Moments, Random Walks, and Limits for Spectrum Approximation

Yujia Jin
Stanford University
yujia@stanford.edu

Christopher Musco
New York University
cmusco@nyu.edu

Aaron Sidford
Stanford University
sidford@stanford.edu

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

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.,

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

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. $p(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta(x - \lambda_i)$ for indicator function δ (see Section 2 for notation).

[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 [BGKS20]. Methods typically used in practice run in time $O(n^3 + n^2 \log(\frac{1}{\varepsilon}))$ [Wil68].

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 [Ski89; SR94; Wan94; WWAF06]. 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 [CTU21].

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 [PSG18; MM19; Pap18], in optimization [GKX19; Sag+17], and in network science [DBB19; CKSV18].

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 [KLS11]. It was recently shown that the $O(n^2/\varepsilon)$ runtime of general purpose algorithms like stochastic Lanczos quadrature can be improved to $\tilde{O}(n/\mathsf{poly}(\varepsilon))$ [BKM22].[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. [CKSV18] 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. [CKSV18] unfortunately makes the algorithm impractical for any reasonable level of accuracy. As 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 1.1. *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

[2] If $\tilde{A}$ is the unnormalized adjacency matrix of G and D is its diagonal degree matrix, we can equivalently consider the asymmetric matrix, $D^{-1}\tilde{A}$ or the symmetric one, $D^{-1/2}\tilde{A}D^{-1/2}$, as they have the same eigenvalues.

[3] We use $\tilde{O}(m)$ to denote $O(m \log m)$. The runtime in [BKM22] 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. [CKSV18] 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.

work, we consider distributions supported on the the $[-1, 1]$, in which case these moments are:

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

Recent work of Kong and Valiant [KV17] 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. [CKSV18] 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. [CKSV18] 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. [CKSV18]. So where does the $2^{O(1/\varepsilon)}$ runtime dependence come from? The issue is that the result of Kong and Valiant [KV17] 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 [KV17] 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 1.1 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 [Hut90; MMMW21] 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 [BKM22]. 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 1.1 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 1.2. *For any $\varepsilon \in (0, 1/4]$, there exist weighted graphs G_1 and G_2 (see Definition 3.1) 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* $m_j(p_1) = \int_{-1}^{1} x^j p_1(x) \mathsf{d}x$ *and* $m_j(p_2) = \int_{-1}^{1} x^j p_2(x) \mathsf{d}x$ *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 1.2 shows that Kong and Valiant [KV17]'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. [CKSV18]. 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 1.2 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 1.2 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 1.2 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. [BKM22]. However, by Theorem 1.2, this is impossible.

While Theorem 1.2 rules out direct improvements to the moment-based method of Cohen-Steiner et al. [CKSV18], 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.

[6] When run for $O(1/\delta^2)$ steps, the random walk method of Cohen-Steiner et al. [CKSV18] 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 $[-1, 1]$.

Theorem 1.3. *For any $\varepsilon \leq 1/2$, there exists a distribution over weighted graphs $\mathcal{D}$ so that, given the transcript of m, length T randoms walks initiated at m uniformly random nodes from $G \sim \mathcal{D}$, no algorithm can estimate the spectral density of G's normalized adjacency matrix to accuracy ε in Wasserstein-1 distance with probability $> 3/4$ unless $m \cdot T > (16e)^{-1} \cdot 2^{1/2\varepsilon}$.*

While more technical, the proof of Theorem 1.3 is based on the same hard instance as Theorem 1.2. 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 1.3 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 [BKM22]. 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 1.2. In Section 4 we prove lower bound for spectrum estimation based on random walks, establishing Theorem 1.3. 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 [KV17]. 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 $\delta(0) \stackrel{\text{def}}{=} 1$ and $\delta(x) \stackrel{\text{def}}{=} 0$ 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 $\mathbb{P}[E^c] = 1 - \mathbb{P}[E^c]$.

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 $\tilde{A} \in \mathbb{R}_{\geq 0}^{V \times V}$ to denote the weighted adjacency matrix of G where $\tilde{A}(v, v') = w_e$ if $e = (v, v') \in E$ and $\tilde{A}(v, v') = 0$ 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 $D(v, v) \stackrel{\text{def}}{=} \sum_{e=(v,v') \in E} w_e$ for all $v \in V$. We let $A(G) \in \mathbb{R}^{V \times V}$ denote the normalized adjacency matrix of G, i.e. $A(G) \stackrel{\text{def}}{=} D^{-1/2} \tilde{A} D^{-1/2}$. We refer to $D^{-1}\tilde{A}$ as the random walk matrix and note that, for degree-regular graphs, $A(G) = D^{-1}\tilde{A}$.

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 2.1. *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 2.2 (Kantorovich-Rubinstein Duality [Kan40; Kan42])**.**

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

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 $W_1(G_1, G_2) \stackrel{\text{def}}{=} W_1(p_1, p_2)$ 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. [KV17]) 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, $\tilde{m}_1, \ldots, \tilde{m}_j$, to the first j moments of G's spectral density p, $m_1(p), \ldots, m_j(p)$. Specifically, we have that $|\tilde{m}_j - m_j(p)| \leq \delta \cdot m_j(p)$.

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 [CKSV18]. 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. The algorithm can see the entire sequence of vertices in each T step random walk, which we call the walk "transcripts" and denote 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 $D^{-1}\tilde{A}$. We use $\mathbb{P}_G(\cdot)$ to denote the probability of some event happening when interacting with some round(s) of non-adaptive random walk on graph G, by abusing notation we also use that to represent in particular the probability distribution function of m random walks with length T when clear from context.

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 2.3 (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 [Spi19]. The Wasserstein distance immediately follows since we have:

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

Remark 2.4. *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.

