \mathsf{d}K = \int_{r=1}^{m\sqrt{d}} \int_{\Omega} \frac{1}{r} \cdot r^{d-1} \mathsf{d}\Omega \mathsf{d}r\,,$$

where Ω is the *angular domain* in spherical coordinates. Separating the terms in the integral gives,

$$\int_{r=1}^{m\sqrt{d}} \int_{\Omega} \frac{1}{r} \cdot r^{d-1} \mathsf{d}\Omega \mathsf{d}r = \int_{r=1}^{m\sqrt{d}} r^{d-2} \mathsf{d}r \int_{\Omega} \mathsf{d}\Omega\,.$$

Using the fact that $\int_\Omega \mathsf{d}\Omega$ is the surface area of the unit sphere in d-dimensions, we can use the equation for the surface area of the unit sphere, see e.g. (Boyd and Vandenberghe, 2004), to evaluate the integral as

$$\int_{r=1}^{m\sqrt{d}} r^{d-2} \mathsf{d}r \cdot \int_{\Omega} \mathsf{d}\Omega = \frac{(m\sqrt{d})^{d-1} - 1}{d-1} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)}\,,$$

where Γ is the gamma function. Combining all the terms, we get that

$$\sum_{K\in\{0,\ldots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2} \leq d + \frac{1}{2^d} \cdot \frac{(m\sqrt{d})^{d-1} - 1}{d-1} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)}\,.$$

By Stirling's approximation $\Gamma(d/2) \geq \sqrt{2}(d/2)^{d/2-1/2}e^{-d/2}$, and observing that the second term dominates in the RHS above, we have

$$\sum_{K\in\{0,\ldots,m\}^d\setminus\mathbf{0}} \frac{1}{\|K\|_2} \leq \frac{4(\pi e)^{d/2}}{2^d} \cdot \frac{m^{d-1}}{d}\,.$$

■

Appendix F. Accuracy of Generic Moment Matching Algorithm

In this section, we give the full proof of Corollary 2, which establishes the accuracy of the generic Chebyshev moment regression algorithm (Algorithm 1). We require the following basic property about the Chebyshev nodes:

Lemma 44 (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 44 in place, we are ready to prove Corollary 2.

Proof [Proof of Corollary 2] Let $\mathcal{C}$ and $r_{\mathcal{C}} : [-1, 1] \to \mathcal{C}$ be as in Lemma 44. For $i \in \{1, \ldots, g\}$, 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 $j \in 1, \ldots, k$ we have:

$$\begin{aligned}
\left|\langle p, \bar{T}_j\rangle - \langle \tilde{p}, \bar{T}_j\rangle\right| &= \left|\sum_{i=1}^{g} \int_{Y_i} \bar{T}_j(x)p(x)\, dx - \left(\int_{Y_i} p(x)\, dx\right)\bar{T}_j(x_i)\right| \\
&= \left|\sum_{i=1}^{g} \left(\int_{Y_i} p(x)\, dx\right)\bar{T}_j(y_i) - \left(\int_{Y_i} p(x)\, dx\right)\bar{T}_j(x_i)\right| \qquad \text{(for some } y_i \in Y_i\text{)} \\
&\leq \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\, dx\right)\left|\bar{T}_j(y_i) - \bar{T}_j(x_i)\right| \\
&= \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\, dx\right) \cdot \sqrt{\frac{2}{\pi}} \cdot \left|\cos(j\cos^{-1}(y_i)) - \cos(j\cos^{-1}(x_i))\right| \\
&\leq \sum_{i=1}^{g} \left(\int_{Y_i} p(x)\, dx\right) \cdot \sqrt{\frac{2}{\pi}} \cdot \frac{j\pi}{2g} = \frac{j\sqrt{\pi/2}}{g} \qquad (37)
\end{aligned}$$

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 44 and the fact that the derivative of $\cos(jx)$ is bounded by j. The bound in (37) then yields:

$$\left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\langle p, \bar{T}_j\rangle - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \leq \frac{\sqrt{\pi k/2}}{g}. \qquad (38)$$

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:

$$\left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle q, \bar{T}_j\rangle\right)^2\right)^{1/2} \leq \left(\sum_{j=1}^{k} \frac{1}{j^2}\left(\hat{m}_j - \langle \tilde{p}, \bar{T}_j\rangle\right)^2\right)^{1/2} \qquad (39)$$

Applying triangle inequality, followed by (39), triangle inequality again, and finally (38), we have:

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

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

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

■

Appendix G. High Probability Bound for Private Synthetic Data

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 (Wainwright, 2019). 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 45 (Sub-Exponential Parameters (Wainwright, 2019, 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 46 (Wainwright (2019, 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 [Proof of high-probability bound of Theorem 4] Recalling the proof of the expectation bound of Theorem 4 from Section A, 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 45, γ_j is a sub-exponential random variable with parameter $\left(2\sigma^2/j, 4\sigma^2/j\right)$. We can then apply Fact 46, 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:

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

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

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

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

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

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

■

Appendix H. 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, showing that our upper bound in Theorem 5 is optimal up to logarithmic factors. We first need the following theorem from Woodruff et al. (2022), which shows that estimating the trace of a positive semi-definite matrix A to within a multiplicative error of $(1 \pm \varepsilon)$ requires $\Omega(1/\varepsilon)$ matrix-vector multiplications with A.

Theorem 47 (Restated (Woodruff et al., 2022, Theorem 4.2)) *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 $2/3$ in outputting an estimate $\tilde{t}$ such that $\left|\tilde{t} - \mathrm{Tr}(A)\right| \leq \varepsilon \cdot \mathrm{Tr}(A)$ requires $\Omega\left(\frac{1}{\varepsilon}\right)$ 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 48 *Any algorithm that is given matrix-vector multiplication access to a symmetric matrix $A \in \mathbb{R}^{n\times n}$ with spectral density p and $\|A\|_2 \leq 1$ requires $\Omega\left(\frac{1}{\varepsilon}\right)$ matrix-vector multiplications with A to output a distribution q such that $W_1(p, q) \leq \epsilon$ with probability at least $2/3$.*

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 $W_1(p,q) \leq \varepsilon/4$. 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 $\Omega(1/\varepsilon)$ matrix-vector products by Theorem 47.

In particular, applying Kantorovich-Rubinstein duality (Fact 9) with the 1-Lipschitz functions $f(x) = x$ and $f(x) = -x$, we have that:

$$\int_{-1}^{1} xp(x)\,dx - \int_{-1}^{1} xq(x)\,dx \leq \varepsilon/4 \qquad \text{and} \qquad \int_{-1}^{1} xq(x)\,dx - \int_{-1}^{1} xp(x)\,dx \leq \varepsilon/4. \tag{40}$$

We have that $\int_{-1}^{1} xp(x)\,dx = \frac{1}{n}\mathrm{Tr}(A)$. So (40) implies that $\tilde{t} = n \cdot \int_{-1}^{1} xq(x)\,dx$ satisfies:

$$|\tilde{t} - \mathrm{Tr}(A)| \leq n \cdot \varepsilon/4 \leq \varepsilon \cdot \mathrm{Tr}(A).$$

■