Understanding the Generalization Performance of Spectral Clustering Algorithms

Shaojie Li, Sheng Ouyang, Yong Liu*

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China
{lishaojie95, ouyangsheng, liuyonggsai}@ruc.edu.cn

Abstract

The theoretical analysis of spectral clustering is mainly devoted to consistency, while there is little research on its generalization performance. In this paper, we study the excess risk bounds of the popular spectral clustering algorithms: relaxed RatioCut and relaxed NCut. Our analysis follows the two practical steps of spectral clustering algorithms: continuous solution and discrete solution. Firstly, we provide the convergence rate of the excess risk bounds between the empirical continuous optimal solution and the population-level continuous optimal solution. Secondly, we show the fundamental quantity influencing the excess risk between the empirical discrete optimal solution and the population-level discrete optimal solution. At the empirical level, algorithms can be designed to reduce this quantity. Based on our theoretical analysis, we propose two novel algorithms that can penalize this quantity and, additionally, can cluster the out-of-sample data without re-eigendecomposition on the overall samples. Numerical experiments on toy and real datasets confirm the effectiveness of our proposed algorithms.

Introduction

Spectral clustering is one of the most popular algorithms in unsupervised learning and has been widely used for many machine learning applications (Von Luxburg 2007; Dhillon 2001; Kannan, Vempala, and Vetta 2004; Shaham et al. 2018; Liu et al. 2018). Given a set of data points independently sampled from an unknown underlying probability distribution, also referred to as population distribution, spectral clustering algorithms aim to divide all data points into several disjoint sets based on some notion of similarity. Spectral clustering originates from the spectral graph partitioning (Fiedler 1973). One way to understand spectral clustering is to view it as a relaxation of searching for the best graph-cut since the latter is known as an NP-hard problem (Von Luxburg 2007). The core method of spectral clustering is the eigendecomposition on the graph Laplacian, and the matrix composed of eigenvectors can be interpreted as a lower-dimensional representation that preserves the grouping relationships among data points as much as possible. Subsequently, various methods such as k-means (Ng, Jordan, and Weiss 2001; Shi and Malik 2000), dynamic programming (Alpert and Kahng 1995), or orthonormal transform (Stella and Shi 2003) can be employed to get the discrete solution on the matrix and therefore the final group partitions.

However, compared with the prosperous development of the design and application, the generalization performance analysis of spectral clustering algorithms is scarce. Hitherto, the theoretical analysis of spectral clustering almost focuses on consistency (Von Luxburg, Belkin, and Bousquet 2008; Von Luxburg, Bousquet, and Belkin 2004; Cao and Chen 2011; Trillos and Slepčev 2018; Trillos et al. 2016; Schiebinger et al. 2015; Terada and Yamamoto 2019). Consistency means that if it is true that as the sample size collected goes to infinity, the partitioning of the data constructed by spectral clustering converges to a certain meaningful partitioning on the population level (Von Luxburg, Belkin, and Bousquet 2008), but consistency alone does not reveal the sample complexity (Vapnik 1999). To our best knowledge, there is only one research that investigates the generalization performance of kernel NCut (Terada and Yamamoto 2019). They adopt the relationship between NCut and the weighted kernel k-means (Dhillon, Guan, and Kulis 2007), based on which they establish the excess risk bounds for kernel NCut. However, Terada and Yamamoto (2019) study the graph-cut solution, not the solution of spectral clustering that we used in practice.

In this paper, we are interested in the generalization performance of the practical solution. We investigate the excess risk bound of two popular spectral clustering algorithms: *relaxed* RatioCut and *relaxed* NCut. Notably, to compare with the RatioCut and NCut without relaxation, we refer to spectral clustering as *relaxed* RatioCut and *relaxed* NCut in this paper. It is known that spectral clustering typically consists of two steps (Von Luxburg 2007): (1) to obtain the optimal continuous solution by the eigendecomposition on the graph Laplacian; (2) to obtain the optimal discrete solution, also referred to as discretization, from the continuous solution by some heuristic algorithms, such as k-means and orthonormal transform. Consistent with the two steps, we first investigate the excess risk bound between the empirical continuous optimal solution and the population-level continuous optimal solution. In deriving this bound, an immediate emerging difficulty is that the empirical continuous solution

*Corresponding author.

and the population-level continuous solution are in different dimensional spaces, making the empirical solution impossible to substitute into the expected error formula. To overcome this difficulty, we define integral operators, and use the spectral relationship between the integral operator and the graph Laplacian to extend the finite-dimensional eigenvector to the infinite-dimensional eigenfunction. Thus the analysis can be continued. We show that for both *relaxed* RatioCut and *relaxed* NCut, their excess risk bounds have a convergence rate of the order $\mathcal{O}(1/\sqrt{n})$. Secondly, we investigate the excess risk bound between the empirical discrete optimal solution and the population-level discrete optimal solution. We observe the fundamental quantity in influencing this excess risk, whose presence is caused by the heuristic algorithms used in step (2). This fundamental quantity motivates us to design algorithms to penalize it from the empirical perspective, reducing it as small as possible. Meanwhile, we observe that the orthonormal transform (Stella and Shi 2003) is an effective algorithm for penalizing this term, whose optimization objective corresponds to the empirical form of this fundamental quantity. Additionally, an obvious drawback of spectral clustering algorithms (*relaxed* NCut and *relaxed* RatioCut) is that they fail to generalize to the out-of-sample data points, requiring re-eigendecomposition on the overall data points. Based on our theoretical analysis, we propose two algorithms, corresponding to *relaxed* NCut and *relaxed* RatioCut, respectively, which can cluster the unseen samples without the eigendecomposition on the overall samples, largely reducing the time complexity. Moreover, when clustering the unseen samples, the proposed algorithms will penalize the fundamental quantity for searching for the optimal discrete solution, decreasing the excess risk. We have numerical experiments on the two algorithms, and the experimental results verify the effectiveness of our proposed algorithms. Our contributions are summarized as follows:

- We provide the convergence rate of the excess risk bounds for the continuous solution of spectral clustering.
- We show the fundamental quantity influencing the excess risk for the discrete solution of spectral clustering. We then propose two algorithms that can penalize this term and, additionally, can generalize to the new samples.
- Numerical experiments on toy and real datasets demonstrate the effectiveness of our proposed algorithms.