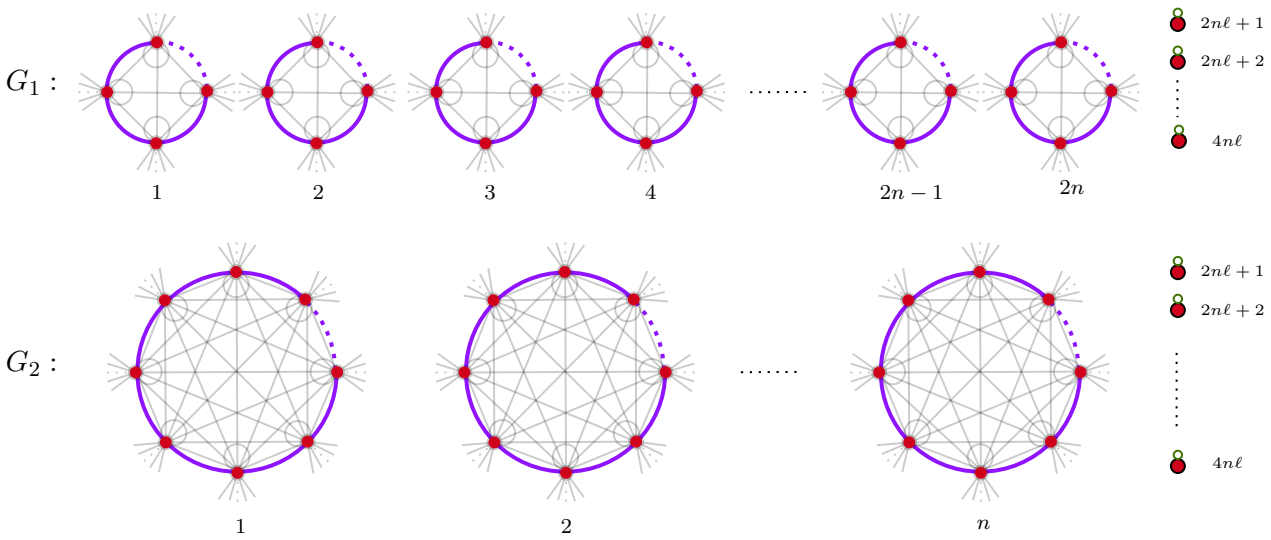

Figure 1: Diagram depicting the construction of graphs G_1 and G_2 in Definition 3.1. G_1 depicts $2n$ cycles of length ℓ and $2n\ell$ isolated vertices, and G_2 depicts n cycles of length 2ℓ and $2n\ell$ isolated vertices. In both the graphs, the purple lines represent the edges of the cycle, each with weight $1/4$, and the grey lines represent the edges of the complete graph (including self-loops) over the $2n\ell$ vertices of the cycles, each with weight $1/(4n\ell)$, and green lines represent self-loops of weight 1.

Definition 3.1. *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 3.1 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 3.2. *For weighted graphs G_1, G_2 constructed in Definition 3.1, $W_1(G_1, G_2) = 1/(4\ell)$.*

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

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

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 $\mathcal{R}_1 \stackrel{\text{def}}{=} R_\ell^{2n}$, $\mathcal{R}_2 = R_{2\ell}^n$. For $i \in [2]$ we have:

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

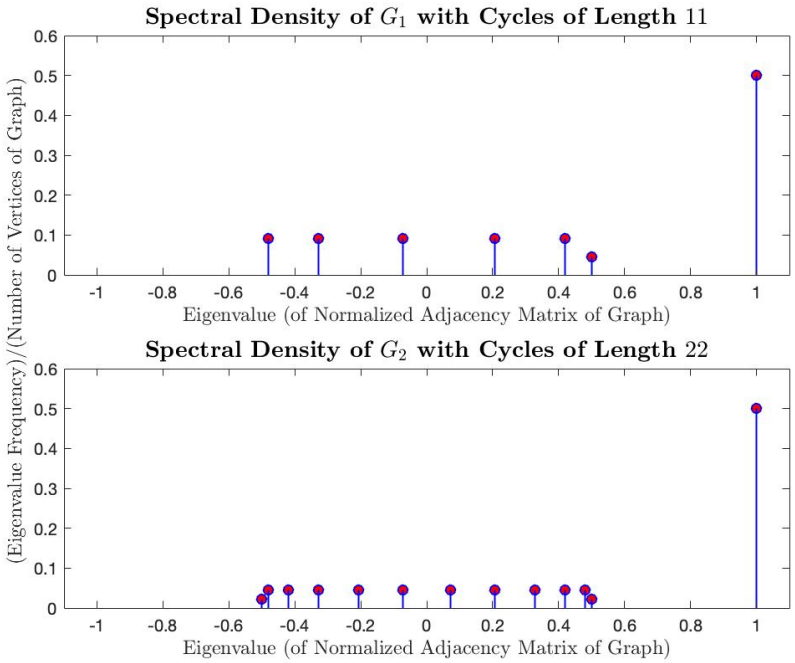

Figure 2: Spectral Density of G_1 and G_2 as defined in Definition 3.1, with cycles of length 11 and 22, respectively.

This implies $W_1(G_1, G_2) = \frac{1}{4n\ell} \|\boldsymbol{\lambda}(G_1) - \boldsymbol{\lambda}(G_2)\|_1 = \frac{1}{2} \cdot \frac{1}{4n\ell} \|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\|_1$ by the characterization of Wasserstein distance given in (3).

Thus it suffices to calculate $\|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\|_1$. 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 2.3, we get $\|\boldsymbol{\lambda}(\mathcal{R}_1) - \boldsymbol{\lambda}(\mathcal{R}_2)\|_1 = n \cdot 2\ell \cdot W_1(R_\ell^2, R_{2\ell}) = 2n$. 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 3.3. *Let G_1 and G_2 be weighted graphs as constructed in Definition 3.1. 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 $m_j(p_i) = \frac{1}{n}\mathrm{tr}(A(G_i)^j)$. Using (4) we can calculate:

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

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 2.4), we get from (5) that

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

We briefly remark that the proof of Lemma 3.3 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 1.2. *For any $\varepsilon \in (0, 1/4]$, there exist weighted graphs G_1 and G_2 (see Definition 3.1) 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 $m_j(p_1) = \int_{-1}^{1} x^j p_1(x)\mathsf{d}x$ and $m_j(p_2) = \int_{-1}^{1} x^j p_2(x)\mathsf{d}x$ 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 3.2. Next, we know that for all j, $m_j(p_1) \in [1/2, 1]$. So, by Lemma 3.3, $|m_j(p_1) - m_j(p_2)| \leq 2^{-\ell+2} m_j(p_1)$. The statement holds since we have $\ell \geq 1/(4\varepsilon) - 2$. $\blacksquare$

4 Limits on Random Walk Methods

In this section we prove Theorem 1.3. We construct a hard distribution that is uniform over two different graphs, G_1 and G_2, which are close in Wasserstein distance and hard to distinguish based on random walks. We define these graphs below. The construction is similar to Section 3.

