Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts

Matthias Hein
Saarland University, Saarbrücken, Germany
hein@cs.uni-saarland.de

Simon Setzer
Saarland University, Saarbrücken, Germany
setzer@mia.uni-saarland.de

Abstract

Spectral clustering is based on the spectral relaxation of the normalized/ratio graph cut criterion. While the spectral relaxation is known to be loose, it has been shown recently that a non-linear eigenproblem yields a tight relaxation of the Cheeger cut. In this paper, we extend this result considerably by providing a characterization of all balanced graph cuts which allow for a tight relaxation. Although the resulting optimization problems are non-convex and non-smooth, we provide an efficient first-order scheme which scales to large graphs. Moreover, our approach comes with the quality guarantee that given any partition as initialization the algorithm either outputs a better partition or it stops immediately.

1 Introduction

The problem of finding the best balanced cut of a graph is an important problem in computer science [9, 27, 13]. It has been used for minimizing the communication cost in parallel computing, reordering of sparse matrices, image segmentation and clustering. In particular, in machine learning spectral clustering is one of the most popular graph-based clustering methods as it can be applied to any graph-based data or to data where similarity information is available so that one can build a neighborhood graph. Spectral clustering is originally based on a relaxation of the combinatorial normalized/ratio graph cut problem, see [31]. The relaxation with the best known worst case approximation guarantee yields a semi-definite program, see [3]. However, it is practically infeasible for graphs with more than 100 vertices due to the presence of $O(n^3)$ constraints where n is the number of vertices in the graph. In contrast, the computation of eigenvectors of a sparse graph scales easily to large graphs. In a line of recent work [6, 29, 14] it has been shown that relaxation based on the nonlinear graph p-Laplacian lead to similar runtime performance while providing much better cuts. In particular, for $p = 1$ one obtains a tight relaxation of the Cheeger cut, see [8, 29, 14].

In this work, we generalize this result considerably. Namely, we provide for almost any balanced graph cut problem a tight relaxation into a continuous problem. This allows flexible modeling of different graph cut criteria. The resulting non-convex, non-smooth continuous optimization problem can be efficiently solved by our new method for the minimization of ratios of differences of convex functions, called RatioDCA. Moreover, compared to [14], we also provide a more efficient way how to solve the resulting convex inner problems by transferring recent methods from total variation denoising, cf. [7], to the graph setting. In first experiments, we illustrate the effect of different balancing terms and show improved clustering results of USPS and MNIST compared to [14].

2 Set Functions, Submodularity, Convexity and the Lovasz Extension

In this section we gather some material from the literature on set functions, submodularity and the Lovasz extension, which we need in the next section. We refer the reader to [11, 4] for a more detailed exposition. We work on weighted, undirected graphs $G = (V, W)$ with vertex set V and a symmetric, non-negative weight matrix W. We define $n := |V|$ and denote by $\overline{A} = V \backslash A$ the complement of A in V, set functions are denoted with a hat, $\hat{S}$, whereas the corresponding Lovasz extension is simply S. The indicator vector of a set A is written as $\mathbf{1}_A$. In the following we always assume that for any considered set function $\hat{S}$ it holds $\hat{S}(\emptyset) = 0$. The Lovasz extension is a way to extend a set function from 2^V to $\mathbb{R}^V$.

Definition 2.1 *Let $\hat{S} : 2^V \to \mathbb{R}$ be a set function with $\hat{S}(\emptyset) = 0$. Let $f \in \mathbb{R}^V$ be ordered in increasing order $f_1 \leq f_2 \leq \ldots \leq f_n$ and define $C_i = \{j \in V \,|\, f_j > f_i\}$ where $C_0 = V$. Then $S : \mathbb{R}^V \to \mathbb{R}$ given by*

$$S(f) \;=\; \sum_{i=1}^{n} f_i \Big(\hat{S}(C_{i-1}) - \hat{S}(C_i) \Big) \;=\; \sum_{i=1}^{n-1} \hat{S}(C_i)(f_{i+1} - f_i) + f_1 \hat{S}(V)$$

*is called the **Lovasz extension** of $\hat{S}$. Note that $S(\mathbf{1}_A) = \hat{S}(A)$ for all $A \subset V$.*

Note that for symmetric set functions $\hat{S}$, that is $\hat{S}(A) = \hat{S}(\overline{A})$ for all $A \subset V$, the property $\hat{S}(\emptyset) = 0$ implies $\hat{S}(V) = 0$. A particular interesting class of set functions are the submodular set functions as their Lovasz extension is convex.

Definition 2.2 *A set function, $\hat{F} : 2^V \to \mathbb{R}$ is **submodular** if for all $A, B \subset V$,*

$$\hat{F}(A \cup B) + \hat{F}(A \cap B) \;\leq\; \hat{F}(A) + \hat{F}(B).$$

*$\hat{F}$ is called **strictly submodular** if the inequality is strict whenever $A \nsubseteq B$ or $B \nsubseteq A$.*

Note that symmetric submodular set functions are always non-negative as for all $A \subset V$,

$$2\hat{F}(A) = \hat{F}(A) + \hat{F}(\overline{A}) \geq \hat{F}(A \cup \overline{A}) + \hat{F}(A \cap \overline{A}) = \hat{F}(V) + \hat{F}(\emptyset) = 0.$$

An important class of set functions for clustering are cardinality-based set functions.

Proposition 2.1 ([4]) *Let $e \in \mathbb{R}^V_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$ is a concave function, then $\hat{F} : A \mapsto g(s(A))$ is submodular. If $\hat{F} : A \mapsto g(s(A))$ is submodular for all $s \in \mathbb{R}^V_+$, then g is concave.*

The following properties hold for the Lovasz extension.