Related Work

This section introduces related work on the theoretical analysis of spectral clustering algorithms. Existing theoretical research of spectral clustering almost focuses on consistency. Specifically, Von Luxburg, Belkin, and Bousquet (2008) establish consistency for the embedding by proving that as much as the eigenvectors of the Laplacian matrix converge uniformly to the eigenfunctions of the Laplacian operator. Rosasco, Belkin, and De Vito (2010) provide the simpler proof of this convergence. Cao and Chen (2011) construct the consistency of regularized spectral clustering. Rohe et al. (2011) analyze the consistency for stochastic block models, Ting, Huang, and Jordan (2011) focus on the spectral convergence, Pelletier and Pudlo (2011) study the convergence of graph Laplacian, and Singer and Wu (2017) analyze the convergence of the connection graph Laplacian. Trillos et al. (2016) propose a framework and improves the results in (Arias-Castro, Pelletier, and Pudlo 2012) by minimizing the discrete functionals over all possible partitions of the data points, while the latter just minimizes a specific family of subsets of the data points. Based on the framework in (Trillos et al. 2016), Trillos and Slepčev (2018) provide a variational approach known as Γ-convergence, proving the convergence of the spectrum of the graph Laplacian towards the spectrum of a corresponding continuous operator. Terada and Yamamoto (2019) investigate the kernel normalized cut. They establish the consistency by the weighted k-means on the reproducing kernel Hilbert space (RKHS) and derive the excess risk bound for kernel NCut. However, as we discussed before, they study the graph-cut solution, not the practical solution of spectral clustering. Unlike the above research, we study the excess risk bound of the popular spectral clustering algorithms (*relaxed* RatioCut and *relaxed* NCut), instead of consistency. Our analysis is based on the practical steps of spectral clustering and spans two perspectives: the continuous solution and discrete solution.

Preliminaries

In this section, we introduce some notations and have a brief introduction to spectral clustering. For more spectral clustering's details, please refer to Von Luxburg (2007).

We consider the real space in this paper. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$, ρ be a probability measure on $\mathcal{X}$, and ρ_n be the empirical measure. Given a set of samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ independently drawn from the population distribution ρ, the weighted graph constructed on $\mathbf{X}$ can be specified by $\mathcal{G} = (\mathbb{V}, \mathbb{E}, \mathbf{W})$, where $\mathbb{V}$ denotes the set of all nodes, $\mathbb{E}$ denotes the set of all edges connecting the nodes, and $\mathbf{W} := (\mathbf{W}_{i,j})_{n\times n} = (\frac{1}{n}W(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$ is a weight matrix calculated by the weight function $W(x, y)$. Let $|\mathbb{V}| = n$ denotes the number of all data points to be grouped. To cluster n points into K groups is to decompose $\mathbb{V}$ into K disjoint sets, i.e., $\mathbb{V} = \cup_{l=1}^{K}\mathbb{V}_l$ and $\mathbb{V}_k \cap \mathbb{V}_l = \varnothing$, $\forall k \neq l$. We define the degree matrix $\mathbf{D}$ to be a diagonal matrix with entries $d_i = \sum_{j=1}^{n}\mathbf{W}_{i,j}$. Then, the unnormalized graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, and the asymmetric normalized graph Laplacian is defined as $\mathbf{L}_{rw} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$.

We now present some facts about spectral clustering. Let $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_K) \in \mathbb{R}^{n\times K}$, where $\mathbf{u}_1, ..., \mathbf{u}_K$ are K vectors. We define the following empirical error:

$$\hat{F}(\mathbf{U}) := \frac{1}{2n(n-1)}\sum_{k=1}^{K}\sum_{i,j=1, i\neq j}^{n}\mathbf{W}_{i,j}(\mathbf{u}_{k,i} - \mathbf{u}_{k,j})^2, \tag{1}$$

where $\mathbf{u}_{k,i}$ means the i-th component of the k-th vector $\mathbf{u}_k$. The optimization objective of RatioCut can be written as:

$$\min_{\mathbf{U}} \hat{F}(\mathbf{U}) \text{ s.t. } \left\{\mathbf{u}_{i,j} = \frac{1}{\sqrt{|\mathbb{V}_j|}} \text{ if } v_i \in \mathbb{V}_j, \text{otherwise } 0\right\}, \tag{2}$$

where $|\mathbb{V}_j|$ denotes the number of vertices of a subset $\mathbb{V}_j$ of a graph. The optimization objective of NCut can be written

as:

$$\min_{\mathbf{U}} \hat{F}(\mathbf{U}), \text{s.t.} \left\{ \mathbf{u}_{i,j} = \frac{1}{\sqrt{\text{vol}(\mathbb{V}_j)}} \text{if } v_i \in \mathbb{V}_j, \text{otherwise } 0 \right\}, \tag{3}$$

where $\text{vol}(\mathbb{V}_j)$ denotes the summing weights of edges of a subset $\mathbb{V}_j$ of a graph. Since searching for the optimal solution of RatioCut and NCut is known as an NP-hard problem (Von Luxburg 2007), spectral clustering often involves a relaxation operation, which allows the entries of $\mathbf{U}$ to take arbitrary real values (Von Luxburg 2007). Thus the optimization objective of *relaxed* RatioCut can be written as:

$$\min_{\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_K)} \hat{F}(\mathbf{U}), \text{ s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}, \tag{4}$$

where $\mathbf{I}$ is the identity matrix. The optimal solution of *relaxed* RatioCut is given by choosing $\mathbf{U}$ as the matrix which contains the first K eigenvectors of $\mathbf{L}$ as columns (Von Luxburg 2007). Similarly, the optimization objective of *relaxed* NCut can be written as:

$$\min_{\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_K)} \hat{F}(\mathbf{U}), \text{ s.t. } \mathbf{U}^T\mathbf{D}\mathbf{U} = \mathbf{I}. \tag{5}$$

The optimal solution of *relaxed* NCut is given by choosing the matrix $\mathbf{U}$ which contains the first K eigenvectors of $\mathbf{L}_{rw}$ as columns (Von Luxburg 2007).