Definition 4.1. *G_1 is defined as two disjoint copies of a graph G_1'. G_1' is constructed by starting with a collection of n disjoint cycles of length ℓ, where ℓ is an odd integer and $n = 2^{10\ell}$. The edges in the cycle have weight $1/4$ and every vertex in a cycle is then connected to all other vertices with weight $1/(2n\ell)$ (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 in a cycle is then connected to all other vertices with weight $1/(4n\ell)$ (including a self-loop).*

We first bound the Wasserstein distance between the spectra of G_1 and G_2. We remark that the same analysis would allow us to compare G_1' to G_2. The reason that we consider G_1 to be two disjoint copies of G_1' is to ensure that the number of vertices in the two instances are equal.

Otherwise, for some fixed labelling of vertices $1, 2, \ldots, |V|$, one could distinguish between G_1' and G_2 by querying vertices and checking if any vertex label is larger than $2nl$. Doing so would only take $O(\log(n\ell)) = \mathsf{poly}(\ell)$ queries with high probability.

Lemma 4.2. *For weighted graphs G_1, G_2 constructed in Definition 4.1, $W_1(G_1, G_2) > 1/(2\ell)$.*

Proof. The normalized adjacency matrices of the two graphs are:

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

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

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

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

Using that the top eigenvalue of R_ℓ^{2n} equals 1, we have that $W_1(G_1, G_2) = \frac{1}{2n\ell}\left\|\boldsymbol{\lambda}(G_1) - \boldsymbol{\lambda}(G_2)\right\|_1 = \frac{1}{2n\ell}\cdot\left(\frac{1}{2}\cdot\left\|\boldsymbol{\lambda}(R_\ell^{2n}) - \boldsymbol{\lambda}(R_{2\ell}^n)\right\|_1 + \frac{1}{2}\right)$. Thus it suffices to calculate $\left\|\boldsymbol{\lambda}(R_\ell^{2n}) - \boldsymbol{\lambda}(R_{2\ell}^n)\right\|_1$. 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 2.3, we get $\left\|\boldsymbol{\lambda}(R_\ell^{2n}) - \boldsymbol{\lambda}(R_{2\ell}^n)\right\|_1 = n\cdot 2\ell\cdot W_1(R_\ell^2, R_{2\ell}) = 2n$. Plugging this back we get the claimed Wasserstein distance $W_1(G_1, G_2) = (n+1)/(2n\ell) > 1/(2\ell)$. ■

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

Lemma 4.3. *Consider conducting m non-adaptive random walks, each with length T in G_1 or G_2. Let $S = \{S_1, \ldots, S_m\}$ denote the set of m transcripts of these random walks and let $\mathbb{P}_{G_1}[S]$ and $\mathbb{P}_{G_2}[S]$ denote probability of observing S when conducting the walks in G_1 and G_2, respectively. Let $\mathbb{P}_{G_1}$ and $\mathbb{P}_{G_2}$ denote the corresponding probability distributions over the set of random walk transcripts. We can bound the total variation distance between $\mathbb{P}_{G_1}$ and $\mathbb{P}_{G_2}$ by $\mathrm{d_{TV}}(\mathbb{P}_{G_1}, \mathbb{P}_{G_2}) \le e\cdot m\cdot T/2^\ell + 1/2^{4\ell}$.*

To prove Lemma 4.3, we show a good event $\mathcal{E}$ (Definition 4.7) under which $\mathbb{P}_{G_1}[S|\mathcal{E}] = \mathbb{P}_{G_2}[S|\mathcal{E}]$. We then show that the probability of $\mathcal{E}$ occurring is very close to 1.

To begin with, we first define the following labeling and correspondence between graphs to give a more explicit characterization of the random walks on both graphs.

Definition 4.4 (Correspondence of random walks in G_1 and G_2)**.** *For the first connected component in G_1, we use $r_1^i, i \in [n]$ to label the n cycles in it and denote each node in a cycle (in clock-wise direction) as $1, 2, \cdots, \ell$. For the other connected component, we use $r_1^i, i \in [n]$ to label the nodes in each cycle (in clock-wise direction) as $-1, -2, \cdots, -\ell$. In G_2 we similarly label the cycles by $r_2^i, i \in [n]$, and we label the nodes in the corresponding cycle (clock-wise direction) by $1, \cdots, \ell, -1, \cdots, -\ell$. Note any pair (r_1^i, j) or (r_2^i, j), for $i \in [n]$, $j \in \{-\ell, \cdots, -1, 1, \cdots, \ell\}$ corresponds to a unique node in G_1 or G_2. Fixing such label, we define a mapping from any fixed path walk in G_2 to a fixed-length path in G_1 (or more concretely G_1'), formally as follows.*

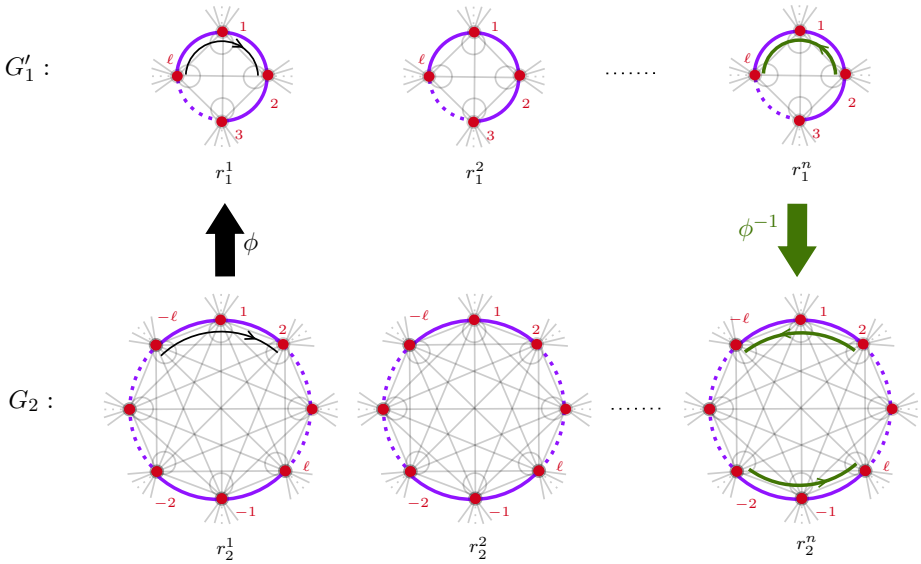

Figure 3: Depicting the graphs G_1' and G_2 and their labeling as in Definition 4.1. G_1' depicts a collection of n cycles of length ℓ each, and G_2 depicts n cycles of length 2ℓ each. In both the graphs, the purple lines represent the edges of the cycle, each with weight $1/4$, and the grey lines represent the edges of the complete graph (including self-loops) over the vertices of the cycles, each with weight less than or equal to $1/(2n\ell)$. The mappings of random-walk paths with edge-weights defined in Definition 4.4 are also depicted. The mapping ϕ from a random-walk path in G_2 to a random-walk path in G_1' is represented in black, and the mapping ϕ^{-1} from a random-walk path in G_1 to paths in G_2 is represented in green.

