Moments, Random Walks, and Limits for Spectrum Approximation

Yujia Jin YUJIA@STANFORD.EDU
Stanford University

Christopher Musco CMUSCO@NYU.EDU
New York University

Aaron Sidford SIDFORD@STANFORD.EDU
Stanford University

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

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study lower bounds for the problem of approximating a one dimensional distribution given (noisy) measurements of its moments. We show that there are distributions on $[-1, 1]$ that cannot be approximated to accuracy ε in Wasserstein-1 distance even if we know *all* of their moments to multiplicative accuracy $(1 \pm 2^{-\Omega(1/\varepsilon)})$; this result matches an upper bound of Kong and Valiant [Annals of Statistics, 2017]. To obtain our result, we provide a hard instance involving distributions induced by the eigenvalue spectra of carefully constructed graph adjacency matrices. Efficiently approximating such spectra in Wasserstein-1 distance is a well-studied algorithmic problem, and a recent result of Cohen-Steiner et al. [KDD 2018] gives a method based on accurately approximating spectral moments using $2^{O(1/\varepsilon)}$ random walks initiated at uniformly random nodes in the graph.

As a strengthening of our main result, we show that improving the dependence on $1/\varepsilon$ in this result would require a new algorithmic approach. Specifically, no algorithm can compute an ε-accurate approximation to the spectrum of a normalized graph adjacency matrix with constant probability, even when given the transcript of $2^{\Omega(1/\varepsilon)}$ random walks of length $2^{\Omega(1/\varepsilon)}$ started at random nodes.

Keywords: spectral density estimation, moment methods, random walks, sublinear algorithm

1. Introduction

A fundamental problem in linear algebra is to approximate the full list of eigenvalues, $\lambda_1 \leq \ldots \leq \lambda_n \in \mathbb{R}$, of a symmetric matrix $A \in \mathbb{R}^{n\times n}$, ideally in less time than it takes to compute a full eigendecomposition.[1] We focus on the particular problem of *spectral density estimation* where given $\varepsilon \in (0, 1)$ and the assumption that $\|A\|_2 \leq 1$, the goal is find approximate eigenvalues $\lambda'_1 \leq \ldots \leq \lambda'_n$ such that their average absolute error is bounded by ε, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} |\lambda_i - \lambda'_i| \leq \varepsilon. \tag{1}$$

1. All eigenvalues can be computed to precision ε in $O(n^{\omega+\eta}\mathsf{polylog}(\frac{n}{\varepsilon}))$ time, where $\omega \approx 2.373$ is the matrix multiplication constant (Banks et al., 2020). Methods typically used in practice run in time $O(n^3 + n^2 \log(\frac{1}{\varepsilon}))$ (Wilkinson, 1968).

This problem is equivalent to that of computing an ε-approximation in Wasserstein-1 distance to the distribution on $[-1, 1]$ induced by the *spectral density (function)* of A, i.e. $p(x) := \frac{1}{n}\sum_{i=1}^{n}\delta(x-\lambda_i)$ for indicator function δ (see Section 2 for notation).

Spectral density estimation is distinct from and in many ways more challenging than related problems like low-rank approximation, where we only seek to approximate the *largest magnitude* eigenvalues of A. Nevertheless, efficient randomized algorithms for spectral density estimation were developed in the early 1990s and have been applied widely in computational physics and chemistry (Skilling, 1989; Silver and Röder, 1994; Wang, 1994; Weiße et al., 2006). These algorithms, which include the kernel polynomial and stochastic Lanczos quadrature methods, achieve ε accuracy with high probability in roughly $O(n^2/\varepsilon)$ time, improving on the $\Omega(n^\omega)$ cost of a full eigendecomposition for moderate values of ε (Chen et al., 2021).

More recently, there has been a resurgence of interest in spectral density estimation within the machine learning and data science communities. Research activity in this area has been fueled by emerging applications in analyzing and understanding deep neural networks (Pennington et al., 2018; Mahoney and Martin, 2019; Papyan, 2018), in optimization (Ghorbani et al., 2019; Sagun et al., 2017), and in network science (Dong et al., 2019; Cohen-Steiner et al., 2018).

1.1. Spectral Density Estimation for Graphs

Interestingly, when A is the normalized adjacency matrix[2] of an undirected graph G, there are faster spectral density estimation algorithms than for general matrices. Specifically, assume that we can randomly sample a node from G and, given a node, randomly sample a neighbor, both in $O(1)$ time. This is possible, for example, in the word RAM model when given arrays containing the neighbors for each node in G, and is also a commonly assumed access for computing on extremely large implicit networks (Katzir et al., 2011). It was recently shown that the $O(n^2/\varepsilon)$ runtime of general purpose algorithms like stochastic Lanczos quadrature can be improved to $\widetilde{O}(n/\mathsf{poly}(\varepsilon))$ (Braverman et al., 2022).[3] This runtime is sublinear in the size of A, e.g., when the matrix has $\Omega(n^2)$ non-zero entries.

Perhaps even more surprisingly, it is possible to solve spectral density estimation for normalized adjacency matrices without any dependence on n. Suppose that we are given a *weighted* graph G, and again that we can randomly sample a node from G in $O(1)$ time. Also assume that, for any given node, we can randomly sample a neighbor with probability proportional to its edge weight in $O(1)$ time. In other words, we can initialize and take steps of an edge-weighted random walk in G in $O(1)$ time.[4] Then Cohen-Steiner et al. (2018) gives an algorithm for any weighted undirected graph that solves the spectral density estimation problem with high probabilty in $2^{O(1/\varepsilon)}$ time[5]. While completely independent of the graph size, the poor dependence on ε in the result of Cohen-Steiner et al. (2018) unfortunately makes the algorithm impractical for any reasonable level of accuracy. As

2. If $\widetilde{A}$ is the unnormalized adjacency matrix of G and D is its diagonal degree matrix, we can equivalently consider the asymmetric matrix, $D^{-1}\widetilde{A}$ or the symmetric one, $D^{-1/2}\widetilde{A}D^{-1/2}$, as they have the same eigenvalues.
3. We use $\widetilde{O}(m)$ to denote $O(m \log m)$. The runtime in (Braverman et al., 2022) can be improved by a logarithmic factor to $O(n/\mathsf{poly}(\varepsilon))$ if we have access to a precomputed list of the degrees of nodes in G.
4. To be more concrete, if a node x is connected to neighbors $y_1, \ldots, y_d$ with edge weights $w_1, \ldots, w_d$, then the walk steps from x to y_i with probability $w_i / \sum_j w_j$.
5. Note that Cohen-Steiner et al. (2018) output a list of approximation eigenvalues $\lambda'_1, \ldots, \lambda'_n$ with only $O(1/\varepsilon)$ distinct values that can be stored and returned in time independent of n.

such, an interesting question is whether the exponential dependence on ε can be improved (maybe even to polynomial), while still avoiding any dependence on the graph size n.

Question 1 *Can we solve the spectral density estimation problem for a normalized adjacency matrix A given access to $2^{o(1/\varepsilon)}$ steps of random walks in the associated graph?*

Central to this question is the connection between spectral density estimation and the problem of learning a one dimensional distribution p given noisy measurements of p's (raw) moments. In this work, we consider distributions supported on the the $[-1, 1]$, in which case these moments are:

$$\int_{-1}^{1} xp(x)dx, \quad \int_{-1}^{1} x^2p(x)dx, \quad \int_{-1}^{1} x^3p(x)dx, \ldots .$$

Recent work of Kong and Valiant (2017) shows that, for a fixed constant c, if the first $\ell = c/\varepsilon$ moments of any two distributions p and q supported on $[-1, 1]$ match *exactly*, then the Wasserstein-1 distance between those distributions is at most ε. Given that the left hand size of (1) exactly equals the Wasserstein-1 distance $W_1(p, q)$ between the discrete distributions $p(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - \lambda_i)$ and $q(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x-\lambda'_i)$, the approach in Cohen-Steiner et al. (2018) is to approximate the first ℓ moments of p, and then to find a set of approximate eigenvalues and eigenvalue multiplicities that correspond to a discrete distribution q with the same moments. Given the approximate moments, finding q can be done in $\mathsf{poly}(\ell)$ time using linear programming algorithms.

Computing the estimates of p's moments is more challenging. Cohen-Steiner et al. (2018) take advantage of the fact that for any $j \leq \ell$, the j^{th} moment of p is equal to $\frac{1}{n}\sum_{i=1}^{n} \lambda_i^j = \frac{1}{n}\mathrm{tr}(A^j)$. This trace can in turn be estimated by random walks of length j in A: if we start a random walk at a random node v, the probability that we return to v at the j^{th} step is exactly equal to $\frac{1}{n}\mathrm{tr}(A^j)$. So, we can obtain an unbiased estimate for the j^{th} moment by simply running random walks from random starting nodes and calculating the empirical frequency that we return to our starting point.

This approach leads to the remarkably simple algorithm of Cohen-Steiner et al. (2018). So where does the $2^{O(1/\varepsilon)}$ runtime dependence come from? The issue is that the result of Kong and Valiant (2017) is brittle to noise. In particular, if the sum of squared distances between p's moments and q's moments differ by Δ, the bound from Kong and Valiant (2017) weakens, only showing that the Wasserstein-1 distance is bounded by $O(\frac{1}{\ell} + \Delta \cdot 3^{\ell})$. To obtain accuracy ε, it is necessary to set $\ell = O(1/\varepsilon)$ and thus Δ equal to $2^{-O(1/\varepsilon)}$. By standard concentration inequalities, to obtain such an accurate estimate to p's moments, we need to run an exponential number of random walks of length $1, \ldots, \ell$. Accordingly, an important step towards answering Question 1 is to understand if such extremely accurate estimates of the moments is necessary for spectral density estimation.