Excess Risk Bounds

Let $W : \mathcal{X} \times \mathcal{X} \to R$ be a symmetric continuous weight function such that

$$0 < W(x, y) \leq C \quad x, y \in \mathcal{X}, \tag{6}$$

measuring the similarities between pairs of data points $x, y \in \mathcal{X}$. Since $W : \mathcal{X} \times \mathcal{X} \to R$ is not necessary to be positive definite and positive W is more common in practice, we assume that W to be positive in this paper. We now define the degree function as $m(x) = \int_{\mathcal{X}} W(x, y)d\rho(y)$, and then define the function: $p(x, y) = m(x)$ if $x = y$ and 0 otherwise, which is the population counterpart of the degree matrix. Let $L^2(\mathcal{X}, \rho)$ denotes the space of square integrable functions with norm $\|f\|_\rho^2 = \langle f, f\rangle_\rho = \int_{\mathcal{X}} |f(x)|^2 d\rho(x)$.

Relaxed RatioCut

Based on the weight function W, we define the function $L : \mathcal{X} \times \mathcal{X} \to R$

$$L(x, y) = p(x, y) - W(x, y) \quad x, y \in \mathcal{X},$$

which is symmetric. When L is restricted to samples $\forall \mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ for any positive integer n, the corresponding matrix $\mathbf{L}$ is positive semi-definite (refer to proposition 1 in (Von Luxburg 2007)), thus $L(x, y)$ is a kernel function and associated with a RKHS $\mathcal{H}$ with scalar product (norm) $\langle\cdot,\cdot\rangle$ ($\|\cdot\|$). For *relaxed* RatioCut, we assume

$$\kappa = \sup_{x\in\mathcal{X}} L(x, x) \tag{7}$$

and $L(x, y)$ to be continuous, which are common assumptions in spectral clustering. The elements in $\mathcal{H}$ are thus bounded continuous functions, and the corresponding integral operator $L_K : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$

$$(L_K f)(x) = \int_{\mathcal{X}} L(x, y)f(y)d\rho(y)$$

is thus a bounded operator. The operator L_K is the limit version of the Laplacian $\mathbf{L}$ (Rosasco, Belkin, and De Vito 2010). In other words, the matrix $\mathbf{L}$ is an empirical version of the operator L_K.

To study the excess risk bound, we need to define the population-level error, a limit version of Eq. (1):

$$F(U) := \frac{1}{2}\sum_{k=1}^{K} \iint W(x, y)(u_k(x) - u_k(y))^2 d\rho(x)d\rho(y),$$

where $U = (u_1, ..., u_K)$ consists of K functions u_k. Further, the optimization objective of the population-level error of *relaxed* RatioCut, analogous to Eq. (4), can be defined as:

$$\min_U F(U) \text{ s.t. } \langle u_i, u_j\rangle_\rho = 1 \text{ if } i = j, \text{ otherwise } 0. \tag{8}$$

Let $\widetilde{U}^* = (\widetilde{u}_1^*, ..., \widetilde{u}_K^*)$ be the optimal solution of Eq. (8). Actually, $\widetilde{u}_1^*, ..., \widetilde{u}_K^*$ are eigenfunctions of the operator L_K (Rosasco, Belkin, and De Vito 2010), that is $L_K\widetilde{u}_k^* = \lambda_k(L_K)\widetilde{u}_k^*$ for $k = 1, ..., K$, where $\lambda_k(L_K)$ is an eigenvalue of the operator L_K, $k = 1, ..., K$.

With the population-level error, we begin to analyze the excess risk bound. Excess risk measures on the population-level how the difference between the error of the empirical solution and the error of the population optima performs (Biau, Devroye, and Lugosi 2008; Liu 2021; Li and Liu 2021), formalized as

$$F(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*),$$

where $\widetilde{\mathbf{U}}^* = (\widetilde{\mathbf{u}}_1^*, ..., \widetilde{\mathbf{u}}_K^*)$ is the optimal solution of the empirical error of *relaxed* RatioCut, i.e., Eq. (4), and, actually, $\widetilde{\mathbf{u}}_1^*, ..., \widetilde{\mathbf{u}}_K^*$ are the eigenvectors of Laplacian $\mathbf{L}$ (Von Luxburg 2007).

However, an immediate difficulty to derive the bound of $F(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)$ is that $\widetilde{\mathbf{U}}^*$ and $\widetilde{U}^*$ are in different spaces. Specifically, $\widetilde{\mathbf{U}}^* \in \mathbb{R}^{n\times K}$ related to sample size n is in finite-dimensional space, while $\widetilde{U}^*$ is in infinite-dimensional function space. The fact that for different sample size n, the elements in $\widetilde{\mathbf{U}}^*$ live in different spaces, making the term $F(\widetilde{\mathbf{U}}^*)$ impossible to be calculated. To overcome this challenge, we define operator $T_n : \mathcal{H} \to \mathcal{H}$:

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \langle\cdot, L_{\mathbf{x}_i}\rangle L_{\mathbf{x}_i},$$

where $L_{\mathbf{x}_i} = L(\cdot, \mathbf{x}_i)$. And we denote $\breve{U} = (\breve{u}_1, ..., \breve{u}_K)$ as the first K eigenfunctions of the operator T_n. Rosasco, Belkin, and De Vito (2010) show that T_n and $\mathbf{L}$ have the same eigenvalues (up to zero eigenvalues) and their corresponding eigenfunctions and eigenvectors are closely related. If λ_k is a nonzero eigenvalue and $\widetilde{\mathbf{u}}_k^*$, $\breve{u}_k$ are the cor-

responding eigenvector and eigenfunction of $\mathbf{L}$ and T_n (normalized to norm 1 in $\mathbb{R}^n$ and $\mathcal{H}$) respectively, then

$$\begin{aligned}\widetilde{\mathbf{u}}_k^* &= \frac{1}{\sqrt{\lambda_k}}\left(\breve{u}_k(\mathbf{x}_1),...,\breve{u}_k(\mathbf{x}_n)\right);\\ \breve{u}_k(x) &= \frac{1}{\sqrt{\lambda_k}}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{u}}_k^{*i}L(x,\mathbf{x}_i)\right),\end{aligned} \tag{9}$$

where $\widetilde{\mathbf{u}}_k^{*i}$ is the i-th component of $\widetilde{\mathbf{u}}_k^*$.

