A Matrix Chernoff Bound for Strongly Rayleigh Distributions and Spectral Sparsifiers from a few Random Spanning Trees

Rasmus Kyng
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA
kyng@seas.harvard.edu

Zhao Song
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA
zhaos@g.harvard.edu

Abstract—
Strongly Rayleigh distributions are a class of negatively dependent distributions of binary-valued random variables [Borcea, Brändén, Liggett JAMS 09]. Recently, these distributions have played a crucial role in the analysis of algorithms for fundamental graph problems, e.g. Traveling Salesman Problem [Gharan, Saberi, Singh FOCS 11]. We prove a new matrix Chernoff bound for Strongly Rayleigh distributions.

As an immediate application, we show that adding together the Laplacians of $\epsilon^{-2} \log^2 n$ random spanning trees gives an $(1 \pm \epsilon)$ spectral sparsifiers of graph Laplacians with high probability. Thus, we positively answer an open question posted in [Baston, Spielman, Srivastava, Teng JACM 13]. Our number of spanning trees for spectral sparsifier matches the number of spanning trees required to obtain a cut sparsifier in [Fung, Hariharan, Harvey, Panigraphi STOC 11]. The previous best result was by naively applying a classical matrix Chernoff bound which requires $\epsilon^{-2} n \log n$ spanning trees. For the tree averaging procedure to agree with the original graph Laplacian in expectation, each edge of the tree should be reweighted by the inverse of the edge leverage score in the original graph. We also show that when using this reweighting of the edges, the Laplacian of single random tree is bounded above in the PSD order by the original graph Laplacian times a factor $\log n$ with high probability, i.e. $L_T \preceq O(\log n) L_G$.

We show a lower bound that almost matches our last result, namely that in some graphs, with high probability, the random spanning tree is *not* bounded above in the spectral order by $\frac{\log n}{\log \log n}$ times the original graph Laplacian.

We also show a lower bound that in $\epsilon^{-2} \log n$ spanning trees are necessary to get a $(1 \pm \epsilon)$ spectral sparsifier.

Keywords-Matrix Chernoff bounds; random spanning trees; Strongly Rayleigh; Spectral sparsifier;

I. INTRODUCTION

The idea of concentration of sums of random variables dates back to Central Limit Theorems, and hence de Moivre and Laplace [Tij], while modern concentration bounds for sums of random variables were perhaps first established by Bernstein [Ber24], and a popular variant now known as Chernoff bounds was introduced by Rubin and published by Chernoff [Che52].

Concentration of measure for matrix-valued variables is the phenomenon that many matrix valued distributions are to close their mean with high probability, closeness usually being measured by spectral norm. Modern quantitative bounds of the form often used in theoretical computer science were derived by for example by Rudelson [Rud99], while Ahlswede and Winter [AW02] established a useful matrix-version of the Laplace transform that plays a central role in scalar concentration results such as those of Bernstein. [AW02] combined this with the Golden-Thompson trace inequality to prove matrix concentration results. Tropp refined this approach, and by replacing the use of Golden-Thompson with deep a theorem on concavity of certain trace functions due to Lieb, Tropp was able to recover strong versions of a wide range of scalar concentration results, including matrix Chernoff bounds, Azuma and Freedman's inequalities for matrix martingales [Tro12].

Matrix concentration results have had an enormous range of applications in computer science, and are ubiquitous throughout spectral graph theory [ST04], [SS11], [CKP$^+$17], sketching [Coh16], approximation algorithms [HSSS16], and deep learning [ZSJ$^+$17], [ZSD17]. Most applications are based on results for independent random matrices, but more flexible bounds, such as Tropp's Matrix Freedman Inquality [Tro11a], have been used to greatly simplify algorithms, e.g. for solving Laplacian linear equations [KS16] and for semi-streaming graph sparsification [AG09], [KPPS17]. Matrix concentration results are also closely related to other popular tools sampling tools, such as Karger's techniques for generating sparse graphs that approximately preserve the cuts of denser graphs [BK96].

Negative dependence of random variables is an appealing property that intuition suggests should help with concentration of measure. Notions of negative dependence can be formalized in many ways. Roughly speaking, these notions characterize distributions where where some event occurring ensures that other events of interest become less likely. A simple example is the distribution of a sequence of coin flips, conditioned on the total number of heads in the outcome. In this distribution, conditioning on some coin coming out heads makes all other coins less likely to come out heads. Unfortunately, negative dependence phenomena are not as robust as positive association which can be established from

IEEE
computer
society

local conditions using the powerful FKG theorem [FKG71].

Strongly Rayleigh distributions were introduced recently by Borcea, Brändén, and Liggett [BBL09] as a class of negatively dependent distributions of binary-valued random variables with many useful properties. Strongly Rayleigh distributions satisfy useful negative dependence properties, and retain these properties under natural conditioning operations. Strongly Rayleigh distributions also satisfy a powerful stability property under conditioning known as *Stochastic Covering* [PP14], which is useful for analyzing them through martingale techniques. A measure on $\{0,1\}^n$ is said to be Strongly Rayleigh if its generating polynomial is real stable [BBL09]. There are many interesting examples of Strongly Rayleigh distributions [PP14]: The example mentioned earlier of heads of independent coin flips conditional on the total number of heads in the outcome; symmetric exclusion processes; determinental point processes and determinental measures on a boolean lattice. An example of particular interest to us is the edges of uniform or weighted random spanning trees, which form a Strongly Rayleigh distribution.

We prove a Matrix Chernoff bound for the case of k-homogeneous Strongly Rayleigh distributions. Our bound is slightly weaker than the bound for independent variables, but we can show that it is tight in some regimes. We use our bound to show new concentration results related to random spanning trees of graphs. An open question is to find other interesting applications of our concentration result, e.g. by analyzing concentration for matrices generated by exclusion processes.

Random spanning trees are one among the most well-studied probabilistic objects in graph theory, going back to the work of Kirchoff [Kir47] in 1847, who gave formula relating the number of spanning trees in a graph to the determinant of the Laplacian of the graph.

Algorithms for sampling of random spanning trees have been studied extensively, [Gue83], [Bro89], [Ald90], [Kul90], [Wil96], [CMN96], [KM09], [MST15], [HX16], [DKP+17], [DPPR17], [Sch18], and a random spanning tree can now be sampled in almost linear time [Sch18].

In theoretical computer science, random spanning trees have found a number of applications, most notably in breakthrough results on approximating the traveling salesperson problem with symmetric [GSS11] and asymmetric costs [AGM+10]. Goyal et al. [GRV09] demonstrated that adding just two random spanning trees sampled from a bounded degree graph gives a $O(\log n)$ cut sparsifier with probability $1 - o(1)$. Later, it was shown by Fung, Hariharan, Harvey, Panigraphi [FHHP11], that if we sample $O(\epsilon^{-2} \log^2 n)$ random spanning trees from a graph, reweight the tree edges by the inverse of their leverage scores in the original graph, and average them together, then whp. we get a graph where every the weight of edges crossing every cut is approximately the same in as in the original graph, up to a factor $(1 \pm \epsilon)$. We refer to this as an ϵ-cut sparsifier.

The techniques of Fung et al. unfortunately do not extend to proving spectral sparsifiers.

Spectral graph sparsifiers were introduced by Spielman and Teng [ST04], who for any graph G showed how to construct a another graph H with $\epsilon^{-2} n \operatorname{poly} \log n$ edges s.t. $(1 - \epsilon)L_G \preceq L_H \preceq (1 + \epsilon)L_G$, which we refer to as an ϵ-spectral sparsifier. The construction was refined by Spielman and Srivastava [SS11], who suggested sampling edges independently[1] with probability proportional to their *leverage scores*, and brought the number of required samples down to $\epsilon^{-2} n \log n$. This analysis is tight in the sense that if fewer than $o(\epsilon^{-2} n \log n)$ samples are used, there will be at least a $1/\operatorname{poly}(n)$ probability of failure. Meanwhile, $\epsilon^{-2} n \frac{\log n}{\log \log n}$ independent samples in a union of cliques can be shown whp. to fail to give a cut sparsifier. This can be observed directly from the degree distribution of a single vertex in the complete graph. For a variant of [SS11] sampling based on flipping a single coin for each edge to decide whether to keep it or not, it can also be shown that when the expected number of edges is $\epsilon^{-2} n \frac{\log n}{\log \log n}$, whp. the procedure fails to give a cut sparsifier. For arbitrary sparsification schemes, bounds in [BSS12] show that $\Theta(\epsilon^{-2} n)$ edges are necessary and sufficient to give an ϵ-spectral sparsifier.