Note that many other spectral density estimation algorithms for general matrices are also based on moment-matching. A common approach is to use randomized trace estimation methods (Hutchinson, 1990; Meyer et al., 2021) to estimate moments of the form $\int_{-1}^{1} T_j(x)p(x)dx = \frac{1}{n}\mathrm{tr}(T_j(A))$, where $T_j(x)$ is a degree j polynomial, not equal to x^j. If T_j is the j^{th} Chebyshev or Legendre polynomial, then it can be shown that only $\mathsf{poly}(1/\varepsilon)$ accurate estimates of the first $\ell = c/\varepsilon$ moments are needed to approximate the spectral density to ε error in Wasserstein-1 distance (Braverman et al., 2022). A natural question then is, can these general polynomial moments be estimated using random walks in time independent of n for graph adjacency matrices? Unfortunately, it is not known how to do so: the challenge is that the ℓ^{th} Legendre polynomial or Chebyshev polynomial has coefficients exponentially large in ℓ, so $\mathrm{tr}(T_j(A))$ cannot be effectively approximated given a routine for approximating $\mathrm{tr}(A^j)$ for different powers j.

1.2. Our Contributions

In this paper, we answer Question 1 negatively. First, we show that exponentially accurate moments are necessary for estimating a distribution in Wasserstein-1 distance, even in the special case of distributions that arise as the spectral density of a graph adjacency matrix.

Theorem 1 *For any $\varepsilon \in (0, 1/4]$, there exist weighted graphs G_1 and G_2 (see Definition 7) with spectral densities p_1 and p_2, such that:*

- *The densities are far in Wasserstein-1 distance: $W_1(p_1, p_2) \geq \varepsilon$.*
- *For* all *positive integers j, moments $m_j(p_1) = \int_{-1}^{1} x^j p_1(x)\mathrm{d}x$ and $m_j(p_2) = \int_{-1}^{1} x^j p_2(x)\mathrm{d}x$ are exponentially close: $(1-\delta)m_j(p_1) \leq m_j(p_2) \leq (1+\delta)m_j(p_1)$ for some $\delta \leq 16 \cdot 2^{-1/4\varepsilon}$.*

Theorem 1 shows that Kong and Valiant (2017)'s requirement that each moment be estimated to accuracy $2^{-O(1/\varepsilon)}$ cannot be avoided if we want an ε accurate approximation in Wasserstein distance. It thus rules out a direct improvement to the analysis of the spectral density estimation algorithm from of Cohen-Steiner et al. (2018). In particular, even if we had a procedure that returned exponentially accurate multiplicative estimates to the moments of a graph's spectral density,[6] and even if it returns such estimates for *all* of the moments (not just the first $O(1/\varepsilon)$), then we would not be able to distinguish between G_1 and G_2.

Our proof of Theorem 1 is based on a hard instance built using cycle graphs. It is not hard to show that the spectral densities of two disjoint cycles of length $1/\varepsilon$ and of one cycle of length $2/\varepsilon$ differ by ε in Wasserstein-1 distance. Additionally, it can be shown that the first c/ε moments of these graphs are exponentially close. This example would thus prove Theorem 1 if we restricted our attention to moments of degree $j \leq c/\varepsilon$. However, for the cycle graph, higher moments can be more informative: for example, the j^{th} moment for $j = O(1/\varepsilon^2)$ can be shown to distinguish the cycles of different length, even when only estimated to polynomial additive accuracy. To see why this is the case, note that, since a random walk of length $O(1/\varepsilon^2)$ mixes on the cycle, the probability of it returning in the shorter cycle is roughly twice that as in the longer cycle.

To avoid this issue, we modify the cycle graph to diminish the value of higher degree moments. In particular, we force all high moments close to zero by creating a graph that consists of many disjoint cycles, either of length $1/\varepsilon$ or $2/\varepsilon$, joined by a lightweight complete graph on all nodes. If weighted correctly, then any walk of length $\Omega(1/\varepsilon)$ will exit the cycle it starts in (via the complete graph) with high probability, and the chance of returning to its starting point can be made extremely low by making the graph large enough. At the same time, the lower moments are not effected significantly, so we can show that the graphs remain far in Wasserstein-1 distance.

Theorem 1 has potentially interesting implications beyond showing a limitation for graph spectrum estimation. For example, related to the discussion about generalized moment methods above, it immediately implies that for any ℓ, the ℓ^{th} Chebyshev polynomial cannot be approximated to accuracy $1/\mathsf{poly}(\ell)$ with a polynomial (of any degree!) whose maximum coefficient is $\leq 2^{\ell}$. If it could, we could use less than exponentially accurate measures of the raw moments to approximate the Chebyshev moments, and then use these moments to approximate the spectral density, following Braverman et al. (2022). However, by Theorem 1, this is impossible.

6. When run for $O(1/\delta^2)$ steps, the random walk method of Cohen-Steiner et al. (2018) actually achieves a weaker moment approximation with additive rror δ. This is always greater than $\delta m_\ell(p_1)$ because all of p_1's moments are upper bounded by 1 since it is supported on $[-1, 1]$.

While Theorem 1 rules out direct improvements to the moment-based method of Cohen-Steiner et al. (2018), it does not rule out the possibility of *some other algorithm* that can estimate the spectral density to ε accuracy using fewer random walk steps. For example, we could consider methods that use more information about each random walk than checking whether or not the last step returns to the starting node. However, our next theorem shows that, in fact, *no such algorithm* can beat the exponential dependence on $1/\varepsilon$; we show that, information theoretically, $2^{\Omega(1/\varepsilon)}$ samples from random walks started from random nodes are necessary to estimate the spectral density accurately in Wasserstein-1 distance.

Theorem 2 *For any $\varepsilon < 1/2$, no algorithm that is given access to the transcript of m, length T random walks initiated at m uniformly random nodes in a given graph G can approximate G's spectral density to ε accuracy in the Wasserstein-1 distance with probability $> 3/4$, unless $m \cdot T > \frac{1}{16} \cdot 2^{1/4\varepsilon}$.*

While more technical, the proof of Theorem 2 is based on the same hard instance as Theorem 1. The distribution $\mathcal{D}$ is supported on two graphs that are ε far in Wasserstein distance: a collection of cycles of length $1/\varepsilon$ added to a lightweight complete graph, and a collection of cycles of length $2/\varepsilon$ added to a lightweight complete graph. We establish that, if node labels are assigned at random, the only way to distinguish between these graphs is to complete a walk around one of the cycles. We show that event happens with exponentially small probability for a random walk of any length.

1.3. Open Problems and Outlook

Our main results open a number of interesting directions for future inquiry. Most directly, the bound from Theorem 2 is based on an instance involving *weighted graphs*. It would be great to extend the lower bound to unweighted graphs, which are common in practice. While we believe the same lower bound should hold, such an extension is surprisingly tricky: for example, replacing the lightweight complete graph in our hard instances with, e.g., an unweighted expander graph significantly impacts the spectra of both graphs, making them more challenging to analyze.

A bigger open question is to extend our lower bounds to what we call the *adaptive random walk model*, which means that the algorithm is allowed to start a random walk either at a random node, or at any other node it wishes. Since this model allows for e.g. sampling random neighbors of any node, it is closely related to other access models. For example, up to logarithmic factors, the number of random walk steps required in the adaptive model is equal to the number of memory accesses needed when given access to data structure storing an array of neighbors for each node in the graph Braverman et al. (2022). Currently, the best lower bound we can prove in the adaptive random walk model is that just $\Omega(1/\varepsilon^2)$ steps are necessary; we show this result in Appendix A. Proving a lower bound exponential in $1/\varepsilon$ or finding a faster algorithm that runs in this model would be a nice contribution. Even a conjectured hard instance would be nice – currently we don't have any.

Finally, we note that our graph-based lower bounds show that, with non-adaptive random walks, it is impossible to distinguish if the spectral densities of two graphs are identical or ε-far away in Wasserstein-1 distance with $2^{o(1/\varepsilon)}$ steps. Consequently this result constitutes a particular type of hardness for comparing graphs. However, one might consider other notions of graph comparison. For example, in Appendix C, we consider estimating the spectrum of the difference $A_1 - A_2$ between two normalized adjaceny matrices A_1 graphs A_2 corresponding to graphs G_1 and G_2 with the same

node degress. We show that an $2^{O(1/\varepsilon)}$ upper bound is obtainable. Seeking matching upper and lower bounds for this and related problems is another interesting direction for future work.

1.4. Paper Organization

In Section 2 we introduce notation and preliminaries. In Section 3 we prove a lower bound for spectrum estimation based on moments, establishing Theorem 1. In Section 4 we prove lower bound for spectrum estimation based on random walks, establishing Theorem 2. In Appendix A, we give an $\Omega(1/\varepsilon^2)$ lower bound for approximating graph spectra in the (stronger) adaptive random walk model. In Appendix B, we use cycle graphs to construct distributions that are $2/\ell$ far in Wasserstein-1 distance and have the same first $\ell - 1$ moments, slightly strengthening a result from Kong and Valiant (2017). In Appendix C, we show a new algorithm that uses alternating random walks to estimate the spectrum of the difference of two normalized adjacency matrices.

2. Preliminaries

General notation. We use $\delta : \mathbb{R} \to \mathbb{R}$ to denote the indicator function with $\delta(0) := 1$ and $\delta(x) := 0$ for all $x \neq 0$. We use $\mathbf{1} \in \mathbb{R}^n$ to denote the all ones vector when n is clear from context. We use $\mathbb{P}[E]$ to denote the probability of an event E. We let E^c denote the complement of a random event E, so $\mathbb{P}[E^c] = 1 - \mathbb{P}[E]$.

Graphs and graph spectra. We consider undirected graphs $G = (V, E)$ where each edge $e \in E$ has a non-negative weight $w_e \in \mathbb{R}_{\geq 0}$. We call G unweighted when $w_e = 1$ for all $e \in E$. We use $\widetilde{A} \in \mathbb{R}_{\geq 0}^{V \times V}$ to denote the weighted adjacency matrix of G where $\widetilde{A}(v, v') = w_e$ if $e = (v, v') \in E$ and $\widetilde{A}(v, v') = 0$ otherwise. We use $D \in \mathbb{R}_{\geq 0}^{V \times V}$ to denote the diagonal degree matrix of G where D is diagonal with $D(v, v) := \sum_{e=(v,v') \in E} w_e$ for all $v \in V$. We let $A(G) \in \mathbb{R}^{V \times V}$ denote the normalized adjacency matrix of G, i.e. $A(G) := D^{-1/2} \widetilde{A} D^{-1/2}$. We refer to $D^{-1} \widetilde{A}$ as the random walk matrix and note that, for degree-regular graphs, $A(G) = D^{-1} \widetilde{A}$.

