A Note On Spectral Clustering*†

Pavel Kolev[1] and Kurt Mehlhorn[2]

1 **Max-Planck-Institut für Informatik, Saarbrücken, Germany**
pkolev@mpi-inf.mpg.de
2 **Max-Planck-Institut für Informatik, Saarbrücken, Germany**
mehlhorn@mpi-inf.mpg.de

Abstract

Spectral clustering is a popular and successful approach for partitioning the nodes of a graph into clusters for which the ratio of outside connections compared to the volume (sum of degrees) is small. In order to partition into k clusters, one first computes an approximation of the bottom k eigenvectors of the (normalized) Laplacian of G, uses it to embed the vertices of G into k-dimensional Euclidean space $\mathbb{R}^k$, and then partitions the resulting points via a k-means clustering algorithm. It is an important task for theory to explain the success of spectral clustering.

Peng et al. (COLT, 2015) made an important step in this direction. They showed that spectral clustering provably works if the gap between the $(k+1)$-th and the k-th eigenvalue of the normalized Laplacian is sufficiently large. They proved a structural and an algorithmic result. The algorithmic result needs a considerably stronger gap assumption and does not analyze the standard spectral clustering paradigm; it replaces spectral embedding by heat kernel embedding and k-means clustering by locality sensitive hashing.

We extend their work in two directions. Structurally, we improve the quality guarantee for spectral clustering by a factor of k and simultaneously weaken the gap assumption. Algorithmically, we show that the standard paradigm for spectral clustering works. Moreover, it even works with the same gap assumption as required for the structural result.

1998 ACM Subject Classification G.2.2 [Discrete Mathematics] Graph Theory, Graph Algorithms, G.3 [Probability and Statistics] Probabilistic Algorithms (including Monte Carlo)

Keywords and phrases spectral embedding, k-means clustering, power method, gap assumption

Digital Object Identifier 10.4230/LIPIcs.ESA.2016.57

1 Introduction

A *cluster* in an undirected graph $G = (V, E)$ is a set S of nodes whose volume is large compared to the number of outside connections. Formally, we define the *conductance* of S by $\phi(S) = \left|E(S, \overline{S})\right| / \mu(S)$, where $\mu(S) = \sum_{v \in S} \deg(v)$ is the *volume* of S. The k-way partitioning problem for graphs asks to partition the vertices of a graph such that the conductance of each block of the partition is small (formal definition below). This problem arises in many applications, e.g., image segmentation and exploratory data analysis. We refer to the survey [10] for additional information. A popular and very successful approach to clustering [4, 8, 10] is *spectral clustering*. One first computes an approximation of the bottom k eigenvectors of the (normalized) Laplacian of G, uses it to embed the vertices of G into k-dimensional Euclidean space $\mathbb{R}^k$, and then partitions the resulting points via a

* The full version of the paper is available at http://arxiv.org/abs/1509.09188.

† This work has been funded by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

k-means clustering algorithm. It is an important task for theory to explain the success of spectral clustering. Recently, Peng et al. [7] made an important step in this direction. They showed that spectral clustering provably works if the $(k+1)$-th and the k-th eigenvalue of the normalized Laplacian differ sufficiently. In order to explain their result, we need some notation.

Let $\mathcal{L}_G = I - D^{-1/2}AD^{-1/2}$ be the normalized Laplacian matrix of G, where D is the diagonal degree matrix and A is the adjacency matrix, and let $f_j \in \mathbb{R}^V$ be the eigenvector corresponding to the j-th smallest eigenvalue λ_j of $\mathcal{L}_G$. The *spectral embedding map* $F : V \to \mathbb{R}^k$ is defined by

$$F(u) = \frac{1}{\sqrt{d_u}} \left(f_1(u), \ldots, f_k(u)\right)^{\mathrm{T}}, \quad \text{for all vertices } u \in V. \tag{1}$$

Peng et al. [7] construct a k-means instance $\mathcal{X}_V$ by inserting d_u many copies of the vector $F(u)$ into $\mathcal{X}_V$, for every vertex $u \in V$.

Let $\mathcal{X}$ be a set of vectors of the same dimension. Then

$$\triangle_k(\mathcal{X}) \triangleq \min_{\text{partition } (X_1,\ldots,X_k) \text{ of } \mathcal{X}} \sum_{i=1}^{k} \sum_{x \in X_i} \|x - c_i\|^2, \quad \text{where} \quad c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x,$$

is the optimal cost of clustering $\mathcal{X}$ into k sets. An α-approximate clustering algorithm returns a k-way partition $(A_1, \ldots, A_k)$ and centers $c_1, \ldots, c_k$ such that

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) \triangleq \sum_{i=1}^{k} \sum_{x \in A_i} \|x - c_i\|^2 \leqslant \alpha \cdot \triangle_k(\mathcal{X}). \tag{2}$$

The *order k conductance constant* $\rho(k)$ is a well studied worst case guarantee for k-way partitioning that is defined by

$$\rho(k) = \min_{\text{disjoint nonempty } Z_1,\ldots,Z_k} \Phi(Z_1, \ldots, Z_k), \quad \text{where} \quad \Phi(Z_1, \ldots, Z_k) = \max_{i \in [1:k]} \phi(Z_i). \tag{3}$$

Lee et al. [3] connected $\rho(k)$ and the k-th smallest eigenvalue of the normalized Laplacian matrix $\mathcal{L}_G$ through the relation, also known as higher order Cheeger inequality,

$$\lambda_k/2 \leqslant \rho(k) \leqslant O(k^2)\sqrt{\lambda_k}. \tag{4}$$

In this work, we focus attention on the *order k partition constant* $\widehat{\rho}(k)$ of G, defined by

$$\widehat{\rho}(k) \triangleq \min_{\text{partition } (P_1,\ldots,P_k) \text{ of } V} \Phi(P_1, \ldots P_k), \quad \text{where} \quad \Phi(Z_1, \ldots, Z_k) = \max_{i \in [1:k]} \phi(Z_i).$$

In a consecutive work, inspired by partitioning graphs into expanders, Oveis Gharan and Trevisan [6] proved the following relation

$$\rho(k) \leqslant \widehat{\rho}(k) \leqslant k\rho(k). \tag{5}$$

We are now ready to state the main structural result by Peng et al [7].

▶ **Theorem 1.1** ([7, Theorem 1.2]). *Let $k \geqslant 3$ and $(P_1, \ldots, P_k)$ be a k-way partition of V with $\Phi(P_1, \ldots, P_k) = \widehat{\rho}(k)$. Let G be a graph that satisfies the gap assumption*[1]

$$\delta = \frac{2 \cdot 10^5 \cdot k^3}{\Upsilon} \in (0, 1/2], \quad \textit{where} \quad \Upsilon \triangleq \frac{\lambda_{k+1}}{\widehat{\rho}(k)}. \tag{6}$$

[1] Note that $\lambda_k/2 \leqslant \widehat{\rho}(k)$, see (4,5). Thus the assumption implies $\lambda_k/2 \leqslant \widehat{\rho}(k) = \delta\lambda_{k+1}/(2 \cdot 10^5 \cdot k^3)$, i.e., there is a substantial gap between the k-th and the $(k+1)$-th eigenvalue.