Let ϕ be a mapping between same-length walks from G_2 to G_1'. Given any T-length path in G_2 which is expressed as $\{(r_1^{i_t}, j_t), w_t\}_{t\in[T]}$ where $w_t \in \{0,1\}$ indicates whether the path at step t takes a heavy edge (weight $= 1/4$) or light edge (weight $\leq 1/2n\ell$), we have

$$\phi(\{(r_2^{i_t}, j_t), w_t\}_{t\in[T]}) = \{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]}.$$

We also use $\phi^{-1}(\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]})$ to denote all valid random walks with length T in G_2 that map to the the random walk of form $\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]}$ in G_1'.

Remark 4.5. *By the definition of the mapping ϕ, the following facts about ϕ are immediate:*

1. *For a valid random walk path $\{(r_2^{i_t}, j_t), w_t\}_{t\in[T]}$ in G_2, its image $\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]}$ must be a valid random walk in G_1'.*
2. *ϕ may map different paths in G_2 to a same path in G_1'.*

For each random walk path in G_1', we view it in G_1 as the two corresponding random walk paths in each of the disjoint component. We now show that the graphs G_1 and G_2 have the same probability of generating the random walk paths under the mapping ϕ.

Lemma 4.6. *Given fixed label of G_1, G_2 and mapping ϕ as in Definition 4.4, we have*

$$\mathbb{P}_{G_2}\left[\phi^{-1}(\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]})\right] = \mathbb{P}_{G_1}\left[\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]} \text{ or } \{(r_1^{i_t}, -|j_t|), w_t\}_{t\in[T]}\right].$$

Proof. Overloading notation, we let ϕ^{-1} also define a node-to-node mapping from G_1' to G_2, i.e., $\phi^{-1}((r_1^i, |j|)) = \{(r_2^i, j), (r_2^i, -j)\}$. We first note that for any two consecutive vertices in the random walk $\phi^{-1}((r_1^{i_t}, |j_t|))$ and $\phi^{-1}((r_1^{i_{t+1}}, |j_{t+1}|))$, by dividing the cases based on $w_t \in \{0, 1\}$ we can conclude that the one-step transition probability

$$\begin{aligned}\mathbb{P}_{G_2}\left[\phi^{-1}((r_1^{i_t}, |j_t|)) \xrightarrow{w_t} \phi^{-1}((r_1^{i_{t+1}}, |j_{t+1}|))\right] &= \mathbb{P}_{G_1}\left[(r_1^{i_t}, |j_t|) \xrightarrow{w_t} (r_1^{i_{t+1}}, |j_{t+1}|)\right] \\ &= \mathbb{P}_{G_1}\left[(r_1^{i_t}, -|j_t|) \xrightarrow{w_t} (r_1^{i_{t+1}}, -|j_{t+1}|)\right].\end{aligned}$$

Additionally, the probability of starting at $(r_1^{i_1}, |j_1|)$ for G_1 is the half of the probability of starting at $(r_2^{i_1}, j_1)$ or $(r_2^{i_1}, -j_1)$ for G_2. Putting these together, we have,

$$\mathbb{P}_{G_2}\left[\phi^{-1}(\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]})\right] = 2\mathbb{P}_{G_1}\left[\{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]}\right] = 2\mathbb{P}_{G_1}\left[\{(r_1^{i_t}, -|j_t|), w_t\}_{t\in[T]}\right],$$

which by rearranging gives the claim. ■

Now we define good event $\mathcal{E}$. Recall $S = \{S_1, \cdots, S_m\}$ is a collection of random walk paths.

Definition 4.7 (Good event $\mathcal{E}$)**.** *$\mathcal{E}$ is the event that if S is generated from G_1 and mapping all paths on $G_1 \setminus G_1'$ to G_1' by flipping the sign of the index, i.e. mapping $(r_1^i, -j)$ to (r_1^i, j), no two paths in S visit same nodes, no path completes a cycle, and when a path leaves a cycle it never comes back to the same cycle; or if S is generated from G_2 it maps under ϕ to a collection of paths in $G_1' \subseteq G_1$ satisfying the aforementioned properties.*

Lemma 4.8. *Conditioning on event $\mathcal{E}$ (Definition 4.7), the probability that the non-adaptive random walk model generates m length T transcripts $S = \{S_1, \cdots, S_m\}$ satisfies $\mathbb{P}_{G_1}[S|\mathcal{E}] = \mathbb{P}_{G_2}[S|\mathcal{E}]$.*

Proof. For a length T path in G_1, let the fixed labelled transcript be $L_1 = \{(r_1^{i_t}, |j_t|), w_t\}_{t\in[T]}$ and $-L_1 = \{(r_1^{i_t}, -|j_t|), w_t\}_{t\in[T]}$. Then we must have $\sum_{L_1' \in \phi^{-1}(L_1)} \mathbb{P}_{G_2}[L_1'] = \mathbb{P}_{G_1}[L_1 \text{ or } -L_1]$. Consequently by independence over trajectories we have $\sum_{L_k' \in \phi^{-1}(L_k)} \mathbb{P}_{G_2}[L_1', \cdots, L_m'] = \mathbb{P}_{G_1}[L_1 \text{ or } -L_1, L_2 \text{ or } -L_2, \cdots, L_m \text{ or } -L_m]$. For any collection of paths $\{L_1', \cdots, L_m'\}$ in G_2 and the corresponding $\{L_1 \text{ or } -L_1, \cdots, L_m \text{ or } -L_m\}$ in G_2, conditioning on $\mathcal{E}$ they must have the same transcript (since information theoretically we cannot tell them apart). Consequently,

$$\begin{aligned}\mathbb{P}_{G_2}[S|\mathcal{E}] &= \sum_{\{L_1,\cdots,L_m\} \text{ aligns with } S} \; \sum_{L_k' \in \phi^{-1}(L_k)} \mathbb{P}_{G_2}[L_1', \cdots, L_m'|\mathcal{E}] \\ &= \sum_{\{L_1,\cdots,L_m\} \text{ aligns with } S} \mathbb{P}_{G_1}[L_1 \text{ or } -L_1, L_2 \text{ or } -L_2, \cdots, L_m \text{ or } -L_m|\mathcal{E}] \\ &= \mathbb{P}_{G_1}[S|\mathcal{E}].\end{aligned}$$

by summing up all possible collections of labeled paths corresponding to the given transcript S, which concludes the proof. ■

Now we provide the full proof of Lemma 4.3, it mainly follows from Lemma 4.8 and the observation that the probability of the good event not happening $\mathcal{E}^c$ is exponentially small.