The marginal probability of an edge being present in a random spanning tree is exactly the leverage score of the edge. This seems to suggest that combining $\epsilon^{-2} \operatorname{poly} \log n$ spanning trees might give a spectral sparsifier, but the lack of independence between the sampled edges means the process cannot be analyzed using existing techniques. Observing this, Baston, Spielman, Srivastava, Teng [BSST13] in their excellent 2013 survey on sparsification noted that *"it remains to be seen if the union of a small number of random spanning trees can produce a spectral sparsifier."* We answer this question in the affirmative. In particular, we show that adding together $O(\epsilon^{-2} \log^2 n)$ spanning trees with edges scaled proportional to inverse leverage scores in the original graph leads to a ϵ-spectral sparsifier. This matches the bound obtained for cut sparsifiers in [FHHP11]. Our result also implies their earlier bound since a spectral sparsifier is always a cut sparsifier with the same approximation quality. Before our result, only a trivial bound on the number of spanning trees required to build a spectral sparsifier was known. In particular standard matrix concentration arguments like those in [SS11] prove that $O(\epsilon^{-2} n \log n)$ spanning trees suffice. Lower bounds in [FHHP11] show that whp. $\Omega(\log n)$ random spanning trees are required to give a constant factor spectral sparsifier.

We show that whp. $\epsilon^{-2} \frac{\log n}{\log \log n}$ random spanning trees do not give an ϵ-spectral sparsifier. We also show that the Laplacian of a single random tree with edges weighted as

---

[1][SS11] analyzed sampling with replacement, but based on [Tro12], a folklore result shows the same behavior can be obtained by doing independent coin flips for every edge with low leverage score, again with inclusion probabilities proportional to leverage scores.

above satisfies $L_T \preceq O(\log n) L_G$ whp., and we give an almost matching lower bound, showing that in some graphs whp. $L_T \npreceq O(\frac{\log n}{\log \log n}) L_G$. Before our work, the main result known about approximating graphs using $O(1)$ random spanning trees is due to Goyal, Rademacher, Vempala [GRV09], who showed that surprisingly, when the original graph has bounded degree, adding two random spanning trees gives a graph whose cuts approximate the cuts in the original graph up to a factor $O(\log n)$ with good probability. Our result for a single tree establishes only a one-sided bound an interesting open question remains: Does sampling $O(1)$ random spanning trees give a $\log n$-factor spectral sparsifier with, say, constant probability?

A. Previous work

Chernoff-type bound for matrices.: Chernoff-like bounds for matrices appear in Rudelson [Rud99] and Ahlswede and Winter [AW02]. The latter introduced a useful matrix-variant of the Laplace transform that is central in concentration bounds for scalar-valued matrices. Their bounds restricted to iid random matrices, an artifact of their use of the Golden-Thompson inequality for bounding traces. In contrast, Tropp obtained more flexible concentration bounds for random matrices by using a result of Lieb to bound the expected trace of various operators [Tro12], including bounds for matrix martingales [Tro11b].

In a recent work by Garg, Lee, Song and Srivastava [GLSS18], they show a Chernoff bound for sums of matrix-valued random variables sampled via a random walk on an expander graph. This work confirms a conjecture due Wigderson and Xiao. The proof of Garg et al. is also concerned with matrices that are not fully independent. In this case the matrices are generated from random walks on an expander graph. The main idea to deal with dependence issue is using a new multi-matrix extension of the Golden-Thompson inequality and an adaptation of Healy's proof of the expander Chernoff bound in the scalar's case [Hea08] to matrix case. Their techniques deal with fairly generic types of dependence, and cannot leverage the very strong stability properties that arise from the negative dependence and stochastic covering properties of Strongly Rayleigh distributions. Harvey and Olver [HO14] proved a matrix concentration result for randomized pipage rounding, which can be used to show concentration results for random spanning trees obtain from pipage rounding, but not for (weighted) uniformly random spaning trees. The central technical element of their proof is a new variant of a theorem of Lieb on concavity of certain matrix trace functions.

Matrix martingales have played a central role in a number of algorithmic results in theoretical computer science [KS16], [CMP16], [KPPS17], but beyond a reliance on Tropp's Matrix Freedman Inequality, these works have little in common with our approach. However, our bound does share a technical similarity with [KS16], namely that a sequence of increasingly restricted random choices in a martingale process lead to a $\log n$ factor in a variance bound.

Strongly Rayleigh Distributions in Theoretical Computer Science.: Perhaps the most prominent result on Strongly Rayleigh distributions in theoretical computer science is the generalization of [MSS13] to Strongly Rayleigh distributions.

The central technical result of [MSS13] essentially shows that given a collection of independent random vectors $v_1, \ldots, v_m$ with finite support in $\mathbb{C}^n$ s.t. $\sum_{i=1}^m \mathbb{E}[v_i v_i^*] = I$ and for all i, $\|v_i\|^2 \leq \epsilon$, then $\Pr[\|\sum_{i=1}^m v_i v_i^*\| \leq (1+\sqrt{\epsilon})^2] > 0$. [AG15] establishes a related result for k-homogeneous Strongly Rayleigh distributions, though they require an additional constraint on the marginal probability that any given random variable is non-zero being bounded above by δ, and then establish $\Pr[\|\sum_{i=1}^m v_i v_i^*\| \leq 4(\epsilon + \delta) + 2(\epsilon + \delta)^2] > 0$. Based on this, [AG15] shows[2] that given an unweighted k-edge connected graph G where every edge has leverage score at most ϵ, there exists an unweighted spanning tree s.t. $L_T \preceq O(\frac{1}{k} + \epsilon) \cdot L_G$. This is referred to as a spectrally thin tree with parameter $O(\frac{1}{k} + \epsilon)$.

[AGR16] showed how to algorithmically sample from k-homogeneous Determinental Point Process in time $\text{poly}(k) n \log(n/\epsilon)$, where n is the dimension of matrix giving rise to the determinental point process and ϵ is the allowed total variation distance. Their techniques are based on generalization proofs of expansion in the base graph associated with a balanced matroid, a result first established by [FM92].

Random spanning trees.: Algorithms for sampling random spanning trees have a long history, but only recently have they explicitly used matrix concentration [DKP+17], [DPPR17], [Sch18]. The matrix concentration arguments in these papers, however, deal mostly with how modifying a graph results in changes to the distribution of random spanning trees in the graph. We instead study how closely random spanning trees resemble the graph they were initially sampled from. Whether our result in turn has applications for improving sampling algorithms for random spanning trees is unclear.

The fact that spanning tree edges exhibit negative dependence has been used strikingly in concentration arguments by Goyal et al. [GRV09] to show that two random spanning trees gives $O(\log n)$-factor approximate cut sparsifier in bounded degree graphs, with good probability. This is clearly false when sampling the same number of edges independently, because this graph has large probability of having isolated vertices. Goyal et al. improve over independent sampling by leveraging the fact that for a fixed tree, in some sense, very few cuts of a given size exist. This is a variant of Karger's famous cut-counting techniques [Kar93],

---

[2]Their full statement is more general, see [AG15] Corollary 1.9.

[KS96] specialized to unweighted trees.

Uses of negatively dependent Chernoff bounds applied to tree edges also appeared in works on approximation algorithms for TSP problems [GSS11], [AGM$^+$10], where additionally the connectivity properties of the tree play an important role

In contrast, the techniques of Fung et al. [FHHP11] show that $O(\epsilon^{-2}\log^2 n)$ spanning trees suffice to give a $(1 \pm \epsilon)$-cut sparsifier, but they do not show that tree-based sparsifiers improve over independent sampling, The focus of their paper is to establish that wide range of different techniques for choosing sampling probabilities all give cut sparsifiers, by establishing a more flexible framework than the original cut-sparsifier results of Benczur-Karger [BK96], using related cut-counting techniques (see [Kar93], [KS96]). To extend their results to spanning trees, they simply observe that the (scalar-valued) Chernoff bounds use immediately apply to negatively dependent variables, and hence edges in spanning trees.