Let $(A_1, \ldots, A_k)$ be the k-way partition[2] of V returned by an α-approximate k-means algorithm applied to $\mathcal{X}_V$. Then the following statements hold (after suitable renumbering of one of the partitions):
1. $\mu(A_i \triangle P_i) \leqslant \alpha\delta \cdot \mu(P_i)$, *and*
2. $\phi(A_i) \leqslant (1 + 2\alpha\delta) \cdot \phi(P_i) + 2\alpha\delta$,
where the symmetric difference $A_i \triangle P_i = (A_i \setminus P_i) \cup (P_i \setminus A_i)$.

Under the stronger gap assumption $\delta = 2 \cdot 10^5 \cdot k^5 / \Upsilon \in (0, 1/2]$, they showed how to obtain a partition in time $O(m \cdot \operatorname{poly}\log(n))$ with essentially the guarantees stated in Theorem 1.1, where $m = |E|$ is the number of edges in G and $n = |V|$ is the number of nodes.

However, their algorithmic result does not analyze the standard spectral clustering paradigm, since it replaces spectral embedding by heat kernel embedding and k-means clustering by locality sensitive hashing. Therefore, their algorithmic result does not explain the success of the standard spectral clustering paradigm.

Our Results

We strengthen the approximation guarantees in Theorem 1.1 by a factor of k and simultaneously weaken the gap assumption. As a consequence, the variant of Lloyd's k-means algorithm analyzed by Ostrovsky et al. [5] applied to[3] $\widetilde{\mathcal{X}_V}$ achieves the improved approximation guarantees in time $O(m(k^2 + \frac{\ln n}{\lambda_{k+1}}))$ with constant probability. Table 1 summarizes these results.

Let $\mathcal{O}$ be the set of all k-way partitions $(P_1, \ldots, P_k)$ with $\Phi(P_1, \ldots, P_k) = \widehat{\rho}(k)$, i.e., the set of all partitions that achieve the order k partition constant. Let

$$\widehat{\rho}_{\mathrm{avr}}(k) \triangleq \min_{(P_1,\ldots,P_k)\in\mathcal{O}} \frac{1}{k} \sum_{i=1}^{k} \phi(P_i)$$

be the *minimal average conductance* over all k-way partitions in $\mathcal{O}$. The minimal average conductance can be considerably smaller than the order k partition constant. Consider a graph consisting of one clique of size $S_k = f(n) = o(n/k^{3/2})$, $k-1$ cliques of size $(n - f(n))/(k-1)$ each, and k additional edges that connect the cliques in the form of a ring. Then $\phi(S_i) \approx k^2/n^2$ for $1 \leqslant i \leqslant k-1$ and $\phi(S_k) \approx 1/f(n)^2$. Thus $\widehat{\rho}(k) = \max_i \phi(S_i) = \phi(S_k) \approx 1/f(n)^2$ and $\widehat{\rho}_{\mathrm{avr}}(k) = (1/k) \sum_{1\leqslant i\leqslant k} \phi(S_k) \approx k^2/n^2 + (1/k) \cdot (1/f(n)^2) \approx \widehat{\rho}(k)/k$.

For the remainder of this paper we denote by $(P_1, \ldots, P_k)$ a k-way partition of V that achieves $\widehat{\rho}_{\mathrm{avr}}(k)$. In the full version of the paper, we give an analogous relation to (5) for $\widehat{\rho}_{\mathrm{avr}}(k)$. We state now our main result.

▶ **Theorem 1.2** (Main Theorem).
(a) (Existence of a Good Clustering) *Let $k \geqslant 3$. Let G be a graph satisfying the gap assumption*

$$\delta = \frac{20^4 \cdot k^3}{\Psi} \in (0, 1/2], \quad \text{where} \quad \Psi \triangleq \frac{\lambda_{k+1}}{\widehat{\rho}_{\mathrm{avr}}(k)}. \tag{7}$$

Let $(A_1, \ldots, A_k)$ be the k-way partition returned by an α-approximate clustering algorithm applied to the spectral embedding $\mathcal{X}_V$. Then for every $i \in [1 : k]$ the following two statements hold (after suitable renumbering of one of the partitions):

[2] The k-means algorithm returns a partition of $\mathcal{X}_V$. One may assume w.l.o.g. that all copies of $F(u)$ are put into the same cluster of $\mathcal{X}_V$. Thus the algorithm also partitions V.
[3] $\widetilde{\mathcal{X}_V}$ is defined as $\mathcal{X}_V$ but in terms of approximate eigenvectors, see Subsection 2.3.

Table 1 A comparison of the results in Peng et al. [7] and our results. The parameter $\delta \in (0, 1/2]$ relates the approximation guarantees with the gap assumption.

	Gap Assumption	Partition Quality	Running Time
Peng et al. [7]	$\delta = 2 \cdot 10^5 \cdot k^3/\Upsilon$	$\mu(A_i \triangle P_i) \leqslant \alpha\delta \cdot \mu(P_i)$ $\phi(A_i) \leqslant (1 + 2\alpha\delta)\,\phi(P_i) + 2\alpha\delta$	Existential result
This paper	$\delta = 20^4 \cdot k^3/\Psi$	$\mu(A_i \triangle P_i) \leqslant \frac{\alpha\delta}{10^3 k} \cdot \mu(P_i)$ $\phi(A_i) \leqslant \left(1 + \frac{2\alpha\delta}{10^3 k}\right)\phi(P_i) + \frac{2\alpha\delta}{10^3 k}$	Existential result
Peng et al. [7]	$\delta = 2 \cdot 10^5 \cdot k^5/\Upsilon$	$\mu(A_i \triangle P_i) \leqslant \frac{\delta \log^2 k}{k^2} \cdot \mu(P_i)$ $\phi(A_i) \leqslant \left(1 + \frac{2\delta \log^2 k}{k^2}\right)\phi(P_i) + \frac{2\delta \log^2 k}{k^2}$	$O\left(m \cdot \operatorname{poly}\log(n)\right)$
This paper	$\delta = 20^4 \cdot k^3/\Psi$ $\delta \leqslant k/10^9$ $\triangle_k(\mathcal{X}_V) \geqslant n^{-O(1)}$	$\mu(A_i \triangle P_i) \leqslant \frac{2\delta}{10^3 k} \cdot \mu(P_i)$ $\phi(A_i) \leqslant \left(1 + \frac{4\delta}{10^3 k}\right)\phi(P_i) + \frac{4\delta}{10^3 k}$	$O\left(m\left(k^2 + \frac{\ln n}{\lambda_{k+1}}\right)\right)$

1. $\mu(A_i \triangle P_i) \leqslant \frac{\alpha\delta}{10^3 k} \cdot \mu(P_i)$, *and*
2. $\phi(A_i) \leqslant \left(1 + \frac{2\alpha\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta}{10^3 k}$.

(b) (An Efficient Algorithm) *If in addition* $\delta \leqslant k/10^9$ *and*[4] $\triangle_k(\mathcal{X}_V) \geqslant n^{-O(1)}$, *then the variant of Lloyd's algorithm analyzed by Ostrovsky et al. [5] applied to* $\widetilde{\mathcal{X}_V}$ *returns in time* $O(m(k^2 + \frac{\ln n}{\lambda_{k+1}}))$ *with constant probability a partition* $(A_1, \ldots, A_k)$ *such that for every* $i \in [1 : k]$ *the following two statements hold (after suitable renumbering of one of the partitions):*

3. $\mu(A_i \triangle P_i) \leqslant \frac{2\delta}{10^3 k} \cdot \mu(P_i)$, *and*
4. $\phi(A_i) \leqslant \left(1 + \frac{4\delta}{10^3 k}\right) \cdot \phi(P_i) + \frac{4\delta}{10^3 k}$.