Proof of Lemma 4.3. We first bound the probability of the event $\mathbb{P}_{G_1}[\mathcal{E}^c]$ (see Definition 4.7). Consider a T-step random walk on G_1' where $T \le 2^\ell$. The probability that a T-step random walk

visits a full cycle or goes back to some cycle it leaves can be upper bounded as

$$\sum_{t\in[T]} \mathbb{P}_{G_1}[\text{at } t, \text{ visit a full cycle for } \ell \text{ consecutive steps or go back to the same cycle after leaving}]$$
$$\overset{(i)}{\leq} T\cdot\left(\left(\frac{1}{2}+\frac{1}{2n}\right)^{\ell}+\frac{T}{2n}\right)\leq \frac{T}{2^{\ell}}\cdot\left(1+\frac{1}{n}\right)^{\ell}+\frac{T^2}{2n}\overset{(ii)}{\leq}\frac{eT}{2^{\ell}}+\frac{1}{2^{5\ell}}.$$

Here we use (i) the fact that the probability of staying on the cycle for ℓ steps is bounded by $(1/2+1/(2n))^{\ell}$ and the probability of going back after leaving is bounded by $T\cdot\ell/(2n\ell)$. We also use (ii) the inequality of $(1+1/n)^{\ell}\leq e$ given $\ell\leq n$ and choice of $n=2^{10\ell}$.

Since we restart the random walk of length T for $m<2^{\ell}$ times. The probability that any later random walk visits a node visited in the previous walk can be bounded by $m^2\cdot T^2/(n\ell)\leq 1/2^{(5\ell)}$. Therefore applying a union bound we get that $\mathbb{P}_{G_1}[\mathcal{E}^c]$ can be bounded by

$$\mathbb{P}_{G_1}[\mathcal{E}^c]\leq\frac{emT}{2^{\ell}}+\frac{m}{2^{5\ell}}+\frac{1}{2^{5\ell}}\,.$$

Following from the definition of $\mathcal{E}$, mapping ϕ and Lemma 4.6, we also have $\mathbb{P}_{G_1}[\mathcal{E}^c]=\mathbb{P}_{G_2}[\mathcal{E}^c]$. By abuse of notation, let $\mathbb{P}_{G_1}$ and $\mathbb{P}_{G_2}$ denote the probability distribution over all possible transcripts $S=\{S_1,\ldots,S_m\}$. Therefore, above equation combined with Lemma 4.8, gives that $\mathrm{d_{TV}}(\mathbb{P}_{G_1},\mathbb{P}_{G_2})\leq \sup_E|\mathbb{P}_{G_1}[E]-\mathbb{P}_{G_2}[E]|\leq\sup_{E:E\subseteq\mathcal{E}^c}\max\left(\mathbb{P}_{G_1}[E],\mathbb{P}_{G_2}[E]\right)\leq\frac{emT}{2^{\ell}}+\frac{1}{2^{4\ell}}$. ■

Lower Bound. Given m transcripts $S=\{S_1,\cdots,S_m\}$ of random walk, each with length T. We show we need at least an exponential number of total steps $(m\cdot T)$ in random walks to identify whether the transcript was generated from graph G_1 or G_2.

Theorem 1.3. *For any $\varepsilon\leq 1/2$, there exists a distribution over weighted graphs $\mathcal{D}$ so that, given the transcript of m, length T randoms walks initiated at m uniformly random nodes from $G\sim\mathcal{D}$, no algorithm can estimate the spectral density of G's normalized adjacency matrix to accuracy ε in Wasserstein-1 distance with probability $>3/4$ unless $m\cdot T>(16e)^{-1}\cdot 2^{1/2\varepsilon}$.*

Proof. We construct the distribution $\mathcal{D}$ as follows. Let G_1 and G_2 be as defined in Definition 4.1. Let $G\sim\mathcal{D}$ be a random random graph such that $\mathbb{P}[G=G_1^{(\pi)}]=\mathbb{P}[G=G_2^{(\pi)}]=1/2$, where $G_i^{(\pi)}$ is a random permutation of node labels of the graph G_i, for $i\in[2]$. Let $S_1,\ldots,S_m$ be a transcript of the m restarts of the random walk on G of length T.

We consider the transcript is labelled in a lazy fashion without loss of generality[7]: we label every unique nodes visited with increasing integers starting from 1.

Now let ℓ be the largest odd integer smaller than $1/2\varepsilon$ so that $W_1(G_1,G_2)\geq\varepsilon$ due to Lemma 4.2. For any $m\cdot T<2^{\ell}\cdot\frac{1}{4e}$, we have $\mathrm{d_{TV}}(\mathbb{P}_{G_1},\mathbb{P}_{G_2})\leq\frac{1}{4}+\frac{1}{4}\leq\frac{1}{2}$ due to Lemma 4.3, where $\mathbb{P}_G$ represents the probability distributions over random walk transcripts on graph G. Then by Le Cam's inequality, any algorithm that takes in the input S and outputs either G_1 or G_2 will make a wrong prediction with probability at least $\frac{1}{2}(1-\mathrm{d_{TV}}(\mathbb{P}_{G_1},\mathbb{P}_{G_2}))\geq\frac{1}{4}$. ■

[7]Since labels are permuted uniformly at random, we lose no information by lazy labeling.

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

[AI19] Myron B. Allen and Eli L. Isaacson. "Chebyshev Polynomials". In: *Numerical Analysis for Applied Science*. John Wiley & Sons, Ltd, 2019, pp. 555–557. DOI: 10.1002/9781119245476.app3.

[BGKS20] 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)*. 2020, pp. 529–540. arXiv: 1912.08805.

[BKM22] Vladimir Braverman, Aditya Krishnan, and Christopher Musco. "Sublinear Time Spectral Density Estimation". In: *Proceedings of the 54th Annual ACM Symposium on Theory of Computing (STOC)*. 2022, pp. 1144–1157. arXiv: 2104.03461.

