Semidefinite spectral clustering

Jaehwan Kim [a], Seungjin Choi [b,*]

[a] *Digital Content Research Division,*
Electronics and Telecommunications Research Institute,
161 Gajeong-dong, Yusung-gu, Daejeon 305-350, Republic of Korea

[b] *Department of Computer Science,*
Pohang University of Science and Technology,
San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Republic Korea

Abstract

Multi-way partitioning of an undirected weighted graph where pairwise similarities are assigned as edge weights, provides an important tool for data clustering, but is an NP-hard problem. Spectral relaxation is a popular way of relaxation, leading to *spectral clustering* where the clustering is performed by the eigen-decomposition of the (normalized) graph Laplacian. On the other hand, *semidefinite relaxation*, is an alternative way of relaxing a combinatorial optimization, leading to a convex optimization. In this paper we employ a semidefinite programming (SDP) approach to the graph equipartitioning for clustering, where sufficient conditions for strong duality hold. The method is referred to as *semidefinite spectral clustering*, where the clustering is based on the eigen-decomposition of the optimal feasible matrix computed by SDP. Numerical experiments with several data sets, demonstrate the useful behavior of our semidefinite spectral clustering, compared to existing spectral clustering methods.

Key words: Clustering, Convex optimization, Multi-way graph equipartitioning, Semidefinite programming, Spectral clustering.

* Corresponding author. Tel.: +82-54-279-2259; Fax: +82-54-279-2299
Email: jh.kim@etri.re.kr (J. Kim), seungjin@postech.ac.kr (S. Choi)
URL: http://www.postech.ac.kr/∼seungjin (S. Choi)

1 Introduction

Partitioning a graph with its edges assigned by pairwise similarities, serves as an important tool for data clustering which is a common goal of considerable research in machine learning, data mining, and pattern recognition communities. In general, the graph partitioning is an NP-hard problem. Spectral relaxation is a popular way of solving the graph partitioning problem, which leads to spectral clustering methods. These methods use eigenvectors of an affinity (similarity) matrix derived from the data, in order to cluster data points. Successful applications of spectral clustering methods are found in image segmentation, perceptual grouping, circuit layout design, biological sequence clustering, document clustering and so on [1, 2, 3, 4, 5]. However, despite its popularity, these methods are not directly related to an optimization and are relied on some heuristics so that it is difficult to obtain a globally optimal clustering result.

Formulating a combinatorial problem as a convex optimization, if possible, provides a useful relaxation such that the original combinatorial problem is solved. The convex optimization consisting of a linear objective function and generalized inequalities, is known as *cone program*. If the generalized equalities are defined over positive semidefinite cones, the associated cone program becomes *semidefinite program*. The semidefinite program (SDP) as well as linear program (LP) can be viewed as special cases of cone programs. The semidefinite relaxation (based on SDP) provides an useful way to relax combinatorial optimization problems such as max-cut [6], assignment and graph partitioning problem [7, 8], graph coloring and has shown to provide tighter relaxation bounds. In contrast to the spectral relaxation, the semidefinite relaxation leads to convex optimization and its optimal solution can be obtained by semidefinite programming.

In this paper we employ the semidefinite relaxation method in [8] where strong duality and Slater's condition qualification are met, and present a semidefinite spectral clustering method where the eigenvectors [1] of the optimal feasible matrix (determined by SDP) are used, whereas spectral clustering methods use the eigenvectors of the graph Laplacian (or normalized Laplacian matrix). We pay our attention to the equipartitioning, although numerical experiments were carried out for both balanced and unbalanced cases. Empirically we show that the optimal feasible matrix better suits for spectral clustering, compared to the Laplacian matrix.

[1] Strictly speaking, eigenvectors scaled by the singular values, obtained from the sum of diagonal blocks of optimal feasible matrix, are used.

2 Graph partitioning

We consider a connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ and $\mathcal{E}$ denote a set of vertices and a set of edges, respectively, with pairwise similarity values being assigned as edge weights. Denote by W_{ij} an edge weight between the ith and jth vertices (corresponding the similarity value between the ith and jth data points). We briefly overview a problem of graph partitioning problem, since it provides a basic method for our semidefinite spectral clustering method. This section starts with the classical binary partitioning (bipartitioning) and presents a cut criterion for multi-way graph equipartitioning.

2.1 Classical graph bipartitioning

Two-way graph partitioning (bipartitioning) involves taking the set $\mathcal{V}$ apart into two coherent groups, $\mathcal{S}_1$ and $\mathcal{S}_2$, satisfying $\mathcal{V} = \mathcal{S}_1 \cup \mathcal{S}_2$, ($|\mathcal{V}| = n$), and $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, by simply cutting edges connecting the two parts. The degree of dissimilarity between $\mathcal{S}_1$ and $\mathcal{S}_2$ can be computed by the total weights of edges that have been removed. The real-valued adjacency matrix, $W = [W_{ij}] \in \mathbb{R}^{n \times n}$, is symmetric and is also referred to as *affinity* or *similarity* matrix. The degree of node i is defined by $d_i = \sum_{j \in \mathcal{V}} W_{ij}$ and the diagonal matrix $D \in \mathbb{R}^{n \times n}$ consists of diagonal elements d_i. The volume of a set $\mathcal{S}_1$, $\mathrm{vol}(\mathcal{S}_1)$, is computed as $\mathrm{vol}(\mathcal{S}_1) = \sum_{i \in \mathcal{S}_1} d_i$.

In a graph theoretic term, the cut criterion is given by

$$\begin{aligned}
\mathrm{Cut}(\mathcal{S}_1, \mathcal{S}_2) &= \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_2} W_{ij} \\
&= \frac{1}{2} \left\{ \sum_{i \in \mathcal{S}_1} d_i + \sum_{j \in \mathcal{S}_2} d_j - \sum_{i \in \mathcal{S}_1} \sum_{j \in \mathcal{S}_1} W_{ij} - \sum_{i \in \mathcal{S}_2} \sum_{j \in \mathcal{S}_2} W_{ij} \right\} \\
&= \frac{1}{2} \left\{ \vec{x}_1^{\top} (D - W) \vec{x}_1 + \vec{x}_2^{\top} (D - W) \vec{x}_2 \right\} \\
&= \frac{1}{2} \left\{ \sum_{i=1}^{2} \vec{x}_i^{\top} L \vec{x}_i \right\}, \qquad (1)
\end{aligned}$$

where $L = D - W$ is the graph Laplacian (or Laplacian matrix) and $\vec{x}_j = [x_{1j} \cdots x_{nj}]^{\top} \in \mathbb{R}^n$ is the indicator vector which represents partitions,

$$x_{ij} = \begin{cases} +1, & \text{if } i \in \mathcal{S}_j \\ 0, & \text{if } i \notin \mathcal{S}_j \end{cases}, \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, 2. \qquad (2)$$

Note that $\vec{x}_1$ and $\vec{x}_2$ are orthogonal, i.e., $\vec{x}_1^\top \vec{x}_2 = 0$.