Proposition 2.2 ([11, 4]) *Let $S : \mathbb{R}^V \to \mathbb{R}$ be the Lovasz extension of $\hat{S} : 2^V \to \mathbb{R}$ with $\hat{S}(\emptyset) = 0$.*

- *$\hat{S}$ is submodular if and only if S is convex,*
- *S is positively one-homogeneous,*
- *$S(f) \geq 0, \; \forall\, f \in \mathbb{R}^V$ and $S(\mathbf{1}) = 0$ if and only if $\hat{S}(A) \geq 0, \; \forall\, A \subset V$ and $\hat{S}(V) = 0$.*
- *$S(f + \alpha \mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V, \alpha \in \mathbb{R}$ if and only if $\hat{S}(V) = 0$,*
- *S is even, if $\hat{S}$ is symmetric.*

Proof: All except the fourth property can be found in [11, 4]. The non-negativity of the Lovasz extension S follows from the non-negativity of $\hat{S}$ and $f_{i+1} - f_i \geq 0$ for $i = 1, \ldots, n-1$ and $\hat{S}(V) = 0$. One has $\hat{S}(A) = S(\mathbf{1}_A)$ and thus the other direction follows immediately. □

One might wonder if the Lovasz extension of all submodular set functions generates the set of all positively one-homogeneous convex functions. This is not the case, as already Lovasz [21] gave a counter-example. In the next section we will be interested in the class of positively one-homogeneous, even, convex functions S with $S(f + \alpha \mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V$. From the above proposition we deduce that these properties are fulfilled for the Lovasz extension of any symmetric, submodular set function. However, also for this special class there exists a counter-example. Take

$$S(f) = \left\| f - \frac{1}{|V|} \langle f, \mathbf{1} \rangle \mathbf{1} \right\|_\infty .$$

It fulfills all the stated conditions but it induces the set function $\hat{S}(A) := S(\mathbf{1}_A)$ given as

$$\hat{S}(A) = \frac{1}{|V|} \begin{cases} \max\{|A|, |V \backslash A|\}, & 0 < |A| < |V| \\ 0, & \text{else} \end{cases}$$

It is easy to check that this function is not submodular. Thus different convex one-homogeneous functions can induce the same set function via $\hat{S}(A) := S(\mathbf{1}_A)$.

It is known [15] that a large class of functions e.g. every $f \in C^2(\mathbb{R}^n)$ can be written as a difference of convex functions. As submodular functions correspond to convex functions in the sense of the Lovasz extension, one can ask if the same result holds for set functions: Is every set function a difference of submodular set functions ? The following result has been reported in [24]. As some properties assumed in the proof in [24] do not hold, we give an alternative constructive proof.

Proposition 2.3 *Every set function $\hat{S} : 2^V \to \mathbb{R}$ can be written as the difference of two submodular functions. The corresponding Lovasz extension $S : \mathbb{R}^V \to \mathbb{R}$ can be written as a difference of convex functions.*

Proof: We introduce for every set function $\hat{S}$ the gap-function $\Delta_{\hat{S}} : 2^V \times 2^V \to \mathbb{R}$ defined as

$$\Delta_{\hat{S}}(A, B) := \hat{S}(A) + \hat{S}(B) - \hat{S}(A \cup B) - \hat{S}(A \cap B).$$

$\hat{S}$ is submodular if and only if $\Delta_{\hat{S}}(A, B) \geq 0$ for all $A, B \subset V$. Note that for any set function $\Delta_{\hat{S}}(A, B) = 0$ if $A \subset B$ or $B \subset A$. By adding a strictly submodular function $\hat{T}$ to $\hat{S}$, we can make the gap of $\hat{S} + \lambda\hat{T}$ for some $\lambda \geq 0$ positive, so that $\hat{S} = \hat{S} + \lambda\hat{T} - \lambda\hat{T}$ can be written as a difference of submodular functions. We use $\hat{T}(A) := |A||\overline{A}|$. A straightforward calculation shows for all $A, B \subset V$ with $A \not\subseteq B$ and $B \not\subseteq A$,

$$\Delta_{\hat{T}}(A, B) = 2|A||B| - 2|A \cap B|(|A| + |B|) + 2|A \cap B|^2 = 2\,|A \backslash B|\,|B \backslash A| \geq 2.$$

As $\Delta_{\hat{S}}(A, B) \geq -2(\max_A \hat{S}(A) - \min_A \hat{S}(A))$ we get that with $\lambda = (\max_A \hat{S}(A) - \min_A \hat{S}(A))$ it suffices to add $\lambda\hat{T}$ to $\hat{S}$ in order that $\hat{S} + \lambda\hat{T}$ is submodular. □

Note that the proof of Proposition 2.3 is constructive. Thus we can always find the decomposition of the set function into a difference of two submodular functions and thus also the decomposition of its Lovasz extension into a difference of convex functions.

Corollary 2.1 *Every symmetric set function $\hat{S} : 2^V \to \mathbb{R}$ can be written as a difference of two non-negative submodular set functions.*

Proof: Every symmetric submodular function $\hat{T}$ is non-negative. Thus $\hat{S} + \lambda\hat{T}$ (in the notation of the proof of Proposition 2.3) as a symmetric and submodular set function is non-negative and $\hat{T}$ is as well symmetric and submodular. □

3 Tight Relaxations of Balanced Graph Cuts

In graph-based clustering a popular criterion to partition the graph is to minimize the cut $\mathrm{cut}(A, \overline{A})$, defined as

$$\mathrm{cut}(A, \overline{A}) = \sum_{i \in A, j \in \overline{A}} w_{ij},$$

where $(w_{ij}) \in \mathbb{R}^{|V| \times |V|}$ are the non-negative, symmetric weights of the undirected graph $G = (V, W)$ usually interpreted as similarities of vertices i and j. Direct minimization of the cut leads typically to very unbalanced partitions, where often just a single vertex is split off. Therefore one has to introduce a balancing term which biases the criterion towards balanced partitions. Two popular balanced graph cut criterion are the Cheeger cut $\mathrm{RCC}(A, \overline{A})$ and the ratio cut $\mathrm{RCut}(A, \overline{A})$

$$\mathrm{RCC}(A, \overline{A}) = \frac{\mathrm{cut}(A, \overline{A})}{\min\{|A|, |\overline{A}|\}}, \qquad \mathrm{RCut}(A, \overline{A}) = |V| \frac{\mathrm{cut}(A, \overline{A})}{|A||\overline{A}|} = \mathrm{cut}(A, \overline{A}) \Big(\frac{1}{|A|} + \frac{1}{|\overline{A}|} \Big).$$

We consider later on also their normalized versions. Spectral clustering is derived as relaxation of the ratio cut criterion based on the second eigenvector of the graph Laplacian. While the second eigenvector can be efficiently computed, it is well-known that this relaxation is far from being tight. In particular there exist graphs where the spectral relaxation is as bad [12] as the isoperimetric inequality suggests [1]. In a recent line of work [6, 29, 14] it has been shown that a tight relaxation for the Cheeger cut can be achieved by moving from the linear eigenproblem to a nonlinear eigenproblem associated to the nonlinear graph 1-Laplacian [14].