For an n-vertex graph G, we let $-1 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n \leq 1$ be the eigenvalues of the normalized adjacency matrix $A(G)$, and use $\boldsymbol{\lambda} = \boldsymbol{\lambda}(G)$ to denote this sorted (in ascending order) eigenvalue list. We let $p(x) : [-1, 1] \to [0, 1]$ denote the spectral density of G, i.e., $p(x) = \frac{1}{n} \sum_{i \in [n]} \delta(x - \lambda_i)$, which is the density of the distribution on $[-1, 1]$ induced by λ_i (for brevity, we do not distinguish between spectral density and the distribution it induces). We use $m_j(p)$ to denote the j^{th} moment of p, i.e., $m_j(p) = \frac{1}{n} \text{tr}(A(G)^j)$.

Wasserstein distance. In this work, we consider the standard Wasserstein-1 distance between distributions, which we may simply refer to as the Wasserstein distance for brevity.

Definition 3 *The Wasserstein-1 distance $W_1(p_1, p_2)$ between two distributions, p_1 and p_2, supported on the real line is defined as the minimum cost of moving probability mass in p_1 to p_2, where the cost of moving probability mass from value a to b is $|a - b|$. Concretely, let Ψ be the set of all couplings $\psi(x, y)$ between p_1 and p_2, i.e., Ψ contains all joint distributions $\psi(x, y)$ over $x \in \mathbb{R}$ and $y \in \mathbb{R}$ with marginals equal to p_1 and p_2. Then:*

$$W_1(p_1, p_2) = \min_{\psi \in \Psi} \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| \cdot \psi(x, y) \, \mathsf{d}x \, \mathsf{d}y \, .$$

A well known fact is that the Wasserstein-1 distance has a dual characterization. Specifically,

Fact 4 (Kantorovich-Rubinstein Duality Kantorovich (1940, 1942))

$$W_1(p_1, p_2) = \sup_{f:1-\text{Lipschitz}} \int_{\mathbb{R}} f(x) \cdot (p_1(x) - p_2(x)) \mathsf{d}x. \tag{2}$$

Above, the supremum is taken over all 1-Lipschitz functions f, i.e., that satify $|f(a)-f(b)| \leq |a-b|$ for all $a, b \in \mathbb{R}$. Overloading notation, for graphs G_1 and G_2 with spectral densities p_1 and p_2 respectively, we let $W_1(G_1, G_2) := W_1(p_1, p_2)$ to denote the Wasserstein-1 distance between p_1 and p_2. We note that, for any two n-vertex graphs G_1 and G_2, it can be checked (see, e.g. Kong and Valiant (2017)) that:

$$W_1(G_1, G_2) = \frac{1}{n} \left\| \boldsymbol{\lambda}(G_1) - \boldsymbol{\lambda}(G_2) \right\|_1 . \tag{3}$$

Access models. As discussed in the introduction, we consider several possible data access models for estimating the spectral density of a normalized graph adjacency matrix, $A(G)$ for $G = (V, E, w)$. First, we consider algorithms that, for some integer $j \geq 0$ and accuracy parameter δ, have access to δ-accurate approximations, $\widetilde{m}_1, \ldots, \widetilde{m}_j$, to the first j moments of G's spectral density p, $m_1(p), \ldots, m_j(p)$. Specifically, we have that $|\widetilde{m}_j - m_j(p)| \leq \delta \cdot m_j(p)$.

A natural generalization of the setting where approximate moments are available is to consider algorithms that access G via random walks, since repeated random walks can be used to approximate moments Cohen-Steiner et al. (2018). In this work, we primarily consider a *non-adaptive random walk model*, where the algorithm can run m random walks each of length $T \geq 1$, starting at m vertices $v_0^{(1)}, \ldots, v_0^{(m)}$ chosen uniformly at random from G. For each walk, the algorithm can observe the entire sequence of vertex labels visited in order We call this information the the walk "transcript" and denote the set of transcripts by $S = \{S_1, \cdots, S_m\}$. Note that, at vertex v, the probability that the next vertex in the random walk is equal to v' is the (v, v') entry of $D^{-1}\widetilde{A}$.

In Appendix A, we also consider the richer random walk model that we refer to as the *adaptive random walk model* where the algorithm can choose the starting node $v_0^{(1)}, \ldots, v_0^{(m)}$. This is in contrast to the *non-adaptive random walk model* where starting nodes are uniformly random.

Cycle spectra. Our lower bound instances in this paper involve collections of cycle graphs. We let R_c denote an undirected cycle graph of length c, and we let R_c^k denote a collection of k such cycles. Recall that we use $A(R_c^k)$ to denote the normalized adjacency matrix and $\boldsymbol{\lambda}(R_c^k)$ for a sorted list of eigenvalues for the normalized adjacency matrix. We leverage the following basic lemma on the spectrum of cycle graphs.

Lemma 5 (Eigenvalues of cycle graph) *For any odd integer ℓ, the eigenvalues of $A(R_\ell)$ are $\cos(\frac{2k}{\ell}\pi)$ with multiplicity 2 for $0 < k < \frac{\ell}{2}$ and 1 with multiplicity 1. The eigenvalues of $A(R_{2\ell})$ are $\cos(\frac{k}{\ell}\pi)$ with multiplicity 2 for $0 < k < \ell$ and ± 1 each with multiplicity 1. Further, we have $W_1(R_\ell^2, R_{2\ell}) = 1/\ell$.*

Proof The eigenvalues of the normalized adjacency matrix of cycle graphs are well known and can be found, e.g., in Spielman (2019). The Wasserstein distance immediately follows since we have:

$$\begin{aligned}\|\boldsymbol{\lambda}(R_\ell^2) - \boldsymbol{\lambda}(R_{2\ell})\|_1 &= |1 - \cos(\pi/\ell)| + |\cos(2\pi/\ell) - \cos(\pi/\ell)| + \cdots + |-1 - \cos(\pi(\ell-1)/\ell)| \\ &= 1 - \cos(\pi/\ell) + \cos(\pi/\ell) - \cos(2\pi/\ell) + \cdots + \cos(\pi(\ell-1)/\ell) - (-1) = 2\,.\end{aligned}$$

■

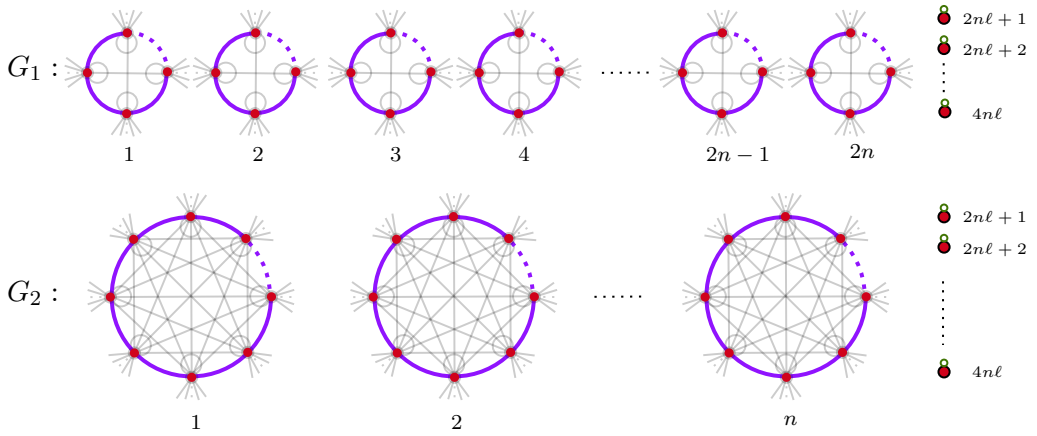

Figure 1: Diagram depicting graphs G_1 and G_2 from Definition 7. G_1 contains $2n$ cycles of length ℓ and $2n\ell$ isolated vertices, and G_2 contains n cycles of length 2ℓ and $2n\ell$ isolated vertices. In both graphs, the purple edges have weight $1/4 + 1/(4n\ell)$, the grey edges connect all the vertices not connected by purple edges (including self-loops), with weight $1/(4n\ell)$, and the green edges are self-loops of weight 1.

Remark 6 *The first $j < \ell$ moments of the spectral density of R_ℓ^2 and $R_{2\ell}$ are the same. This is true because the number of ways a walk of length $j < \ell$ can return to its starting node is the same in both R_ℓ^2 and $R_{2\ell}$: $2 \cdot \binom{j}{j/2}$ for even j and 0 for odd j.*

3. Limits on Moment Estimation Methods

In this section, we construct two weighted graphs G_1, G_2 with a same number of vertices, i.e., $|V_1| = |V_2|$, that we prove are ε-far in Wasserstein distance but have exponentially close moments. We detail the construction in the definition below.

Definition 7 *G_1 is constructed by starting with a collection of $2n\ell$ isolated vertices and $2n$ disjoint cycles, each of size ℓ. G_2 is constructed by starting with a collection $2n\ell$ isolated vertices and n disjoint cycles, each of size 2ℓ. In both graphs, the edges in the cycle have weight $1/4$ and every vertex in a cycle is then connected to all other cycle vertices with weight $1/(4n\ell)$ (including a self-loop); the isolated vertices only have self-loop with weight 1. We choose ℓ to be an odd number and let $n = \lceil 2^\ell/4 \rceil$. Note that each graph has $4n\ell$ vertices.*

See Figure 1 for a visual representation of the construction from Definition 7 and Figure 2 for a plot of the spectra of G_1 and G_2. We bound the Wasserstein distance between these spectra below.