From Eq. (9), one can see that the eigenvectors of $\mathbf{L}$ are the empirical version of the eigenfunctions of T_n. In other words, if the eigenfunction $\breve{u}_k(x)$ is restricted to the dataset $\mathbf{X}$, it can be mapped into the eigenvector $\widetilde{\mathbf{u}}_k^*$. Meanwhile, the eigenfunctions of T_n are the extensions of the eigenvectors of $\mathbf{L}$, which are infinite-dimensional. Back to the term $F(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)$, we can replace the vectors in $\widetilde{\mathbf{U}}^*$ by its corresponding extended eigenfunctions in $\breve{U}$. Therefore, we now can investigate the excess risk bound between the empirical continuous optimal solution and the population-level continuous optimal solution by bounding the term $F(\breve{U}) - F(\widetilde{U}^*)$. Additionally, the relations between the eigenvectors in $\widetilde{\mathbf{U}}^*$ and the eigenfunctions in $\breve{U}$ can be applied to cluster out-of-sample data points. One can approximately calculate the eigenvectors of the out-of-sample data by the eigenfunctions in $\breve{U}$. We will show the details in the Algorithms Section.

We now present the first risk bound for *relaxed* RatioCut.

Theorem 1. *Suppose for any $\breve{u} \in \mathcal{H}$ such that $\|\breve{u}\|_\infty \leq \sqrt{B}$ with an absolute constant $B > 0$, then for any $\delta > 0$, with probability at least $1 - 2\delta$, we have*

$$\begin{aligned}&F(\breve{U}) - F(\widetilde{U}^*)\\ \leq& 8CBK\left(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + K\frac{2\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt{n}},\end{aligned}$$

where C and κ are positive constants, and where K is the clustering number.

Remark 1. Theorem 1 suggests that the excess risk bound of *relaxed* RatioCut between the empirical continuous optimal solution and the population-level continuous optimal solution has a convergence rate of the order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ if we assume that the eigenfunctions $\breve{u} \in \mathcal{H}$ of operator T_n are bounded, i.e., $\|\breve{u}\|_\infty \leq \sqrt{B}$. This assumption is mild. Since we assume the kernel function $L(x,y) \leq \kappa$ and is continuous, the elements in $\mathcal{H}$ associated with $L(x,y)$ are bounded. The definition of operator T_n is: $\mathcal{H} \to \mathcal{H}$, so it is reasonable to assume the eigenfunctions of T_n are bounded, i.e., $\|\breve{u}\|_\infty \leq \sqrt{B}$. C and κ are constants in Eqs. (6) and (7, respectively. We provide the proof of Theorem 1 in the Appendix[1].

Remark 2. We highlight that we study the excess risk of spectral clustering. Compared with the generalization error bound $\hat{F}(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)$ that measures the difference between the empirical error of the empirical solution and the population-level error of the population-level solution, excess risk analysis is much more difficult because $\widetilde{\mathbf{U}}^*$ can not be calculated in expectation $F(\widetilde{\mathbf{U}}^*)$. The generalization error bound of *relaxed* RatioCut is easier to obtain since $\widetilde{\mathbf{U}}^*$ can be directly substituted into $\hat{F}(\cdot)$ to calculate, and its proof indeed is included in the proof of Theorem 1. We show the generalization error bound as a corollary below.

Corollary 1. *Under the above assumptions, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\hat{F}(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*) \leq K\frac{2\sqrt{2}\kappa\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}},$$

where κ is a positive constant, and where K is the clustering number.

In practice, after obtaining eigenvectors of the Laplacian $\mathbf{L}$, spectral clustering uses the heuristic algorithms on the eigenvectors to obtain the discrete solution. In analogy to this empirical process, we define the population-level discrete solution $\ddot{U} = (\ddot{u}_1,...,\ddot{u}_K)$, which are K functions in RKHS $\mathcal{H}$ and are sought through $\mathcal{H}$ by the population-level continuous solution $\breve{U}$. Let $U^* = (u_1^*,...,u_K^*)$ be the optimal solution of the minimal population-level error of RatioCut, i.e., optimal solution of the population-level version of Eq. (2). We then study the excess risk between the empirical discrete optimal solution and the population-level discrete optimal solution by bounding the term $F(\ddot{U}) - F(U^*)$.

Theorem 2. *Denoted by $\epsilon := \sum_{k=1}^K \|\ddot{u}_k - \breve{u}_k\|_2$. Suppose for any $\breve{u} \in \mathcal{H}$ such that $\|\breve{u}\|_\infty \leq \sqrt{B}$ with an absolute constant $B > 0$, then for any $\delta > 0$, with probability at least $1 - 2\delta$, we have*

$$\begin{aligned}&F(\ddot{U}) - F(U^*)\\ \leq& 4C\epsilon + 8CBK\left(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + \frac{2K\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt{n}},\end{aligned}$$

where C and κ are positive constants, and where K is the clustering number.

Remark 3. In the proof of Theorem 2, we make an error decomposition: $F(\ddot{U}) - F(U^*) = \underbrace{F(\ddot{U}) - F(\breve{U})}_{\mathcal{A}} + \underbrace{F(\breve{U}) - \hat{F}(\widetilde{\mathbf{U}}^*)}_{\mathcal{B}} + \underbrace{\hat{F}(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)}_{\mathcal{C}} + \underbrace{F(\widetilde{U}^*) - F(U^*)}_{\mathcal{D}}$. Term $\mathcal{B}$ is proved by the empirical process theory, term $\mathcal{C}$ is proved by spectral properties of the integral operator and the operator theory, while term $\mathcal{D} \leq 0$ can be revealed easily. Combining the bounds of terms $\mathcal{B}$ and $\mathcal{C}$ gives the result of Theorem 1. For term $\mathcal{A}$, we show that it can be bounded by $4C\sum_{k=1}^K \|\ddot{u}_k - \breve{u}_k\|_2$ (The proof is provided in the Appendix). We denote this quantity as ϵ, and the upper bound implies that $\sum_{k=1}^K \|\ddot{u}_k - \breve{u}_k\|_2$ is a fundamental quantity in influencing the excess risk between the empirical discrete

[1] https://arxiv.org/abs/2205.00281v2.

optimal solution and the population-level discrete optimal solution, which motivates us to penalize it as much as possible at the empirical level. We thus propose new algorithms in the next Section. Additionally, since searching for the best graph-cut is known as an NP-hard problem (Von Luxburg 2007), we investigate the generalization performance of the discrete solution obtained from the continuous solution, conducted in the practical spectral clustering process rather than the agnostic graph-cut solution. We hope that the theoretical study on such a kind of discrete solution can guide the design of novel spectral clustering algorithms.