In this work we generalize this result considerably by showing in Theorem 3.1 that a tight relaxation exists for every balanced graph cut measure which is of the form cut divided by balancing term. More precisely, let $\hat{S}: 2^V \to \mathbb{R}$ be a symmetric non-negative set function. Then a **balanced graph cut criterion** $\phi : 2^V \to \mathbb{R}_+$ of a partition $(A, \overline{A})$ has the form,

$$\phi(A) := \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)}. \tag{1}$$

As we consider undirected graphs, the cut is a symmetric set function and thus $\phi(A) = \phi(\overline{A})$. In order to get a balanced graph cut, $\hat{S}$ is typically chosen as a function of $|A|$ (or some other type of volume) which is monotonically increasing on $[0, |V|/2]$. The first part of the theorem showing the equivalence of combinatorial and continuous problem is motivated by a result derived by Rothaus in [28] in the context of isoperimetric inequalities on Riemannian manifolds. It has been transferred to graphs by Tillich and independently by Houdre in [30, 17]. We generalize their result further so that it now holds for all possible non-negative symmetric set functions. In order to establish the link to the result of Rothaus, we first state the following characterization

Lemma 3.1 *A function $S : V \to \mathbb{R}$ is positively one-homogeneous, even, convex and $S(f + \alpha\mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V, \alpha \in \mathbb{R}$ if and only if $S(f) = \sup_{u \in U} \langle u, f \rangle$ where $U \subset \mathbb{R}^n$ is a closed symmetric convex set and $\langle u, \mathbf{1} \rangle = 0$ for any $u \in U$.*

Proof: Note that every convex, positively one-homogeneous function, $S : \mathbb{R}^V \to \mathbb{R}$, has the form,

$$S(f) = \sup_{u \in U} \langle u, f \rangle ,$$

where U is a convex set, see e.g. [16]. If U is symmetric, we have for any $f \in \mathbb{R}^V$,

$$S(f) = \sup_{u \in U} \langle u, f \rangle = \sup_{u \in U} \langle -u, f \rangle = \sup_{u \in U} \langle u, -f \rangle = S(-f),$$

and thus S is even. Moreover, if S is even we have for all f,

$$S(f) = \sup_{u \in U} \langle u, f \rangle = \sup_{u \in U} \langle -u, -f \rangle = \sup_{u \in U} \langle u, -f \rangle = S(-f),$$

which implies that U is symmetric. Second, if $S(f + \alpha\mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V$ and $\alpha \in \mathbb{R}$, we have in particular

$$S(\mathbf{1}) = \sup_{u \in U} \langle u, \mathbf{1} \rangle = 0.$$

As U is symmetric, this is only possible if for all $u \in U$, $\langle u, \mathbf{1} \rangle = 0$. The other direction follows easily. □

Theorem 3.1 *Let $G = (V, E)$ be a finite, weighted undirected graph and $S : \mathbb{R}^V \to \mathbb{R}$ and let $\hat{S} : 2^V \to \mathbb{R}$ be symmetric with $\hat{S}(\emptyset) = 0$, then*

$$\inf_{f \in \mathbb{R}^V} \frac{\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}|f_i - f_j|}{S(f)} = \inf_{A \subset V} \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)},$$

if either one of the following two conditions holds

1. *S is positively one-homogeneous, even, convex and $S(f + \alpha\mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V$, $\alpha \in \mathbb{R}$ and $\hat{S}$ is defined as $\hat{S}(A) := S(\mathbf{1}_A)$ for all $A \subset V$.*

2. *S is the Lovasz extension of the non-negative, symmetric set function $\hat{S}$ with $\hat{S}(\emptyset) = 0$.*

Let $f \in \mathbb{R}^V$ and denote by $C_t := \{i \in V \mid f_i > t\}$, then it holds under both conditions,

$$\min_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \leq \frac{\frac{1}{2}\sum_{i,j=1}^{n} w_{ij}|f_i - f_j|}{S(f)}.$$

Proof: Let us define $R(f) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}|f_i - f_j|$. Then with $C_t := \{i \in V \mid f_i > t\}$ one has,

$$R(f) = \sum_{f_i > f_j} w_{ij}\,(f_i - f_j) = \sum_{f_i > f_j} w_{ij} \int_{f_j}^{f_i} 1dt = \int_{-\infty}^{\infty} \sum_{f_i > t \geq f_j} w_{ij}\; dt = \int_{-\infty}^{\infty} \mathrm{cut}(C_t, \overline{C_t})\, dt,$$

where we use that $\mathrm{cut}(\emptyset, V) = 0$. Let f be ordered in increasing order $f_1 \leq f_2 \leq \ldots \leq f_n$. With Lemma 3.1 we have $S(f) = \sup_{u\in U} \langle u, f\rangle$, where U is a symmetric and convex set with $\langle u, \mathbf{1}\rangle = 0$ for all $u \in U$. We have for any $u \in U$, $\langle u, f\rangle \leq S(f)$ and thus

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathrm{cut}(C_t, \overline{C_t})\, dt &= \int_{\min_i f_i}^{\max_i f_i} \mathrm{cut}(C_t, \overline{C_t})\, dt \\
&\geq \int_{\min_i f_i}^{\max_i f_i} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \langle u, \mathbf{1}_{C_t}\rangle\; dt \\
&\geq \inf_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \int_{\min_i f_i}^{\max_i f_i} \langle u, \mathbf{1}_{C_t}\rangle\; dt.
\end{aligned}$$

Let $C_i := C_{f_i}$. Note that $C_n = \emptyset$ and we define $C_0 = V$ and use the convention $0/0 = \infty$. Then with $\langle u, \mathbf{1}_V\rangle = \langle u, \mathbf{1}_{C_0}\rangle = 0$,

$$\begin{aligned}
\int_{\min_i f_i}^{\max_i f_i} \langle u, \mathbf{1}_{C_i}\rangle\, dt &= \sum_{i=1}^{n-1} \langle u, \mathbf{1}_{C_i}\rangle \big(f_{i+1} - f_i\big) \\
&= \sum_{i=1}^{n} f_i \Big(\langle u, \mathbf{1}_{C_{i-1}}\rangle - \langle u, \mathbf{1}_{C_i}\rangle \Big) = \sum_{i=1}^{n} f_i u_i
\end{aligned}$$