Introducing a bipolar indicator vector, $\vec{q} = \vec{x}_1 - \vec{x}_2 \in \{+1, -1\}^n$, then the indicator vectors $\vec{x}_1$ and $\vec{x}_2$ can be written as

$$\vec{x}_1 = \frac{1+\vec{q}}{2},$$
$$\vec{x}_2 = \frac{1-\vec{q}}{2}.$$

Using these relations, the cut criterion (1) is simplified as

$$\mathrm{Cut}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{4}\vec{q}^\top L \vec{q}. \tag{3}$$

The optimal binary partitioning of a graph involves minimizing the cut criterion (3). Minimizing the cut criterion (3), is likely to produce extremely unbalanced partition, $\vec{e}_n = [1 \cdots 1]^\top \in \mathbb{R}^n$ which is the eigenvector of the Laplacian matrix L associated with eigenvalue 0. A way to avoid this in the conventional graph partitioning, is to add a constraint, $\vec{e}_n^\top \vec{q} = 0$, leading to the following combinatorial problem:

$$\begin{aligned} &\arg\min_{\vec{q}} \vec{q}^\top L \vec{q}, \\ &\text{subject to } \vec{e}_n^\top \vec{q} = 0,\ \vec{q} \in \{+1, -1\}^n. \end{aligned} \tag{4}$$

Dropping the integer constraint in (4), the constraint $\vec{e}_n^\top \vec{q} = 0$ forces the solution to be orthogonal to $\vec{e}_n$. Then the problem reduces to the symmetric eigenvalue problem and the solution is the second smallest eigenvector of L which is known as the *Fiedler vector* [9]. The integer constraint is taken into account by rounding the eigenvector by a suitably chosen threshold value. A way of relaxing the combinatorial problem (4) by dropping the integer constraint, is known as *spectral relaxation*, which produced spectral clustering methods [2, 1, 3, 4, 10], where normalized Laplacian matrix was considered.

2.2 Multi-way graph equipartitioning

The multi-way graph equipartioning aims at taking the set $\mathcal{V}$ apart into k disjoint groups, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$, ($|\mathcal{S}_i| = m$, $i = 1, 2, \ldots, k$), satisfying $\mathcal{V} = \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots \cup \mathcal{S}_k$, ($|\mathcal{V}| = n$, i.e., $n = m \times k$) and $\mathcal{S}_1 \cap \mathcal{S}_2 \cap \cdots \cap \mathcal{S}_k = \emptyset$. The

cut criterion for the multi-way equipartioning, 'MECut', is a straightforward extension of the bipartioning criterion (1),

$$\begin{aligned}\mathrm{MECut}(\mathcal{S}_1,\ldots,\mathcal{S}_k) &= \frac{1}{2}\sum_{i=1}^{k}\vec{e}_m^{\top}W_{\mathcal{S}_i\overline{\mathcal{S}}_i}\vec{e}_m,\\ &= \frac{1}{2}\sum_{i=1}^{k}\vec{e}_m^{\top}\left(D_{\mathcal{S}_i}-W_{\mathcal{S}_i\mathcal{S}_i}\right)\ \vec{e}_m\\ &= \frac{1}{2}\sum_{i=1}^{k}\vec{x}_i^{\top}(D-W)\vec{x}_i. \end{aligned} \tag{5}$$

where $\vec{e}_m = [1\ \cdots\ 1]^{\top} \in \mathbb{R}^m$, $W_{\mathcal{S}_p\mathcal{S}_q}$ is the adjacency matrix for nodes belonging to $\mathcal{S}_p$ and $\mathcal{S}_q$, i.e., $W_{\mathcal{S}_p\mathcal{S}_q} = [W_{ij}]_{i\in\mathcal{S}_p, j\in\mathcal{S}_q} \in \mathbb{R}^{m\times m}$, and $D_{\mathcal{S}_i}$ is the degree matrix of nodes in $\mathcal{S}_i$, $D_{\mathcal{S}_i} = \sum_{j=1}^{k} W_{\mathcal{S}_i\mathcal{S}_j} = \vec{e}_m^{\top}(W_{\mathcal{S}_i\mathcal{V}})\vec{e}_m$. The affinity matrix W and the degree matrix D (block diagonal matrix) have the form

$$W = \begin{bmatrix} W_{\mathcal{S}_1\mathcal{S}_1} & W_{\mathcal{S}_1\mathcal{S}_2} & \cdots & W_{\mathcal{S}_1\mathcal{S}_k}\\ W_{\mathcal{S}_2\mathcal{S}_1} & W_{\mathcal{S}_2\mathcal{S}_2} & \cdots & W_{\mathcal{S}_2\mathcal{S}_k}\\ \vdots & \vdots & \ddots & \vdots\\ W_{\mathcal{S}_k\mathcal{S}_1} & W_{\mathcal{S}_k\mathcal{S}_2} & \cdots & W_{\mathcal{S}_k\mathcal{S}_k}\end{bmatrix}, \tag{6}$$

and

$$D = \begin{bmatrix} D_{\mathcal{S}_1} & 0 & \cdots & 0\\ 0 & D_{\mathcal{S}_2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & D_{\mathcal{S}_k}\end{bmatrix}. \tag{7}$$

Note that $\vec{x}_i$ are orthogonal, i.e., $\vec{x}_i^{\top}\vec{x}_j = 0$ for $i \neq j$. One can easily see that the MECut criterion (5) for the case of $k = 2$, becomes equivalent to the bipartitioning criterion (1).

Define a partition matrix $X = [\vec{x}_1\ \cdots\ \vec{x}_k] \in \mathbb{R}^{n\times k}$, collecting indicator vectors $\vec{x}_i$. With this partition matrix, the MECut criterion can be written in a compact form:

$$\arg\min_{X} \mathrm{tr}(X^{\top}LX), \tag{8}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator.

The minimum is achieved when X is taken to be any orthogonal basis for the subspace spanned by the eigenvalues corresponding to the k smallest eigenvalues of L. Note that the MECut problem (8) is NP-hard as well as non-convex problem such as (4). Spectral relaxation has been widely used to solve this difficult problems, for example see [1, 3, 4]. Next section explains a semidefinite relaxation method, which was recently employed in a problem of bipartitioning [11]. Motivated by existing work [11, 4, 6, 8, 7], we present a semidefinite programming method for clustering where we use the eigen-decomposition of the optimal feasible matrix resulting from a semidefinite relexation of the multi-way graph equipartitioning.