Relaxed NCut

We consider *relaxed* NCut related to the asymmetric normalized Laplacian $\mathbf{L}_{rw}$. The analysis of *relaxed* NCut follows the similar pattern to *relaxed* RatioCut. Bound (6) implies the corresponding integral operator $\mathbb{L} : L^2(\mathcal{X}, \rho) \to L^2(\mathcal{X}, \rho)$

$$(\mathbb{L}f)(x) = f(x) - \int_{\mathcal{X}} \frac{W(x,y)f(y)}{m(x)} d\rho(y)$$

is well defined and continuous. To avoid the notation abuse, we use symbols provided in *relaxed* RatioCut. Corresponding minimal population-level error similar to Eq. (8) can be easily written from the empirical version of Eq. (5). For brevity, we omit it and just give some notations here. Let $\widetilde{U}^* = (\widetilde{u}_1^*, ..., \widetilde{u}_K^*)$ be the optimal solution of the minimal population-level error of *relaxed* NCut, which are eigenfunctions of the operator $\mathbb{L}$ (Rosasco, Belkin, and De Vito 2010). We denote $\widetilde{\mathbf{U}}^* = (\widetilde{\mathbf{u}}_1^*, ..., \widetilde{\mathbf{u}}_K^*)$ as the optimal solution of minimal empirical error of *relaxed* NCut, i.e., Eq. (5), which actually are eigenvectors of the Laplacian $\mathbf{L}_{rw}$(Von Luxburg 2007).

Firstly, we bound the term $F(\check{U}) - F(\widetilde{U}^*)$. However, another immediate difficulty is that the methods described in *relaxed* RatioCut are not directly applicable for *relaxed* NCut. The operator corresponding to T_n in the previous subsection appears to be impossible to be defined for *relaxed* NCut because W is not necessarily positive definite, so there is no RKHS associated with it. Moreover, even if W is positive definite, the operator $\mathbb{L}$ involves a division by a function, so there may not be a map from the RKHS $\mathcal{H}$ to itself. To overcome this challenge, we use an assumption on W introduced in (Rosasco, Belkin, and De Vito 2010) to construct an auxiliary RKHS $\mathcal{H}$ associated with a continuous real-valued bounded kernel $\mathcal{K}$. Here is the assumption:

Assumption 1. *Assume that $W : \mathcal{X} \times \mathcal{X} \to R$ is a positive, symmetric function such that*

$$W(x,y) \geq c > 0 \quad x, y \in \mathcal{X}; \quad W \in C_b^{d+1}(\mathcal{X} \times \mathcal{X}),$$

where $C_b^{d+1}(\mathcal{X}\times\mathcal{X})$ is a family of continuous bounded functions such that all the (standard) deviations of orders exist and are continuous bounded functions.

According to (Rosasco, Belkin, and De Vito 2010), Assumption 1 implies that there exists a RKHS $\mathcal{H}$ with bounded continuous kernel $\mathcal{K}$ such that: $W_x, \frac{1}{m_n} W_x \in \mathcal{H}$, where $W_x = W(\cdot, x)$ and $m_n = \frac{1}{n}\sum_{i=1}^n W_{x_i}$. This allows us to define the following empirical operators $\mathbb{L}_n, A_n : \mathcal{H} \to \mathcal{H}$

$$A_n = \frac{1}{n}\sum_{i=1}^n \langle \cdot, \mathcal{K}_{\mathbf{x}_i}\rangle_{\mathcal{H}} \frac{1}{m_n} W_{\mathbf{x}_i}; \quad \mathbb{L}_n = I - A_n,$$

where $\mathcal{K}_x = \mathcal{K}(\cdot, x)$. Let $\check{U} = (\check{u}_1, ..., \check{u}_K)$ be the first K eigenfunctions of the operator $\mathbb{L}_n$. Rosasco, Belkin, and De Vito (2010) show that $\mathbb{L}_n$, A_n and $\mathbf{L}_{rw}$ have closely related eigenvalues and eigenfunctions. The spectra of $\mathbf{L}_{rw}$ and $\mathbb{L}_n$ are the same up to the eigenvalue 1. Moreover, if $\lambda_k \neq 1$ is an eigenvalue and $\widetilde{\mathbf{u}}_k^*$, $\check{u}_k$ are the eigenvector and eigenfunction of $\mathbf{L}_{rw}$ and $\mathbb{L}_n$, respectively, then

$$\begin{aligned} \widetilde{\mathbf{u}}_k^* &= (\check{u}_k(\mathbf{x}_1), ..., \check{u}_k(\mathbf{x}_n)); \\ \check{u}_k(x) &= \frac{1}{1-\lambda_k}\frac{1}{n}\sum_{i=1}^n \frac{W(x, \mathbf{x}_i)}{m_n(x)} \widetilde{\mathbf{u}}_k^{*i}, \end{aligned} \tag{10}$$

where $\widetilde{\mathbf{u}}_k^{*i}$ is the i-th component of the eigenvector $\widetilde{\mathbf{u}}_k^*$. From Eq. (10), one can observe that the eigenvectors of $\mathbf{L}_{rw}$ are the empirical version of the eigenfunctions of $\mathbb{L}_n$. Moreover, the eigenfunctions of $\mathbb{L}_n$ are the extensions of the eigenvectors of $\mathbf{L}_{rw}$, which are infinite-dimensional. Therefore, given the eigenvectors of $\mathbf{L}_{rw}$, we can extend it to the corresponding eigenfunctions. With this relationship, we can now investigate the excess risk between the empirical continuous optimal solution and the population-level continuous optimal solution by bounding the term $F(\check{U}) - F(\widetilde{U}^*)$. The following is the first theorem of *relaxed* NCut.