Thus we get,

$$\begin{aligned}
R(f) = \int_{-\infty}^{\infty} \mathrm{cut}(C_t, \overline{C_t})\, dt &\geq \inf_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \sup_{u\in U} \langle u, f\rangle \\
&= \inf_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} S(f).
\end{aligned}$$

For the other case we note that with the definition of the Lovasz extension we get

$$\begin{aligned}
R(f) = \int_{-\infty}^{\infty} \mathrm{cut}(C_t, \overline{C_t})\, dt &\geq \inf_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \int_{\min_i f_i}^{\max_i f_i} \hat{S}(C_t)\, dt \\
&= \inf_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)}\; S(f)
\end{aligned}$$

□

Theorem 3.1 can be generalized by replacing the cut with an arbitrary other set function. However, the emphasis of this paper is to use the new degree of freedom for balanced graph clustering. The more general approach will be discussed elsewhere. Note that the first condition in Theorem 3.1 implies that $\hat{S}$ is symmetric as

$$\hat{S}(A) = S(\mathbf{1}_A) = S(-\mathbf{1}_A) = S(\mathbf{1} - \mathbf{1}_A) = S(\mathbf{1}_{\overline{A}}) = \hat{S}(\overline{A}).$$

Moreover, $\hat{S}$ is non-negative with $\hat{S}(\emptyset) = \hat{S}(V) = 0$ as S is even, convex and positively one-homogeneous. For the second condition note that by Proposition 2.3 the Lovasz extension of any set function can be written as a difference of convex (d.c.) functions. As the total variation term in the enumerator is convex, we thus have to minimize a ratio of a convex and a d.c. function. The efficient minimization of such problems will be the topic of the next section.

We would like to point out a related line of work for the case where the balancing term $\hat{S}$ is submodular and the balanced graph cut measure is directly optimized using submodular minimization techniques. In [26] this idea is proposed for the ratio cut and subsequently generalized [25, 18] so that every submodular balancing function $\hat{S}$ can be used. While the general framework is appealing, it is unclear if the minimization can be done efficiently. Moreover, note that Theorem 3.1 goes well beyond the case where $\hat{S}$ is submodular.

3.1 Examples of Balancing Set Functions

Theorem 3.1 opens up new modeling possibilities for clustering based on balanced graph cuts. We discuss in the experiments differences and properties of the individual balancing terms. However, it is out of the scope of this paper to answer the question which balancing term is the "best". An answer to such a question is likely to be application-dependent. However, for a given random graph model it might be possible to suggest a suitable balancing term given one knows how cut and volume behave. A first step in this direction has been done in [22] where the limit of cut and volume has been discussed for different neighborhood graph types.

In the following we assume that we work with graphs which have non-negative edge weights $W = (w_{ij})$ and non-negative vertex weights $e : V \to \mathbb{R}_+$. The volume $\mathrm{vol}(A)$ of a set $A \subset V$ is defined as $\mathrm{vol}(A) = \sum_{i \in A} e_i$. The volume reduces to the cardinality if $e_i = 1$ for all $i \in V$ (unnormalized case) or to the volume considered in the normalized cut, $\mathrm{vol}(A) = \sum_{i \in A} d_i$ for $e_i = d_i$ for all $i \in V$ (normalized case), where d_i is the degree of vertex i. We denote by E the diagonal matrix with $E_{ii} = e_i, i = i, \ldots, n$. Using general vertex weights allows us to present the unnormalized and normalized case in a unified framework. Moreover, general vertex weights allow more modeling freedom e.g. one can give two different vertices very large vertex weights and so implicitly enforce that they will be in different partitions.

We report here the Lovasz extension of two important set functions which will be needed in the sequel. For that we define the functions $g_{\max,\alpha}$ and $g_{\min,\alpha}$ as:

$$g_{\max,\alpha}(f) = \max \Big\{ \langle \rho, f \rangle \;\Big|\; 0 \le \rho_i \le e_i, \; \forall\, i = 1, \ldots, n, \sum_{i=1}^n \rho_i = \alpha \, \mathrm{vol}(V) \Big\},$$

$$g_{\min,\alpha}(f) = \min \Big\{ \langle \rho, f \rangle \;\Big|\; 0 \le \rho_i \le e_i, \; \forall\, i = 1, \ldots, n, \sum_{i=1}^n \rho_i = \alpha \, \mathrm{vol}(V) \Big\}$$

and the weighted p-mean $\mathrm{wmean}_p(f)$ is defined as $\mathrm{wmean}_p(f) = \inf_{a \in \mathbb{R}} \sum_{i=1}^n e_i |f_i - a|^p$. Note that $g_{\max,\alpha}$ is convex, whereas $g_{\min,\alpha}$ is concave. Both functions can be easily evaluated by sorting the componentwise product $e_i f_i$.

Proposition 3.1 *Let $\hat{S} : 2^V \to \mathbb{R}$, $\hat{S}(A) := \min\{\mathrm{vol}(A), \mathrm{vol}(\overline{A})\}$. Then the Lovasz extension $S : V \to \mathbb{R}$ is given by $S(f) = \left\| E(f - \mathrm{wmean}_1(f)\mathbf{1}) \right\|_1$.*
Let $e_i = 1, \forall i \in V$ and $\hat{S} : 2^V \to \mathbb{R}$, $\hat{S}(A) := \begin{cases} \min\{|A|, |\overline{A}|\}, & \text{if } \min\{|A|, |\overline{A}|\} \le K, \\ K, & \text{else.} \end{cases}$. Then the Lovasz extension $S : V \to \mathbb{R}$ is given as $S(f) = g_{\max, \frac{K}{|V|}}(f) - g_{\min, \frac{K}{|V|}}(f)$.