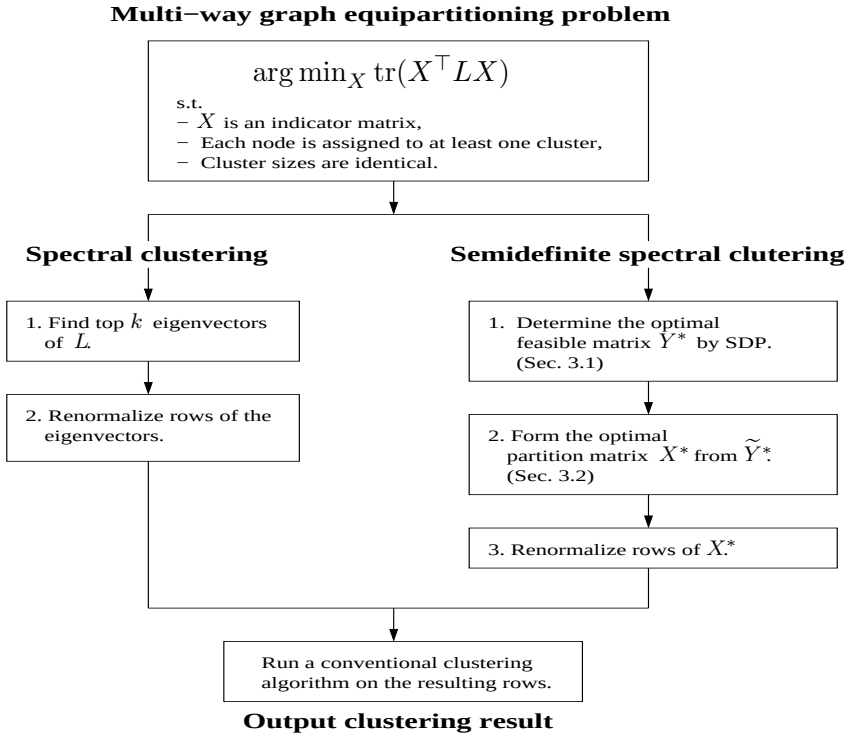

Fig. 1. A schematic diagram for the (standard) spectral clustering and the semidefinite spectral clustering.

3 The proposed method

The schematic diagram (see Fig. 1) summarizes the standard spectral clustering as well as our semidefinite spectral clustering, where one can easily find a distinction between them. Both methods are based on the graph partitioning, however, relaxation methods are different. In the spectral clustering, eigenvectors of the (normalized) graph Laplacian are used for grouping data points into cohererent clusters. In contrast, the semidefinite spectral clustering first

finds the optimal feasible matrix through the projected semidefinite relaxation. The clustering is carried out using the eigenvectors of the optimal feasible matrix. Subsequent two sections will illustrate details on the semidefinite spectral clustering.

3.1 Projected semidefinite relaxation

We employ the semidefinite relaxation method, in order to solve the problem in (8), following the main idea in [7, 8] where the method was developed for the quadratic assignment problem. The semidefinite relaxation allows us to formulate the non-convex combinatorial problem (8) as a convex optimization problem. Here we sketch the outline of the semidefinite relaxation to the problem (8) and more details are left in Appendix.

For semidefinite relaxation, we first write (5) as a constrained quadratic programming, consisting of a quadratic objective function as well as inequality or equality constraints,

$$\begin{aligned} \arg\min_{X} \quad & \operatorname{tr}(X^\top L X) \\ \text{s.t.} \quad & X \circ X = X \quad \Longleftrightarrow \{X : x_{ij}^2 = x_{ij},\, \forall i, j\}, \\ & \left.\begin{aligned} & X\vec{e}_k = \vec{e}_n \\ & X^\top \vec{e}_n = m\vec{e}_k \end{aligned}\right\} \Longleftrightarrow \{X : \|X\vec{e}_k - \vec{e}_n\|^2 + \|X^\top \vec{e}_n - m\vec{e}_k\|^2 = 0\}, \end{aligned} \tag{9}$$

where the operator $\circ$ denotes the Hadamard product (element-wise product). The first constraint indicates that the partition matrix X is an indicator matrix. Second constraints reflect that each node (or data point) should be assigned to one of disjoint groups with identical size m.

Following the results in [8, 7], we relax the problem (9) to a standard SDP that is of the form

$$\begin{aligned} \pi_r^s := \min_{Y} \; & \operatorname{tr}(L_e Y) \\ \text{s.t.} \; & \operatorname{diag}(Y) = [1, Y_{1,2:nk+1}]^\top \in \mathbb{R}^{nk+1} \\ & \operatorname{tr}(CY) = 0, \\ & Y \succeq 0, \end{aligned} \tag{10}$$

where $Y \in \mathcal{S}_+^{nk+1}$ is the dual variable matrix, where $\mathcal{S}_+^{nk+1}$ is the *self-dual positive semidefinite cone*, i.e., $\mathcal{S}_+^{nk+1} = \{Y \mid \operatorname{tr}(ZY) \geq 0 \text{ for all } Z \in \mathcal{S}_+^{nk+1}\}$, and

$$L_e = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & I_k \otimes L \end{array}\right],$$

$$C = \left[\begin{array}{c|c} n + m^2 \vec{e}_k^\top \vec{e}_k & -\vec{e}_k^\top \otimes \vec{e}_n^\top - m\vec{e}_k^\top \otimes \vec{e}_n^\top \\ \hline -\vec{e}_k \otimes \vec{e}_n - m\vec{e}_k \otimes \vec{e}_n & (\vec{e}_k \vec{e}_k^\top) \otimes I_n + I_k \otimes (\vec{e}_n \vec{e}_n^\top) \end{array}\right],$$

where $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, $\vec{e}_n = [1 \cdots 1]^\top \in \mathbb{R}^n$, and the operator $\otimes$ denotes the Kronecker product. The detailed illustration on the SDP relaxation (10) is given in Appendix A.

Note that C is positive definite, implying that Y is singular in order to satisfy the constraint, $\mathrm{tr}(CY) = 0$. This means that the feasible set of the semidefinite relaxation (10) is not strictly feasible. In other words, Slater's constraint qualification for the problem (10) fails, since Y is singular. Hence, a direct application of an interior-point method is not possible to solve (10). In order to overcome this problem, we exploit the geometrical structure of the feasible set $\mathcal{F}$ of the semidefinite relaxation (10) and develop a barycenter-based method which is a slight modification of existing methods in [8, 7]. The basic idea of the projected semidefinite relaxation, is to project the semidefinite relaxation (10) onto the minimal face of the feasible set $\mathcal{F}$.

We define $\vec{x} = \mathrm{vec}(X)$ where $\mathrm{vec}(\cdot)$ is the *vec-function* which stacks the columns of the given matrix into one long vector. The matrices $[1, \vec{x}^\top]^\top [1, \vec{x}^\top]$ for $X \in \Pi_x$ are feasible points of $\mathcal{F}$, where Π_x is a set which contains every possible partitions. Moreover, these rank-one matrices are contained in the set of extreme points of $\mathcal{F}$ [12]. It is known that if an optimal solution for convex optimization problem is unique, it must be an extreme point of the feasible region. There may be infinitely many faces containing such extreme points. Therefore, we need to characterize the minimal face including all of these extreme points through the *barycenter* of the convex hull of the partition matrices.