Theorem 3. *Under Assumption 1, suppose for any $\check{u} \in \mathcal{H}$ such that $\|\check{u}\|_\infty \leq \sqrt{B}$ with an absolute constant $B > 0$, then for any $\delta > 0$, with probability at least $1 - 2\delta$, we have*

$$\begin{aligned} &F(\check{U}) - F(\widetilde{U}^*) \\ \leq &8CBK\left(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + KC\sqrt{\frac{\log\frac{2}{\delta}}{n}}. \end{aligned}$$

where C is a positive constant, and where K is the clustering number.

Remark 4. Theorem 3 suggests that the excess risk of *relaxed* NCut has a convergence rate of the order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. The proof techniques used in Theorem 3 conclude spectral properties of integral operators, operator theory, and empirical processes. C in Theorem 3 comes from Eq. (6). We provide the proof of Theorem 3 in the Appendix. Moreover, the generalization error bound of *relaxed* NCut is shown below.

Corollary 2. *Under the above assumptions, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\hat{F}(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*) \leq KC\sqrt{\frac{\log\frac{2}{\delta}}{n}},$$

where C is a positive constant, and where K is the clustering number.

Remark 5. For the lower bound of *relaxed* NCut, Dhillon, Guan, and Kulis (2007) have constructed the relationship between NCut and the weight kernel k-means. The lower bound of k-means is of order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ (Bartlett, Linder, and Lugosi 1998). Building a connection between *relaxed* NCut and weight kernel k-means is probably a method to investigate the lower bound of *relaxed* NCut, whose lower bound may be of the order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ as well.

As discussed before, the continuous solution of spectral clustering typically involves a discretization process, thus we then study the excess risk bound between the empirical discrete optimal solution and the population-level discrete optimal solution for *relaxed* NCut. In analogy to the previous subsection, we investigate $F(\ddot{U}) - F(U^*)$, where $\ddot{U} = (\ddot{u}_1, ..., \ddot{u}_K)$ are K functions in RKHS $\mathcal{H}$ and are sought through $\mathcal{H}$ by the continuous eigenfunctions $\breve{U}$, and where $U^* = (u_1^*, ..., u_K^*)$ is the optimal solution of the minimal population-level error of NCut, i.e., optimal solution of the population-level version of Eq. (3).

Theorem 4. *Denoted by $\epsilon := \sum_{k=1}^{K} \|\ddot{u}_k - \breve{u}_k\|_2$. Under Assumption 1, suppose for any $\breve{u} \in \mathcal{H}$ such that $\|\breve{u}\|_\infty \leq \sqrt{B}$ with an absolute constant $B > 0$, then for any $\delta > 0$, with probability at least $1 - 2\delta$, we have*

$$\begin{aligned} &F(\ddot{U}) - F(U^*) \\ \leq &4C\epsilon + 8CBK\left(\sqrt{\frac{1}{n}} + 2\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\right) + KC\sqrt{\frac{\log\frac{2}{\delta}}{n}}. \end{aligned}$$

where C is a positive constant, and where K is the clustering number.

Remark 6. From Theorem 4, one can see that the term $\sum_{k=1}^{K} \|\ddot{u}_k - \breve{u}_k\|_2$ is also a fundamental quantity in influencing the excess risk of *relaxed* NCut between the empirical discrete optimal solution and the population-level discrete optimal solution, which motivates us to propose algorithms in the next section to penalize this term to make the risk bound as small as possible. In addition to the difficulties mentioned above, proving excess risk bounds also has the following difficulties: (1) The objective function of spectral clustering (see Eq (1)) is a pairwise function, which can not be written as a summation of independent and identically distributed (i.i.d.) random variables so that the standard techniques in the i.i.d. case can not apply to it. We use the U-process technique introduced in (Clémençon et al. 2008) to overcome this difficulty. (2) The operator $\mathbb{L}$ involves a division by a function, thus the term $\mathcal{C}$ can not be bounded directly by the proof technique of Theorem 2. We must introduce equivalent probability measures to construct equivalent vector space (please refer to the proof).

Remark 7. A positive point of Theorems 1-4 is that these bounds are dimension-independent, which allows the results to be applicable to high-dimensional problems. Although dimension-independent excess risk bounds have been given for some clustering algorithms, e.g., k-means (Biau, Devroye, and Lugosi 2008; Liu 2021) and kernel-NCut (Terada and Yamamoto 2019), the results of this paper are novel for spectral clustering.

Remark 8. This remark discusses why we use the asymmetric normalized Laplacian, not the symmetric normalized Laplacian. Using the asymmetric normalized graph Laplacian, we can analyze *relaxed* NCut in a unified form of the empirical error (i.e., Eq. (1)). While for the normalized symmetric Laplacian, we need to transform Eq. (1) to

$$\hat{F}(\mathbf{U}) := \frac{1}{2n(n-1)} \sum_{k=1}^{K} \sum_{i,j=1, i\neq j}^{n} \mathbf{W}_{i,j} \left(\frac{\mathbf{u}_{k,i}}{\sqrt{d_i}} - \frac{\mathbf{u}_{k,j}}{\sqrt{d_j}}\right)^2.$$

Please refer to Proposition 3 and Eq. (11) in (Von Luxburg 2007) for details.

Remark 9. This remark discusses the relationship between this paper and (Li and Liu 2021). Li and Liu (2021) study the clustering algorithm through a general framework and then give excess risk bounds based on this framework. Specifically, the excess risk in (Li and Liu 2021) is of the form $F(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)$. However, we have discussed that $F(\widetilde{\mathbf{U}}^*)$ is impossible to be calculated for spectral clustering due to the dimensional issue. Thus, the bounds of $F(\widetilde{\mathbf{U}}^*) - F(\widetilde{U}^*)$ established in (Li and Liu 2021) do not hold for the specific spectral clustering problem, and that's also the reason why we introduce the integral operator tool to revisit the spectral clustering problem. Hence, we highlight that the results of this paper, both the bounds and the algorithms, are novel compared to (Li and Liu 2021).