Part (b) of Theorem 1.2 gives a theoretical support for the practical success of spectral clustering based on approximate spectral embedding followed by k-means clustering. Moreover, if $k \leqslant \operatorname{poly}(\log n)$ and $\lambda_{k+1} \geqslant \operatorname{poly}(\log n)$, our algorithm works in nearly linear time. Previous papers [3, 7, 9] replaced k-means clustering by other techniques for their algorithmic results.

The k-means algorithm in [5] is efficient only for inputs $\mathcal{X}$ for which some partition into k clusters is much better than any partition into $k-1$ clusters. The authors proved that the algorithm is efficient for inputs $\mathcal{X}$ satisfying $\triangle_k(\mathcal{X}) \leqslant \varepsilon^2 \triangle_{k-1}(\mathcal{X})$ for some $\varepsilon \in (0, \varepsilon_0]$, where $\varepsilon_0 = 6/10^7$, stated that the result should also hold for a larger ε_0, and mentioned that they did not attempt to maximize ε_0. For the proof of part (b) of Theorem 1.2, we show in Section 5 that $\widetilde{\mathcal{X}_V}$ satisfies this assumption. In this proof, we need $\delta \leqslant k \cdot \varepsilon_0/600 = k/10^9$.

One of the reviewers suggested to include a numerical example. Consider a graph consisting of k cliques of size n/k each plus k additional edges that connect the cliques in the form of a ring. Such a graph is about the easiest input for a clustering algorithm. Then $\widehat{\rho}_{\text{avr}}(k) = \widehat{\rho}(k) \approx (k/n)^2$. For the gap assumption to hold we need $\lambda_{k+1} \geqslant 2 \cdot 20^4 \cdot k^3 \cdot \widehat{\rho}_{\text{avr}}(k)$. Since $\lambda_{k+1} \leqslant 2$, this implies $n \geqslant 400 \cdot k^{2.5}$. For small k, this is a modest requirement on the size of the graph.

[4] The case $\triangle_k(\mathcal{X}_V) \leqslant n^{-O(1)}$ constitutes a trivial clustering problem. For technical reasons, we have to exclude too easy inputs.

For the algorithmic result, we need in addition $\delta \leqslant k \cdot \varepsilon_0/600$. For the gap condition to hold, we need $2 \geqslant \lambda_{k+1} \geqslant (600/\varepsilon_0 k) \cdot 20^4 \cdot k^3 \cdot (k^2/n^2)$ or $n \geqslant 4\sqrt{3} \cdot 10^3 \cdot k^2/\sqrt{\varepsilon_0}$. For $\varepsilon_0 = 6/10^7$, this amounts to $n \geqslant 4\sqrt{5} \cdot 10^6 \cdot k^2$, a quite large lower bound on n.

Our statement above that Part (b) of Theorem 1.2 gives a theoretical support for the practical success of spectral clustering based on approximate spectral embedding followed by k-means clustering therefore has to be taken with a grain of salt. It is only an asymptotic statement and does not explain the good behavior on small graphs.

2 Highlights of Our Technical Contribution

2.1 Exact Spectral Embedding – Notation

We use the notation adopted by Peng et al. [7]. Let $f_j \in \mathbb{R}^V$ be the eigenvector corresponding to the j-th smallest eigenvalue λ_j of $\mathcal{L}_G$, and let $\overline{g_i} = \frac{D^{1/2}\chi_{P_i}}{\|D^{1/2}\chi_{P_i}\|}$ be the normalized indicator vector associated with the i-th optimal cluster $P_i \subset V$.

Since the eigenvectors $\{f_i\}_{i=1}^n$ form an orthonormal basis of $\mathbb{R}^n$, each normalized indicator vector $\overline{g_i}$ can be expressed as $\overline{g_i} = \sum_{j=1}^n \alpha_j^{(i)} f_j$, for all $i \in [1:k]$. Its *projection* into the subspace spanned by the bottom k eigenvectors is given by $\widehat{f}_i = \sum_{j=1}^k \alpha_j^{(i)} f_j$. Peng et al. [7] proved that if the gap parameter Υ is large enough then $\mathrm{span}(\{\widehat{f}_i\}_{i=1}^k) = \mathrm{span}(\{f_i\}_{i=1}^k)$ and hence the bottom k eigenvectors can be expressed by $f_i = \sum_{j=1}^k \beta_j^{(i)} \widehat{f}_j$, for all $i \in [1:k]$. We show that similar statements hold with substituted gap parameter Ψ.

A corner stone in the analysis of spectral clustering is to prove the existence of exactly k directions near which all spectrally embedded vectors are closely concentrated. These estimation centers are defined by

$$p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}} \left(\beta_i^{(1)}, \ldots, \beta_i^{(k)}\right)^{\mathrm{T}}. \tag{8}$$

Our analysis crucially relies on the isometric properties of the following square matrix. Let $\mathbf{B} \in \mathbb{R}^{k \times k}$ be a matrix defined by $\mathbf{B}_{j,i} = \beta_j^{(i)}$, for every $i, j \in [1:k]$.

2.2 Exact Spectral Embedding – Structural Results

The proof of Theorem 1.2 (a) follows the proof-structure of [7, Theorem 1.2] in Peng et al., but improves upon it in essential ways.

Our key technical insight is that the matrices $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ are close to the identity matrix. The proof of Theorem 2.1 appears in the full version of the paper.

▶ **Theorem 2.1** (Matrix $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ is Close to Identity Matrix). *If $\Psi \geqslant 10^4 \cdot k^3/\varepsilon^2$ and $\varepsilon \in (0,1)$ then for all distinct $i, j \in [1:k]$ it holds*

$$1 - \varepsilon \leqslant \langle \mathbf{B}_{i,:}, \mathbf{B}_{i,:} \rangle \leqslant 1 + \varepsilon \quad \textit{and} \quad |\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle| \leqslant \sqrt{\varepsilon}.$$

Informally, Theorem 2.1 implies that each normalized indicator vector $\overline{g_i}$ is close to the corresponding eigenvector f_i. This gives a simple and intuitive explanation for the success of spectral clustering.

To see this, let $\mathbf{F}_k, \widehat{\mathbf{F}}_k \in \mathbb{R}^{k \times k}$ be matrices whose i-th column is f_i and $\widehat{f}_i$, respectively. The projection matrix $\mathbf{P}_k$ into the k-th principle subspace of $\mathcal{L}_G$ is given by $\mathbf{P}_k = \mathbf{F}_k \mathbf{F}_k^{\mathrm{T}}$ and since $\widehat{\mathbf{F}}_k \mathbf{B} = \mathbf{F}_k$, by Theorem 2.1 it follows that $\widehat{\mathbf{F}}_k \widehat{\mathbf{F}}_k^{\mathrm{T}} \approx \mathbf{F}_k \mathbf{F}_k^{\mathrm{T}}$. Therefore, each projected indicator vector satisfies $\widehat{f}_i \approx f_i$. This implies $\alpha^{(i)} \approx \boldsymbol{\chi}_i$ and hence we have $\overline{g_i} \approx f_i$.

Formally, Theorem 2.1 allows us to improve the separation guarantee between any pair of estimation centers by a factor of k over [7, Lemma 4.3], measured in terms of the Euclidean distance.