[CTU21] Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. "Analysis of stochastic Lanczos quadrature for spectrum approximation". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021. arXiv: 2105.06595.

[CKSV18] 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)*. 2018, pp. 1263–1271. arXiv: 1712.01725.

[DBB19] Kun Dong, Austin R Benson, and David Bindel. "Network density of states". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2019, pp. 1152–1161. arXiv: 1905.09758.

[GKX19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. "An Investigation into Neural Net Optimization via Hessian Eigenvalue Density". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. 2019, pp. 2232–2241. arXiv: 1901.10159.

[Hut90] Michael F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines". In: *Communications in Statistics-Simulation and Computation* 19.2 (1990), pp. 433–450. DOI: 10.1080/03610919008812866.

[Kan40] Leonid V Kantorovich. "On an effective method of solving certain classes of extremal problems". In: *Dokl. Akad. Nauk., USSR* 28 (1940), pp. 212–215.

[Kan42] Leonid V Kantorovich. "On the translocation of masses". In: *Dokl. Akad. Nauk., USSR* 37 (1942). English translation in J. Math. Sci. 133, 4 (2006), 1381–1382., pp. 199–201. DOI: 10.1007/s10958-006-0049-2.

[KLS11] 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)*. 2011, pp. 597–606. DOI: 10.1145/1963405.1963489.

[KV17] Weihao Kong and Gregory Valiant. "Spectrum estimation from samples". In: *The Annals of Statistics* 45.5 (2017), pp. 2218–2247. arXiv: 1602.00061.

[Le 12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012. DOI: 10.1007/978-1-4612-4946-7.

[MM19] Michael Mahoney and Charles Martin. "Traditional and Heavy Tailed Self Regularization in Neural Network Models". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. 2019, pp. 4284–4293. arXiv: 1901.08276.

[MMMW21] Raphael A. Meyer, Cameron Musco, Christopher Musco, and David Woodruff. "Hutch++: Optimal Stochastic Trace Estimation". In: *Proceedings of the 4th Symposium on Simplicity in Algorithms (SOSA)* (2021). arXiv: 2010.09649.

[Pap18] Vardan Papyan. "The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size". In: *arXiv* (2018). arXiv: 1811.07062.

[PSG18] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. "The emergence of spectral universality in deep networks". In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2018, pp. 1924–1932. arXiv: 1802.09979.

[Sag+17] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. "Empirical Analysis of the Hessian of Over-Parametrized Neural Networks". In: *arXiv* (2017). arXiv: 1706.04454.

[SR94] R.N. Silver and H. Röder. "Densities of States of Mega-Dimensional Hamiltonian Matrices". In: *International Journal of Modern Physics C* 5.4 (1994), pp. 735–753. DOI: 10.1142/S0129183194000842.

[Ski89] John Skilling. "The Eigenvalues of Mega-dimensional Matrices". In: *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*. Springer Netherlands, 1989, pp. 455–466. DOI: 10.1007/978-94-015-7860-8_48.

[Spi19] Daniel Spielman. *Spectral and Algebraic Graph Theory*. Unpublished Manuscript, 2019. Chap. Fundamental Grpahs. URL: http://cs-www.cs.yale.edu/homes/spielman/sagt/.

[Wan94] Lin-Wang Wang. "Calculating the density of states and optical-absorption spectra of large quantum systems by the plane-wave moments method". In: *Phys. Rev. B* 49 (15 1994), pp. 10154–10158. DOI: 10.1103/PhysRevB.49.10154.

[WWAF06] Alexander Weiße, Gerhard Wellein, Andreas Alvermann, and Holger Fehske. "The kernel polynomial method". In: *Reviews of modern physics* 78.1 (2006), p. 275. arXiv: cond-mat/0504627.

[Wil68] J.H. Wilkinson. "Global convergene of tridiagonal QR algorithm with origin shifts". In: *Linear Algebra and its Applications* 1.3 (1968), pp. 409–420. DOI: 10.1016/0024-3795(68)90017-7.

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. It also has the ability to uniformly sample nodes from the graph, as in the standard random walk model considered for Theorem 1.3. **⋘ Sidford: I'm not sure about the relationship of this paragraph to the first sentence of the next; is it giving an intuitive view or re-explaining? I think I might just add a sentence to this paragraph emphasizing what adaptivity means and make the "off track" comment if needed and then delete the first paragraph of the next paragaph. ⋙**

The adaptive model allows for sampling random neighbors of nodes, restarting random walks that have gone "off track", and more. Interestingly, adaptive algorithms can solve the hard instance

from Theorem 1.3 using roughly $O(\log(1/\varepsilon)/\varepsilon)$ random walk steps. Specifically, for any node,⋘ **Sidford: Maybe change the "we" to "the algorithm" or something like that?** ⋙ we 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 us 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. In comparison to the hard instance in Theorem 1.3, the lower bound instance here is also based on collection of cycles. The main difference is that here weconsider two graphs that are collections of disjoint cycles of size 2ℓ and cycles of size ℓ, respectively. In order to distinguish the two graphs, any algorithm needs to tell the fraction of longer cycles rather than the cycle length.⋘ **Sidford: Might be good to say what we will end up setting ℓ to be, e.g., "Interestingly, we choose ℓ ...** ⋙

Definition A.1. *Given some 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.*

First, we compute the Wasserstein distance between the spectra of the two graphs.