The barycenter is the point representing the mean position of any mass. The *barycenter*-based method was first introduced in [8, 7] which well fit in our problem. We slightly modify the *barycenter* introduced in [8, 7], in order to match the multi-way equipartitioning constraints. The our modified *barycenter* is the value with its *barycentric coordinates* as the number of ways of partitioning a set of n data into k equipartitions with m elements. The following theorem defines the barycenter and describe some results associated with it. This is a direct consequence of the result in [8, 7].

Theorem 1 *The barycenter $\widehat{Y}$ is defined as*

$$\widehat{Y} \doteq \frac{(m!)^k}{n!} \sum_{\Pi_x} \left[\begin{array}{c|c} 1 & \vec{x}^\top \\ \hline \vec{x} & \vec{x}\vec{x}^\top \end{array}\right], \tag{11}$$

where, for $\widehat{Y} \in \mathrm{conv}(\Pi_x)$ *where the convex hull of a set* Π_x *is denoted* $\mathrm{conv}(\Pi_x)$*, the barycentric coordinate of* $\widehat{Y}$ *with respect to all possible partition matrices is the number of ways of partitioning a set of* n *data into* k *equipartitions with* m *elements;* ($\frac{n!}{(m!)^k}$ *cases*)*. Then we have:*

1. *The barycenter has the form*

$$\widehat{Y} = \left[\begin{array}{c|ccc} 1 & \frac{m}{n}\vec{e}_n^\top & \cdots & \frac{m}{n}\vec{e}_n^\top \\ \hline \frac{m}{n}\vec{e}_n & \left\{\frac{m}{n}I_n + \frac{m(m-1)}{n(n-1)}(E_n - I_n)\right\} & \cdots & \left\{\frac{m^2}{n(n-1)}\right\}(E_n - I_n) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{m}{n}\vec{e}_n & \left\{\frac{m^2}{n(n-1)}\right\}(E_n - I_n) & \cdots & \left\{\frac{m}{n}I_n + \frac{m(m-1)}{n(n-1)}\right\}(E_n - I_n) \end{array}\right],$$

where $E_n = \vec{e}_n\vec{e}_n^\top$.

2. *The columns of* $\widehat{V}$ *form basis vectors for the range space of* $\widehat{Y}$*, where*

$$\widehat{V} = \left[\begin{array}{c|c} 1 & 0 \\ \hline \frac{m}{n}\vec{e}_k \otimes \vec{e}_n & V_k \otimes V_n \end{array}\right], \quad \textit{where} \quad V_s := \left[\begin{array}{c} I_{s-1} \\ -\vec{e}_{s-1}^\top \end{array}\right] \in \mathbb{R}^{s\times(s-1)}.$$

Results in Theorem 1 can be proved in a similar way as in [8]. It was shown in [8] that points in minimal faces's relative interior, $\widehat{Y}$, can be expressed as $\widehat{V}S\widehat{V}^\top$ for $S \in \mathcal{S}_+^{(n-1)(k-1)+1}$, if $\mathcal{R}(\widehat{V}) = \mathcal{R}(\widehat{Y})$, where $\mathcal{R}(\cdot)$ represents the range space. Using these results, we project the semidefinite relaxation (A.7) onto the minimal face, replacing Y and $\mathrm{arrow}(Y)$ by $\widehat{V}S\widehat{V}^\top$ and $\mathrm{diag}(\widehat{V}S\widehat{V}^\top)$, respectively. The projected semidefinite relaxation and its dual are of the form

$$\begin{aligned} \pi_r^p := \min_S\ & \mathrm{tr}(\widehat{V}^\top L_e \widehat{V} S) \\ \text{s.t. }\ & \mathrm{diag}(\widehat{V}S\widehat{V}^\top) = [1, (1/k)\vec{e}_{nk}^\top]^\top, \\ & S \in \mathcal{S}_+^{(n-1)(k-1)+1}, \end{aligned} \tag{12}$$

$$\begin{aligned} \pi_d^p := \max_\rho\ & a^\top \rho \\ \text{s.t. }\ & H = \widehat{V}^\top(L_e - \mathrm{Diag}(\rho))\widehat{V}, \\ & H \succeq 0, \end{aligned} \tag{13}$$

where $\rho \in \mathbb{R}^{nk+1}$ is Lagrange multiplier and $a = [1, (1/k)\vec{e}_{nk}^\top]^\top$.

The primal (12) and dual (13), which were attainted by the projected semidefinite relaxation, satisfy Slater's condition, implying that the strong duality,

$\pi_r^p = \pi_d^p$, is guaranteed. In practice, the iteration runs until the duality gap falls below a pre-specified tolerance level ϵ (extremely near to zero).

We use an an interior-point method to solve these projected semidefinite relaxation problems. To this end, we consider the dual barrier problem for (13)

$$\begin{aligned} \min_{H,\rho} \; & -a^\top \rho - \mu \log\det H \\ \text{s.t. } & H = \widehat{V}^\top (L_e - \mathrm{Diag}(\rho))\widehat{V}, \\ & H \succeq 0, \end{aligned} \tag{14}$$

where $\mu > 0$. The Lagrangian with the strictly feasible matrix S is given by

$$\mathcal{L}_\mu(H, S, \rho) = -a^\top \rho - \mu \log\det H + \{H - \widehat{V}^\top (L_e - \mathrm{Diag}(\rho))\widehat{V}\} \bullet S.$$

The cost function of the barrier problem is strictly convex. Thus an unique minimizer, (H_μ, S_μ, ρ_μ), is characterized by the first order optimality conditions of the Lagrangian, given by

$$\nabla_S \mathcal{L}_\mu = H - \widehat{V}^\top (L_e - \mathrm{Diag}(\rho))\widehat{V} = 0, \tag{15}$$
$$\nabla_H \mathcal{L}_\mu = -\mu H^{-1} + S = 0, \tag{16}$$
$$\nabla_\rho \mathcal{L}_\mu = -a + \mathrm{diag}(\widehat{V} S \widehat{V}^\top) = 0. \tag{17}$$

The basic idea to find a solution of (15)-(17), is to use the Newton's method. In order to find a search direction, $(\delta H, \delta S, \delta\rho)$, we solve the following equations:

$$\delta H = -\widehat{V}^\top \mathrm{Diag}(\delta\rho)\widehat{V}, \tag{18}$$
$$(\delta H)S + H(\delta S) = -HS + \mu I, \tag{19}$$
$$\mathrm{diag}(\widehat{V}(\delta S)\widehat{V}^\top) = 0. \tag{20}$$

It follows from (18) and (19) that we have With the first and second equations, we obtain the

$$\delta S = H^{-1}\widehat{V}^\top \mathrm{Diag}(\delta\rho)\widehat{V} S - S + \mu H^{-1}. \tag{21}$$

Substituting (21) into (20) yields the search direction for ρ,

$$\delta\rho = (\bar{H} \circ \bar{S})^{-1} \mathrm{diag}(\bar{S} - \mu\bar{H}), \tag{22}$$

where

$$\bar{H} = \widehat{V} H^{-1} \widehat{V}^\top$$
$$\bar{S} = \widehat{V} S \widehat{V}^\top. \tag{23}$$

Therefore, the search direction, $(\delta H, \delta S, \delta\rho)$, is found through Eqs. (18), (21), and (22).