▶ **Lemma 2.2.** *If $\delta = 20^4 \cdot k^3/\Psi \in (0,1]$ then for every $i \in [1:k]$ it holds that $\left\|p^{(i)}\right\|^2 \in \left[1 \pm \sqrt{\delta}/4\right] \frac{1}{\mu(P_i)}$.*

Proof. By definition $p^{(i)} = \frac{1}{\sqrt{\mu(P_i)}} \cdot \mathbf{B}_{i,:}$ and Theorem 2.1 yields $\|\mathbf{B}_{i,:}\|^2 \in [1 \pm \sqrt{\delta}/4]$. ◀

▶ **Lemma 2.3** (Larger Distance Between Estimation Centers)**.** *If $\delta = 20^4 \cdot k^3/\Psi \in (0,1/2]$ then for any distinct $i,j \in [1:k]$ it holds that $\left\|p^{(i)} - p^{(j)}\right\|^2 \geqslant [2 \cdot \min\{\mu(P_i), \mu(P_j)\}]^{-1}$.*

Proof. Since $p^{(i)}$ is a row of matrix B, Theorem 2.1 with $\varepsilon = \sqrt{\delta}/4$ yields

$$\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle = \frac{\langle \mathbf{B}_{i,:}, \mathbf{B}_{j,:} \rangle}{\|\mathbf{B}_{i,:}\| \, \|\mathbf{B}_{j,:}\|} \leqslant \frac{\sqrt{\varepsilon}}{1-\varepsilon} = \frac{2\delta^{1/4}}{3}.$$

W.l.o.g. assume that $\left\|p^{(i)}\right\|^2 \geqslant \left\|p^{(j)}\right\|^2$, say $\left\|p^{(j)}\right\| = \alpha \cdot \left\|p^{(i)}\right\|$ for some $\alpha \in (0,1]$. Then by Lemma 2.2 we have $\left\|p^{(i)}\right\|^2 \geqslant (1-\sqrt{\delta}/4) \cdot [\min\{\mu(P_i), \mu(P_j)\}]^{-1}$, and hence

$$\begin{aligned}\left\|p^{(i)} - p^{(j)}\right\|^2 &= \left\|p^{(i)}\right\|^2 + \left\|p^{(j)}\right\|^2 - 2\left\langle \frac{p^{(i)}}{\left\|p^{(i)}\right\|}, \frac{p^{(j)}}{\left\|p^{(j)}\right\|} \right\rangle \left\|p^{(i)}\right\| \left\|p^{(j)}\right\| \\ &\geqslant \left(\alpha^2 - \frac{4\delta^{1/4}}{3} \cdot \alpha + 1\right) \left\|p^{(i)}\right\|^2 \geqslant [2 \cdot \min\{\mu(P_i), \mu(P_j)\}]^{-1}. \end{aligned}$$ ◀

The observation that Υ can be replaced by Ψ in all statements in [7] is technically easy. However, this is crucial for Theorem 1.2 (b), since it yields an improved version of [7, Lemma 4.5] showing that a weaker by a factor of k assumption is sufficient. Due to space limitation, we defer the proof of Lemma 2.4 to the full version of the paper.

▶ **Lemma 2.4.** *Let $(P_1, \ldots, P_k)$ and $(A_1, \ldots, A_k)$ are partitions of the vector set. Suppose for every permutation $\pi : [1:k] \to [1:k]$ there is an index $i \in [1:k]$ such that*

$$\mu(A_i \triangle P_{\pi(i)}) \geqslant \frac{2\varepsilon}{k} \cdot \mu(P_{\pi(i)}), \tag{9}$$

where $\varepsilon \in (0,1)$ is a parameter. If $\delta = 20^4 \cdot k^3/\Psi \in (0,1/2]$ and $\varepsilon \geqslant 64\alpha \cdot k^3/\Psi$ then

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) > \frac{2k^2}{\Psi}\alpha.$$

With the above Lemmas in place, the proof of Theorem 1.2 (a) is then completed as in [7]. We give more details in Section 3.

Before we turn to Theorem 1.2 (b), we consider the variant of Lloyd's algorithm analyzed by Ostrovsky et al. [5] applied to $\mathcal{X}_V$. This algorithm is efficient for inputs $\mathcal{X}$ satisfying: some partition into k clusters is much better than any partition into $k-1$ clusters.

▶ **Theorem 2.5.** *[5, Theorem 4.15] Assuming that $\triangle_k(\mathcal{X}) \leqslant \varepsilon^2 \triangle_{k-1}(\mathcal{X})$ for $\varepsilon \in (0, 6/10^7]$, there is an algorithm that returns a solution of cost at most $[(1-\varepsilon^2)/(1-37\varepsilon^2)]\triangle_k(\mathcal{X})$ with probability at least $1 - O(\sqrt{\varepsilon})$ in time $O(nkd + k^3 d)$.*

In Section 4, we establish the assumption of Ostrovsky et al. [5] for $\mathcal{X}_V$.

▶ **Theorem 2.6** (Normalized Spectral Embedding is ε-separated). *Let G be a graph that satisfies the gap assumption $\delta = 20^4 \cdot k^3/\Psi \in (0, 1/2]$ and $\delta \leqslant k \cdot \varepsilon/600$, where $\varepsilon = 6/10^7$ is the Ostrovsky et al.'s constant. Then it holds*

$$\triangle_k(\mathcal{X}_V) \leqslant \varepsilon^2 \triangle_{k-1}(\mathcal{X}_V). \tag{10}$$

However, Theorem 2.6 is insufficient for Theorem 1.2 (b), since we need a similar result for the set $\widetilde{\mathcal{X}_V}$ formed by approximate eigenvectors. To overcome this issue we build upon the recent work by Boutsidis et al. [1] which shows that running an approximate k-means clustering algorithm on approximate eigenvectors obtained via the power method, yields an additive approximation to solving the k-means clustering problem on exact eigenvectors.

In order to state the connection, we need to introduce some of their notation.

2.3 Approximate Spectral Embedding – Notation

Let $Z \in \mathbb{R}^{n\times k}$ be a matrix whose rows represent n vectors that are to be partitioned into k clusters. For every k-way partition we associate an indicator matrix $X \in \mathbb{R}^{n\times k}$ that satisfies $X_{ij} = 1/\sqrt{|C_j|}$ if the i-th row $Z_{i,:}$ belongs to the j-th cluster C_j, and $X_{ij} = 0$ otherwise. We denote the optimal indicator matrix X_{opt} by

$$X_{\text{opt}} = \arg\min_{X\in\mathbb{R}^{n\times k}} \left\|Z - XX^{\mathrm{T}}Z\right\|_F^2 = \arg\min_{X\in\mathbb{R}^{n\times k}} \sum_{j=1}^{k} \sum_{u\in X_j} \|Z_{u,:} - c_j\|_2^2\,, \tag{11}$$

where $c_j = (1/|X_j|)\sum_{u\in X_j} Z_{u,:}$ is the center point of cluster C_j.

The normalized Laplacian matrix $\mathcal{L}_G \in \mathbb{R}^{n\times n}$ of a graph G is define by $\mathcal{L}_G = I - \mathcal{A}$, where $\mathcal{A} = D^{-1/2}AD^{-1/2}$ is the normalized adjacency matrix. Let $U_k \in \mathbb{R}^{n\times k}$ be a matrix composed of the bottom k orthonormal eigenvectors of $\mathcal{L}_G$ corresponding to the smallest eigenvalues $\lambda_1, \ldots, \lambda_k$. We define by $Y \triangleq U_k$ the canonical spectral embedding.