Algorithms

According to Theorems 2 and 4, the imperative is to penalize $\sum_{k=1}^{K} \|\ddot{u}_k - \breve{u}_k\|_2$ to make it as small as possible. Towards this aim, we should solve the following optimization objective to find the optimal discrete solution $\ddot{U}$:

$$\begin{aligned} \ddot{U} := &\mathop{\arg\min}_{U=(u_1,...,u_K)} \sum_{k=1}^{K} \|u_k - \breve{u}_k\|_2 \\ &\text{s.t. } u_k(x) \in \{0, 1\}, \end{aligned} \tag{11}$$

where $U = (u_1, ..., u_K)$ is any set of K functions in RKHS $\mathcal{H}$. In empirical clustering process, (11) implies that we should optimize this term $\sum_{k=1}^{K} \|\ddot{\mathbf{u}}_k - \widetilde{\mathbf{u}}_k^*\|_2$. It can be roughly equivalent to optimize $\|\ddot{\mathbf{U}} - \widetilde{\mathbf{U}}^*\|_F$, to find the optimal discrete solution $\ddot{\mathbf{U}} = (\ddot{\mathbf{u}}_1, ..., \ddot{\mathbf{u}}_K)$, where F denotes the Frobenius norm. Stella and Shi (2003) propose an iterative fashion to optimize $\|\ddot{\mathbf{U}} - \widetilde{\mathbf{U}}^*\|_F$ to get the discrete solution closest to the continuous optimal solution $\widetilde{\mathbf{U}}^*$. At a high level, this paper provides a theoretical explanation on (Stella and Shi 2003) from the population view.

The key intuition in (Stella and Shi 2003) is that the continuous optimal solutions consist of not only the eigenvectors but of a whole family spanned by the eigenvectors through orthonormal transform. Thus the discrete optimal solution can be searched by orthonormal transform. With this idea, we can solve the following optimization objective to find the optimal discrete solution $\ddot{\mathbf{U}}$ and orthonormal

transform:

$$(\ddot{\mathbf{U}}, \mathbf{R}^*) := \arg\min_{\mathbf{U},\mathbf{R}} \|\mathbf{U} - \widetilde{\mathbf{U}}^*\mathbf{R}\|$$
$$\text{s.t.} \quad \mathbf{U} \in \{0,1\}^{n\times K}, \mathbf{U}\mathbf{1}_K = \mathbf{1}_n, \mathbf{R}\mathbf{R}^T = \mathbf{I}_K,$$

where $\mathbf{1}_n$ is a vector with all one elements, $\mathbf{U}$ is any set of K discrete vectors in the eigenspace, and $\mathbf{R} \in \mathbb{R}^{K\times K}$ is an orthonormal matrix. The orthonormal transform program finds the optimal discrete solution in an iterative fashion. This iterative fashion is shown below:
(1) given $\mathbf{R}^*$, solving the following optimization objective:

$$\arg\min_{\mathbf{U}} \|\mathbf{U} - \widetilde{\mathbf{U}}^*\mathbf{R}^*\|,$$
$$\text{s.t. } \mathbf{U} \in \{0,1\}^{n\times K}, \mathbf{U}\mathbf{1}_K = \mathbf{1}_n.$$

(2) given $\ddot{\mathbf{U}}$, solving the following optimization objective:

$$\arg\min_{\mathbf{R}} \|\ddot{\mathbf{U}} - \widetilde{\mathbf{U}}^*\mathbf{R}\|,$$
$$\text{s.t. } \mathbf{R}\mathbf{R}^T = \mathbf{I}_K.$$

We denote this iterative fashion in (Stella and Shi 2003) as POD (Program of Optimal Discretization). The pseudocode of POD is shown in Algorithm 3 in the Appendix.

GPOD

We now introduce our proposed algorithms, called *Generalized* POD (GPOD) algorithm, which can not only penalize the fundamental quantity in influencing the excess risk of the discrete solution but also allow clustering the unseen data points.

Firstly, for the samples $\mathbf{X}$, we can use the eigenvectors $\widetilde{\mathbf{U}}^*$ of $\mathbf{L}$ (or $\mathbf{L}_{rw}$) to obtain its extensions based on Eq. (9) (or Eq. (10)), that is to obtain the eigenfunctions $\breve{U}$ of T_n (or $\mathbb{L}_n$). Secondly, when the new data points $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_m\}$ come, we can calculate its eigenvectors $\bar{\mathbf{U}} = \{\bar{\mathbf{u}}_1, ..., \bar{\mathbf{u}}_K\} \in \mathbb{R}^{m\times K}$ with the help of the eigenfunctions $\breve{U} = (\breve{u}_1, ..., \breve{u}_K)$. By mapping the eigenfunctions into finite dimensional space, we can approximately obtain the eigenvectors of the new samples $\bar{\mathbf{X}}$. Specifically, we can use formula

$$\bar{\mathbf{u}}_k = \frac{1}{\sqrt{\lambda_k}}(\breve{u}_k(\bar{\mathbf{x}}_1), ..., \breve{u}_k(\bar{\mathbf{x}}_m))$$

to obtain the eigenvectors of $\bar{\mathbf{X}}$ for *relaxed* RatioCut and use

$$\bar{\mathbf{u}}_k = (\breve{u}_k(\bar{\mathbf{x}}_1), ..., \breve{u}_k(\bar{\mathbf{x}}_m)$$

for *relaxed* NCut. Note that for *relaxed* RatioCut, since the underlying ρ is unknown, the term $L(x, \mathbf{x}_i)$ can be empirically approximated by

$$\frac{1}{n}\sum_{i=1}^{n} W(\cdot, \mathbf{x}_i) - W(\cdot, \mathbf{x}_i).$$

After obtaining the eigenvectors of the out-of-sample data points $\bar{\mathbf{X}}$, we can use the POD iterative fashion to optimize the following optimization problem to seek the empirical optimal discrete solution:

$$(\ddot{\mathbf{U}}, \mathbf{R}^*) := \arg\min_{\mathbf{U},\mathbf{R}} \|\mathbf{U} - \bar{\mathbf{U}}\mathbf{R}\|$$
$$\text{s.t.} \quad \mathbf{U} \in \{0,1\}^{m\times K}, \mathbf{U}\mathbf{1}_K = \mathbf{1}_m, \mathbf{R}\mathbf{R}^T = \mathbf{I}_K.$$

This optimization process can penalize the fundamental quantity for the out-of-sample data points.

The ability of our proposed algorithm in clustering unseen data points without the eigendecomposition on the overall data points makes the spectral clustering more applicable, largely reducing the time complexity. The concrete algorithm steps are presented in the Appendix, where we also analyze how the time complexity of our proposed algorithm is significantly improved in Remark 1. Overall, the proposed algorithms can not only penalize the fundamental quantity but also cluster the out-of-sample data points.