Fung et at. [FHHP11] also establish a lower bound, showing that for any constant c, there exists a graph for which obtaining a factor c-cut sparsifier by averaging trees requires using at least $\Omega(\log n)$ trees to succeed with constant probability.

B. Our results and techniques

Theorem I.1. (First main result, a Matrix Chernoff Bound k-homogeneous Strongly Rayleigh Distributions)**.** *Suppose* $(\xi_1, \ldots, \xi_m) \in \{0,1\}^m$ *is a random vector of* $\{0,1\}$ *variables whose distribution is k-homogeneous and Strongly Rayleigh.*

Given a collection of PSD matrices $A_1, \ldots A_m \in \mathbb{R}^{n \times n}$ s.t. for all $e \in [m]$ we have $\|A_e\| \leq R$ and $\|\mathbb{E}[\sum_e \xi_e A_e]\| \leq \mu$.

Then for any $\epsilon > 0$,

$$\Pr\left[\left\|\sum_e \xi_e A_e - \mathbb{E}\left[\sum_e \xi_e A_e\right]\right\| \geq \epsilon\mu\right]$$
$$\leq n\exp\left(-\frac{\epsilon^2\mu}{R(\log k + \epsilon)}\Theta(1)\right)$$

This Matrix Chernoff bound matches the bounds due to Tropp [Tro12], up to the $\log k$ factor in the exponent. We know that the bound is essentially tight in some cases when $\epsilon \approx \frac{\log k}{\log\log k}$, since if we could replace the $\log k$ factor in Theorem I.1 with a factor $\frac{\log k}{\log\log k}$, then applying this strengthened version to improve the concentration in Theorem I.5 would lead to a contradiction of the lower bound in Theorem I.6. We do not know if the theorem is tight in other regimes, and it seems plausible that it should be possible to improve Theorem I.3 to show that $O(\epsilon^{-2}\log n)$ suffice to get an ϵ-spectral sparsifier. This perhaps suggests there are regimes where our Theorem I.1 is not tight.

Remark I.2. *When the Strongly Rayleigh distribution is in fact a product distribution on a collection of l Strongly Rayleigh distributions that are each t-homogeneous, then the joint distribution is lt-homogeneous Strongly Rayleigh, but in Theorem I.1, the factor $\log(lt)$ can be replaced by a factor $\log t$. This applies to the case of independent random spanning trees, but only gives a constant factor improvement in the number of trees required.*

Our work is related to the concentration inequality of Peres and Pemantle [PP14], who showed a concentration result for scalar-valued Lipschitz functions of Strongly Rayleigh distributions. They used Doob martingales (martingales constructed from sequences of conditional expectations) to prove their result. We use a similar approach for matrices, constructing Doob matrix martingales from our Strongly Rayleigh distributions. In addition, we use the stochastic covering property of Strongly Rayleigh distributions observed by Peres and Pemantle, but implicitly derived in [BBL09]. This property leads to bounded differences in Doob martingale sequences for scalars. As in the scalar setting, it is possible to show concentration results for matrix-valued martingales. We use the Matrix Freedman inequality[3] of Tropp. This inequality allows makes it possible to establish strong concentration bounds based on control of sample norms and control of the *predictable quadratic variation process* of the martingale, a matrix-valued object that is used to measure variance (see [Tro11b]). We show that as in the scalar setting, the stochastic covering property of Strongly Rayleigh distributions leads to bounded differences for Doob matrix martingales. But, we also combine the stochastic covering property with deceptively simple matrix martingale properties and a negative dependence condition to derive additional bounds on the predictable quadratic variation process of the martingale. The key negative dependence property we use is a simple observation that generalizes, to k-homogeneous Strongly Rayleigh distributions, the fact that in a random spanning tree, conditioning on the presence of a set of edges lowers the marginal probability of every other graph edge being in the tree (see Lemma I.10). While we frame it differently, it is essentially an immediate consequence of statements in [BBL09]. The surprise here is how useful this simple observation is for removing issues with characterizing conditional k-homogeneous Strongly Rayleigh distributions. As a corollary we get our second main result.

Theorem I.3. (Second main result, concentration bound of a batch of independent random spanning trees)**.** *Given as input*

---

[3]Note, however, that we are able to prove deterministic bounds on the *predictable quadratic variation process*, which means the bound we use is more analogous to a matrix version Bernstein's inequality, adapted to martingales. We resort to the more complicated Freedman's inequality only because it gives a directly applicable statement that is known in the literature.

a weighted graph G with n vertices and a parameter $\epsilon > 0$, let $T_1, T_2, \cdots, T_t$ denote t independent inverse leverage score weighted random spanning trees, if we choose $t = O(\epsilon^{-2} \log^2 n)$ then with probability $1 - 1/\operatorname{poly}(n)$,

$$(1 - \epsilon) L_G \preceq \frac{1}{t} \sum_{i=1}^{t} L_{T_i} \preceq (1 + \epsilon) L_G.$$

Prior to our work, only a trivial bound on the number of spanning trees required to build a spectral sparsifier was known, namely that standard matrix concentration arguments like those in [SS11] prove that $O(\epsilon^{-2} n \log n)$ spanning trees suffice. Note that, the number of spanning trees required to build spectral sparsifier in our Theorem I.3 matches the number of spanning trees required to construct cut sparsifier in previous best result [FHHP11]. The total edge count we require is $\Theta(\epsilon^{-2} n \log^2 n)$, worse by a factor $\log n$ than the bound for independent edge sampling obtained in [SS11]. It is not clear whether this factor in necessary.

Remark I.4. *Suppose we apply our Theorem I.5 to show that any single random spanning tree satisfies $L_T \preceq O(\log n) \cdot L_G$ whp. This is tight up a $\log \log n$ factor. Then, one can from use this to derive Theorem I.3 based on a standard (and tight) Matrix Chernoff bound, and a (laborious) combination of Doob martingales and stopping time arguments similar to those found in [Tro12], [KS16]. This line of reasoning will lead to the same bounds as Theorem I.5. Thus, unless one proves more than just a norm bound for each individual tree, it is not possible improve over our result, except for $\log \log n$ factors.*

Like the work of Fung et al. our results for spanning trees do not improve over the independent case. Fung et al. achieved their result by combining cut counting techniques with Chernoff bounds for scalar-valued negatively dependent variables. In our random matrix setting, there are no clear candidates for a Chernoff bound for negatively dependent random matrices that we can adopt, and this type of bound is exactly what we develop in the Strongly Rayleigh case.

We establish a one-sided concentration result for a single tree, namely that whp. $L_T \preceq O(\log n) \cdot L_G$. Again, this is a direct application of Theorem I.1.

Theorem I.5. (Third main result, upper bound for the concentration of one random spanning tree)**.** *Given a graph G, let T be a random spanning tree, then with probability at least $1 - 1/\operatorname{poly}(n)$*

$$L_T \preceq O(\log n) \cdot L_G.$$

This upper bound is tight up to a factor $\log \log n$ as shown by our almost matching lower bound stated below.

Theorem I.6 (Lower bound for the concentration of one random spanning tree)**.** *There is an unweighted graph G,*

s.t. if we sample an inverse leverage score weighted random spanning tree T, then with probability at least $1 - e^{-n^{.99}}$,

$$L_T \not\preceq O(\log n / \log \log n) \cdot L_G.$$

Trivially, the presence of degree one nodes in L_T means that in a complete graph, $L_G \not\preceq c L_T$ for any $c < 1$. So choosing any other scaling of the tree will make at least one of the inequalities $L_T \not\preceq O(\log n / \log \log n) \cdot L_G$ and $L_G \not\preceq 0.99 L_T$ true with a larger gap. Note that in the complete unweighted graph the trees we consider have weight $\Theta(n)$ on each edge. A random spanning tree in the complete graph has diameter about $\sqrt{n}$ [RS67]. This can be shown to imply that for an *unweighted* random tree $\widehat{T}$, $L_G \not\preceq \sqrt{n} L_{\widehat{T}}$. But once we scale up every edge of the tree by a factor $\Theta(n)$, the diameter bound no longer directly implies a spectral gap of the form $L_G \not\preceq \alpha L_T$ for some α. In a ring graph, we get $L_G \not\preceq (n - 2) L_T$ and $L_T \not\preceq L_G$.