Our approximate spectral embedding is computed by the so called "**Power method**". Let $S \in \mathbb{R}^{n\times k}$ be a matrix whose entries are i.i.d. samples from the standard Gaussian distribution $N(0, 1)$ and p be a positive integer. Then the approximate spectral embedding $\widetilde{Y}$ is defined by the following process:

$$1)\ \mathcal{B} \triangleq I + \mathcal{A}; \quad 2)\ \text{Let } \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}} \text{ be the SVD of } \mathcal{B}^p S; \quad \text{and} \quad 3)\ \widetilde{Y} \triangleq \widetilde{U} \in \mathbb{R}^{n\times k}. \tag{12}$$

We proceed by defining the normalized (approximate) spectral embedding. We construct a matrix $Y' \in \mathbb{R}^{m\times k}$ such that for every vertex $u \in V$ we add $\deg(u)$ many copies of the normalized row $U_k(u,:)/\sqrt{\deg(u)}$ to Y'. Formally, the normalized (approximate) spectral embedding Y' ($\widetilde{Y'}$) is defined by

$$Y' = \begin{pmatrix} \mathbf{1}_{\deg(1)}\frac{U_k(1,:)}{\sqrt{\deg(1)}} \\ \cdots \\ \mathbf{1}_{\deg(n)}\frac{U_k(n,:)}{\sqrt{\deg(n)}} \end{pmatrix}_{m\times k} \quad \text{and} \quad \widetilde{Y'} = \begin{pmatrix} \mathbf{1}_{\deg(1)}\frac{\widetilde{U}(1,:)}{\sqrt{\deg(1)}} \\ \cdots \\ \mathbf{1}_{\deg(n)}\frac{\widetilde{U}(n,:)}{\sqrt{\deg(n)}} \end{pmatrix}_{m\times k}, \tag{13}$$

where $\mathbf{1}_{\deg(i)}$ is all-one column vector with dimension $\deg(i)$.

Similarly to (11) we associate to Y' ($\widetilde{Y'}$) an indicator matrix X' ($\widetilde{X'}$) that satisfies $X'_{ij} = 1/\sqrt{\mu(C_j)}$ if the i-th row $Y'_{i,:}$ belongs to the j-th cluster C_j, and $X'_{ij} = 0$ otherwise. We may assume w.l.o.g. that a k-means algorithm outputs an indicator matrix X' such that all copies of row $U_k(v,:)/\sqrt{\deg(v)}$ belong to the same cluster, for every vertex $v \in V$.

We associate to matrices Y' and $\widetilde{Y'}$ the sets of points $\mathcal{X}_V$ and $\widetilde{\mathcal{X}_V}$ respectively. We present now a key connection between the spectral embedding map $F(\cdot)$, the optimal k-means cost $\triangle_k(\mathcal{X}_V)$ and matrices Y', X'_{opt}:

$$\left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}}\right)^{\mathrm{T}} Y' \right\|_F^2 = \sum_{j=1}^{k} \sum_{v \in C_j^\star} \deg(v) \left\| F(v) - c_j^\star \right\|_F^2 = \triangle_k(\mathcal{X}_V), \tag{14}$$

where each center satisfies $c_j^\star = \mu(C_j^\star)^{-1} \cdot \sum_{v \in C_j^\star} \deg(v) F(v)$ and $F(v) = Y_{v,:}/\sqrt{\deg(v)}$.

2.4 Approximate Spectral Embedding – Algorithmic Results

Our analysis relies on the proof techniques developed in [1, 2]. By adjusting these techniques (c.f. [1, Lemma 5] and [2, Lemma 7]) to our setting, we prove (in the full version of the paper) the following result for the symmetric positive semi-definite matrix $\mathcal{B}$ whose largest k singular values (eigenvalues) correspond to the eigenvectors $u_1, \ldots, u_k$ of $\mathcal{L}_G$.

▶ **Lemma 2.7.** *Let $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}$ be the SVD of $\mathcal{B}^p S \in \mathbb{R}^{n \times k}$, where $p \geqslant 1$ and S is an $n \times k$ matrix of i.i.d. standard Gaussians. Let $\gamma_k = \frac{2-\lambda_{k+1}}{2-\lambda_k} < 1$ and fix $\delta, \epsilon \in (0,1)$. Then for any $p \geqslant \ln(8nk/\epsilon\delta)/\ln(1/\gamma_k)$ with probability at least $1 - 2e^{-2n} - 3\delta$ it holds*

$$\left\| U_k U_k^{\mathrm{T}} - \widetilde{U}\widetilde{U}^{\mathrm{T}} \right\|_F \leqslant \epsilon.$$

We establish several technical Lemmas that combined with Lemma 2.7 allow us to apply the proof techniques in [1, Theorem 6]. More precisely, we prove in Subsection 5.1 that running an approximate k-means algorithm on a normalized approximate spectral embedding $\widetilde{Y'}$ computed by the power method, yields an approximate clustering of the normalized spectral embedding Y'.

▶ **Theorem 2.8.** *Compute matrix $\widetilde{Y'}$ via the power method with $p \geqslant \ln(8nk/\epsilon\delta)/\ln(1/\gamma_k)$, where $\gamma_k = (2-\lambda_{k+1})/(2-\lambda_k) < 1$. Run on the rows of $\widetilde{Y'}$ an α-approximate k-means algorithm with failure probability δ_α. Let the outcome be a clustering indicator matrix $\widetilde{X'_\alpha} \in \mathbb{R}^{n \times k}$. Then with probability at least $1 - 2e^{-2n} - 3\delta_p - \delta_\alpha$ it holds*

$$\left\| Y' - \widetilde{X'_\alpha} \left(\widetilde{X'_\alpha}\right)^{\mathrm{T}} Y' \right\|_F^2 \leqslant (1+4\varepsilon) \cdot \alpha \cdot \left\| Y' - X'_{\text{opt}} \left(X'_{\text{opt}}\right)^{\mathrm{T}} Y' \right\|_F^2 + 4\varepsilon^2.$$

Our main technical contribution is to prove, in Subsection 5.2, that $\widetilde{\mathcal{X}_V}$ satisfies the assumption of Ostrovsky et al. [5]. Our analysis builds upon Theorem 2.6 and Theorem 2.8.

▶ **Theorem 2.9** (Approximate Normalized Spectral Embedding is ε-separated)**.** *Let G be a graph that satisfies the gap assumption $\delta = 20^4 \cdot k^3/\Psi \in (0, 1/2]$ and $\delta \leqslant k \cdot \varepsilon/600$, where $\varepsilon = 6/10^7$ is the Ostrovsky et al.'s constant. If the optimum cost[5] $\left\| Y' - X'_{\text{opt}}(X'_{\text{opt}})^{\mathrm{T}} Y' \right\|_F \geqslant n^{-O(1)}$ and the matrix $\widetilde{Y'}$ is constructed via the power method with $p \geqslant \Omega(\frac{\ln n}{\lambda_{k+1}})$, then w.h.p it holds $\triangle_k\left(\widetilde{\mathcal{X}_V}\right) < 5\varepsilon^2 \cdot \triangle_{k-1}\left(\widetilde{\mathcal{X}_V}\right)$.*

Based on the preceding results, we prove Theorem 1.2 (b) in Subsection 5.3.