3.2 Semidefinite spectral clustering algorithm

In this section, we illustrate our semidefinite spectral clustering algorithm which incorporates the projected semidefinite relaxation (described in previous section) into the spectral clustering method. To this end, we first define the *optimal feasible matrix*, Y^*, as $Y^* = \widehat{V} S^* \widehat{V}^\top$ where S^* is obtained by the interior point method, through updating rules in (18), (21), and (22). Clustering involves determining the optimal partition matrix X^* from the optimal feasible matrix Y^*.

We denote by $\widetilde{Y}^*$ the sum of diagonal blocks of $Y^*_{2:nk+1,2:nk+1}$. Then, it follows from Theorem 1 that we can interpret $\widetilde{Y}^*$ as a relaxation of $X^* {X^*}^\top$. Hence, we restore the optimal partition matrix from the rank k approximation of $\widetilde{Y}^*$. Suppose that the spectral decomposition of $\widetilde{Y}^*$ is given by $\widetilde{Y}^* = U \Lambda U^\top$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k, \ldots, \lambda_n)$ is the diagonal matrix containing eigenvalues in a descending order in its diagonal entries and $U = [\vec{u}_1, \ldots, \vec{u}_k, \ldots, \vec{u}_n]$ contains the eigenvectors associated with λ_i's for $i = 1, \ldots, n$. The rank k approximation leads to

$$\widetilde{Y}^*_k = U_k \Lambda_k U_k^\top, \tag{24}$$

where U_k and Λ_k contain only first k components. Hence, the optimal partition matrix X^* is determined by the spectral decomposition of $\widetilde{Y}^*$. Our semidefinite spectral clustering algorithm is summarized below.

Algorithm outline: Semidefinite spectral clustering

---

1. Given a set of vertices (associated with data points), $\mathcal{V} = \{v_1, \ldots, v_n\}$, construct the graph Laplacian $L = W - D$ where W is the affinity matrix, defined by $W_{ij} = \exp\left\{-\frac{\|v_i - v_j\|^2}{2\sigma^2}\right\}$, and D is the diagonal matrix with its diagonal entries representing the degree of nodes, i.e., $D_{ii} = \sum_j W_{ij}$.

2. Find the optimal feasible matrix $Y^* \in \mathcal{S}_+^{nk+1}$ by solving the projected semidefinite relaxation using the interior point method.

3. Form $\widetilde{Y}^*$, defined by the sum of diagonal blocks of $Y^*_{2:nk+1,2:nk+1}$.

4. Construct the partition matrix $X^* = \left[\sqrt{\lambda_1}\vec{u}_1, \ldots, \sqrt{\lambda_k}\vec{u}_k\right]$ where $\{\vec{u}_i\}$ and $\{\lambda_i\}$ $(i = 1, \ldots, k)$ are k largest eigenvectors and eigenvalues of $\widetilde{Y}^*$.

5. Renormalize each row of X^* to obtain

$$[\widetilde{X}^*]_{ij} = \frac{X^*_{ij}}{\left\{\sum_j \left(X^*_{ij}\right)^2\right\}^{\frac{1}{2}}}.$$

6. Do grouping row vectors of $\widetilde{X}^*$ into k clusters using a conventional clustering algorithm (e.g., k-means).

7. Assign the original vertice v_i to cluster j if the ith row of $\widetilde{X}^*$ was assigned to cluster j.

---

4 Numerical experiments

In this section, we show the useful behavior of the semidefinite spectral clustering algorithm, through the empirical comparison to the spectral clustering [4, 3], with artificial data sets as well as internet newsgroup data sets. The baseline spectral clustering algorithms [4, 3] which are considered here, are referred to as 'NJWCut' following the first letter of each author and as 'MNCut(Modified Normalized Cut)', respectively. Clustering performance of spectral-related methods is directly affected from the fact whether their affinity or graph Laplacian has the well-formed block diagonal structure or not. Because only thing we use is the eigenvectors (or scaled eigenvectors) obtained from those matrices for clustering. An ideal performance of the spectral clustering methods is based on the assumption that infinitly distant subgroups yield an affinity matrix which has an exact block diagonal [4], but this is an ideal case. Therefore, it is impossible to obtain the exact block diagonal affinity or graph Laplacian in the practical clustering problems. In this paper, we try to overcome such a limitation through a barycenter-based technique developed in [8, 7]. As the above mentioned explanation, the *barycenter* encodes the information about constraints for the multi-way equipartitioning; the partition matrix is an indicator matrix, each node should be assigned to one of disjoint groups with identical size. The *barycenter* gets an affinity or graph Laplacian having almost block diagonal, which also provides an outstanding clustering result, because eigenvectors of the almost block diagonal matrix can be expected to be *piecewise-constant*.

In empirical study, we consider balanced cases as well as unbalanced data sets, although the semidefinite spectral clustering algorithm is developed with focusing on an equipartitioning problem. The algorithm was implemented with Matlab.

We consider three different artificial data sets, for which clustering results are shown in Figs. 2 and 3. The first artificial data set is 'Two Moons' data where each moon consists of 100 data points (balanced data). Fig. 2 shows the results of applying our semidefinite spectral clustering algorithm to Two Moons data.

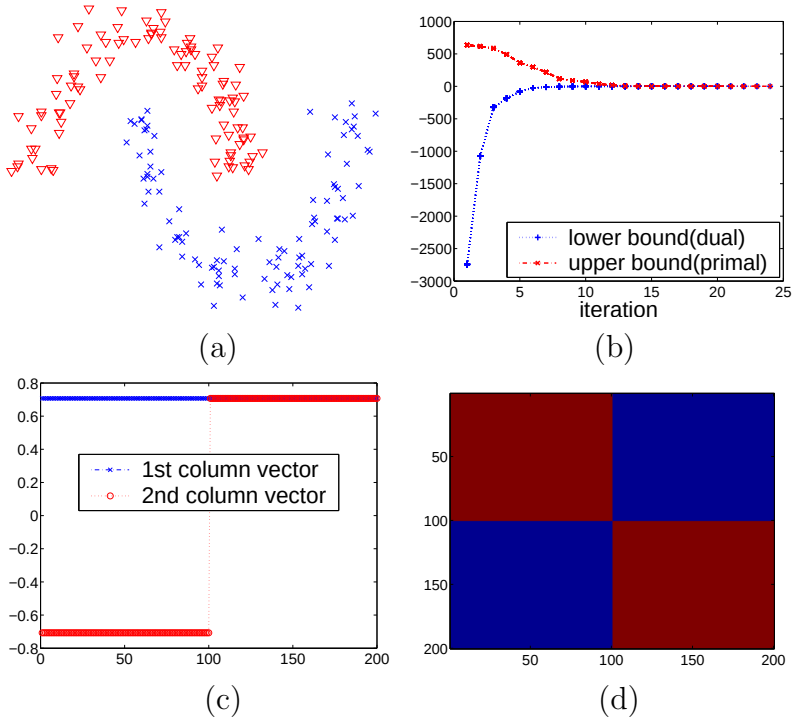

Fig. 2. The results of applying the semidefinite spectral clustering to Two Moons data (with $\sigma = 0.7$) are shown. Data points belonging to each moon are corrected grouped (see (a)). Objective values in the primal and dual are plotted, showing that the strong duality is satisfied (see (b)). The column vectors of $\widetilde{X}^*$ constructed from the optimal feasible matrix, are piecewise-constant vectors (see (c)). The rank k approximation of the optimal feasible matrix, $\widetilde{Y}_k^*$, shown the block diagonal structure, implying that correct grouping is expected (see (d)).