Proof: One has,

$$\begin{aligned} \hat{S}(C_{i-1}) - \hat{S}(C_i) &= \min\{\mathrm{vol}(C_{i-1}), \mathrm{vol}(\overline{C_{i-1}})\} - \min\{\mathrm{vol}(C_i), \mathrm{vol}(\overline{C_i})\} \\ &= \begin{cases} e_i, & \text{if } \mathrm{vol}(C_i) \le \frac{\mathrm{vol}(V)}{2} - e_i, \\ -e_i, & \text{if } \mathrm{vol}(C_i) > \frac{\mathrm{vol}(V)}{2}, \\ \mathrm{vol}(V) - 2\,\mathrm{vol}(C_i) - e_i, & \text{if } \frac{\mathrm{vol}(V)}{2} - e_i < \mathrm{vol}(C_i) \le \frac{\mathrm{vol}(V)}{2}. \end{cases} \end{aligned}$$

There can be only one unique element i^* where the third condition holds as it has to fulfill,

$$\mathrm{vol}(C_{i^*}) \le \frac{\mathrm{vol}(V)}{2}, \text{ and } \mathrm{vol}(\overline{C_{i^*-1}}) \le \frac{\mathrm{vol}(V)}{2}.$$

Name	$S(f)$	$\hat{S}(A)$
Cheeger p-cut	$\Big(\sum_{i=1}^{n} e_i \lvert f_i - \mathrm{wmean}_p(f)\rvert^p\Big)^{\frac{1}{p}}$	$\frac{\big(\mathrm{vol}(A)\,\mathrm{vol}(\overline{A})\big)^{\frac{1}{p}}}{\big(\mathrm{vol}(A)^{\frac{1}{p-1}}+\mathrm{vol}(\overline{A})^{\frac{1}{p-1}}\big)^{1-\frac{1}{p}}}$
Normalized p-cut	$\Big(\sum_{i=1}^{n} e_i \lvert f_i - \frac{\langle e,f\rangle}{\mathrm{vol}(V)}\rvert^p\Big)^{\frac{1}{p}}$	$\frac{\big(\mathrm{vol}(A)\,\mathrm{vol}(\overline{A})^p+\mathrm{vol}(A)^p\,\mathrm{vol}(\overline{A})\big)^{\frac{1}{p}}}{\mathrm{vol}(V)}$
Trunc. Cheeger cut	$g_{\max,\alpha}(f) - g_{\min,\alpha}(f)$	$\begin{cases} \mathrm{vol}(A), & \text{if } \mathrm{vol}(A) \leq \alpha\,\mathrm{vol}(V), \\ \mathrm{vol}(\overline{A}), & \text{if } \mathrm{vol}(\overline{A}) \leq \alpha\,\mathrm{vol}(V), \\ \alpha\,\mathrm{vol}(V), & \text{else.} \end{cases}$
Hard balanced cut	$\big(g_{\max,\frac{K}{\lvert V\rvert}}(f) - g_{\min,\frac{K}{\lvert V\rvert}}(f)\big)$ $-\big(g_{\max,\frac{K-1}{\lvert V\rvert}}(f) - g_{\min,\frac{K-1}{\lvert V\rvert}}(f)\big)$	$\begin{cases} 1, & \text{if } \min\{\lvert A\rvert, \lvert\overline{A}\rvert\} \geq K \\ 0, & \text{else.} \end{cases}$
Hard Cheeger cut	$\lVert f - \mathrm{median}(f)\mathbf{1}\rVert_1$ $-\big(g_{\max,\frac{K-1}{\lvert V\rvert}}(f) - g_{\min,\frac{K-1}{\lvert V\rvert}}(f)\big)$	$\begin{cases} 0, & \text{if } \min\{\lvert A\rvert, \lvert\overline{A}\rvert\} < K, \\ \min\{\lvert A\rvert, \overline{\lvert A\rvert}\} \\ -(K-1), & \text{else.} \end{cases}$

Table 1: Examples of balancing set functions and their continuous counterpart. For the hard balanced and hard Cheeger cut we have unit vertex weights, that is $e_i \equiv 1$.

We will show in the following that this element is equal to the weighted median. We define $S_+ = \{i \in V \mid f_i > \mathrm{wmean}_1(f)\}$, $S_- = \{i \in V \mid f_i < \mathrm{wmean}_1(f)\}$ and $S_= = \{i \in V \mid f_i = \mathrm{wmean}_1(f)\}$. The optimality condition for the weighted median reads,

$$0 \in E^T \operatorname{sign}(E(f - \mathrm{wmean}_1(f))),$$

and summing up yields

$$0 = \sum_{i \in S_+} e_i - \sum_{i \in S_-} e_i + \sum_{i \in S_=} \alpha_i e_i,$$

where $\alpha_i \in [-1, 1]$. Thus $\mathrm{vol}(S_=) \geq \lvert\,\mathrm{vol}(S_-) - \mathrm{vol}(S_+)\rvert$ which implies with $\mathrm{vol}(S_-) + \mathrm{vol}(S_+) + \mathrm{vol}(S_=) = \mathrm{vol}(V)$ that $\mathrm{vol}(S_-) \leq \frac{\mathrm{vol}(V)}{2}$ and $\mathrm{vol}(S_+) \leq \frac{\mathrm{vol}(V)}{2}$. As this condition can only be fulfilled for one element the weighted median has to be equal to the element f_{i^*} where the condition above is fulfilled. We decompose $S_=$ as $S_= = S_{>i^*} \cup S_{<i^*} \cup i^*$ according to the order of f used in the Lovasz extension. Using $C_{i^*} = S_+ \cup S_{>i^*}$ we obtain,

$$\begin{aligned}
&\lVert E(f - \mathrm{wmean}_1(f)\mathbf{1})\rVert_1 \\
&= \sum_{i \in S_+} e_i(f_i - c) - \sum_{i \in S_-} e_i(f_i - c) \\
&= \sum_{i \in S_+} e_i f_i - \sum_{i \in S_-} e_i f_i + c\Big(\sum_{i \in S_-} e_i - \sum_{i \in S_+} e_i\Big) \\
&= \sum_{i \in S_+} e_i f_i - \sum_{i \in S_-} e_i f_i + c\big(\mathrm{vol}(V) - 2\,\mathrm{vol}(S_+) - \mathrm{vol}(S_=)\big) \\
&= \sum_{i \in S_+} e_i f_i - \sum_{i \in S_-} e_i f_i + c\big(\mathrm{vol}(V) - 2\,\mathrm{vol}(C_{f_{i^*}}) + 2\,\mathrm{vol}(S_{>i^*}) - \mathrm{vol}(S_{>i^*}) - e_{i^*} - \mathrm{vol}(S_{<i^*})\big) \\
&= \sum_{i \in S_+ \cup S_{>i^*}} e_i f_i - \sum_{i \in S_- \cup S_{<i^*}} e_i f_i + c\big(\mathrm{vol}(V) - 2\,\mathrm{vol}(C_{f_{i^*}}) - e_{i^*}\big) \\
&= \sum_{i=1}^{n} f_i\big(\hat{S}(C_{i-1}) - \hat{S}(C_i)\big)
\end{aligned}$$