Lemma 8 *For weighted graphs G_1, G_2 constructed in Definition 7, $W_1(G_1, G_2) = 1/(4\ell)$.*

Proof Let $\mathbf{I}$ denote a $2n\ell \times 2n\ell$ identity matrix. The normalized adjacency matrices of the two graphs are

$$A(G_1) = \begin{bmatrix} \frac{1}{2} \cdot A(R_\ell^{2n}) + \frac{1}{2} \cdot \frac{1}{2n\ell} \cdot \mathbf{1}\mathbf{1}^\top & 0 \\ 0 & \mathbf{I} \end{bmatrix}, \text{ and } A(G_2) = \begin{bmatrix} \frac{1}{2} \cdot A(R_{2\ell}^{n}) + \frac{1}{2} \cdot \frac{1}{2n\ell} \cdot \mathbf{1}\mathbf{1}^\top & 0 \\ 0 & \mathbf{I} \end{bmatrix}.$$

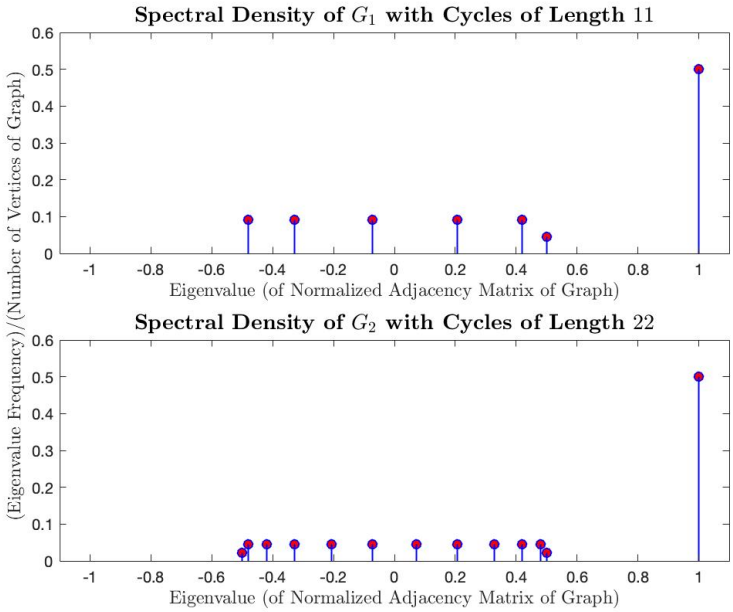

Figure 2: Spectral Density of G_1 and G_2 as defined in Definition 7, with cycles of length 11 and 22, respectively.

Recall that we use R_ℓ^{2n} to denote the graph of $2n$ disjoint cycles of size ℓ, and $R_{2\ell}^n$ to denote the graph of n disjoint cycles of size 2ℓ, respectively. Additionally, recall we use $\boldsymbol{\lambda}(G_1)$ and $\boldsymbol{\lambda}(G_2)$ to denote the sorted (in ascending order) eigenvalues of $A(G_1)$ and $A(G_2)$, and $\boldsymbol{\lambda}(R_\ell^{2n})$ and $\boldsymbol{\lambda}(R_{2\ell}^n)$ for the sorted eigenvalue list of $A(R_\ell^{2n})$ and $A(R_{2\ell}^n)$, respectively.

Since $A(R_\ell^{2n})$ and $A(R_{2\ell}^n)$ are regular graphs and both commute with $\mathbf{1}\mathbf{1}^\top$, they both share the same eigenvectors with $\mathbf{1}\mathbf{1}^\top$. For simplicity of notation we let $\mathcal{R}_1 := R_\ell^{2n}$, $\mathcal{R}_2 = R_{2\ell}^n$. For $i \in [2]$ we have:

$$\boldsymbol{\lambda}_j(G_i) = \begin{cases} \frac{1}{2}\boldsymbol{\lambda}_j(\mathcal{R}_i) & \text{for } j \in \{1, 2, \cdots, 2n\ell - 1\} \\ 1 & \text{for } j \in \{2n\ell, 2n\ell + 1, \cdots, 4n\ell\}\,. \end{cases} \tag{4}$$

This implies $W_1(G_1, G_2) = \frac{1}{4n\ell} \left\|\boldsymbol{\lambda}(G_1) - \boldsymbol{\lambda}(G_2)\right\|_1 = \frac{1}{2} \cdot \frac{1}{4n\ell} \left\|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\right\|_1$ by the characterization of Wasserstein distance given in (3). Thus it suffices to calculate $\left\|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\right\|_1$. Since these are disjoint cycles, we only need to focus on the Wasserstein distance between a cycle of size 2ℓ and 2 disjoint cycles of size ℓ. Applying Lemma 5, we get $\left\|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\right\|_1 = n \cdot 2\ell \cdot W_1(R_\ell^2, R_{2\ell}) = 2n$. Plugging this back we get the claimed Wasserstein distance $W_1(G_1, G_2) = 1/(4\ell)$. ■

Next we show that the moments of the constructed graphs G_1 and G_2 are exponentially close.

Lemma 9 *Let G_1 and G_2 be weighted graphs as constructed in Definition 7. Let p_1, p_2 be the spectral density of G_1, G_2 respectively. It holds that $m_j(p_i) \in [1/2, 1]$ for all $j \geq 0, i = 1, 2$ and*

also

$$|m_j(p_1) - m_j(p_2)| = 0 \text{ for } j < \ell \qquad \text{and} \qquad |m_j(p_1) - m_j(p_2)| \leq 2^{-\ell+1} \text{ for } j \geq \ell.$$

Proof For the first claim, we note that $m_j(p_i) \geq \frac{2n\ell}{4n\ell} \cdot 1^j \geq \frac{1}{2}$. The upper bound of 1 follows trivially given boundedness of all eigenvalues of normalized adjacency matrices.

For $j \geq \ell$, and $i \in [2]$, we also have

$$m_j(p_i) \leq \frac{2n\ell + 1}{4n\ell} \cdot 1^j + \frac{2n\ell - 1}{4n\ell} \cdot \left(\frac{1}{2}\right)^j \leq \frac{1}{2} + \frac{1}{4n\ell} + \frac{1}{2^{j+1}}.$$

Thus, we can immediately conclude that $m_j(p_i) \in [\frac{1}{2}, \frac{1}{2} + \frac{1}{2^{j+1}} + \frac{1}{4n\ell}]$ and obtain the claimed bounds for $j \geq \ell$ by plugging in the choice of $n \geq 2^\ell/4$.

For $j < \ell$, we use the fact that $m_j(p_i) = \frac{1}{n}\text{tr}(A(G_i)^j)$. Using (4) we can calculate:

$$\begin{aligned} |m_j(p_1) - m_j(p_2)| &= \frac{1}{4n\ell} \cdot \left| \sum_{i=1}^{2n\ell-1} \frac{\boldsymbol{\lambda}_i^j\left(R_\ell^{2n}\right)}{2^j} + (2n\ell+1) - \sum_{i=1}^{2n\ell-1} \frac{\boldsymbol{\lambda}_i^j\left(R_{2\ell}^{n}\right)}{2^j} - (2n\ell+1) \right| \\ &= \frac{1}{4n\ell} \cdot \left| \sum_{i=1}^{2n\ell} \frac{\boldsymbol{\lambda}_i^j\left(R_\ell^{2n}\right) - \boldsymbol{\lambda}_i^j\left(R_{2\ell}^{n}\right)}{2^j} \right|. \end{aligned} \tag{5}$$

Since R_ℓ^{2n} and $R_{2\ell}^n$ are disjoint cycles, the moments of the spectral density of R_ℓ^{2n} and R_ℓ^{2n} are the same as the moments of the spectral density of R_ℓ^2 and $R_{2\ell}$. This is true because the eigenvalues of the disjoint copies of $A(R_\ell^{2n})$ and $A(R_\ell^{2n})$ are the same as the eigenvalues of the disjoint copies of $A(R_\ell^2)$ and $A(R_{2\ell})$ with increased multiplicity, which is scaled by the size of the respective graphs. Since the first $j < \ell$ moments of R_ℓ^2 and $R_{2\ell}$ are the same (see Remark 6), we get from (5) that

$$|m_j(p_1) - m_j(p_2)| = \frac{1}{4n\ell} \cdot \left| \sum_{i=1}^{2n\ell} \frac{\boldsymbol{\lambda}_i\left(R_\ell^{2n}\right)^j - \boldsymbol{\lambda}_i\left(R_{2\ell}^{n}\right)^j}{2^j} \right| = 0\,.$$

■

We briefly remark that the proof of Lemma 9 required picking a value of n that is exponentially large in ℓ to ensure that when a random walk leaves the cycle it started from, it only comes back to the same cycle with a very low probability. Otherwise, we would not have been able to show that the higher moments of G_1 and G_2 ($j \geq \ell$) are close.

Theorem 1 *For any $\varepsilon \in (0, 1/4]$, there exist weighted graphs G_1 and G_2 (see Definition 7) with spectral densities p_1 and p_2, such that:*

- *The densities are far in Wasserstein-1 distance: $W_1(p_1, p_2) \geq \varepsilon$.*
- *For* all *positive integers j, moments $m_j(p_1) = \int_{-1}^{1} x^j p_1(x)\mathrm{d}x$ and $m_j(p_2) = \int_{-1}^{1} x^j p_2(x)\mathrm{d}x$ are exponentially close: $(1-\delta)m_j(p_1) \leq m_j(p_2) \leq (1+\delta)m_j(p_1)$ for some $\delta \leq 16 \cdot 2^{-1/4\varepsilon}$.*

Proof The proof of the first statement follows by substituting ℓ with the largest odd integer smaller than $1/(4\varepsilon)$ in Lemma 8. Next, we know that for all j, $m_j(p_1) \in [1/2, 1]$. So, by Lemma 9, $|m_j(p_1) - m_j(p_2)| \leq 2^{-\ell+2} m_j(p_1)$. The statement holds since we have $\ell \geq 1/(4\varepsilon) - 2$. ■

4. Limits on Random Walk Methods

In this section, we prove Theorem 2, which can be viewed as a strengthening of Theorem 1. While Theorem 1 rules out directly improving SDE algorithms like that of Cohen-Steiner et al. (2018) based on estimating moments, Theorem 2 shows that no method that performs less than $2^{O(1/\varepsilon)}$ steps of non-adaptive random walks in a graph can reliably estimate the spectral density to error ε in Wasserstein distance, whether or not the algorithm is based on moment estimation or not.