Eigenvectors of the Laplacian matrix (in spectral clustering) or the optimal feasible matrix (in semidefinite spectral clustering), are expected to be *piecewise-constant*, in order to provide desirable discrete solutions such that nodes belonging to the same group have an identical value [4, 3]. However, in practice, eigenvectors have oscilating values, which deteriorates the clustering performance since it is not easy to choose an appropriate threshold value for rounding such eigenvectors. In contrast to spectral clustering where the eigenvectors of the (normalized) Laplacian matrix are used, the eigenvectors in the semidefinite spectral clustering algorithm are very close to piecewise-constant vectors (see Fig. 2 (c)). This results from the well-formed block diagonal structure of the optimal feasible matrix (see Fig. 2 (d)). Also, Table.1 for

an experimental comparison to the NJWCut for Two moons data, according to various kernel sizes, σ, to construct an affinity matrix, which shows our semidefinite spectral clustering is very robust to determine a kernel size, while spectral clustering methods are very susceptible to value of the kernel size.

The following two more examples, more clearly show the benefit of the semidefinite spectral clustering (see Fig. 3).

Table 1
Results of Two moons data clustering according to various kernel sizes, σ, in terms of the classification accuracy(%) are summarized for two methods(semidefinite spectral clustering and NJWCut).

Methods \ σ	1.18	0.88	0.71	0.59	$\cdots$	0.35	0.32	0.29	0.27	0.25
Ours (%)	91.5	98.5	**100**	**100**	$\cdots$	**100**	**100**	**100**	**100**	**100**
NJWCut (%)	89.5	91.5	93.0	**100**	$\cdots$	**100**	58.0	57.5	52.0	51.0

Empirical comparison to the NJWCut for two artificial data sets, are shown in Fig. 3. The second artificial data set (see Fig. 3 (a)-(d)) consists of two groups, one of which is a dense cluster in the center (100 data points) and the other is a scattered-background cluster (100 data points). The third artifical data set (see Fig. 3 (e)-(h)) consists of three coherent groups (a dense cluster in the center and two outer rings), each of which contains 130 data points. Centroid-based clustering algorithms, including k-means and mixture of Gaussians, have difficulty in grouping these exemplary data sets. Grouping by NJWCut is not satisfactory in these examples, since the eigenvectors of the Laplacian matrix in NJWCut are not piecewise-constant (in fact, very osciallting). In contrast, the semidefinte spectral clustering shows the successful clustering result and column vectors of $\widetilde{X}^*$ are piecewise-constant.

4.2 Experiment 2: Document clustering

We apply our semidefinite spectral clustering algorithm to the task of document clustering that plays an important role in text information analysis. To this end, we use the '20 newsgroup' [2] data set which contains about 20,000 articles (see Table.2).

Some of newsgroups invlove similar topics. For instance, NG18 and NG 19 are related each other. In our numerical study, we selected top 1000 words by ranking the values of mutual information between terms and documents. These words are used to construct term-document matrices where entries of

[2] Dataset and the bow toolkit required to construct a term-document matrix, are available online [13].