The second case follows directly from the definition of the Lovasz extension. □

In Table 1 we collect a set of interesting set functions enforcing different levels of balancing. For the Cheeger and Normalized p-cut family and the truncated Cheeger cut the functions S are convex and not necessarily the Lovasz extension of the induced set functions $\hat{S}$ (first case in Theorem 3.1). In the case of hard balanced and hard Cheeger cut the set function $\hat{S}$ is not submodular. However, in both cases we know an explicit decomposition of the set function $\hat{S}$ into a difference of submodular functions and thus their Lovasz extension S can be written as a difference of the convex functions.

The RatioDCA algorithm in Section 4 requires a subgradient of a part of the balancing function. Note that the subdifferential of a convex function S in the setting of Lemma 3.1 is given for any $f \neq 0$,

$$\partial S(f) = \{u \in \mathbb{R}^n \,|\, u \in U, \quad \langle u, f\rangle = S(f)\},$$

so that one can always determine a subgradient using that description. However, in order to enable a direct implementation we give an explicit description of the subdifferentials below.

Cheeger p-cut (CC_p): The weighted p-mean $\mathrm{wmean}_p(f)$ of $f \in \mathbb{R}^V$ is defined as

$$\mathrm{wmean}_p(f) = \inf_{a \in \mathbb{R}} \sum_{i=1}^{n} e_i |f_i - a|^p.$$

It reduces to the weighted median for $p = 1$. The convex function CC_p induces the set function $\widehat{\mathrm{CC}}_p$

$$\mathrm{CC}_p(f) = \Big(\sum_{i=1}^{n} e_i |f_i - \mathrm{wmean}_p(f)|^p\Big)^{\frac{1}{p}} \implies \widehat{\mathrm{CC}}_p(A) = \frac{\big(\mathrm{vol}(A)\,\mathrm{vol}(\overline{A})\big)^{\frac{1}{p}}}{\big(\mathrm{vol}(A)^{\frac{1}{p-1}} + \mathrm{vol}(\overline{A})^{\frac{1}{p-1}}\big)^{1-\frac{1}{p}}}.$$

For $p = 1$ we get the balance function $\widehat{\mathrm{CC}}_1(A) = \min\{\mathrm{vol}(A), \mathrm{vol}(\overline{A})\}$ of the Cheeger cut considered in [14, 29, 18].
An element u of the subdifferential $\partial \mathrm{CC}_p(f)$ is given as

$$u_i = \frac{e_i a_i |f_i - \mathrm{wmean}_p(f)|^{p-1}}{\Big(\sum_{i=1}^{n} e_i |f_i - \mathrm{wmean}_p(f)|^p\Big)^{\frac{p-1}{p}}}$$

where[1] $a_i \in \mathrm{sign}(f_i - \mathrm{wmean}_p(f))$ with $\langle a, \mathbf{1}\rangle = 0$. Using the optimality condition of $\mathrm{wmean}_p(f)$ it is not difficult to see that such a choice is always possible.

Normalized p-cut (NC_p): The convex function NC_p induces the set function $\widehat{\mathrm{NC}}_p$,

$$\mathrm{NC}_p(f) = \Big(\sum_{i=1}^{n} e_i |f_i - \frac{\langle e, f\rangle}{\mathrm{vol}(V)}|^p\Big)^{\frac{1}{p}} \implies \widehat{\mathrm{NC}}_p(A) = \frac{\big(\mathrm{vol}(A)\,\mathrm{vol}(\overline{A})^p + \mathrm{vol}(A)^p\,\mathrm{vol}(\overline{A})\big)^{\frac{1}{p}}}{\mathrm{vol}(V)}.$$

For $p = 1$ we get the balance function $\widehat{\mathrm{NC}}_1(A) = \frac{2\,\mathrm{vol}(A)\,\mathrm{vol}(\overline{A})}{\mathrm{vol}(V)}$ of the ratio/normalized cut (up to a constant factor).
The subdifferential of NC_p at f is given as

$$\partial \mathrm{NC}_p(f) = (E - \frac{1}{\mathrm{vol}(V)} e e^T) \frac{\mathrm{sign}\,\big(E(f - \frac{1}{\mathrm{vol}(V)}\langle e, f\rangle \mathbf{1})\big)\big|E(f - \frac{1}{\mathrm{vol}(V)}\langle e, f\rangle \mathbf{1})\big|^{p-1}}{\Big(\sum_{i=1}^{n} e_i |f_i - \frac{1}{\mathrm{vol}(V)}\langle e, f\rangle\,|^p\Big)^{\frac{p-1}{p}}},$$

where the power on the right hand side is taken componentwise.