[5] $\left\| Y' - X'_{\text{opt}}(X'_{\text{opt}})^{\mathrm{T}} Y' \right\|_F \geqslant n^{-O(1)}$ asserts a multiplicative approximation guarantee in Theorem 2.8.

3 The Proof of Part (a) of Theorem 1.2

The proof of part $(a.1)$ builds upon the following Lemmas. Recall that $\mathcal{X}_V$ contains d_u copies of $F(u)$ for each $u \in V$. W.l.o.g. we may restrict attention to clusterings of $\mathcal{X}_V$ that put all copies of $F(u)$ into the same cluster and hence induce a clustering of V. Let $(A_1, \ldots, A_k)$ with cluster centers c_1 to c_k be a clustering of V. Its k-means cost is

$$\mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) = \sum_{i=1}^{k} \sum_{u \in A_i} d_u \left\| F(u) - c_i \right\|^2 .$$

The proofs of Lemma 3.1 and Lemma 3.2 appear in the full version of the paper.

▶ **Lemma 3.1** ($(P_1, \ldots, P_k)$ is a good k-means partition). *If $\Psi > 4 \cdot k^{3/2}$ then there are vectors $\{p^{(i)}\}_{i=1}^k$ such that $\mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leqslant (1 + \frac{3k}{\Psi}) \cdot \frac{k^2}{\Psi}$.*

▶ **Lemma 3.2** (Only partitions close to $(P_1, \ldots, P_k)$ are good). *Under the hypothesis of Theorem 1.2, the following holds. If for every permutation $\sigma : [1:k] \to [1:k]$ there exists an index $i \in [1:k]$ such that*

$$\mu(A_i \triangle P_{\sigma(i)}) \geqslant \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_{\sigma(i)}), \quad \textit{then it holds} \quad \mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) > \frac{2\alpha k^2}{\Psi}.$$

We note that Lemma 3.2 follows directly by applying Lemma 2.4 with $\varepsilon = 64\alpha \cdot k^3/\Psi$. Substituting these bounds into (2) yields a contradiction, since

$$\frac{2\alpha k^2}{\Psi} < \mathrm{Cost}(\{A_i, c_i\}_{i=1}^k) \leqslant \alpha \cdot \triangle_k(\mathcal{X}_V) \leqslant \alpha \cdot \mathrm{Cost}(\{P_i, p^{(i)}\}_{i=1}^k) \leqslant \left(1 + \frac{3k}{\Psi}\right) \cdot \frac{\alpha k^2}{\Psi}.$$

Therefore, there exists a permutation π (the identity after suitable renumbering of one of the partitions) such that $\mu(A_i \triangle P_i) < \frac{8\alpha\delta}{10^4 k} \cdot \mu(P_i)$ for all $i \in [1:k]$.

Part $(a.2)$ follows from part $(a.1)$. Indeed, for $\delta' = 8\delta/10^4$ we have

$$\mu(A_i) \geqslant \mu(P_i \cap A_i) = \mu(P_i) - \mu(P_i \setminus A_i) \geqslant \mu(P_i) - \mu(A_i \triangle P_i) \geqslant \left(1 - \frac{\alpha\delta'}{k}\right) \cdot \mu(P_i)$$

and $|E(A_i, \overline{A_i})| \leqslant |E(P_i, \overline{P_i})| + \mu(A_i \Delta P_i)$ since every edge that is counted in $|E(A_i, \overline{A_i})|$ but not in $|E(P_i, \overline{P_i})|$ must have an endpoint in $A_i \Delta P_i$. Thus

$$\Phi(A_i) = \frac{|E(A_i, \overline{A_i})|}{\mu(A_i)} \leqslant \frac{|E(P_i, \overline{P_i})| + \frac{\alpha\delta'}{k} \cdot \mu(P_i)}{(1 - \frac{\alpha \cdot \delta'}{k}) \cdot \mu(P_i)} \leqslant \left(1 + \frac{2\alpha\delta'}{k}\right) \cdot \phi(P_i) + \frac{2\alpha\delta'}{k}.$$

This completes the proof of Theorem 1.2 (a).

4 The Normalized Spectral Embedding is ε-separated

In this section, we prove that the normalized spectral embedding $\mathcal{X}_V$ is ε-separated.

Proof of Theorem 2.6

We establish first a lower bound on $\triangle_{k-1}(\mathcal{X}_V)$.

▶ **Lemma 4.1.** *Let G be a graph that satisfies the gap assumption $\delta = 20^4 \cdot k^3/\Psi \in (0, 1/2]$. Then for $\delta' = 2\delta/20^4$ it holds $\triangle_{k-1}(\mathcal{X}_V) \geqslant 1/12 - \delta'/k$.*

Before we prove Lemma 4.1 we show that it implies (10). By Lemma 3.1 we have $\triangle_k(\mathcal{X}_V) \leqslant 2k^2/\Psi = \delta'/k$. Then, we apply Lemma 4.1 with $\delta \leqslant k \cdot \varepsilon/600$, where $\varepsilon = 6/10^7$ is the Ostrovsky et al.'s constant, yielding

$$\triangle_{k-1}(\mathcal{X}_V) \geqslant \frac{1}{12} - \frac{\delta'}{k} = \frac{1}{12} - \frac{2}{20^4} \cdot \frac{\delta}{k} \geqslant \frac{10^{10}}{9 \cdot 2^5} \cdot \frac{\delta}{k} = \frac{1}{\varepsilon^2} \cdot \frac{\delta'}{k} \geqslant \frac{1}{\varepsilon^2} \cdot \triangle_k(\mathcal{X}_V).$$

Proof of Lemma 4.1

Let $(P_1, \ldots, P_k)$ and $(Z_1, \ldots, Z_{k-1})$ be partitions of V. We define a mapping $\sigma : [1 : k-1] \mapsto [1 : k]$ by

$$\sigma(i) = \arg\max_{j \in [1:k]} \frac{\mu(Z_i \cap P_j)}{\mu(P_j)}, \quad \text{for every } i \in [1 : k-1].$$

We lower bound now the clusters overlapping in terms of the volume between any k-way and $(k-1)$-way partitions of V.

▶ **Lemma 4.2.** *Suppose $(P_1, \ldots, P_k)$ and $(Z_1, \ldots, Z_{k-1})$ are partitions of V. Then for any index $\ell \in [1 : k] \setminus \{\sigma(1), \ldots, \sigma(k-1)\}$ (there is at least one such ℓ) and for every $i \in [1 : k-1]$ it holds*

$$\left\{\mu(Z_i \cap P_{\sigma(i)}), \mu(Z_i \cap P_\ell)\right\} \geqslant \tau_i \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\},$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geqslant 0$.