Remark 10. Eqs. (9) and (10) hold when the denominator is not 0. This remark discusses the case when the denominator is 0, i.e., the 0 or 1 eigenvalue. According to the spectral projection view, for the unnormalized Laplacian, respectively the asymmetric graph Laplacian, the 0-eigenvalue, respectively the 1 eigenvalue, doesn't affect the performance of spectral clustering, see Proposition 9 and Proposition 14 in (Rosasco, Belkin, and De Vito 2010), respectively. Thus, the 0 or 1 eigenvalue doesn't influence the performance of GPOD in clustering the out-of-sample data.

Remark 11. The excess risk bounds and algorithms provided in this paper are fundamental, thus it is applicable to other variants of spectral clustering, e.g., multiview spectral clustering (Yang et al. 2022b,a). We leave it to the interested readers.

Numerical Experiments

We have made numerical experiments on both toy and real datasets for the two proposed algorithms. Considering the length limit, we leave the experimental settings and results in the Appendix. The experimental results show that the proposed algorithms can cluster the out-of-sample data points, verifying their effectiveness.

Conclusions

In this paper, we investigate the generalization performance of popular spectral clustering algorithms: *relaxed* RatioCut and *relaxed* Ncut, and provide the excess risk bounds. According to the two steps of practical spectral clustering algorithms, we first provide a convergence rate of the order $\mathcal{O}(1/\sqrt{n})$ for the continuous solution for both *relaxed* RatioCut and *relaxed* Ncut. We then show the fundamental quantity in influencing the excess risk of the discrete solution. Theoretical analysis inspires us to propose two novel algorithms that can not only cluster the out-of-sample data, largely reducing the time complexity, but also penalize this fundamental quantity to be as small as possible. By numerical experiments, we verify the effectiveness of the proposed algorithms. One limitation of this paper is that we don't provide a true convergence rate for the excess risk of the empirical discrete solution. We believe that this problem is pretty important and worthy of further study.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (No. 62076234, No.61703396, No. 62106257), the Beijing Outstanding Young Scientist Program NO.BJJWZYJH012019100020098, the Beijing Natural Science Foundation (No. 4222029), Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" initiative, Renmin University of China, China Unicom Innovation Ecological Cooperation Plan, and Public Computing Cloud of Renmin University of China.

References

Alpert, C. J.; and Kahng, A. B. 1995. Multiway partitioning via geometric embeddings, orderings, and dynamic programming. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, 14(11): 1342–1358.

Arias-Castro, E.; Pelletier, B.; and Pudlo, P. 2012. The normalized graph cut and Cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44(4): 907–937.

Bartlett, P. L.; Linder, T.; and Lugosi, G. 1998. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5): 1802–1813.

Biau, G.; Devroye, L.; and Lugosi, G. 2008. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 54(2): 781–790.

Cao, Y.; and Chen, D.-R. 2011. Consistency of regularized spectral clustering. *Applied and Computational Harmonic Analysis*, 30(3): 319–336.

Clémençon, S.; Lugosi, G.; Vayatis, N.; et al. 2008. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2): 844–874.

Dhillon, I. S. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 269–274.

Dhillon, I. S.; Guan, Y.; and Kulis, B. 2007. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(11): 1944–1957.

Fiedler, M. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2): 298–305.

Kannan, R.; Vempala, S.; and Vetta, A. 2004. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3): 497–515.

Li, S.; and Liu, Y. 2021. Sharper generalization bounds for clustering. In *International Conference on Machine Learning*, 6392–6402.

Liu, F.; Choi, D.; Xie, L.; and Roeder, K. 2018. Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5): 927–932.

Liu, Y. 2021. Refined Learning Bounds for Kernel and Approximate k-Means. In *Advances in Neural Information Processing Systems*.

Ng, A.; Jordan, M.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, 849–856.

Pelletier, B.; and Pudlo, P. 2011. Operator norm convergence of spectral clustering on level sets. *The Journal of Machine Learning Research*, 12: 385–416.

Rohe, K.; Chatterjee, S.; Yu, B.; et al. 2011. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4): 1878–1915.

Rosasco, L.; Belkin, M.; and De Vito, E. 2010. On learning with integral operators. *Journal of Machine Learning Research*, 11(2).

Schiebinger, G.; Wainwright, M. J.; Yu, B.; et al. 2015. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2): 819–846.

Shaham, U.; Stanton, K.; Li, H.; Nadler, B.; Basri, R.; and Kluger, Y. 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8): 888–905.

Singer, A.; and Wu, H.-T. 2017. Spectral convergence of the connection laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1): 58–123.

Stella, X. Y.; and Shi, J. 2003. Multiclass spectral clustering. In *Computer Vision, IEEE International Conference on*, 313.

Terada, Y.; and Yamamoto, M. 2019. Kernel Normalized Cut: a Theoretical Revisit. In *International Conference on Machine Learning*, 6206–6214.

Ting, D.; Huang, L.; and Jordan, M. 2011. An analysis of the convergence of graph Laplacians. *arXiv preprint arXiv:1101.5435*.

Trillos, N. G.; and Slepčev, D. 2018. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2): 239–281.

Trillos, N. G.; Slepčev, D.; Von Brecht, J.; Laurent, T.; and Bresson, X. 2016. Consistency of Cheeger and ratio graph cuts. *The Journal of Machine Learning Research*, 17(1): 6268–6313.

Vapnik, V. 1999. *The nature of statistical learning theory*. Springer science & business media.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416.

Von Luxburg, U.; Belkin, M.; and Bousquet, O. 2008. Consistency of spectral clustering. *The Annals of Statistics*, 555–586.

Von Luxburg, U.; Bousquet, O.; and Belkin, M. 2004. On the convergence of spectral clustering on random samples: the normalized case. In *International Conference on Computational Learning Theory*, 457–471.

Yang, B.; Zhang, X.; Nie, F.; and Wang, F. 2022a. Fast Multi-view Clustering with Spectral Embedding. *IEEE Transactions on Image Processing*.

Yang, H.; Gao, Q.; Xia, W.; Yang, M.; and Gao, X. 2022b. Multi-view Spectral Clustering with Bipartite Graph. *IEEE Transactions on Image Processing*.