Truncated Cheeger cut (TCC_α): The truncated Cheeger cut of a partition $(A, \overline{A})$ is defined as

$$\frac{\mathrm{cut}(A, \overline{A})}{\min\{\mathrm{vol}(A), \mathrm{vol}(\overline{A}), \alpha\,\mathrm{vol}(V)\}}.$$

Lemma 3.2 *The one-homogeneous convex function $TCC_\alpha(f)$ defined as*

$$\mathrm{TCC}_\alpha(f) = g_{\max,\alpha}(f) - g_{\min,\alpha}(f).$$

[1]The sign function sign is set-valued and defined as $\mathrm{sign}(x) = 1$ if $x > 0$, $\mathrm{sign}(x) = -1$ if $x < 0$ and $\mathrm{sign}(x) \in [-1, 1]$ if $x = 0$.

induces for $\alpha \le \frac{1}{2}$ *the balancing function*

$$\widehat{\mathrm{TCC}}_\alpha(A) = \begin{cases} \mathrm{vol}(A), & \textit{if } \mathrm{vol}(A) \le \alpha\,\mathrm{vol}(V), \\ \alpha\,\mathrm{vol}(V), & \textit{if } \mathrm{vol}(A) > \alpha\,\mathrm{vol}(A) \textit{ and } \mathrm{vol}(\overline{A}) > \alpha\,\mathrm{vol}(A), \\ \mathrm{vol}(\overline{A}), & \textit{if } \mathrm{vol}(\overline{A}) \le \alpha\,\mathrm{vol}(V). \end{cases}$$

Proof: It is easy to check that

$$g_{\max,\alpha}(\mathbf{1}_A) = \begin{cases} \mathrm{vol}(A), & \text{if } \mathrm{vol}(A) \le \alpha\,\mathrm{vol}(V), \\ \alpha\,\mathrm{vol}(V), & \text{else.} \end{cases}$$

$$g_{\min,\alpha}(\mathbf{1}_A) = \begin{cases} \alpha\,\mathrm{vol}(V) - \mathrm{vol}(\overline{A}), & \text{if } \mathrm{vol}(\overline{A}) \le \alpha\,\mathrm{vol}(V), \\ 0, & \text{else.} \end{cases}$$

□

As $\widehat{\mathrm{TCC}}_\alpha(A)$ is a concave function of $\mathrm{vol}(A)$ it is submodular by Proposition 2.1, see also [23]. In the case where $e_i = 1$, $\forall i = 1, \ldots, n$ it has been shown in Proposition 3.1 that TCC_α is the Lovasz-extension of $\widehat{\mathrm{TCC}}_\alpha$. This fact will be used for the hard-balanced cut considered next. Let f be ordered in decreasing order of $u_i = e_i f_i$. Then a subgradient $v \in \partial \mathrm{TCC}_\alpha(f)$ is given as

$$v_r = \begin{cases} e_r, & \text{if } \sum_{i=1}^r e_i < \alpha\,\mathrm{vol}(V), \\ \alpha\,\mathrm{vol}(V) - \sum_{i=1}^r e_i, & \text{if } \sum_{i=1}^r e_i < \alpha\,\mathrm{vol}(V) \le \sum_{i=1}^{r+1} e_i, \\ -\alpha\,\mathrm{vol}(V) + \sum_{i=r}^n e_i, & \text{if } \sum_{i=r-1}^n d_i \ge \alpha\,\mathrm{vol}(V) > \sum_{i=r}^n e_i, \\ -e_r, & \text{if } \sum_{i=r}^n e_i < \alpha\,\mathrm{vol}(V), \\ 0, & \text{else} \end{cases}$$

Hard balanced cut (HBC_K)**:** The hard balanced cut of a partition $(A, \overline{A})$ is defined for $K < |V|/2$,

$$\mathrm{HardCut}_K(A, \overline{A}) = \begin{cases} \mathrm{cut}(A, \overline{A}) & \text{if } \min\{|A|, |\overline{A}|\} \ge K \\ \infty & \text{else} \end{cases}$$

Thus the hard balanced cut problem is

$$\min_{A \subset V} \mathrm{HardCut}_K(A, \overline{A}) = \min_{\min\{|A|,|\overline{A}|\} \ge K} \mathrm{cut}(A, \overline{A}), \tag{2}$$

where one just minimizes the cut subject to hard constraints on the cardinality.

Lemma 3.3 *Let* $e_i = 1$, $i = 1, \ldots, n$ *then for* $K < |V|/2$,

$$\mathrm{HBC}_K(f) = \Big[\big(g_{\max,\frac{K}{|V|}}(f) - g_{\min,\frac{K}{|V|}}(f)\big) - \big(g_{\max,\frac{K-1}{|V|}}(f) - g_{\min,\frac{K-1}{|V|}}(f)\big)\Big],$$

is a non-negative difference of Lovasz extensions of submodular functions and thus a difference of convex functions. $\mathrm{HBC}_K(f)$ *induces the set function*

$$\widehat{\mathrm{HBC}}_K(A) = \begin{cases} 0 & \textit{if } \min\{|A|, |\overline{A}|\} < K, \\ 1 & \textit{else.} \end{cases}.$$

Proof: The result follows from the statements for the truncated Cheeger cut. □

Thus we can write the constrained hard-balanced cut as $\mathrm{HardCut}_K(A, \overline{A}) = \mathrm{cut}(A, \overline{A}) / \widehat{\mathrm{HBC}}(A)$ and can minimize without constraints on the set A.

Hard Cheeger Cut (HCC_K)**:** The hard Cheeger cut has hard constraints on the size of the smallest part of the partition and induces stronger balancing towards balanced partitions than the Cheeger Cut. It is defined via the balancing function

$$\widehat{\mathrm{HCC}}_K(A) = \begin{cases} 0, & \text{if } \min\{|A|, |\overline{A}|\} < K, \\ \min\{|A|, |\overline{A}|\} - (K-1) & \text{else.} \end{cases}.$$

One can write $\mathrm{HCC}_K(A)$ as

$$\widehat{\mathrm{HCC}}_K(A) = \widehat{\mathrm{CC}}_1(A) - \widehat{\mathrm{TCC}}_{\frac{K-1}{n}}(A),$$

As $\widehat{\mathrm{CC}}_1$ and $\widehat{\mathrm{TCC}}_{\frac{K-1}{n}}$ are submodular and their convex Lovasz extension is provided in Proposition 3.1, we can write down directly the corresponding function $\mathrm{HCC}_K(f)$ as a difference of convex functions,

$$\mathrm{HCC}_K(f) = \|f - \mathrm{median}(f)\mathbf{1}\|_1 - \big(g_{\max,\frac{K-1}{|V|}}(f) - g_{\min,\frac{K-1}{|V|}}(f)\big).$$

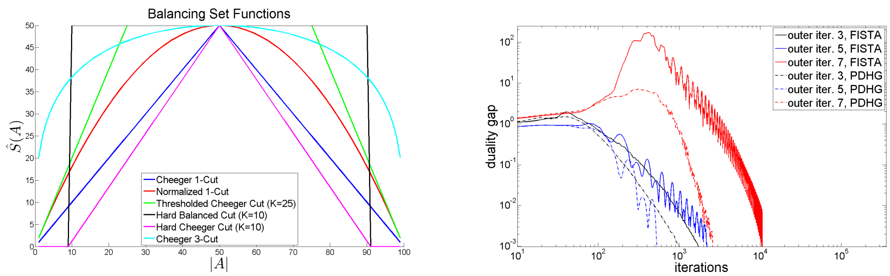

Figure 1: Left: Illustration of different balancing functions (rescaled so that they attain value $|V|/2$ at $|V|/2$). Right: Log-log plot of the duality gap of the inner problem vs. the number of iterations of PDHG (dashed) and FISTA (solid) in outer iterations 3 (black), 5 (blue) and 7 (red) of RatioDCA corresponding to increasing difficulty of the problem. PDHG significantly outperforms FISTA.