We can also show in general that $L_T \not\preceq O(\log n) \cdot L_G$, but with a much smaller probability.

Theorem I.7 (Lower bound for the concentration of one random spanning tree)**.** *There is an unweighted graph G, s.t. if we sample an inverse leverage score weighted random spanning tree T, then with probability at least $2^{-\log n \log \log n}$,*

$$L_T \not\preceq \Omega(\log n) \cdot L_G.$$

And we show a lower bound for ϵ-spectral sparsifiers for random spanning trees.

Theorem I.8 (Lower bound for the concentration of multiple random spanning trees)**.** *There is an unweighted graph G, s.t. for any accuracy parameter ϵ, if we sample $t = O(\epsilon^{-2} \log n)$ independent random spanning trees with edges weighted by inverse leverage score, then with probability at least $1 - e^{-n^{.99}}$,*

$$(1 - \epsilon) L_G \not\preceq \frac{1}{t} \sum_{i=1}^{t} L_{T_i} \text{ and } \frac{1}{t} \sum_{i=1}^{t} L_{T_i} \not\preceq (1 + \epsilon) L_G.$$

Our lower bound is incomparable with that of Fung et al.[FHHP11], who showed that for any constant c, there exists a graph for obtaining a factor c-cut sparsifier by averaging trees requires using at least $\Omega(\log n)$ trees to succeed with constant probability. Where Fung et. al [FHHP11] used triangles in their lower bound construction, our bad examples are based on collections of small cliques, which lets us ensure cut differences in even a single tree, by giving longer-tailed degree distributions. All of our lower bounds are based on simple constructions from collections of edge disjoint cliques, and use the fact that the exact distribution of degrees in a random spanning tree of the complete graph is known. Note that a lower bound for cut approximation implies a lower-bound for spectral approximation, because the contrapositive statement is true: spectral approximation implies cut approximation.

Remark I.9. *In fact, all our lower bounds also directly apply for cut approximation, which is a strictly stronger result. For example, there is an unweighted graph G, s.t. if we sample an inverse leverage score weighted random spanning tree T, then with probability at least $1 - e^{-n^{.99}}$, T has a cut which is larger than the corresponding cut in G by a factor $\Omega(\log n / \log \log n)$.*

Connection to Spectrally Thin Trees: Using their MSS-type existence proof for "small norm outcomes" of homogeneous Strongly Rayleigh distributions, [AG15] showed that in an unweighted k-edge connected graph G where every edge has leverage score at most ϵ, there exists an unweighted spanning tree $\widehat{T}$ s.t. $L_{\widehat{T}} \preceq O(\frac{1}{k} + \epsilon) \cdot L_G$. This is referred to a spectrally thin tree with parameter $O(\frac{1}{k} + \epsilon)$.

In contrast, applying our k-homogeneous Strongly Rayleigh Matrix Chernoff bound to an unweighted graph G where every edge has leverage score at most ϵ, we can show that an unweighted *random* spanning tree satisfies $L_T \preceq O(\epsilon \log n) L_G$ with high probability. This follows immediately from our Theorem I.5, because if we let T denote the unweighted spanning tree and $\widehat{T}$ corresponding spanning tree with edges weighted by inverse leverage scores, then

$$
\begin{aligned}
L_{\widehat{T}} &= \sum_{e \in T} w(e) b_e b_e^\intercal \\
&\preceq \epsilon \sum_{e \in T} \frac{1}{l(e)} w(e) b_e b_e^\intercal \\
&= \epsilon L_{\widehat{T}} \\
&\preceq O(\epsilon \log n) L_G
\end{aligned}
$$

whp.

The proof in [AG15] is based on an adaptation of the [MSS13] proof, and does not have clear parallels with our approach. Whereas the key properties of Strongly Rayleigh distributions that we use are stochastic covering (a property that limits change in a distribution under conditioning) and conditional negative dependence, the central element of their approach is a proof that certain mixed characteristic polynomials associated with k-homogeneous distributions are real stable when the original distribution is Strongly Rayleigh.

The following Lemma captures a simple but crucial property of Strongly Rayleigh distributions.

Lemma I.10 (Shrinking Marginals)**.** *Suppose $(\xi_1, \ldots, \xi_m) \in \{0,1\}^m$ is a random vector of $\{0,1\}$ variables whose distribution is k-homogeneous and Strongly Rayleigh, then any set $S \subseteq [m]$ with $|S| \leq k$ for all $j \in [m] \setminus S$*

$$
\Pr[\xi_j = 1 | \xi_S = \mathbf{1}_S] \leq \Pr[\xi_j = 1].
$$

We provide the proof of this lemma in the full version. In the remaining sections, we prove Theorems I.1, I.3 and I.5.

We defer the proofs of our other results to the full version.

II. NOTATION

We use $[n]$ to denote set $\{1, 2, \cdots, n\}$. Given a vector x, we use $\|x\|_0$ to denote the number of non-zero entries in the vector.

Matrix and Norms.: For matrix A, we use $A^\intercal$ to denote the transpose of A. We say matrix A is positive-semidefinite (psd) if $A = A^\intercal$ and $x^\intercal A x \geq 0$ for all $x \in \mathbb{R}$. We use $\succeq, \preceq$ to denote the semidefinite ordering, e.g. $A \succeq 0$ means that A is psd.

For matrix $A \in \mathbb{R}^{n \times n}$, we define $\|A\|$ to be the spectral norm of A, i.e.,

$$
\|A\| = \max_{\|x\|_2 = 1, x \in \mathbb{R}^n} x^\intercal A x.
$$

Let $\mathrm{tr}(A)$ denote the trace of a square matrix A. We use $\lambda_{\max}(A)$ to denote the largest eigenvalue of matrix A. For symmetric matrix $A \in \mathbb{R}^{n \times n}$, $\lambda_{\max}(A) = \|A\|$ and $\mathrm{tr}(A) \leq n\|A\|$.

The Laplacian matrix related definitions.: Let $G = (V, E, w)$ be a connected weighted undirected graph with n vertices and m edges and edge weights $w_e > 0$. If we orient the edges of G arbitrarily, we can write its Laplacian as $L = B^\intercal W B$, where $B \in \mathbb{R}^{m \times n}$ is the signed edge-vertex incidence matrix and defined as follows

$$
B(e, v) = \begin{cases} 1, & \text{if } v \text{ is } e\text{'s head;} \\ -1, & \text{if } v \text{ is } e\text{'s tail;} \\ 0, & \text{otherwise.} \end{cases}
$$

and $W \in \mathbb{R}^{m \times m}$ is the diagonal matrix with $W(e, e) = w_e$.

III. PRELIMINARIES

A. Useful facts and tools

This section, we provide some useful tools. For completeness, we prove the following statement in the full version.

Fact III.1. *For any two symmetric matrices*

$$
(A - B)^2 \preceq 2A^2 + 2B^2.
$$

B. Strongly Rayleigh distributions

This section provides the definition of Strongly Rayleigh distributions. For more details, we refer the readers to [BBL09], [PP14].

Let $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ denote a probability distribution over $2^{[n]}$, and $\sum_{S \subseteq [n]} \mu(S) = 1$.

Let $x_1, x_2, \cdots, x_n$ denote n variables, we use x to denote $(x_1, x_2, \cdots, x_n)$. For each set $S \subseteq [n]$, we define $x_S = \prod_{i \in S} x_i$. We define the **generating polynomial** for μ as follows

$$
f_\mu(x) = \sum_{S \subseteq [n]} \mu(S) \cdot x_S.
$$

We say distribution μ is k-homogeneous if the polynomial f_μ is a homogeneous polynomial of degree k. In other words, for each $S \in \text{supp}(\mu)$, $|S| = k$.

We say a polynomial $p(x_1, x_2, \cdots, x_n)$ is stable, if $\text{Im}(x_i) > 0, \forall i \in [n]$, then $p(x_1, \cdots, x_n) \neq 0$. We say polynomial p is real stable, it is stable and all of its coefficients are real. We say μ is a **Strongly Rayleigh** distribution if f_μ is a real stable polynomial.