To prove Theorem 2, we construct a hard pair of graphs that are ε far in Wasserstein distance, but difficult to distinguish based on random walks. This pair is identical to the hard instance constructed in the previous section, although without isolated nodes. These nodes were necessary to show that even accurate *relative error* moment estimates do not suffice for spectral density estimation. However, they are not needed for the random-walk lower bound, and eliminating them simplifies the analysis. Formally, the construction is as follows:

Definition 10 *G_1 is constructed by starting with a collection of $2n$ disjoint cycles, each having size ℓ for odd integer ℓ. The edges in the cycle have weight $1/4$. After constructing the cycles, we connect every vertex in G_1 to all other vertices with weight $1/(4n\ell)$ (including a self-loop). G_2 is constructed by starting with a collection of n disjoint cycles, each having size 2ℓ. The edges in the cycle have weight $1/4$ and every vertex is then connected to all other vertices with weight $1/(4n\ell)$ (including a self-loop). We choose $n = 2 \cdot 2^{2\ell}$.*

Lemma 11 *For weighted graphs G_1, G_2 constructed in Definition 10, $W_1(G_1, G_2) \geq 1/(2\ell)$.*

Proof The normalized adjacency matrices of the two graphs are:

$$A(G_1) = \frac{1}{2} \cdot A(R_\ell^{2n}) + \frac{1}{4n\ell}\mathbf{1}\mathbf{1}^\top \text{ and } A(G_2) = \frac{1}{2} \cdot A(R_{2\ell}^{n}) + \frac{1}{4n\ell} \cdot \mathbf{1}\mathbf{1}^\top.$$

As before, since $A(R_\ell^{2n})$ and $A(R_{2\ell}^{n})$ are degree-regular graphs and both commute with $\mathbf{1}\mathbf{1}^\top$, we can write the sorted vector of eigenvalues $\boldsymbol{\lambda}(G_1), \boldsymbol{\lambda}(G_2)$ of $A(G_1), A(G_2)$ as

$$\boldsymbol{\lambda}_j(G_1) = \begin{cases} \frac{1}{2}\boldsymbol{\lambda}_j(R_\ell^{2n}) & \text{for } j \in \{1, \cdots, 2n\ell - 1\} \\ 1 & \text{for } j = 2n\ell\,, \end{cases}$$

$$\text{and } \boldsymbol{\lambda}_j(G_2) = \begin{cases} \frac{1}{2}\boldsymbol{\lambda}_j(R_{2\ell}^{n}) & \text{for } j \in \{1, \cdots, 2n\ell - 1\} \\ 1 & \text{for } j = 2n\ell\,. \end{cases}$$

Since the top eigenvalues of $R_{2\ell}^n$ and R_ℓ^{2n} are the same, we conclude that $W_1(G_1, G_2) = \frac{1}{2n\ell} \cdot \left(\frac{1}{2} \cdot \left\|\boldsymbol{\lambda}(R_\ell^{2n}) - \boldsymbol{\lambda}(R_{2\ell}^{n})\right\|_1\right)$. Applying Lemma 5, we have that $\left\|\boldsymbol{\lambda}(R_\ell^{2n}) - \boldsymbol{\lambda}(R_{2\ell}^{n})\right\|_1 = n \cdot 2\ell \cdot W_1(R_\ell^2, R_{2\ell}) = 2n$. Plugging in, we conclude that $W_1(G_1, G_2) = 1/(2\ell)$. ■

We next show that the transcripts of randomly started, non-adaptive random walks generated on G_1 and G_2 have similar distributions.

Definition 12 *For m non-adaptive random walks, each with length T, a random walk transcript S is a collection of m individual walk, $S = \{S_1, \ldots, S_m\}$ where each S_i consists of a list of T node labels $v_{i,1}, \ldots v_{i,T}$ (the nodes visited in the walk). Let $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_2}$ denote probability distribution over random walk transcripts generated when walking in G_1 and G_2, respectively, with nodes labeled using a uniform random permutation of the integers $1, \ldots, 2n\ell$.*

Our main result is as follows:

Lemma 13 *For m non-adaptive walks of length T, the total variation distance between $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_2}$ (Definition 12) is bounded by $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_{G_1}, \mathcal{D}_{G_2}) \leq \frac{2m^2T^2}{n} + \frac{mT}{2^\ell}$.*

To prove Lemma 13, we define a coupling between $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_2}$. We then show that, with high probability, the coupling outputs an identical transcript in both graphs. This establishes closeness in TV distance via the standard coupling lemma, which we state specialized to our setting below:

Fact 14 *Let $\mathcal{D}$ be any distribution over pairs of random walk transcripts S^1 and S^2 such that the marginal distribution of S^1 equals $\mathcal{D}_{G_1}$ and the marginal distribution of S^2 equals $\mathcal{D}_{G_2}$. Then:*

$$\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_{G_1}, \mathcal{D}_{G_2}) \leq \mathbb{P}_{\mathcal{D}}[S^1 \neq S^2].$$

Proof [Proof of Lemma 13] We define a coupling $\mathcal{D}$ by describing a process that explicitly generates two random walk transcripts S^1 and S^2 which are distributed according to $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_1}$. To do so, we use a "lazy labelling" procedure that randomly labels nodes as they are visited in the random walks. To support that labeling, we define two dictionaries, $L_1 : V_1 \to 1, \ldots, 2n\ell$ and $L_2 : V_2 \to 1, \ldots, 2n\ell$ that maps the vertex sets of G_1 and G_2 (denoted as V_1 and V_2) to labels. Initially, $L_i(v)$ returns NULL for any vertex $v \in V_i$. However, if we set $L_i(v) \leftarrow j$ for a label j, then for all future calls to the dictionary, $L_i(v)$ returns j. Additionally, in our description of the coupling we will refer to the "cycle" that a node v lies in (in G_1 or G_2) and to v's "left neighbor" and "right neighbor". Referring to Definition 10, these terms refer to the cycle that v would be in if the lightweight copy of the complete graph had not been added to to the graph, and respectively to v's neighbors in that cycle.

With this notation in place, we describe the coupling procedure below.

1. Choose a random permutation Π of the labels $1, \ldots, 2n\ell$. Let $\Pi(j)$ denote the j^{th} label in the permutation. Initialize $j \leftarrow 1$.

2. For $k = 1, \ldots, m$:

   (a) Choose independent, uniformly random nodes $v^1_{k,1}$ in G_1 and $v^2_{k,1}$ in G_2. If $L_1(v^1_{k,1}) =$ NULL (which means that the node has never been visited before in any of our $k-1$ previous random walks) set $L_1(v^1_{k,1}) \leftarrow \Pi(j)$. Likewise, if $L_2(v^2_{k,1}) =$ NULL, set $L_2(v^2_{k,1}) \leftarrow \Pi(j)$. Increment $j \leftarrow j + 1$.

   (b) For $i = 1, \ldots, T$

      i. With probability $\frac{1}{4}$ let $v^1_{k,i+1}$ be the right neighbor of $v^1_{k,i}$ in G_1 and let $v^2_{k,i+1}$ be the right neighbor of $v^2_{k,i}$ in G_2. With probability $\frac{1}{4}$ let $v^1_{k,i+1}$ be the left neighbor of $v^1_{k,i}$ in G_1 and let $v^2_{k,i+1}$ be the left neighbor of $v^2_{k,i}$ in G_2. With probability $\frac{1}{2}$, let $v^2_{k,i+1}$ and $v^2_{k,i+1}$ be uniformly random nodes in G_1 and G_2, respectively. In this last case, which we refer to as the RESET case, $v^2_{k,i+1}$ and $v^2_{k,i+1}$ can be chosen independently.

      ii. If $L_1(v^1_{k,i+1}) =$ NULL, set $L_1(v^1_{k,i+1}) \leftarrow \Pi(j)$. Likewise, if $L_2(v^2_{k,i+1}) =$ NULL, set $L_2(v^1_{k,i+1}) \leftarrow \Pi(j)$. Increment $j = j + 1$.

3. Return

$$S^1 = \{\{L_1(v^1_{1,1}), \ldots, L_1(v^1_{1,T})\}, \ldots, \{L_1(v^1_{m,1}), \ldots, L_1(v^1_{m,T})\}\},$$
$$S^2 = \{\{L_2(v^1_{2,1}), \ldots, L_2(v^1_{2,T})\}, \ldots, \{L_2(v^2_{m,1}), \ldots, L_2(v^2_{m,T})\}\}.$$

We first observe that the above process is a coupling, as it returns S^1 sampled from $\mathcal{D}_{G_1}$ and S^2 sampled from $\mathcal{D}_{G_2}$. So, we are left to argue that, with high probability, $S^1 = S^2$.

To do so, we use the fact that the transcripts are identical if two events hold. To define these events, note that each walk in each transcript begins in a cycle in G_1 or G_2, and then takes a random number of steps left and right in that cycle until "resetting" with probability $1/2$ to a uniformly random node in the graph (which could bring the walk to a new cycle, the same cycle it is currently in, or a cycle visited previously). For transcript S^1, let $R^1_1, \ldots, R^1_q$ denote the list of cycles visited between each RESET step across all m walks in that transcript. Likewise, let $R^2_1, \ldots, R^2_q$ denote the set of cycles visited in S^2. S^1 and S^2 are always identical if the following events occur:

Event 1: For all $j \neq k$, $R^1_j \neq R^1_k$ and $R^2_j \neq R^2_k$.

Event 2: For all $j \in 1, \ldots, s$, we take fewer than ℓ left/right steps in R^1_j and R^2_j before a RESET.

To see why this is the case, note that, if Event 1 occurs, the only way that S^1 and S^2 would differ is if, while random walking in R^1_j and R^2_j, we move to nodes v^1 and v^2 where $L_1(v^1)$ is defined but $L_2(v^2)$ is NULL, or vice-versa. However, the only way this can happen is if we complete an entire loop around R^1_j, and thus return to a node that was previously labeled. Since each R^1_j has ℓ nodes, such a loop cannot be completed if we always take $< \ell$ steps before resetting to a new cycle.