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

Proof. We can compute the exact eigenvalues of the two graphs by combining Lemma 2.3 with the fact that eigenvalues just increase in multiplicity with repeated components. As in that lemma, recall we use R_ℓ denote a cycle of length ℓ and $R_{2\ell}$ a cycle of length 2ℓ. The Wasserstein distance between R_ℓ^2 and $R_{2\ell}$ is $W_1(R_\ell^2, R_{2\ell}) = 1/\ell$. Due to the dual⋘ **Sidford: for consistency (and other reasons) I would make sure we say "dual characterization" instead of "dual definition" everywhere** ⋙ definition of Wasserstein distance in (2), we can cancel out the common disconnected components⋘ **Sidford: I'm not exactly sure what "disconnected components" refer to technically, can we say something more technical?** ⋙ which lead to common (multiplicity of same) eigenvalues in both graphs, leaving G_1 with $(2\alpha-1)n$ extra $R_{2\ell}$ cycles and G_2 with $(2\alpha-1)n$ extra copies of R_ℓ^2. Thus, we can compute the Wasserstein distance directly using ⋘ **Sidford: If distributoin 1 has α copies of A and $n-\alpha$ copies of B and distribution 2 has β copies of A ad $n-\beta$ copies of B, is it alwasy the case that there difference is $|\alpha-\beta|/n$ times the difference between A and B? Because of the sorting, its not immediatly clear (but I could be missing a trick). I wonder if it is worth proving this (and some of the other Wasserstein distance facts we use, e.g. that disjoint duplication doesn't increase distance)** ⋙ ⋘ **Apoorv: the "missing trick" is we use the actual def of W1 distance, i.e., how much I have to shift one distribution to get the other. Since exactly eigs overlap, we get $|\alpha-\beta|/n$ times the difference between A and B** ⋙⋘ **Yujia: addressed by explaining more details, see if okay then remove** ⋙ ⋘ **Sidford: Thanks! It makes sense and looks better, but I worry that the "Cancel Out" terminology is a bit too informal. Maybe just say that if p_1 and p_2 are the densities originall and $\tilde{p}_1$ and $\tilde{p}_2$ are the ones of just the etra cycle then $\tilde{p}_1(x) - \tilde{p}_2(x) = p_1(x) = p_2(x)$ for all x and by the dual characterization ...** ⋙

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

Lower bound. Our lower bound argument follows from applying the Le Cam's inequality. We consider any testing procedure that takes in m adaptive random walk samples and consider $\ell = 1$

(reducing to a self-cycle for R_1) for simplicity.

Lemma A.3. *When $\ell = 1$, using the adaptive random walk model we sample transcripts $S = \{S_1, S_2, \cdots, S_m\}$, such that $S \sim P_1$ on G_1, $S \sim P_2$ on G_2, then* ⋘ **Sidford: where did we say that the permutation was random? we need this for the lemma to be true right?** ⋙

$$\mathrm{d_{TV}}(P_1, P_2) \le \frac{2m^2}{n} + \sqrt{\frac{m}{2}} \cdot \sqrt{\alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha)\log\frac{1-\alpha}{\alpha}}.$$

.

Proof. We treat S to be unchanged up to permutation of vertex labeling without loss of generality (as we can always label as we go).⋘ **Sidford: I'm not sure I understand this sentence; we have some restrictions on labeling as we go right? I think it might be good to make this a little more formal.** ⋙ Also, we will treat P_i to be generating trajectories from random starting vertices without replacement, as this is the most efficient way to gain information of the two graphs.⋘ **Sidford: I feel this sentence needs to be justified a little more precisely / rigorously (is there a way to just more concretely map the realizatio to the i.i.d. samples or something?** ⋙

For simplicity we first let $\hat{P}_1$ and $\hat{P}_2$ denote the distribution of ⋘ **Sidford: I think there is a typo or gramatical isssue here** ⋙ sampling random walk according to length of $S_1, \cdots, S_m$ from a uniformly random vertex (with replacement), and $\hat{P}_1^m$ and $\hat{P}_2^m$ be the distribution of generating m such samples vertices in an i.i.d. manner. Since each adaptive random walk can identify the size of the current cycle within 1 step (given $\ell = 1$), each random walk trajectory gives a sample of the Bernoulli variable with mean $1-\alpha$ or α, and we have ⋘ **Sidford: What is being used below? it would be good to say** ⋙

$$\begin{aligned}\mathrm{d_{TV}}(\hat{P}_1^m, \hat{P}_2^m) &\le \sqrt{\frac{m}{2}}\sqrt{D_{\mathrm{KL}}(\hat{P}_1, \hat{P}_2)} = \sqrt{\frac{m}{2}}\sqrt{D_{\mathrm{KL}}(\mathsf{Ber}(1-\alpha), \mathsf{Ber}(\alpha))} \\ &= \sqrt{\frac{m}{2}}\sqrt{\alpha\log(\alpha/(1-\alpha)) + (1-\alpha)\log((1-\alpha)/\alpha)}\,.\end{aligned}$$

⋘ **Sidford: Maybe instead; "we next argue that" ... (when I see "note" I think it is just being asserted rather then the proof is about to be given** ⋙ We further note that due to the large size of n, one can show $\hat{P}_i^m$ and P_i are close for $i \in \{1, 2\}$. In particular, we let $\mathcal{E}$ to be the event of the generated samples satisfying the replacement assumptions under P_1, for any $(S_1, S_2, \cdots, S_m)$ one has

$$\begin{aligned}\hat{P}_1^m(S_1, S_2, \cdots, S_m) &= \hat{P}_1^m(S_1, S_2, \cdots, S_m|\mathcal{E}) \cdot P(\mathcal{E}) + \hat{P}_1^m(S_1, S_2, \cdots, S_m|\mathcal{E}^c) \cdot P(\mathcal{E}^c) \\ &= P_1(S_1, S_2, \cdots, S_m) \cdot P(\mathcal{E}) + \hat{P}_1^m(S_1, S_2, \cdots, S_m|\mathcal{E}^c) \cdot P(\mathcal{E}^c).\end{aligned}$$

Since ⋘ **Sidford: might be good to do a find-replace to make sure we are abbreviating "with high probability" the same everywhere** ⋙ with high probability ⋘ **Sidford: maybe instead of "no same nodes" we just say that the elements are "distinct" or something?** ⋙ no same nodes will appear in $\{X_i\}_{i\in[m]}$ we have $P(\mathcal{E}^c) \le m^2\frac{1}{n}$. Consequently, this implies $\mathrm{d_{TV}}(\hat{P}_1^m, P_1) \le \frac{m^2}{n}$ and symmetrically $\mathrm{d_{TV}}(\hat{P}_2^m, P_2) \le \frac{m^2}{n}$.