Fact III.2 (Conditioning on subset of coordinates). *Consider a random vector $(\xi_1, \ldots, \xi_m) \in \{0, 1\}^m$ whose distribution is k-homogeneous Strongly Rayleigh. Suppose we get a binary vector $\mathbf{b} = (b_1, \ldots, b_t) \in \{0, 1\}^t$ with $\|\mathbf{b}\|_0 = l \leq k$, and we get a set $S \subset [m]$ with $|S| = t$. Then conditional on $\xi_S = \mathbf{b}$, the distribution of $\xi_{[m] \setminus S}$ is $(k - l)$-homogeneous Strongly Rayleigh.*

This fact tells us that if we condition on the value of some entries in the vector, the remaining coordinates still have a Strongly Rayleigh distribution.

Fact III.3 (Stochastic Covering Property). *Consider a random vector $(\xi_1, \ldots, \xi_m) \in \{0, 1\}^m$ whose distribution is k-homogeneous Strongly Rayleigh. Suppose we are given an index $i \in [m]$. Let $\xi' = \xi_{[m] \setminus \{i\}}$ be the distribution on entries of ξ except i. Let ξ'' be the distribution of $\xi_{[m] \setminus \{i\}}$ conditional on $\xi_i = 1$. Then, there exists a coupling between ξ' and ξ'' (i.e. a joint distribution the two vectors), s.t. in every outcome of the coupling the value of ξ' can be obtain from the value of ξ'' by either changing a single from 0 to 1 or by leaving all entries unchanged.*

This fact is known as the *Stochastic Covering Property* (see [PP14]). It gives us a convenient tool for relating the conditional distribution of a subset of the coordinates of the vector to the unconditional distribution.

Note that by Fact III.2, the distribution of ξ'' used in Fact III.3 is $k - 1$ homogeneous. In contrast, the outcomes of ξ' may have k or $k-1$ ones. Fact III.3 tells us that we can pair up all the outcomes of the conditional distribution ξ'' with outcome of the unconditional distribution ξ' s.t. only a small change is required to make them equal. This tells us that the distribution is in some sense not changing too quickly under conditioning.

C. Random spanning trees

We provide the formal definition of random spanning tree in this section.

We use the same definitions about spanning trees as [DKP+17]. Let $\mathcal{T}_G$ denote the set of all spanning subtrees of G. We now define a probability distribution on these trees.

Definition III.4 (w-uniform distribution on trees). *Let $\mathcal{D}_G$ be a probability distribution on $\mathcal{T}_G$ such that*

$$\Pr_{X \sim \mathcal{D}_G}[X = T] \propto \prod_{e \in T} w_e.$$

We refer to $\mathcal{D}_G$ as the w-uniform distribution on $\mathcal{T}_G$. When the graph G is unweighted, this corresponds to the uniform distribution on $\mathcal{T}_G$. Crucially, random spanning tree distributions are Strongly Rayleigh, as shown in [BBL09].

Fact III.5 (Spanning Trees are Strongly Rayleigh). *In a connected weighted graph G, the w-uniform distribution on spanning trees is $(n-1)$-homogeneous Strongly Rayleigh.*

Definition III.6 (Effective Resistance). *The effective resistance of a pair of vertices $u, v \in V_G$ is defined as*

$$R_{\text{eff}}(u, v) = b_{u,v}^\top L^\dagger b_{u,v},$$

where $b_{u,v}$ is an all zero vector corresponding to V_G, except for entries of 1 at u and v.

The a reference for following standard fact about random spanning trees can be found in [DKP+17].

Definition III.7 (Leverage Score). *The statistical leverage score, which we will abbreviate to leverage score, of an edge $e = (u, v) \in E_G$ is defined as*

$$l_e = w_e R_{\text{eff}}(u, v).$$

Fact III.8 (Spanning Tree Marginals). *The probability $\Pr[e]$ that an edge $e \in E_G$ appears in a tree sampled w-uniformly randomly from $\mathcal{T}_G$ is given by*

$$\Pr[e] = l_e,$$

where l_e is the leverage score of the edge e.

IV. A MATRIX CHERNOFF BOUND FOR STRONGLY RAYLEIGH DISTRIBUTIONS

We first define a mapping which maps an element into a psd matrix.

Definition IV.1 (Y-operator). *We use Γ to denote $[m]$, we define a mapping $Y : \Gamma \to \mathbb{R}^{n \times n}$ such that Y_e is a psd matrix and $\|Y_e\| \leq R$.*

Throughout this section, we will use $\xi \in \{0, 1\}^m$ to denote a random length m boolean vector whose distribution is k-homogeneous Strongly Rayleigh. For any set $S \subseteq [m]$, we use ξ_S to denote the length $|S|$ vector that only chooses the entry from indices in S.

We will frequently need to work with a different representation of the random variable ξ. We use γ to denote this second representation. The random variable γ is composed of a sequence of k random indices $\gamma_1, \gamma_2, \cdots, \gamma_k$, each of which takes a value $e_1, e_2, \cdots, e_k \in [m]$. The indices give the locations of the ones in ξ, i.e. in an outcome of the two variables (ξ, γ), we always have $\xi_{\{\gamma_1, \gamma_2, \cdots, \gamma_k\}} = \mathbf{1}$ and $\xi_{[m] \setminus \{\gamma_1, \gamma_2, \cdots, \gamma_k\}} = \mathbf{0}$. Additionally, we want to ensure that the distribution of γ is invariant under permutation: This can clearly be achieved by starting with any distribution for γ that satisfies the coupling with ξ and the applying

a uniformly random permutation to reorder the k indices of γ (see [PP14] for a further discussion).

For convenience, for each $i \in [k]$, we define $\gamma_{\leq i}$ and $\gamma_{\geq i}$ as abbreviated notation for

$$\gamma_1, \gamma_2, \cdots, \gamma_i \text{ and } \gamma_i, \gamma_{i+1}, \cdots, \gamma_k$$

respectively. Let $S = \{e_1, e_2, \cdots, e_i\} \subset [m]$ be one possible assignment for indices of a subset of the ones in ξ, (we require $i \leq k$). Then the distribution of $\xi_{[m] \setminus S}$ conditional on $\xi_S = \mathbf{1}$ is the same as the the distribution of $\xi_{[m] \setminus S}$ conditional on $(\gamma_1, \gamma_2, \cdots, \gamma_i) = (e_1, e_2, \cdots, e_i)$. In other words, in terms of the resulting distribution of $\xi_{[m] \setminus S}$, it is equivalent to condition on either $\gamma_{\leq i}$ or $\xi_{\gamma_{\leq i}} = \mathbf{1}$. We define matrix $Z \in \mathbb{R}^{n \times n}$ as follows.

Definition IV.2 (Z)**.** *Let Z denote $\sum_{e \in \Gamma} \xi_e \cdot Y_e$ where $\|\xi\|_0 = k$. Due to the relationship between ξ and γ, we can also write Z as*

$$Z = \sum_{i=1}^{k} Y_{\gamma_i}.$$

For simplicity, for each $i \in [k]$, we define $Z_{\leq i}$ and $Z_{\geq i}$ as follows,

$$Z_{\leq i} = \sum_{j=1}^{i} Y_{\gamma_j} \text{ and } Z_{\geq i} = \sum_{j=i}^{k} Y_{\gamma_j}.$$

We define a series of matrices $M_i \in \mathbb{R}^{n \times n}$ as follows

Definition IV.3 (M_i, martingale)**.** *We define $M_0 = \mathbb{E}[Z]$. For each $i \in \{1, 2, \cdots, k-1\}$, we define M_i as follows*

$$M_i = \mathop{\mathbb{E}}_{\gamma_{\geq i+1}} [Z \mid \gamma_1, \cdots, \gamma_i].$$

It is easy to see that

$$\mathop{\mathbb{E}}_{\gamma_{i+1}} [M_{i+1}] = M_i,$$

which implies

$$\mathop{\mathbb{E}}_{\gamma_{i+1}} [M_{i+1} - M_i \mid \gamma_1, \cdots, \gamma_i] = 0.$$

We can rewrite the Z in M_{i+1} in the following sense,

$$Z = \sum_{j=i+2}^{k} Y_{\gamma_j} + Y_{\gamma_{i+1}} + \sum_{j=1}^{i} Y_{\gamma_j}$$
$$= Z_{\geq i+2} + Y_{\gamma_{i+1}} + Z_{\leq i}. \tag{1}$$