We proceed to show that Event 1 and Event 2 both occur with high probability. First, consider Event 1. We take at most mT RESET steps across all m random walks. At each step, the probability we return to a cycle we had previously visited is at most $\frac{mT}{2n}$ for the walk in G_1 and at most $\frac{mT}{n}$ for the walk in G_2. So, by a union bound, we do not return to any previously visited cycle after a RESET in either walk with probability:

$$\mathbb{P}\left[\text{Event 1}\right] \geq 1 - mT \cdot \frac{mT}{2n} - mT \cdot \frac{mT}{n} \geq 1 - \frac{2m^2T^2}{n}. \tag{6}$$

Next, consider Event 2. Note that, since we take a left/right step with probability $1/2$ (and RESET with probability $1/2$) the chance that we take ℓ steps or more in a given cycle is equal to $(1/2)^\ell$. We visit at most $q = mT$ cycles, so by a union bound, we take less then ℓ left/right steps in each cycle with probability:

$$\mathbb{P}\left[\text{Event 2}\right] \geq 1 - \frac{mT}{2^\ell}. \tag{7}$$

Combining (6) and (7) with a union bound, we conclude that both events hold, and thus $S^1 = S^2$ with probability at least $1 - \frac{2m^2T^2}{n} - \frac{mT}{2^\ell}$. Combined with Fact 14, this proves the lemma. ■

With Lemma 13 in place, we can prove our main lower bound result for non-adaptive random walks:

Theorem 2 *For any $\varepsilon < 1/2$, no algorithm that is given access to the transcript of m, length T random walks initiated at m uniformly random nodes in a given graph G can approximate G's spectral density to ε accuracy in the Wasserstein-1 distance with probability $> 3/4$, unless $m \cdot T > \frac{1}{16} \cdot 2^{1/4\varepsilon}$.*

Proof The theorem follows from Lemma 11 and Lemma 13. In particular, choose ℓ to equal the largest odd integer smaller than $1/4\varepsilon$ and choose $n = 2 \cdot 2^{2\ell}$. Then consider graphs G_1 and G_2 generated as in Definition 10 with random node labels. By Lemma 11, $W_1(G_1, G_2) > 2\varepsilon$. So, there is no distribution p that is ε-close in Wasserstein distance to both the spectral density of G_1 and G_2. Accordingly, any algorithm that estimates the SDE of a graph to error ε with probability $3/4$ can be used to correctly distinguish samples from $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_2}$ with probability $3/4$. However, with n set as above, we have that $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_{G_1}, \mathcal{D}_{G_2}) \leq \frac{2m^2T^2}{n} + \frac{mT}{2^\ell} = \frac{m^2T^2}{2^{2\ell}} + \frac{mT}{2^\ell}$. And thus we can check that $\mathrm{d}_{\mathrm{TV}}(\mathcal{D}_{G_1}, \mathcal{D}_{G_2}) \leq 1/2$ whenever $mT \leq \frac{1}{4}2^{1/4\varepsilon - 2}$. As is standard, no algorithm can distinguish between samples from two distributions with TV distance δ with probability greater than $\frac{1}{2} + \frac{1}{2}\delta$, which establishes the result: any method that correctly distinguishes $\mathcal{D}_{G_1}$ and $\mathcal{D}_{G_2}$ with probability $> 3/4$ must use $mT > \frac{1}{4}2^{1/4\varepsilon - 2}$ random walk steps. ■

Acknowledgements

We would like to thank Aditya Krishnan for early discussions on the questions addressed in this paper. Aaron Sidford was supported by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research award, and a Sloan Research Fellowship. This work was also supported by NSF Award CCF-2045590.

References

Myron B. Allen and Eli L. Isaacson. *Chebyshev Polynomials*, pages 555–557. John Wiley & Sons, Ltd, 2019. ISBN 9781119245476.

Jess Banks, Jorge Garza-Vargas, Archit Kulkarni, and Nikhil Srivastava. Pseudospectral shattering, the sign function, and diagonalization in nearly matrix multiplication time. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 529–540, 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)*, pages 1144–1157, 2022.

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.

David Cohen-Steiner, Weihao Kong, Christian Sohler, and Gregory Valiant. Approximating the spectrum of a graph. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1263–1271, 2018.

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.

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)*, volume 97, pages 2232–2241, 2019.

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.

Leonid V Kantorovich. On an effective method of solving certain classes of extremal problems. *Dokl. Akad. Nauk., USSR*, 28:212–215, 1940.

Leonid V Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk., USSR*, 37:199–201, 1942. English translation in J. Math. Sci. 133, 4 (2006), 1381–1382.

Liran Katzir, Edo Liberty, and Oren Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International World Wide Web Conference (WWW)*, pages 597–606, 2011.

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

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)*, volume 97, pages 4284–4293, 2019.

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.

Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with SGD training and sample size. *arXiv*, 2018.

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.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv*, 2017.

R.N. Silver and H. Röder. Densities of states of mega-dimensional hamiltonian matrices. *International Journal of Modern Physics C*, 5(4):735–753, 1994.

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

Daniel Spielman. *Spectral and Algebraic Graph Theory*. Unpublished Manuscript, 2019.

Lin-Wang Wang. Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method. *Phys. Rev. B*, 49:10154–10158, 1994.

Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. The kernel polynomial method. *Reviews of modern physics*, 78(1):275, 2006.

J.H. Wilkinson. Global convergene of tridiagonal qr algorithm with origin shifts. *Linear Algebra and its Applications*, 1(3):409–420, 1968.

Appendix A. Lower Bound for the Adaptive Random Walk Model

In this section, we consider lower bounds against a possibly richer class of spectral density estimation algorithms that can access graphs via *adaptive* random walks. Specifically, the algorithm is allowed to start random walks (of any length) at any node of its choosing and can store the entire transcript of these walks. In the adaptive model, the algorithm also has the ability to uniformly sample nodes from the graph, as in the non-adaptive random walk model considered for Theorem 2.

Interestingly, an adaptive algorithm can solve the hard instance from Theorem 2 using roughly $O(\log(1/\varepsilon)/\varepsilon)$ random walk steps. Specifically, for any node, the algorithm can identify its adjacent cycle nodes with high probability by taking a logarithmic number of 1-step random walks and identifying the two nodes that are visited most frequently. This allows it to walk one way around the cycle, check its length, and thus distinguish between G_1 and G_2.

Proving a lower bound in the adaptive random walk setting appears to be much harder than the non-adaptive setting, and we do not have any proposed constructions that we conjecture could establish that $2^{O(1/\varepsilon)}$ random walk steps are necessary. However, in this section we give a simple argument for a lower bound of $\Omega(1/\varepsilon^2)$ steps. The lower bound is via a reduction to a natural sampling problem, introduced below.

Problem 15 *For a parameter $\alpha \in (1/2, 1)$ and integer n, suppose we have a jar that contains either $\alpha \cdot n$ red marbles and $(1-\alpha) \cdot n$ blue marbles (Case 1) or contains $(1-\alpha) \cdot n$ red marbles and $\alpha \cdot n$ blue marbles (Case 2). Our goal is to determine if we are in Case 1 or 2 given a sample of s marbles drawn without replacement from the jar.*

Lemma 16 *Let $\varepsilon \in (0, 1/2)$, let $\alpha = (1+\varepsilon)/2$, and let $n = 2/\varepsilon^4$. There is no algorithm that solves Problem 15 with probability $> 3/4$ unless $s > 1/(4\varepsilon^2)$.*

Proof Suppose we draw s marbles from the jar and encode the result in a length s vector (e.g., with a 0 at position i if the i^{th} marble drawn is red, and a 1 if it is blue). Let $X_1^{(s)}$ denote the distribution over vectors observed in Case 1, and let $X_2^{(s)}$ denote the distribution for Case 2. We will show that $\mathrm{d_{TV}}(X_1^{(s)}, X_2^{(s)})$ is small. To do so, we introduce two auxiliary distributions: let $\hat{X}_1^{(s)}$ denote the distribution over vectors observed if we are in Case 1 and draw marbles randomly *with replacement* and let $\hat{X}_2^{(s)}$ denote the distribution if we are in Case 2 and draw marbles *with replacement*.

We first show that $\mathrm{d_{TV}}(X_i^{(s)}, \hat{X}_i^{(s)})$ is small for $i \in \{1, 2\}$ when n is large. To do so, let $\mathcal{E}$ be the event that in s independent draws with replacement, we never pick a previously picked marble. Let $[\hat{X}_i^{(s)}]_{\mathcal{E}}$ denote the distribution $\hat{X}_i^{(s)}$ conditioned on $\mathcal{E}$, and note that $[\hat{X}_i^{(s)}]_{\mathcal{E}} = X_i^{(s)}$. The

probability that $\mathcal{E}$ happens is equal to $1 \cdot (1 - \frac{1}{n}) \cdot (1 - \frac{2}{n}) \cdot (1 - \frac{s}{n}) \geq 1 - \frac{s^2}{n}$. Therefore, we conclude that:

$$\mathrm{d_{TV}}(\hat{X}_i^{(m)}, X_i^{(m)}) \leq s^2/n. \tag{8}$$

Next, we show that $\mathrm{d_{TV}}(\hat{X}_1^{(s)}, \hat{X}_2^{(s)})$ is small. Doing so is equivalent to bounding the total variation distance between s independent draws from a Bernoulli distribution with mean $1 - \alpha$ and s independent draws from Bernoulli distribution with mean α. Let $D_{\mathrm{KL}}(p, q)$ denote the Kullback–Leibler divergence between distributions p and q. Applying Pinsker's inequality, we have:

$$\begin{aligned}
\mathrm{d_{TV}}(\hat{X}_1^{(s)}, \hat{X}_2^{(s)}) &\leq \sqrt{\frac{1}{2}} \sqrt{D_{\mathrm{KL}}(\hat{X}_1^{(s)}, \hat{X}_2^{(s)})} = \sqrt{\frac{s}{2}} \sqrt{D_{\mathrm{KL}}(\mathsf{Ber}(1-\alpha), \mathsf{Ber}(\alpha))} \\
&= \sqrt{\frac{s}{2}} \sqrt{\alpha \log(\alpha/(1-\alpha)) + (1-\alpha)\log((1-\alpha)/\alpha)} \\
&\leq \sqrt{\frac{s}{2}} \cdot \varepsilon.
\end{aligned} \tag{9}$$

The last inequality holds for any α equal to $(1 + \varepsilon)/2$ whenever $\varepsilon \leq 1/2$. Applying triangle inequality to combine (8) and (9), we have that:

$$\mathrm{d_{TV}}(X_1^{(s)}, X_2^{(s)}) \leq \mathrm{d_{TV}}(\hat{X}_1^{(s)}, X_1^{(s)}) + \mathrm{d_{TV}}(\hat{X}_2^{(s)}, X_2^{(s)}) + \mathrm{d_{TV}}(\hat{X}_1^{(s)}, \hat{X}_2^{(s)}) \leq \frac{2s^2}{n} + \sqrt{\frac{s}{2}} \cdot \varepsilon.$$

For any $s \leq 1/(4\varepsilon^2)$ and $n = 2/\varepsilon^4$ we conclude that $\mathrm{d_{TV}}(X_1^{(m)}, X_2^{(m)}) < 1/2$. Accordingly, no algorithm can distinguish between $X_1^{(s)}$ and $X_2^{(s)}$ with probability $\geq 3/4$ unless $s > 1/(4\varepsilon^2)$. ■

With Lemma 16 in place, we are now ready to prove our lower bound for spectral density estimation. To do so, we will show that any adaptive random walk algorithm that can estimate the spectral density of a graph to accuracy ε using s total random walk steps can solve the Problem 15 using $\leq s$ samples. This reduction requires introducing a second pair of "hard graphs" that are close in Wasserstein distance. In comparison to the hard instance in Theorem 2, these graphs are also based on collection of cycles. The main difference is that we consider two graphs that each contain a mixture of cycles of length 2ℓ and ℓ, but in different proportions.