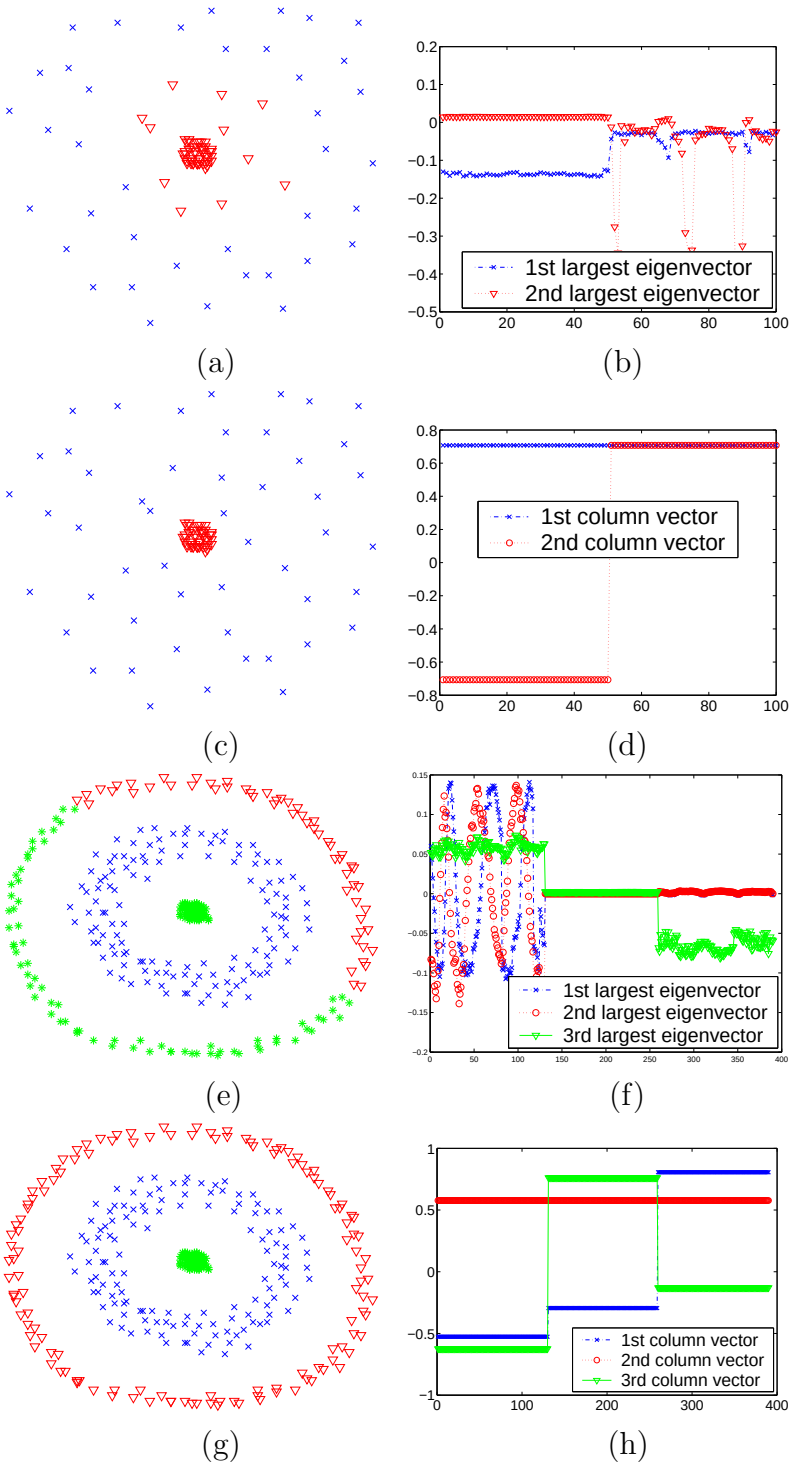

Fig. 3. Comparison of the semidefinite spectral clustering to NJWCut with two data sets is shown ((a)-(d) with $\sigma = 0.8$, (e)-(h) with $\sigma = 1.8$). Clustering results and plots of top two and three eigenvectors of the normalized Laplacian matrix, are shown in (a)-(b) and (e)-(f). The eigenvectors of the normalized Laplacian matrix (in NJWCut, $D^{-1/2}WD^{-1/2}$ replaced L as the normalized Laplacian) contain oscillating values (see (b) and (f)). Thus clustering was not satisfactory. In contrast, the semidefintie spectral clustering provides piecewise-constant vectors (see (d) and (h)) and the clustering was successful (see (c) and (g)).

its column vector indicate the frequencies of words in a document. (the tf.idf scheme was used). Document-document affinities are determined by the inner product of these column vectors. We carried out numerical experiments for two data sets where one data sets consists of NG18 and NG19 and the other contains three groups (NG 17, NG18, NG19). For both balanced and unbalanced cases, we compared the semidefinite spectral clustering with NJWCut and MNCut.

In the first experiment (two groups, NG18/NG19), we chose 500 articles randomly from each newsgroup for the balanced case and randomly selected 350 and 650 articles from NG18 and NG19, respectively for the unbalanced case. In the second experiment (three groups, NG17/NG18/NG19), we chose 260 articles randomly from each newsgroup for the balanced case and randomly selected 200, 320, 260 articles from NG18 and NG19, respectively for the unbalanced case. For each case, we carried out 10 independent experiments and its averaged results are summarized in Table 3. The clustering performance was measured in terms of classification accuracy, indicating how many documents were correctly classified. In balanced cases, our semidefinite spectral clustering outperforms the NJWCut and MNCut. In unbalanced cases, our method still shows some improvement over the NJWCut and MNCut.

Table 2
20 newsgroup data and their indexing.

NG1	alt.atheism	NG11	rec.sport.hockey
NG2	comp.graphics	NG12	sci.crypy
NG3	comp.os.ms-windows.misc	NG13	sci.electronics
NG4	comp.sys.ibm.pc.hardware	NG14	sci.med
NG5	comp.sys.mac.hardware	NG15	sci.space
NG6	comp.windows.x	NG16	soc.religion.christian
NG7	misc.forsale	NG17	talk.politics.guns
NG8	rec.autos	NG18	talk.politics.mideast
NG9	rec.motorcycles	NG19	talk.politics.misc
NG10	rec.sport.baseball	NG20	talk.religion.misc

5 Discussions

We have presented the semidefinite spectral clustering algorithm which jointly exploits semidefinite programming and spectral learning methods. In contrast

Table 3
Results of document clustering in terms of the classification accuracy (%) are summarized for two experiments (two groups and three groups). Values in parenthesis represent standard deviation.

Dataset	Balanced case			Unbalanced case		
	Ours	NJWCut	MNCut	Ours	NJWCut	MNCut
NG18	**96.040**	62.650	60.900	**96.952**	72.000	73.714
	(±0.518)	(±9.202)	(±8.703)	**(±0.437)**	(±4.849)	(±5.656)
NG19	95.280	**97.067**	**97.100**	75.641	**95.230**	**96.461**
	(±0.856)	**(±0.611)**	**(±0.383)**	(±0.235)	**(±0.218)**	**(±0.870)**
Total (%)	**95.660**	79.859	79.000	**86.297**	83.615	85.087
	(±0.666)	(±4.907)	(±4.239)	**(±0.295)**	(±2.533)	(±2.393)
NG17	81.923	**82.692**	**83.173**	74.500	**77.167**	**76.500**
	(±0.666)	**(±3.140)**	**(±3.208)**	(±5.220)	**(±2.754)**	**(±3.535)**
NG18	**85.513**	53.269	51.153	**77.917**	44.271	44.843
	(±1.175)	(±5.640)	(±4.752)	**(±6.885)**	(±1.721)	(±1.988)
NG19	**89.231**	87.212	87.500	**89.616**	86.667	84.230
	(±5.996)	(±4.479)	(±4.757)	**(±5.820)**	(±4.289)	(±1.087)
Total (%)	**85.556**	74.391	73.942	**80.677**	69.368	68.524
	(±2.481)	(±1.859)	(±1.663)	**(±4.275)**	(±1.465)	(±0.153)

to the spectral clustering where the eigenvectors of the (normalized) Laplacian matrix are used, the semidefinite spectral clustering used the eigenvectors of the optimal feasible matrix determined through the projected semidefinite relaxation of the graph equipartitioning. Empirical study showed that the eigenvectors of the optimal feasible matrix are very close to piecewise-constant vectors, implying that the clustering based on those vectors provides successful grouping. Numerical comparison to spectral methods (NJWCut and MECut) with artificial data sets and newsgroup data sets, confirmed the high performance of the semidefinite spectral clustering algorithm.

A major limitation of semidefinite spectral clustering is a considerable amount of computational complexity and memory usage which are required to determine an optimal feasible matrix Y^*. The computational complexity increases exponentially with a heavy memory requirement for the barycenter point of size $(nk + 1)$ by $(nk + 1)$ matrix (n is the number of data points and k is the number of clusters), even though an interior-point method for SDP has a polynomial-time complexity. The development of computationally efficient method will be a future work.

6 Acknowledgments

This work was supported by Korea Ministry of Commerce, Industry, and Energy under Brain Neuroinformatics Program and by Korea MIC under ITRC support program supervised by the IITA (IITA-2005-C1090-0501-0018).

A Appendix: Details on Eq. (10)

The Lagrangian associated with (9) has the form

$$\begin{aligned}\mathcal{L}(X,\Omega,\nu_0) = \operatorname{tr}(X^\top L X) + \sum_{i=1}^{n}\sum_{j=1}^{k}\omega_{ij}\left(x_{ij}^2 - x_{ij}\right) \\ + \nu_0\left(\|X\vec{e}_k - \vec{e}_n\|^2 + \|X^\top\vec{e}_n - m\vec{e}_k\|^2\right),\end{aligned} \tag{A.1}$$

where $\Omega = [\omega_{ij}] \in \mathbb{R}^{n\times k}$ and $\nu_0 \in \mathbb{R}$ are Lagrange multipliers (dual variables).