For the Z in M_i, we can write it as

$$Z = \sum_{j=i+1}^{k} Y_{\widehat{\gamma}_j} + \sum_{j=1}^{i} Y_{\gamma_j}$$
$$= \widehat{Z}_{\geq i+1} + Z_{\leq i}. \tag{2}$$

Note that k-homogeneous Strongly Rayleigh implies the stochastic covering property. By [BBL09], [PP14], stochastic

covering property implies that coupling is exists between $Z_{\geq i+2}$ and $Y_{\widehat{\gamma}_{i+1}}$, thus

$$Z_{\geq i+2} + Y_{\widehat{\gamma}_{i+1}} = \widehat{Z}_{\geq i+1}. \tag{3}$$

We define $X_{i+1} = M_{i+1} - M_i$ and then

$$\mathop{\mathbb{E}}_{\gamma_{i+1}} [X_{i+1} \mid \gamma_1, \cdots, \gamma_i] = 0. \tag{4}$$

Then we can rewrite X_{i+1} in the following way,

Claim IV.4. *Let $\mathcal{D}_{\gamma_{\leq i+1}}$ denote the coupling distribution between $Z_{\geq i+2}$ and $Y_{\widehat{\gamma}_{i+1}}$ such that Eq.* (3) *holds. Then*

$$X_{i+1} = Y_{\gamma_{i+1}} - \mathop{\mathbb{E}}_{(Z_{\geq i+2}, Y_{\widehat{\gamma}_{i+1}}) \sim \mathcal{D}_{\gamma_{\leq i+1}}} [Y_{\widehat{\gamma}_{i+1}} \mid \gamma_{\leq i+1}].$$

We provide the proof of the above Claim in the full version.

Fact IV.5. *We conditioning on $\gamma_{\leq i+1}$ are already fixed. Let $\mathcal{D}_{\gamma_{\leq i+1}}$ denote the coupling distribution such that $Z_{\geq i+2} + Y_{\widehat{\gamma}_{i+1}} = \widehat{Z}_{\geq i+1}$ holds. We define $U_{\gamma_{i+1}}$ as follows*

$$U_{\gamma_{i+1}} = \mathop{\mathbb{E}}_{(Z_{\geq i+2}, Y_{\widehat{\gamma}_{i+1}}) \sim \mathcal{D}_{\gamma_{\leq i+1}}} [Y_{\widehat{\gamma}_{i+1}} \mid \gamma_{\leq i+1}].$$

Then, we have the following four properties,

$$\text{(I)} \quad \mathop{\mathbb{E}}_{\gamma_{i+1}} [U_{\gamma_{i+1}} \mid \gamma_{\leq i}] = \mathop{\mathbb{E}}_{\gamma_{i+1}} [Y_{\gamma_{i+1}} \mid \gamma_{\leq i}],$$
$$\text{(II)} \quad \|Y_{\gamma_{i+1}}\| \leq R, \|U_{\gamma_{i+1}}\| \leq R,$$
$$\text{(III)} \quad \|Y_{\gamma_{i+1}} - U_{\gamma_{i+1}}\| \leq R,$$
$$\text{(IV)} \quad Y_{\gamma_{i+1}}^2 \preceq R \cdot Y_{\gamma_{i+1}}, U_{\gamma_{i+1}}^2 \preceq R \cdot U_{\gamma_{i+1}}.$$

We provide the proof of the above Fact in the full version. We can show

Claim IV.6.

$$\mathop{\mathbb{E}}_{\gamma_{i+1}} \left[(Y_{\gamma_{i+1}} - U_{\gamma_{i+1}})^2 \mid \gamma_{\leq i} \right] \preceq 4R \cdot \mathop{\mathbb{E}}_{\gamma_{i+1}} [Y_{\gamma_{i+1}} \mid \gamma_{\leq i}]$$

Proof:

$$\mathop{\mathbb{E}}_{\gamma_{i+1}} \left[(Y_{\gamma_{i+1}} - U_{\gamma_{i+1}})^2 \mid \gamma_{\leq i} \right]$$
$$\preceq \mathop{\mathbb{E}}_{\gamma_{i+1}} \left[2Y_{\gamma_{i+1}}^2 + 2U_{\gamma_{i+1}}^2 \mid \gamma_{\leq i} \right]$$
$$\preceq \mathop{\mathbb{E}}_{\gamma_{i+1}} \left[2R \cdot Y_{\gamma_{i+1}} + 2R \cdot U_{\gamma_{i+1}} \mid \gamma_{\leq i} \right]$$
$$\preceq \mathop{\mathbb{E}}_{\gamma_{i+1}} \left[4R \cdot Y_{\gamma_{i+1}} \mid \gamma_{\leq i} \right],$$

where the first step follows by Fact III.1, and the second step follows by $U_{\gamma_{i+1}}^2 \preceq R \cdot U_{\gamma_{i+1}}$ and $Y_{\gamma_{i+1}}^2 \preceq R \cdot Y_{\gamma_{i+1}}$. ∎

Lemma IV.7. *Let $\mathbb{E}[\sum_{e \in \Gamma} \xi_e Y_e] \preceq \mu I$. For each $i \in \{1, 2, \cdots, k\}$, we have*

$$\mathop{\mathbb{E}}_{\gamma_i} [Y_{\gamma_i} \mid \gamma_{\leq i-1}] \preceq \frac{1}{k+1-i} \mu I.$$

Proof: We use $\mathbf{1}$ to a length $i - 1$ vector where each entry is one. We can think of $\gamma_{\leq i-1}$ are already chosen to

some $e \in \Gamma$, for example $\gamma_1 = e_1, \cdots, \gamma_{i-1} = e_{i-1}$. Note that, all of $e_1, \cdots, e_{i-1}$ are different. Then we use $\Gamma \backslash \gamma_{\leq i-1}$ to denote $\Gamma \backslash \{e_1, \cdots, e_{i-1}\}$.

$$
\begin{aligned}
\mathbb{E}\left[Y_{\gamma_i} \mid \gamma_{\leq i-1}\right] &= \sum_{e \in \Gamma \backslash \gamma_{\leq i-1}} \Pr[\gamma_i = e \mid \gamma_{\leq i-1}] \cdot Y_e \\
&= \sum_{e \in \Gamma \backslash \gamma_{\leq i-1}} \frac{\Pr[\xi_e = 1 \mid \gamma_{\leq i-1}]}{k - (i-1)} \cdot Y_e \\
&= \sum_{e \in \Gamma \backslash \gamma_{\leq i-1}} \frac{\Pr[\xi_e = 1 \mid \xi_{\gamma_{\leq i-1}} = \mathbf{1}]}{k - (i-1)} \cdot Y_e \\
&\preceq \sum_{e \in \Gamma \backslash \gamma_{\leq i-1}} \frac{\Pr[\xi_e = 1]}{k - (i-1)} \cdot Y_e \\
&\preceq \sum_{e \in \Gamma} \frac{\Pr[\xi_e = 1]}{k - (i-1)} \cdot Y_e \\
&= \frac{1}{k - (i-1)} \mathbb{E}\left[\sum_e \xi_e Y_e\right] \\
&\preceq \frac{1}{k + 1 - i} \mu I
\end{aligned}
$$

where the first step follows by definition of expectation, the second step follows by $\Pr[\gamma_i = e \mid \gamma_{\leq i-1}] = \Pr[\xi_e = 1 \mid \gamma_{\leq i-1}]/(k - (i-1))$, the third step follows $[\cdot|\gamma_{\leq i-1}]$ is equivalent to $[\cdot|\xi_{\gamma_{\leq i-1}} = \mathbf{1}]$, the fourth step follows by $(\Pr[\xi_e = 1 \mid \xi_{\gamma_{\geq i-1}} = \mathbf{1}] \leq \Pr[\xi_e = 1])$ from Shrinking Marginals Lemma I.10, the fifth step follows by relaxing $\Gamma \backslash \gamma_{\leq i-1}$, the sixth step follows by $\Pr[\xi_e = 1] = \mathbb{E}[\xi_e]$ and linearity of expectation, and the last step follows by $\mathbb{E}[\sum_{e \in \Gamma} \xi_e Y_e] \preceq \mu I$. ∎

Lemma IV.8. *For each $i \in \{1, 2, \cdots, k\}$*

$$
\mathbb{E}_{\gamma_i}\left[X_i^2 \mid \gamma_{\leq i-1}\right] \preceq 4\mu R \frac{1}{k + 1 - i} I.
$$

Proof: It follows by combining Claim IV.6 and Lemma IV.7 directly. ∎

The above lemma implies this corollary directly

Corollary IV.9.

$$
\sum_{i=1}^{k} \mathbb{E}_{\gamma_i}\left[X_i^2 \mid \gamma_{\leq i-1}\right] \preceq 10\mu R \log k \cdot I.
$$

A. Main result

Before finally proving our main theorem I.1, we state a useful tool: Freedman's inequality for matrices

We state a version from [Tro11b], and there is also another version can be found in [Oli09].