Definition 17 *For odd integer ℓ and parameter $\alpha \in (0.5, 1)$, let G_1 be a collection of αn disjoint cycles of length 2ℓ and $2(1-\alpha)n$ cycles of size ℓ. Similarly, let G_2 be a collection of $(1-\alpha)n$ cycles of length 2ℓ and $2\alpha n$ cycles of size ℓ. Both graphs have $2n\ell$ vertices in total.*

We use the following expression for the Wasserstein distance between the spectra of the two graphs.

Lemma 18 *Let G_1 and G_2 be unweighted graphs as in Definition 17. $W_1(G_1, G_2) = \frac{(2\alpha-1)}{\ell}$.*

Proof We can compute the exact eigenvalues of the two graphs by combining Lemma 5 with the fact that eigenvalues just increase in multiplicity with repeated components. As in that lemma, recall we use R_ℓ to denote a cycle of length ℓ and $R_{2\ell}$ to denote a cycle of length 2ℓ. The Wasserstein distance between R_ℓ^2 and $R_{2\ell}$ is $W_1(R_\ell^2, R_{2\ell}) = 1/\ell$.

Note that G_1 and G_2 both have $(1-\alpha)n$ cycles of length 2ℓ and $2(1-\alpha)n$ cycles of length ℓ, while G_1 has $(2\alpha-1)n$ extra $R_{2\ell}$ cycles and G_2 has $(2\alpha-1)n$ extra copies of R_ℓ^2. Let $p_1(x)$ and $p_2(x)$ be the spectral density of G_1 and G_2 respectively, and let $\widetilde{p}_1(x)$ and $\widetilde{p}_2(x)$ be the spectral density of $R_{2\ell}^{(2\alpha-1)n}$ and $R_\ell^{2(2\alpha-1)n}$, respectively. We have that $2n\ell \cdot (p_1(x) - p_2(x)) = (2(2\alpha - 1)n\ell \cdot (\widetilde{p}_1(x) - \widetilde{p}_2(x))$ for all $x \in [-1, 1]$. Thus, due to the dual characterization of Wasserstein distance in (2),

$$W_1(G_1, G_2) = (2\alpha-1) \cdot W_1(R_{2\ell}^{(2\alpha-1)n}, R_\ell^{2(2\alpha-1)n}) = (2\alpha-1) \cdot W_1(R_{2\ell}, R_\ell^2) = \frac{(2\alpha-1)}{\ell}.$$

■

We now have all the ingredients in place to prove the main result of this section:

Theorem 19 *For any $\varepsilon < 1/6$, no algorithm that takes s* adaptive *random walks steps in a given graph G can approximate G's spectral density to ε accuracy in the Wasserstein-1 distance with probability $> 3/4$, unless $s \geq 1/(36\varepsilon^2)$.*

Proof Suppose we had such an algorithm (call it $\mathcal{A}$) that uses $s < 1/(4\varepsilon^2)$ random walk steps to output an $\varepsilon/3$-accurate spectral density with probability greater than $3/4$. We will show that the algorithm could be used to solve Problem 15 using $< 1/(4\varepsilon^2)$ samples from the jar with probability greater than $3/4$, which is impossible by Lemma 16.

To prove this reduction we associated an instance of Problem 15 with a hidden graph G that is either isomorphic to G_1 or G_2 as defined in Definition 17. To make the association, every marble will correspond to 2ℓ vertices in the graph with some fixed set of known labels. However, the connections between those nodes is hidden. In particular, if the marble is red, the 2ℓ vertices are arranged in a single cycle of length 2ℓ. Otherwise, they are arranged in two cycles of length ℓ. The ordering of nodes in both cases is known in advance, but we do not know which of the two cases we are in. Also note that there are no other connections between vertices. Observe that if we are in Case 1 for Problem 15, G is isomorphic to G_1 and if we are in Case 2, G is isomorphic to G_2. So in particular, G's spectral density is either equal to the spectral density of G_1 or G_2.

Our main claim is that we can run algorithm $\mathcal{A}$ on the hidden graph G while only accessing s marbles from the jar. To so do, every time the algorithm requests to visit a specific node, we draw the marble from the jar associated with that node's label. In doing so, we learned all edges in the ring containing that node (as well as other edges), so we can perform any future random walk steps initiated from that node. Since $\mathcal{A}$ takes s steps, we at most need to draw s marbles over the course of running the algorithm. At the same time, note that when we choose $\ell = 1$ (considering self loops) and $\alpha = (1+\varepsilon)/2$ as in Lemma 16, Lemma 18 implies that the Wasserstein distance between G_1 and G_2 is equal to ε. So, if $\mathcal{A}$ returns an $\varepsilon/3$-accurate spectral density with probability $3/4$, we can determine if we are in Case 1 or Case 2 with probability $3/4$, violating Lemma 16. We conclude that no such algorithm can exist.

The final statement of the theorem follows by adjusting constants on ε. ■

Appendix B. Wasserstein Distance Bounds via Chebyshev Polynomials

In this section, we give an alternative proof of a lower-bound by Kong and Valiant (2017), which shows that there exist distributions whose first $\ell - 1$ moments match exactly, but the Wasserstein

distance between the distributions is greater than $1/(2(\ell+1))$. Our analysis tightens their result by a factor of ~ 4, showing two such distributions with Wasserstein distance $2/\ell$. Moreover, we prove that the Wasserstein distance is $\Omega(\ell^{-1})$ for *any* distributions p, q whose first $\ell-1$ moments are the same and whose ℓ-th moments differ by $\Omega(2^{-\ell})$.

Lemma 20 (Improvement of (Kong and Valiant, 2017, Proposition 2)) *For any odd ℓ, there exists a pair of distributions p, q, each consisting of $(\ell+1)/2$ point masses, supported within the unit interval $[-1,1]$ such that p and q have identical first $\ell-1$ moments, and the Wasserstein distance $W_1(p,q) \geq 2/\ell$.*

Proof Recall that we use R_ℓ^2 to denote 2 disjoint cycles of length ℓ, and use $R_{2\ell}$ to denote a cycle of length 2ℓ, where ℓ is an odd number. We know the spectrum of R_ℓ^2 and $R_{2\ell}$ from Lemma 5. Let p' and q' denote the spectral density of $A(R_\ell^2)$ and $A(R_{2\ell})$. We first note that the first $\ell-1$ moments of the spectral density of p' and q' are the same because a random walk of length $\ell-1$ cannot distinguish R_ℓ^2 from $R_{2\ell}$ (see Remark 6).

Also, recall we use $\boldsymbol{\lambda}(R_\ell^2)$ to denote the sorted eigenvalue list of $A(R_\ell^2)$ and $\boldsymbol{\lambda}(R_{2\ell})$ to denote the sorted eigenvalue list of $A(R_{2\ell})$. We make the following observations about the spectrum of $A(R_\ell^2)$ and $A(R_{2\ell})$ based on Lemma 5.

1. $A(R_\ell^2)$ has $(\ell+1)/2$ unique eigenvalues, and $A(R_{2\ell})$ has $\ell+1$ unique eigenvalues.
2. All eigenvalues of $A(R_\ell^2)$ overlap with eigenvalues of $A(R_{2\ell})$. In particular, all the eigenvalues of $A(R_\ell^2)$ occur two times more in frequency than the corresponding eigenvalues in $A(R_{2\ell})$. Formally, $\forall \lambda \in \boldsymbol{\lambda}(R_\ell^2)$,
$$\left|\left\{j \,:\, \lambda_j \in \boldsymbol{\lambda}(R_\ell^2), \lambda_j = \lambda, j \in [2\ell]\right\}\right| = 2 \cdot \left|\{j \,:\, \lambda_j \in \boldsymbol{\lambda}(R_{2\ell}), \lambda_j = \lambda, j \in [2\ell]\}\right| . \quad (10)$$
3. All the eigenvalues of $A(R_{2\ell})$ lies in $[-1,1]$.

Let $\boldsymbol{\Lambda}^{(2)}$ denote the sorted list of eigenvalues where we remove all the eigenvalues from $\boldsymbol{\lambda}(R_{2\ell})$ that occurs in $\boldsymbol{\lambda}(R_\ell^2)$. Let $\boldsymbol{\Lambda}^{(1)}$ be the set of removed eigenvalues. The following observations follow from (10). The size of $\boldsymbol{\Lambda}^{(2)}$, and $\boldsymbol{\Lambda}^{(1)}$ is ℓ. Moreover $\boldsymbol{\Lambda}^{(1)}$ has the same eigenvalues as $\boldsymbol{\lambda}(R_\ell^2)$ where the frequency of each unique eigenvalue is $\boldsymbol{\lambda}(R_\ell^2)$ is reduced by a factor of 2. Consequently, we define $p(x) = \frac{1}{\ell}\sum_{j\in[\ell]} \delta\left(x - \boldsymbol{\Lambda}_j^{(1)}\right) = p'(x)$, and $q(x) = \frac{1}{\ell}\sum_{j\in[\ell]} \delta\left(x - \boldsymbol{\Lambda}_j^{(2)}\right) = 2q'(x) - p'(x)$. This ensures that p, q are valid distributions and have a support size of $(\ell+1)/2$.