We define $\vec{x} = \operatorname{vec}(X)$ and $\vec{\omega} = \operatorname{vec}(\Omega)$ where $\operatorname{vec}(\cdot)$ is the *vec-function* which stacks the columns of the given matrix into one long vector. We denote by $\operatorname{Diag}(\vec{\omega})$ the diagonal matrix formed by the elements of $\vec{\omega}$. With these definitions, we introduce a dummy scalar variable ω_0 to homogenize the Lagrangian (A.1), leading to

$$\mathcal{L}(X,\vec{\widetilde{\omega}},\nu_0) = [1,\vec{x}^\top]\left\{L_e + \operatorname{Arrow}(\vec{\widetilde{\omega}}) + \nu_0 C\right\}\begin{bmatrix}1\\ \vec{x}\end{bmatrix} - \omega_0, \tag{A.2}$$

where

$$L_e = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & I_k \otimes L \end{array}\right], \quad \operatorname{Arrow}(\vec{\widetilde{\omega}}) = \left[\begin{array}{c|c} \omega_0 & -\frac{1}{2}\vec{\omega}^\top \\ \hline -\frac{1}{2}\vec{\omega} & \operatorname{Diag}(\vec{\omega}) \end{array}\right],$$

$$C = \left[\begin{array}{c|c} n & -\vec{e}_k^\top \otimes \vec{e}_n^\top \\ \hline -\vec{e}_k \otimes \vec{e}_n & (\vec{e}_k\vec{e}_k^\top)\otimes I_n \end{array}\right] + \left[\begin{array}{c|c} m^2\vec{e}_k^\top\vec{e}_k & -m\vec{e}_k^\top \otimes \vec{e}_n^\top \\ \hline -m\vec{e}_k \otimes \vec{e}_n & I_k \otimes (\vec{e}_n\vec{e}_n^\top) \end{array}\right],$$

where $\vec{\widetilde{\omega}} = [\omega_0, \vec{\omega}^\top]^\top \in \mathbb{R}^{nk+1}$, $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and the operator $\otimes$ denotes the Kronecker product.

The dual function (defined as the minimum value of the Lagrangian over X),

yields lower bounds on the optimal value of the original problem (9). The best lower bound that can be obtained from the dual function is given by

$$\begin{aligned}\pi_d := &\max\ -\omega_0 \\ &\text{s.t. } L_e + \text{Arrow}(\vec{\bar{\omega}}) + \nu_0 C = Z \in \mathcal{S}_+^{nk+1},\end{aligned} \tag{A.3}$$

where $\mathcal{S}_+^{nk+1}$ is the *positive semidefinite cone* which is also *self-dual*, i.e., $\mathcal{S}_+^{nk+1} = \{Y \mid \text{tr}(ZY) \geq 0 \text{ for all } Z \in \mathcal{S}_+^{nk+1}\}$ [14].

For semidefinite relaxation, we derive the Lagrangian dual of (A.3). Choosing a dual variable matrix $Y \in \mathcal{S}_+^{nk+1}$ and considering the self-duality and minimax inequality, leads to

$$\begin{aligned}\pi_d &= \max_{\vec{\bar{\omega}},\, \nu_0} \min_{Y \in \mathcal{S}_+^{nk+1}} -\omega_0 + Y \bullet (L_e + \text{Arrow}(\vec{\bar{\omega}}) + \nu_0 C) \\ &\leq \min_{Y \in \mathcal{S}_+^{nk+1}} \max_{\vec{\bar{\omega}},\, \nu_0} L_e \bullet Y - \omega_0 + \vec{\bar{\omega}}^{\top} \text{arrow}(Y) + \nu_0 \text{tr}(CY) \\ &= \pi_r,\end{aligned} \tag{A.4}$$

where the symbol, $\bullet$, denotes the inner product and the operator 'arrow', that is the adjoint linear operator of Arrow$(\cdot)$ [7], is defined by

$$\text{arrow}(Y) \doteq \text{diag}(Y) - [0, Y_{1,2:nk+1}]^{\top}, \tag{A.5}$$

where diag(Y) is the vector consisting of the diagonal elements of Y and $Y_{1,2:nk+1}$ represents a row vector containing elements of the 1st row in the 2nd column till $(nk+1)$th column. In (A.4), we used the following property of the arrow operator:

$$\text{tr}\left(\text{Arrow}(\vec{\bar{\omega}})Y\right) = \vec{\bar{\omega}}^{\top} \text{arrow}(Y). \tag{A.6}$$

The inner maximization of the 2nd equation in (A.4) is finite if and only if $\vec{\bar{\omega}}^{\top} \text{arrow}(Y) = \omega_0$ and $\text{tr}(CY) = 0$. Note that $\vec{\bar{\omega}}^{\top} \text{arrow}(Y) = \omega_0$ is satisfied when $\text{arrow}(Y) = [1, 0, \ldots, 0]^{\top} \in \mathbb{R}^{nk+1}$. Therefore, our desired semidefinite relaxation of (9), which is the Lagrangian dual of (A.3), is given as follows:

$$\begin{aligned}\pi_r^s := &\min_{Y}\ \text{tr}(L_e Y) \\ &\text{s.t. } \text{arrow}(Y) = [1, 0, \ldots, 0]^{\top} \in \mathbb{R}^{nk+1} \\ &\quad\ \text{tr}(CY) = 0, \\ &\quad\ Y \succeq 0.\end{aligned} \tag{A.7}$$

Incorporating with (A.5), (A.7) leads to (10)

References

[1] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[2] Y. Weiss. Segmentation using eigenvectors: A unifying view. In *Proc. Int'l Conf. Computer Vision*, pages 975–982, 1999.

[3] M. Meila and J. Shi. A random walks view of spectral segmentation. In *Prof. Int'l Workshop on Artificial Intelligence and Statistics*, Key West, FL, 2001.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analaysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

[5] M. Gu, H. Zha, C. Ding, X. He, and H. Simon. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering. Technical Report CSE-01-007, Penn State University, 2001.

[6] F. Rendl. Semidefinite programming and combinatorial optimization. *Applied Numerical Mathematics*, 29(3):255–281, 1999.

[7] Q. Zhao. *Semidefinite Programming for Assignment and Partitioning Problems*. PhD thesis, University of Waterloo, 1996.

[8] Q. Zhao, S. Karisch, F. Rendl, and H. Wolkowicz. Semidefinite programming relaxations for the quadratic assignment problem. *Journal of Combinatorial Optimization*, 2(1):71–109, March 1998.

[9] M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Math. Journal*, 25(2):619–633, 1975.

[10] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. Spectral Min-Max cut for graph partitioning and data clustering. In *Proc. IEEE Int'l Conf. Data Mining*, pages 107–114, San Nose, CA, 2001.

[11] J. Keuchel, C. Schnörr, C. Schellewald, and D. Cremers. Binary partitioning, perceptual grouping, and restoration with semidefinite programming. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(11):1364–1379, 2003.

[12] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23:339–358, 1998.

[13] A. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.