Theorem IV.10 (Matrix Freedman). *Consider a matrix martingale $\{Y_i : i = 0, 1, 2, \cdots\}$ whose values are self-adjoint matrices with dimension n, and let $\{X_i : i = 1, 2, 3, \cdots\}$ be*

the difference sequence. Assume that the difference sequence is uniformly bounded in the sense that

$$
\lambda_{\max}(X_i) \leq R, \text{ almost surely } \text{ for } i = 1, 2, 3, \cdots.
$$

Define the predictable quadratic variation process of the martingale :

$$
W_i = \sum_{j=1}^{i} \mathbb{E}_{j-1}[X_j^2], \text{ for } i = 1, 2, 3, \cdots.
$$

Then, for all $t \geq 0$ and $\sigma^2 > 0$,

$$
\Pr\left[\exists i \geq 0 : \lambda_{\max}(Y_i) \geq t \text{ and } \|W_i\| \leq \sigma^2\right]
$$
$$
\leq n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).
$$

Now, we are ready to prove our main theorem,

Theorem I.1. (First main result, a Matrix Chernoff Bound k-homogeneous Strongly Rayleigh Distributions)**.** *Suppose $(\xi_1, \ldots, \xi_m) \in \{0, 1\}^m$ is a random vector of $\{0, 1\}$ variables whose distribution is k-homogeneous and Strongly Rayleigh.*

Given a collection of PSD matrices $A_1, \ldots A_m \in \mathbb{R}^{n \times n}$ s.t. for all $e \in [m]$ we have $\|A_e\| \leq R$ and $\|\mathbb{E}[\sum_e \xi_e A_e]\| \leq \mu$.

Then for any $\epsilon > 0$,

$$
\Pr\left[\left\|\sum_e \xi_e A_e - \mathbb{E}\left[\sum_e \xi_e A_e\right]\right\| \geq \epsilon\mu\right]
$$
$$
\leq n \exp\left(-\frac{\epsilon^2 \mu}{R(\log k + \epsilon)} \Theta(1)\right)
$$

Proof: We use Y to denote A and Γ to denote $[m]$.

In order to use Theorem IV.10, we first we define W_i as follows

$$
W_i = \sum_{j=1}^{i} \mathbb{E}_{\gamma_i}\left[X_i^2 \mid \gamma_{\leq i-1}\right].
$$

According to definition of M_i, $\{M_0, M_1, M_2 \cdots\}$ is a matrix martingale and $M_k - M_0 = \sum_e \xi_e A_e - \mathbb{E}[\sum_e \xi_e A_e]$.

We have proved the following facts,

The first one is, $\mathbb{E}_{\gamma_i}[X_i|\gamma_{\leq i-1}] = 0$. It follows by Eq. (4)

The second one is

$$
\lambda_{\max}(X_i) \leq R
$$

It follows by combining Property (III) of Fact IV.5 and Claim IV.4.

The third one is

$$
\|W_i\| \leq \sigma^2, \forall i \in [k]
$$

where $\sigma^2 = 10\mu R \log k$. It follows by Corollary IV.9.

Thus,

$$
\Pr\left[\lambda_{\max}(M_k - M_0) \geq \epsilon\mu\right] \leq n \exp\left(-\frac{(\epsilon\mu)^2/2}{\sigma^2 + R(\epsilon\mu)/3}\right).
$$

We have

$$\frac{t^2/2}{\sigma^2 + Rt/3} = \frac{\epsilon^2\mu^2/2}{10\mu R \log k + R\epsilon\mu/3} \quad \text{by choosing } t = \epsilon\mu$$

$$= \frac{3\epsilon^2\mu}{(60\log k + 2\epsilon)R}.$$

Thus we prove one side of the bound. Since $\mathbb{E}_{\gamma_i}[-X_i|\gamma_{\leq i-1}] = 0$ and $\mathbb{E}_{\gamma_i}[(-X_i)^2|\gamma_{\leq i-1}] = \mathbb{E}_{\gamma_i}[X_i^2|\gamma_{\leq i-1}]$, then following the similar procedure as proving $\lambda_{\max}$, we have bound for $\lambda_{\min}$

$$\Pr[\lambda_{\min}(M_k - M_0) \leq -\epsilon\mu] \leq n\exp\left(-\frac{3\epsilon^2\mu}{(60\log k + 2\epsilon)R}\right).$$

Putting two sides of the bound together, we complete the proof. ∎

V. APPLICATIONS TO RANDOM SPANNING TREES

In this section, we show how to use Theorem I.1 to prove the bound for one random spanning and also summation of random spanning trees.

Theorem I.5. (Third main result, upper bound for the concentration of one random spanning tree). *Given a graph G, let T be a random spanning tree, then with probability at least $1 - 1/\operatorname{poly}(n)$*

$$L_T \preceq O(\log n) \cdot L_G.$$

Proof:

Let $G = (V, E, w)$ be a undirected weighted graph, $w : E \to R$, which is connected. The Laplacian of G is $L_G = \sum_{e \in E} w(e)b_e b_e^{\mathsf{T}}$.

Let $T \subseteq E$ be a random spanning tree of G in the sense of Definition III.4. Let the weights of the edges in T be given by $w' : T \to R$ where $w'(e) = w(e)/l_e$, where l_e is the leverage score of e in G. Thus the Laplacian of the tree is $L_T = \sum_{e \in T} w'(e)b_e b_e^{\mathsf{T}} = \sum_{e \in T} \frac{w(e)}{l_e} b_e b_e^{\mathsf{T}}$. Then by Fact III.8, $Pr[e \in T] = l_e$, and hence $\mathbb{E}[L_T] = L_G$.

Note also that for all $e \in E$, $||(L_G^\dagger)^{1/2}w(e)b_e b_e^{\mathsf{T}}(L_G^\dagger)^{1/2}|| = l_e$. Consider the random matrix $(L_G^\dagger)^{1/2}L_T(L_G^\dagger)^{1/2}$. The distribution of edge in the spanning tree can be seen as an $n-1$ homogeneous vector in $\{0,1\}^m$ where $m = |E|$. To apply Theorem I.1, let ξ_e be the eth entry of this random vector, and

$$A_e = (L_G^\dagger)^{1/2}w'(e)b_e b_e^{\mathsf{T}}(L_G^\dagger)^{1/2}$$

Note $A_e \succeq 0$. Now $||A_e|| = 1$ and $\mathbb{E}\left[\sum_e \xi_e A_e\right] = \mathbb{E}[(L_G^\dagger)^{1/2}L_T(L_G^\dagger)^{1/2}] = (L_G^\dagger)^{1/2}L_G(L_G^\dagger)^{1/2} = \Pi = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$, where we used in the last equality that the null space of the Laplacian of a connected graph is the span of the all ones vector. Thus, as each we get $||\mathbb{E}\left[\sum_e \xi_e A_e\right]|| = 1$ This means we can apply Theorem I.1 with $R = 1$, $\mu = 1$ and $\epsilon = 100\log n$ to whp. $||(L_G^\dagger)^{1/2}L_T(L_G^\dagger)^{1/2} - \Pi|| \leq 100\log n$.