Summing the proven inequalities together and using the triangle inequality of total variation distance gives the final claim. ■

Consequently, applying Le Cam's inequality Le Cam [Le 12] we have: Given a distribution over uniformly random vertex permutation under G_1 (with probability 1/2) or G_2 (with probability 1/2), any test Ψ that takes in m total samples of adaptive random walk and output graph G_1 or G_2 will make the wrong prediction with probability at least

$$p = \frac{1}{2}\left(1 - \frac{2m^2}{n} - \sqrt{\frac{m}{2}} \cdot \sqrt{\alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha)\log\frac{1-\alpha}{\alpha}}\right). \tag{6}$$

Theorem A.4. *Given any $\varepsilon \le 1/6$, there exists a distribution over unweighted graphs $G \sim \mathcal{D}$ so that, given the transcript of m adaptive randoms walks no algorithm can estimate the spectral density of G's normalized adjacency matrix to accuracy ε in Wasserstein-1 distance with probability $> 3/4$ unless $m > 1/(32\varepsilon^2)$.*

Proof. We pick $\ell = 1$, $\alpha = (1+\varepsilon)/2$, $m \le 1/(32\varepsilon^2) \le \lfloor (1-\varepsilon)/(16\varepsilon^2) \rfloor$ and $n \ge 8m^2$. Plugging these choice of parameters in Lemma A.2 gives the Wasserstein distance bound and in Eq. (6) gives the failure probability. ■

B Wasserstein Distance Bounds via Legendre Polynomials

In this section, we give an alternative proof of a lower-bound by Kong and Valiant [KV17], 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 whose ℓ-th moments differ by $\Omega(2^{-\ell})$.

Lemma B.1 (Improvement of [KV17, 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 $W_1(p, q) \ge 2/\ell$.*

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 2.3. 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 2.4).

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 2.3.

1. $A(R_\ell^2)$ has $(\ell+1)/2$ unique eigenvalues, and $A(R_{2\ell})$ has $\ell + 1$ unique eigenvalues.
2. All unique eigenvalues of $A(R_\ell^2)$ overlaps 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, $\forall \lambda \in \boldsymbol{\lambda}(R_\ell^2)$,

$$\left|\left\{j \ : \ \lambda_j \in \boldsymbol{\lambda}(R_\ell^2), \lambda_j = \lambda, j \in [2\ell]\right\}\right| = 2 \cdot \left|\{j \ : \ \lambda_j \in \boldsymbol{\lambda}(R_{2\ell}), \lambda_j = \lambda, j \in [2\ell]\}\right| . \tag{7}$$

3. All the eigenvalues of $A(R_{2\ell})$ lies in $[-1, 1]$.

Let $\mathbf{\Lambda}^{(2)}$ 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 $\mathbf{\Lambda}^{(1)}$ be the set of removed eigenvalues. The following observations follow from (7). The size of $\mathbf{\Lambda}^{(2)}$, and $\mathbf{\Lambda}^{(1)}$ is ℓ. Moreover $\mathbf{\Lambda}^{(1)}$ has same eigenvalues as $\boldsymbol{\lambda}(R_\ell^2)$ where frequency of each unique eigenvalue in $\boldsymbol{\lambda}(R_\ell^2)$ is reduced by a factor of 2. Consequently, we define $p(x) = \frac{1}{\ell}\sum_{j\in[\ell]} \delta\left(x - \mathbf{\Lambda}_j^{(1)}\right) = p'(x)$, and $q(x) = \frac{1}{\ell}\sum_{j\in[\ell]} \delta\left(x - \mathbf{\Lambda}_j^{(2)}\right) = 2q'(x) - p'(x)$. 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 definition of Wasserstein distance in (2) and the last equality follows from Lemma 2.3. ■

We complement Proposition B.1 with the following Lemma B.3, 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 [AI19].

Fact B.2. *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 B.3. *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 B.2 we know that $f_\ell(x) = \int T_{\ell-1}(x)\mathsf{d}x$ (with arbitrary constant) is a degree ℓ, 1-Lipschitz polynomial in $[-1, 1]$, with leading coefficient $2^{\ell-2}/\ell$. Define $g_\ell(x)$ as follows:

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

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. ■

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 $\hat{p}_j \leftarrow \hat{p}_j + \hat{p}_{j,x}$
Construct a distribution p on $[-1,1]$ with first k moments equal to $\{\hat{p}_j\}_{j\in[k]}$
Return: p

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 $\tilde{A}_1$ and $\tilde{A}_2$ with a common degree matrix D. Here we consider learning the spectrum of their difference matrix, i.e., $A(G_1) - A(G_2) = D^{-1/2}\tilde{A}_1 D^{-1/2} - D^{-1/2}\tilde{A}_2 D^{-1/2}$, or equivalently $D^{-1}(\tilde{A}_1 - \tilde{A}_2)$. 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. [CKSV18].

We first restate the main theorem in Kong and Valiant [KV17] for completeness.

Theorem C.1 (Kong and Valiant [KV17, 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 $W_1(p, q) \le C(\frac{b-a}{k} + 3^k(b-a)\|\alpha - \beta\|_2)$ for some absolute constant C.*

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 $D^{-1}\tilde{A}_1$ when $x_i = 1$ and $D^{-1}\tilde{A}_2$ when $x_i = 0$. We consider Algorithm 1 for estimating the spectral density of matrix $D^{-1}(\tilde{A}_1 - \tilde{A}_2)$.

Algorithm 1 computes estimates of the moments of difference matrix $D^{-1}(\tilde{A}_1 - \tilde{A}_2)$. Together with the procedure of computing a distribution based on first k moments using linear programming as stated in Cohen-Steiner et al. [CKSV18], we have the following guarantee.

Theorem C.2. *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 $D^{-1}(\tilde{A}_1 - \tilde{A}_2)$, whose j^{th} moment is $\frac{1}{n}\mathrm{tr}((D^{-1}\tilde{A}_1 - D^{-1}\tilde{A}_2)^j)$.

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

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

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 $\frac{1}{n}\mathrm{tr}(\prod_{i=1,\cdots,j}(x_i \cdot D^{-1}\tilde{A}_1 + (1-x_i)\cdot D^{-1}\tilde{A}_2))$ (as in Line 6). By concentration we can estimate each term $\frac{1}{n}\mathrm{tr}(\prod_{i=1,\cdots,j}(x_i \cdot D^{-1}\tilde{A}_1 + (1-x_i)\cdot D^{-1}\tilde{A}_2))$ using $\hat{p}_{j,x}$ (as in Line 7) up to $\theta/2^j$ additive accuracy with high probability $1-\delta/(k2^j)$ using a total of $\frac{1}{2}\theta^{-2}j4^j\log(2k/\delta)$ 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 $O(\theta^{-2}j2^{3j}\log(2k/\delta))$ 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 C.1 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

$$\begin{aligned}&\sum_{j\in[k]} O(\theta^{-2}j2^{3j}\log(2k/\delta)) \text{ calls to } \widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot)\\ &\qquad = \sum_{j\in[O(1/\varepsilon)]} 2^{O(1/\varepsilon)} \text{ calls to } \widetilde{\mathcal{O}}_{\mathsf{NA-RW}}(G_1, G_2, j, \cdot).\end{aligned}$$

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.