Proof. By pigeonhole principle there is an index $\ell \in [1 : k]$ such that $\ell \notin \{\sigma(1), \ldots, \sigma(k-1)\}$. Thus, for every $i \in [1 : k-1]$ we have $\sigma(i) \neq \ell$ and

$$\frac{\mu(Z_i \cap P_{\sigma(i)})}{\mu(P_{\sigma(i)})} \geqslant \frac{\mu(Z_i \cap P_\ell)}{\mu(P_\ell)} \triangleq \tau_i,$$

where $\sum_{i=1}^{k-1} \tau_i = 1$ and $\tau_i \geqslant 0$ for all i. Hence, the statement follows. ◀

Proof of Lemma 4.1. Let $(Z_1, \ldots, Z_{k-1})$ be a $(k-1)$-way partition of V with centers $c'_1, \ldots, c'_{k-1}$ that achieves $\triangle_{k-1}(\mathcal{X}_V)$, and $(P_1, \ldots, P_k)$ be a k-way partition of V achieving $\widehat{\rho}_{\mathrm{avr}}(k)$. Our goal now is to lower bound the optimal $(k-1)$-means cost

$$\triangle_{k-1}(\mathcal{X}_V) = \sum_{i=1}^{k-1} \sum_{j=1}^{k} \sum_{u \in Z_i \cap P_j} d_u \left\|F(u) - c'_i\right\|^2. \tag{15}$$

By Lemma 4.2 there is an index $\ell \in [1 : k] \setminus \{\sigma(1), \ldots, \sigma(k-1)\}$. For $i \in [1 : k-1]$ let

$$p^{\gamma(i)} = \begin{cases} p^\ell & \text{, if } \left\|p^\ell - c'_i\right\| \geqslant \left\|p^{\sigma(i)} - c'_i\right\|; \\ p^{\sigma(i)} & \text{, otherwise.} \end{cases}$$

Then by combining Lemma 2.3 and Lemma 4.2, we have

$$\left\|p^{\gamma(i)} - c'_i\right\|^2 \geqslant \left[8 \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\}\right]^{-1} \text{ and } \mu(Z_i \cap P_{\gamma(i)}) \geqslant \tau_i \cdot \min\left\{\mu(P_\ell), \mu(P_{\sigma(i)})\right\}, \tag{16}$$

where $\sum_{i=1}^{k-1} \tau_i = 1$. We now lower bound the expression in (15). Since

$$\left\|F(u) - c'_i\right\|^2 \geqslant \frac{1}{2}\left\|p^{\gamma(i)} - c'_i\right\|^2 - \left\|F(u) - p^{\gamma(i)}\right\|^2,$$

it follows for $\delta' = 2\delta/20^4$ that

$$\begin{aligned}
\triangle_{k-1}(\mathcal{X}_V) &= \sum_{i=1}^{k-1}\sum_{j=1}^{k}\sum_{u\in Z_i\cap P_j} d_u \left\|F(u)-c'_i\right\|^2 \geqslant \sum_{i=1}^{k-1}\sum_{u\in Z_i\cap P_{\gamma(i)}} d_u \left\|F(u)-c'_i\right\|^2 \\
&\geqslant \frac{1}{2}\sum_{i=1}^{k-1}\sum_{u\in Z_i\cap P_{\gamma(i)}} d_u \left\|p^{\gamma(i)}-c'_i\right\|^2 - \sum_{i=1}^{k-1}\sum_{u\in Z_i\cap P_{\gamma(i)}} d_u \left\|F(u)-p^{\gamma(i)}\right\|^2 \\
&\geqslant \frac{1}{2}\sum_{i=1}^{k-1}\frac{\mu(Z_i\cap P_{\gamma(i)})}{8\cdot\min\left\{\mu(P_{\gamma(i)}),\mu(P_{\sigma(i)})\right\}} - \sum_{i=1}^{k}\sum_{u\in P_i} d_u \left\|F(u)-p^i\right\|^2 \\
&\geqslant \frac{1}{16}-\frac{\delta'}{k},
\end{aligned}$$

where the last inequality holds due to (16) and Lemma 3.1. ◀

5 An Efficient Spectral Clustering Algorithm

Here, we apply the proof techniques developed by Boutsidis et al. [1, 2] to our setting. More precisely, we prove that any α-approximate k-means algorithm that runs on an approximate normalized spectral embedding $\widetilde{Y'}$ computed by the power method, yields an approximate clustering $\widetilde{X'_\alpha}$ of the normalized spectral embedding Y'.

Furthermore, we prove under our gap assumption that $\widetilde{Y'}$ is ε-separated. This allows us to apply the variant of Lloyd's k-means algorithm analyzed by Ostrovsky et al. [5] to efficiently compute $\widetilde{X'_\alpha}$. Then we use Theorem 1.2 (a) to establish the desired statement.

This section is organized as follows. In Subsection 5.1, we prove Theorem 2.8. Then in Subsection 5.2, we present the proof of Theorem 2.9. Based on the results from the preceding two subsections, we prove Theorem 1.2 (b) in Subsection 5.3.

5.1 Proof of Theorem 2.8

Due to space limits, we defer the proofs of the next Lemmas to the full version of the paper.

▶ **Lemma 5.1.** *$X'X'^{\mathrm{T}}$ is a projection matrix.*

▶ **Lemma 5.2.** *It holds that $Y'^{\mathrm{T}}Y' = I_{k\times k} = \widetilde{Y'}^{\mathrm{T}}\widetilde{Y'}$.*

▶ **Lemma 5.3.** *It holds that $\left\|Y'Y'^{\mathrm{T}} - \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}}\right\|_F = \left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F$.*

▶ **Lemma 5.4.** *For any matrix U with orthonormal columns and every matrix A it holds*

$$\left\|UU^{\mathrm{T}} - AA^{\mathrm{T}}UU^{\mathrm{T}}\right\|_F = \left\|U - AA^{\mathrm{T}}U\right\|_F. \tag{17}$$

Proof Sketch of Theorem 2.8. By combining Lemma 2.7 and Lemma 5.3, with probability at least $1-2e^{-2n}-3\delta_p$ we have $\left\|Y'Y'^{\mathrm{T}} - \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}}\right\|_F = \left\|YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}}\right\|_F \leqslant \varepsilon$.

Let $Y'Y'^{\mathrm{T}} = \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} + E$ such that $\|E\|_F \leqslant \varepsilon$. Based on Lemma 5.2 and Lemma 5.4 we have that (17) holds for the matrices Y' and $\widetilde{Y'}$. Hence, by Lemma 5.1 we can apply the proof in [1, Theorem 6] to obtain $\left\|Y' - \widetilde{X'_\alpha}\left(\widetilde{X'_\alpha}\right)^{\mathrm{T}}Y'\right\|_F \leqslant \sqrt{\alpha}\cdot\left(\left\|Y' - X'_{\mathrm{opt}}\left(X'_{\mathrm{opt}}\right)^{\mathrm{T}}Y'\right\|_F + 2\varepsilon\right)$. The desired statement follows by simple algebraic manipulations. ◀

5.2 Proof of Theorem 2.9

In this subsection, we show under our gap assumption that the approximate normalized spectral embedding $\widetilde{Y'}$ is ε-separated, i.e. $\triangle_k(\widetilde{\mathcal{X}_V}) < 5\varepsilon^2 \cdot \triangle_{k-1}(\widetilde{\mathcal{X}_V})$. Our analysis builds upon Theorem 2.6, Theorem 2.8 and the proof techniques in [1, Theorem 6].

We use interchangeably X'_{opt} and $X'^{(k)}_{\text{opt}}$ to denote the optimal indicator matrix for the k-means problem on $\mathcal{X}_V$ that is induced by the rows of matrix Y'. Similarly, we denote by $X'^{(k-1)}_{\text{opt}}$ the optimal indicator matrix for the $(k-1)$-means problem on $\mathcal{X}_V$.