As L_T is a Laplacian, it has $\mathbf{1}$ in the null space, so can conclude that $(L_G^\dagger)^{1/2}L_T(L_G^\dagger)^{1/2} \preceq 100\log n\Pi$. Hence $L_T \preceq \log n L_G$. ∎

Theorem I.3. (Second main result, concentration bound of a batch of independent random spanning trees). *Given as input a weighted graph G with n vertices and a parameter $\epsilon > 0$, let $T_1, T_2, \cdots, T_t$ denote t independent inverse leverage score weighted random spanning trees, if we choose $t = O(\epsilon^{-2}\log^2 n)$ then with probability $1 - 1/\operatorname{poly}(n)$,*

$$(1-\epsilon)L_G \preceq \frac{1}{t}\sum_{i=1}^t L_{T_i} \preceq (1+\epsilon)L_G.$$

Proof: The proof is similar to the proof of Theorem I.5. Now we viewing edges of $t = O(\epsilon^{-2}\log^2 n)$ independent random spanning trees as a $t(n-1)$-homogeneous Strongly Rayleigh Distribution a vector in $\{0,1\}^{t|E|}$. Note that the product of independent Strongly Rayleigh distributions is Strongly Rayleigh [BBL09]. Again we get $||\mathbb{E}\left[\sum_e \xi_e A_e\right]|| = 1$, but now we can take $R = \frac{1}{t}$, and hence we obtain the desired result by plugging into Theorem I.1. ∎

REFERENCES

[AG09]   Kook Jin Ahn and Sudipto Guha. Graph sparsification in the semi-streaming model. In *International Colloquium on Automata, Languages, and Programming*, pages 328–338. Springer, 2009.

[AG15]   Nima Anari and Shayan Oveis Gharan. Effective-resistance-reducing flows, spectrally thin trees, and asymmetric tsp. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 20–39. IEEE, 2015.

[AGM⁺10] Arash Asadpour, Michel X. Goemans, Aleksander Mądry, Shayan Oveis Gharan, and Amin Saberi. An o(log n/ log log n)-approximation algorithm for the asymmetric traveling salesman problem. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 379–389, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[AGR16] Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte carlo markov chain algorithms for sampling strongly rayleigh distributions and determinantal point processes. In *Conference on Learning Theory*, pages 103–115, 2016.

[Ald90] David Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. In *SIAM Journal on Discrete Mathematics*, pages 450–465, 1990.

[AW02] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[BBL09] Julius Borcea, Petter Brändén, and Thomas Liggett. Negative dependence and the geometry of polynomials. *Journal of the American Mathematical Society*, 22(2):521–567, 2009.

[Ber24] Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[BK96] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in Õ(n2) time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 47–55, New York, NY, USA, 1996. ACM.

[Bro89] Andrei Broder. Generating random spanning trees. In *Proceedings of the 30th annual Symposium on Foundations of Computer Science, FOCS 1989*, pages 442–447, 1989.

[BSS12] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

[BSST13] Joshua Batson, Daniel A Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: theory and algorithms. *Communications of the ACM*, 56(8):87–94, 2013.

[Che52] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[CKP⁺17] Michael B Cohen, Jonathan Kelner, John Peebles, Richard Peng, Anup B Rao, Aaron Sidford, and Adrian Vladu. Almost-linear-time algorithms for markov chains and new spectral primitives for directed graphs. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 410–419. ACM, 2017.

[CMN96] Charles J Colbourn, Wendy J Myrvold, and Eugene Neufeld. Two algorithms for unranking arborescences. *Journal of Algorithms*, 20(2):268–281, 1996.

[CMP16] Michael B Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. *arXiv preprint arXiv:1604.05448*, 2016.

[Coh16] Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 278–287. SIAM, 2016.

[DKP⁺17] David Durfee, Rasmus Kyng, John Peebles, Anup B. Rao, and Sushant Sachdeva. Sampling random spanning trees faster than matrix multiplication. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pages 730–742, New York, NY, USA, 2017. ACM.

[DPPR17] David Durfee, John Peebles, Richard Peng, and Anup B. Rao. Determinant-preserving sparsification of SDDM matrices with applications to counting and sampling spanning trees. In *Proceedings of the 2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, FOCS '17, 2017.

[FHHP11] Wai Shing Fung, Ramesh Hariharan, Nicholas JA Harvey, and Debmalya Panigrahi. A general framework for graph sparsification. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 71–80. ACM, 2011.

[FKG71] Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[FM92] Tomás Feder and Milena Mihail. Balanced matroids. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 26–38. ACM, 1992.

[GLSS18] Ankit Garg, Yin-Tat Lee, Zhao Song, and Nikhil Srivastava. A matrix expander chernoff bound. In *STOC 2018*. arXiv preprint arXiv:1704.03864, 2018.

[GRV09] Navin Goyal, Luis Rademacher, and Santosh Vempala. Expanders via random spanning trees. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 576–585. Society for Industrial and Applied Mathematics, 2009.

[GSS11] Shayan Oveis Gharan, Amin Saberi, and Mohit Singh. A randomized rounding approach to the traveling salesman problem. In *Proceedings of the 2011 IEEE 52Nd Annual Symposium on Foundations of Computer Science*, FOCS '11, pages 550–559, Washington, DC, USA, 2011. IEEE Computer Society.

[Gue83] Alain Guenoche. Random spanning tree. *Journal of Algorithms*, 4(3):214–220, 1983.

[Hea08] Alexander D Healy. Randomness-efficient sampling within nc. *Computational Complexity*, 17(1):3–37, 2008.

[HO14]   Nicholas JA Harvey and Neil Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 926–945. Society for Industrial and Applied Mathematics, 2014.

[HSSS16]   Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191. ACM, 2016.

[HX16]   Nicholas J. A. Harvey and Keyulu Xu. Generating random spanning trees via fast matrix multiplication. In *LATIN 2016: Theoretical Informatics*, volume 9644, pages 522–535, 2016.

[Kar93]   David R Karger. Global min-cuts in rnc, and other ramifications of a simple min-cut algorithm. In *SODA*, volume 93, pages 21–30, 1993.

[Kir47]   Gustav Kirchhoff. Uber die auflosung der gliechungen, auf welche man bei der undersuchung der linearen vertheilung galvanischer strome gefuhrt wird. *Poggendorgs Ann. Phys. Chem.*, pages 497–508, 1847.

[KM09]   Jonathan Kelner and Aleksander Madry. Faster generation of random spanning trees. In *Proceedings of the 50th annual Symposium on Foundations of Computer Science, FOCS 2009*, pages 13–21, 2009. Available at https://arxiv.org/abs/0908.1448.

[KPPS17]   Rasmus Kyng, Jakub Pachocki, Richard Peng, and Sushant Sachdeva. A framework for analyzing resparsification algorithms. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2032–2043. SIAM, 2017.

[KS96]   David R Karger and Clifford Stein. A new approach to the minimum cut problem. *Journal of the ACM (JACM)*, 43(4):601–640, 1996.

[KS16]   Rasmus Kyng and Sushant Sachdeva. Approximate gaussian elimination for laplacians-fast, sparse, and simple. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 573–582. IEEE, 2016.

[Kul90]   Vidyadhar G. Kulkarni. Generating random combinatorial objects. *Journal of Algorithms*, 11(2):185–207, 1990.

[MSS13]   Adam Marcus, Daniel A Spielman, and Nikhil Srivastava. Interlacing families ii: Mixed characteristic polynomials and the kadison-singer problem. *arXiv preprint arXiv:1306.3969*, 2013.

[MST15]   Aleksander Madry, Damian Straszak, and Jakub Tarnawski. Fast generation of random spanning trees and the effective resistance metric. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 2019–2036, 2015. Available at http://arxiv.org/pdf/1501.00267v1.pdf.

[Oli09]   Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[PP14]   Robin Pemantle and Yuval Peres. Concentration of lipschitz functionals of determinantal and other strong rayleigh measures. *Combinatorics, Probability and Computing*, 23(1):140–160, 2014.

[RS67]   A Rényi and G Szekeres. On the height of trees. *J. Austral. Math. Soc*, 7(4):497–5, 1967.

[Rud99]   Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

[Sch18]   Aaron Schild. An almost-linear time algorithm for uniform random spanning tree generation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, 2018.

[SS11]   Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

[ST04]   Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 81–90, New York, NY, USA, 2004. ACM.

[Tij]   H Tijms. Understanding probability: Chance rules in everyday life. 2004.

[Tro11a]   Joel A Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

[Tro11b]   Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA, 2011.

[Tro12]   Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[Wil96]   David Bruce Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. ACM.

[ZSD17]   Kai Zhong, Zhao Song, and Inderjit S Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017.

[ZSJ+17]   Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *ICML*. https://arxiv.org/pdf/1706.03175.pdf, 2017.