Since p', and q' have the same first $\ell-1$ moments, we have p and q also have the same first $\ell-1$ moments. Moreover, $W_1(p,q) = W_1(2q'-p', p') = 2W_1(q', p') = \frac{2}{\ell}$, where the penultimate equality follows from the dual characterization of Wasserstein distance in (2) and the last equality follows from Lemma 5. ■

We complement Proposition 20 with the following Lemma 22, which shows that for two distributions p and q such that all their first $\ell-1$ moments are the same and the ℓ-th moment differ only by $\Omega(2^{-\ell})$, even then the Wasserstein distance between p, q is large. The proof follows just by using the fact that there are 1-Lipschitz polynomials with large leading coefficient. We note the following standard facts about the Chebyshev polynomials which can, for example, be found in Allen and Isaacson (2019).

Fact 21 *The Chebyshev polynomials of the first kind of degree $i, (i \in \mathbb{Z}_{\geq 0})$, denoted by $T_i(x)$, satisfy the following properties:*

1. $\forall i \in \mathbb{Z}_{\geq 0}$, $\forall x \in [-1, 1]$, $|T_i(x)| \leq 1$.
2. *The leading coefficient of T_i is 2^{i-1}.*

Lemma 22 *Consider two distributions p and q supported on $[-1, 1]$ such that the difference of their first $\ell - 1$ moments are 0 and the difference of their ℓ-th moment is $c \cdot 2^{-\ell}$. Then, for such a distribution, their Wasserstein distance is bounded by*

$$W_1(p, q) \geq \frac{c}{4\ell}.$$

Proof We use the dual characterization of the Wasserstein distance in Definition 2 and consequently, it suffices to exhibit a 1-Lipschitz function g which has a high inner-product with $p - q$. Let $T_{\ell-1}$ be a degree $\ell - 1$ Chebyshev polynomial. From Fact 21 we know that $f_\ell(x) = \int T_{\ell-1}(x)\mathsf{d}x$ is a degree ℓ, 1-Lipschitz polynomial in $[-1, 1]$, with leading coefficient $2^{\ell-2}/\ell$. Define $g_\ell(x)$ as follows:

$$g_\ell(x) := \begin{cases} f_\ell(-1), & \text{for } x \in (-\infty, -1) \\ f_\ell(x), & \text{for } x \in [-1, 1] \\ f_\ell(1), & \text{for } x \in (1, \infty). \end{cases}$$

From properties of $f_\ell(x)$ and by construction, we know that $g_\ell(x)$ is a 1-Lipschitz function. Therefore,

$$\begin{aligned} W_1(p, q) &\geq \left| \int_{\mathbb{R}} g_\ell(x)(p(x) - q(x))\mathsf{d}x \right| = \left| \int_{-1}^{1} f_\ell(x)(p(x) - q(x))\mathsf{d}x \right| \\ &= \left| \int_{-1}^{1} \frac{2^{\ell-2}}{\ell} x^\ell (p(x) - q(x))\mathsf{d}x \right| = \frac{1}{4\ell} \, 2^\ell \cdot c 2^{-\ell} = c \cdot \frac{1}{4\ell}, \end{aligned}$$

where the first equality holds because $p(x) - q(x) = 0$ outside $[-1, 1]$ and the second equality follows from the fact that the difference of the first $1, \ldots, \ell - 1$ moments are 0. ■

Appendix C. Another Spectral Metric for Graph Comparison

Throughout this section we consider two graphs G_1, G_2 with the same vertex size n and same vertex labeling $V = [n]$, and their un-normalized adjacency matrix $\widetilde{A}_1$ and $\widetilde{A}_2$ with a common degree matrix D. Here we consider learning the spectrum of their difference matrix, i.e., $A(G_1) - A(G_2) = D^{-1/2}\widetilde{A}_1 D^{-1/2} - D^{-1/2}\widetilde{A}_2 D^{-1/2}$, or equivalently $D^{-1}(\widetilde{A}_1 - \widetilde{A}_2)$. We provide a simple proof that $\exp(O(1/\varepsilon))$ number of samples also suffice to estimate this distribution up to ε-Wasserstein distance, using similar techniques as in Cohen-Steiner et al. (2018).

We first restate the main theorem in Kong and Valiant (2017) for completeness.

Theorem 23 (Kong and Valiant (2017, Proposition 1)) *Given two distributions with respective density functions p, q supported on $[a, b]$ whose first k moments are $\alpha = (\alpha_1, \cdots, \alpha_k)$ and $\beta = (\beta_1, \cdots, \beta_k)$, respectively. The Wasserstein distance $W_1(p, q)$ between p, q is bounded by $W_1(p, q) \leq C(\frac{b-a}{k} + 3^k(b - a)\|\alpha - \beta\|_2)$ for some absolute constant C.*

Algorithm 1: Spectral Density of Difference of Adjacency Matrices

Input: Graphs G_1, G_2, oracle $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}$, accuracy ε, probability δ
Parameters: $k \in \mathbb{Z}_+, \theta > 0$
for $j \in [k]$ **do**
  Initialize $\hat{p}_j = 0$
  for $(x_1, x_2, \ldots, x_j) \in \{0,1\}^j$ **do**
    Generate $\frac{1}{2}\theta^{-2} j 4^j \log(2k/\delta)$ independent samples of $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \{x_j\}_{j\in[k]})$
    Let $\hat{p}_{j,x}$ be the fraction of the trajectories which start and end at the same vertex
    Update $\hat{p}_j \leftarrow \hat{p}_j + \hat{p}_{j,x}$
  end
end
Construct a distribution p on $[-1,1]$ with first k moments equal to $\{\hat{p}_j\}_{j\in[k]}$
Return: p

We define a variant of the non-adaptive random walk access model, represented via an oracle $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \{x_i\}_{i\in[j]})$, specifically for this problem, which outputs the random trajectory after taking a length j random walk starting from a uniformly randomly chosen vertex, where at step $i \in [j]$ it the follows probabilistic transition of $D^{-1}\widetilde{A}_1$ when $x_i = 1$ and $D^{-1}\widetilde{A}_2$ when $x_i = 0$. We consider Algorithm 1 for estimating the spectral density of matrix $D^{-1}(\widetilde{A}_1 - \widetilde{A}_2)$.

Algorithm 1 computes estimates of the moments of difference matrix $D^{-1}(\widetilde{A}_1 - \widetilde{A}_2)$. Together with the procedure of computing a distribution based on first k moments using linear programming as stated in Cohen-Steiner et al. (2018), we have the following guarantee.

Theorem 24 *Given any two graphs G_1, G_2 on same set of vertices with a common degree matrix D, Algorithm 1 with $k = 4C/\varepsilon$ and $\theta = \varepsilon/(3^{2k+2})$ outputs a distribution p that is ε-close in Wasserstein-1 distance with the spectral density function of $A(G_1) - A(G_2)$ with probability 0.9, using a total of $2^{O(1/\varepsilon)}$ calls to $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot)$, $j \in [O(1/\varepsilon)]$.*

Proof Note similarity transformation doesn't affect eigenvalues, thus it suffices to estimate the spectral density function of matrix $D^{-1}(\widetilde{A}_1 - \widetilde{A}_2)$, whose j^{th} moment is $\frac{1}{n}\mathrm{tr}((D^{-1}\widetilde{A}_1 - D^{-1}\widetilde{A}_2)^j)$.

For any $j \in \mathbb{Z}_+$, note that

$$\frac{1}{n}\mathrm{tr}((D^{-1}\widetilde{A}_1 - D^{-1}\widetilde{A}_2)^j) = \sum_{x_1,x_2,\cdots,x_j\in\{0,1\}} \frac{1}{n}\mathrm{tr}\left(\prod_{i=1,\cdots,j}(x_i \cdot D^{-1}\widetilde{A}_1 + (1-x_i)\cdot D^{-1}\widetilde{A}_2)\right).$$

Given any $x = (x_1, \cdots, x_j)$, we run an alternating random walk as in $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}$ to generate unbiased samples of term $\frac{1}{n}\mathrm{tr}(\prod_{i=1,\cdots,j}(x_i \cdot D^{-1}\widetilde{A}_1 + (1-x_i)\cdot D^{-1}\widetilde{A}_2))$ (as in Line 1). By concentration we can estimate each term $\frac{1}{n}\mathrm{tr}(\prod_{i=1,\cdots,j}(x_i \cdot D^{-1}\widetilde{A}_1 + (1-x_i)\cdot D^{-1}\widetilde{A}_2))$ using $\hat{p}_{j,x}$ (as in Line 1) up to $\theta/2^j$ additive accuracy with high probability $1 - \delta/(k2^j)$ using a total of $\frac{1}{2}\theta^{-2} j 4^j \log(2k/\delta)$ calls to $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \{x_i\}_{i\in[j]})$. Consequently, using a union bound we have with probability $1 - \delta$, $\hat{p}_j$ estimates the j^{th} moments up to θ additive accuracy, each using a total of $O(\theta^{-2} j 2^{3j} \log(2k/\delta))$ calls to some $\widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot)$ for all $j \in [k]$.

Picking $k = \frac{4C}{\varepsilon}$, $\theta = \frac{\varepsilon}{3^{2k+2}}$, we can apply Theorem 23 to conclude that the constructed distribution p is an ε-approximation in Wasserstein distance to the spectral density function of $A(G_1) - A(G_2)$. Also, the algorithm uses a total of

$$\sum_{j \in [k]} O(\theta^{-2} j 2^{3j} \log(2k/\delta)) \text{ calls to } \widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot)$$
$$= \sum_{j \in [O(1/\varepsilon)]} 2^{O(1/\varepsilon)} \text{ calls to } \widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot).$$

In the above equality we also used that $\delta = 0.1$. ■

An interesting open problem is whether similar algorithms exist for comparing two graphs on the same vertex set without a common degree matrix D.