Proof Sketch of Theorem 2.9. By Theorem 2.6 we have

$$\left\| Y' - X'^{(k)}_{\text{opt}} \left(X'^{(k)}_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F \leqslant \varepsilon \left\| Y' - X'^{(k-1)}_{\text{opt}} \left(X'^{(k-1)}_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F . \tag{18}$$

We set the approximation parameter in Theorem 2.8 to

$$\varepsilon' \triangleq \frac{1}{4}\sqrt{\triangle_k(\mathcal{X}_V)} = \frac{1}{4} \left\| Y' - X'^{(k)}_{\text{opt}} \left(X'^{(k)}_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F \geqslant n^{-O(1)}, \tag{19}$$

and we note that by Theorem 2.6 it holds $\varepsilon' \leqslant \frac{\varepsilon}{4}\sqrt{\triangle_{k-1}(\mathcal{X}_V)}$.

We construct now matrix $\widetilde{Y}$ via the power method with $p \geqslant \Omega(\frac{\ln n}{\lambda_{k+1}})$. By Lemma 5.3 we have $\left\| Y'Y'^{\mathrm{T}} - \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} \right\|_F = \left\| YY^{\mathrm{T}} - \widetilde{Y}\widetilde{Y}^{\mathrm{T}} \right\|_F$ and thus by Lemma 2.7 with high probability it holds that $\left\| Y'(Y')^{\mathrm{T}} - \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} \right\|_F \leqslant \varepsilon'$.

Let $Y'(Y')^{\mathrm{T}} = \widetilde{Y'}\widetilde{Y'}^{\mathrm{T}} + E$ such that $\|E\|_F \leqslant \varepsilon'$. By combining Lemma 5.1, Lemma 5.2, Lemma 5.3 and by applying the proof techniques in [1, Theorem 6] we obtain

$$\sqrt{\triangle_k\left(\widetilde{\mathcal{X}_V}\right)} \leqslant \|E\|_F + \left\| Y' - \widetilde{X'^{(k)}_{\text{opt}}} \left(\widetilde{X'^{(k)}_{\text{opt}}} \right)^{\mathrm{T}} Y' \right\|_F .$$

Furthermore, we can show that $\ln\left(\frac{2-\lambda_k}{2-\lambda_{k+1}}\right) \geqslant \frac{1}{2}\left(1 - \frac{4\delta}{20^4 k^2}\right)\lambda_{k+1}$. Then Theorem 2.8 yields

$$\left\| Y' - \widetilde{X'^{(k)}_{\text{opt}}} \left(\widetilde{X'^{(k)}_{\text{opt}}} \right)^{\mathrm{T}} Y' \right\|_F^2 \leqslant (1+4\varepsilon') \cdot \left\| Y' - X'^{(k)}_{\text{opt}} \left(X'^{(k)}_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 + 4\varepsilon'^2.$$

Also, we can show $\left\| Y' - X'^{(k)}_{\text{opt}} \left(X'^{(k)}_{\text{opt}} \right)^{\mathrm{T}} Y' \right\|_F^2 \leqslant \frac{1}{8\cdot 10^{13}}$ and by the definition of ε' it follows

$$\sqrt{\triangle_k\left(\widetilde{\mathcal{X}_V}\right)} \leqslant 2\sqrt{\triangle_k(\mathcal{X}_V)} \leqslant 2\varepsilon \cdot \sqrt{\triangle_{k-1}(\mathcal{X}_V)}. \tag{20}$$

Moreover, by applying similar arguments as in the proof of [1, Theorem 6] we can prove that

$$\sqrt{\triangle_{k-1}(\mathcal{X}_V)} \leqslant \left(1 + \frac{\varepsilon}{2}\right)\sqrt{\triangle_{k-1}\left(\widetilde{\mathcal{X}_V}\right)}. \tag{21}$$

The statement follows by combining (20) and (21). ◀

5.3 Proof of Part (b) of Theorem 1.2

Let $p = \Theta(\frac{\ln n}{\lambda_{k+1}})$. We can compute the matrix $\mathcal{B}^p S$ in time $O(mkp)$ and its singular value decomposition $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}$ in time $O(nk^2)$. Based on it, we construct in time $O(mk)$ matrix $\widetilde{Y'}$ (c.f. (13)).

By Theorem 2.9, $\widetilde{\mathcal{X}_V}$ is ε-separated for $\varepsilon = 6/10^7$, i.e. $\triangle_k\left(\widetilde{\mathcal{X}_V}\right) < 5\varepsilon^2 \cdot \triangle_{k-1}\left(\widetilde{\mathcal{X}_V}\right)$. Hence, by Theorem 2.5 there is an algorithm that outputs in time $O(mk^2 + k^4)$ a clustering with indicator matrix $\widetilde{X'_\alpha}$ satisfying

$$\left\|\widetilde{Y'} - \widetilde{X'_\alpha}\left(\widetilde{X'_\alpha}\right)^{\mathrm{T}}\widetilde{Y'}\right\|_F^2 \leqslant \left(1 + \frac{1}{10^{10}}\right) \cdot \left\|\widetilde{Y'} - \widetilde{X'_{\mathrm{opt}}}\left(\widetilde{X'_{\mathrm{opt}}}\right)^{\mathrm{T}}\widetilde{Y'}\right\|_F^2$$

with constant probability (close to 1), where $\alpha = 1 + 1/10^{10}$.

Moreover, by applying Theorem 2.8 with $\varepsilon' = \frac{1}{4\cdot 10^3}\left\|Y' - X'_{\mathrm{opt}}\left(X'_{\mathrm{opt}}\right)^{\mathrm{T}}Y'\right\|_F$ we can prove that the indicator matrix $\widetilde{X'_\alpha}$ yields a multiplicative approximation of $\mathcal{X}_V$, i.e.

$$\left\|Y' - \widetilde{X'_\alpha}\left(\widetilde{X'_\alpha}\right)^{\mathrm{T}}Y'\right\|_F^2 \leqslant \left(1 + \frac{1}{10^6}\right)\left\|Y' - X'_{\mathrm{opt}}\left(X'_{\mathrm{opt}}\right)^{\mathrm{T}}Y'\right\|_F^2. \tag{22}$$

The statement follows by Theorem 1.2 (a) applied to the partition $(A_1, \ldots, A_k)$ of V that is induced by the indicator matrix $\widetilde{X'_\alpha}$.

Acknowledgement. We would like to thank the anonymous reviewers for their constructive comments.

References

1. Christos Boutsidis, Prabhanjan Kambadur, and Alex Gittens. Spectral clustering via the power method - provably. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 40–48, 2015.
2. Christos Boutsidis and Malik Magdon-Ismail. Faster svd-truncated regularized least-squares. In *2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, June 29 - July 4, 2014*, pages 1321–1325, 2014.
3. James R. Lee, Shayan Oveis Gharan, and Luca Trevisan. Multi-way spectral partitioning and higher-order Cheeger inequalities. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC'12, pages 1117–1130, New York, NY, USA, 2012. ACM.
4. Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.
5. Rafail Ostrovsky, Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. The effectiveness of Lloyd-type methods for the k-means problem. *J. ACM*, 59(6):28:1–28:22, January 2013.
6. Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1256–1266, 2014.
7. Richard Peng, He Sun, and Luca Zanetti. Partitioning well-clustered graphs: Spectral clustering works! In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1423–1455, 2015.
8. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

9 Ali Kemal Sinop. How to round subspaces: A new spectral clustering algorithm. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1832–1847, 2016.

10 Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, pages 395–416, 2007.