4 Minimization of Ratios of Non-negative Differences of Convex Functions

In [14], the problem of computing the optimal Cheeger cut partition is formulated as a nonlinear eigenproblem. Hein and Bühler show that the second eigenvector of the nonlinear 1-graph Laplacian is equal to the indicator function of the optimal partition. In Theorem 3.1, we have generalized this relation considerably. In this section, we discuss the efficient computation of critical points of the continuous ratios of Theorem 3.1. We propose a general scheme called RatioDCA for minimizing ratios of non-negative differences of convex functions and thus generalizes Algorithm 1 of [14] which could handle only ratios of convex functions. As the optimization problem is non-smooth and non-convex, only convergence to critical points can be guaranteed. However, we will show that for every balanced graph cut criterion our algorithm improves a given partition or it terminates directly. Note that such types of algorithms have been considered for specific graph cut criteria [26, 25, 2].

4.1 General Scheme

The continuous optimization problem in Theorem 3.1 has the form

$$\min_{f\in\mathbb{R}^V} \frac{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j|}{S(f)}, \tag{3}$$

where S is one-homogeneous and either convex or the Lovasz extension of a non-negative symmetric set function. By Proposition 2.3 the Lovasz extension of any set function can be written as a difference of one-homogeneous convex functions. Using the fourth property of Proposition 2.2 the Lovasz extension S is non-negative, that is $S(f) \geq 0$ for all $f \in \mathbb{R}^V$. With the algorithm RatioDCA below, we provide a general scheme for the minimization of a ratio $F(f) := R(f)/S(f)$, where R and S are non-negative and one-homogeneous and each can be written as a difference of convex functions: $R(f) = R_1(f) - R_2(f)$ and $S(f) = S_1(f) - S_2(f)$ with R_1, R_2, S_1, S_2 being convex. In our setting $R(f) = R_1(f) = \frac{1}{2}\sum_{i,j=1}^n w_{i,j}|f_i - f_j|$. Let

$$\Phi_{f^k}(u) := R_1(u) - \langle u, r_2(f^k)\rangle + \lambda^k\left(S_2(u) - \langle u, s_1(f^k)\rangle\right).$$

Then we denote the convex optimization problem

$$\min_{\|u\|_2\leq 1} \Phi_{f^k}(u).$$

which has to be solved at each step in RatioDCA as *inner problem*. Note, that the exact form of the constraint set which appears in the inner problem is irrelevant for the properties shown in the next Proposition 4.1 as long as it is compact and contains the origin.

Algorithm RatioDCA – Minimization of a non-negative ratio of 1-homogeneous d.c. functions

Initialization: $f^0 =$ random with $\|f^0\| = 1$, $\lambda^0 = F(f^0)$
repeat
  $s_1(f^k) \in \partial S_1(f^k)$, $r_2(f^k) \in \partial R_2(f^k)$
  $f^{k+1} = \underset{\|u\|_2 \leq 1}{\arg\min} \big\{ R_1(u) - \big\langle u, r_2(f^k) \big\rangle + \lambda^k \left(S_2(u) - \big\langle u, s_1(f^k) \big\rangle \right) \big\}$
  $\lambda^{k+1} = (R_1(f^{k+1}) - R_2(f^{k+1}))/(S_1(f^{k+1}) - S_2(f^{k+1}))$
until $\frac{|\lambda^{k+1} - \lambda^k|}{\lambda^k} < \epsilon$
Output: eigenvalue λ^{k+1} and eigenvector f^{k+1}.

Proposition 4.1 *The sequence f^k produced by RatioDCA satisfies $F(f^k) > F(f^{k+1})$ for all $k \geq 0$ or the sequence terminates.*

Proof: The optimal value of the inner problem is non-positive as

$$\begin{aligned}\Phi_{f^k}(f^k) &= R_1(f^k) - \big\langle f^k, r_2(f^k) \big\rangle + \lambda^k \big(S_2(f^k) - \big\langle f^k, s_1(f^k) \big\rangle \big) \\ &= R_1(f^k) - R_2(f^k) + \lambda^k \big(S_2(f^k) - S_1(f^k) \big) = 0,\end{aligned}$$

where we used $\big\langle f^k, s_1(f^k) \big\rangle = S_1(f^k)$ and $\big\langle f^k, r_2(f^k) \big\rangle = R_2(f^k)$. Moreover, as Φ_{f^k} is 1-homogeneous, the minimum of Φ_{f^k} is always attained at the boundary of the constraint set. If the optimal value is zero, then f^k is a possible minimizer and the sequence terminates. Otherwise the optimal value is negative and at the optimal point f^{k+1} we get

$$\begin{aligned}0 >& R_1(f^{k+1}) - \big\langle f^{k+1}, r_2(f^k) \big\rangle + \lambda^k \big(S_2(f^{k+1}) - \big\langle f^{k+1}, s_1(f^k) \big\rangle \big) \\ \geq& R_1(f^{k+1}) - R_2(f^{k+1}) + \lambda^k \big(S_2(f^{k+1}) - S_1(f^{k+1}) \big)\end{aligned}$$

where we have used that for a positively 1-homogeneous convex function one has for all $f, g \in \mathbb{R}^V$ and $s \in \partial S(g)$,

$$S(f) \geq S(g) + \langle f - g, s \rangle = \langle f, s \rangle \ ,$$

Thus we obtain

$$F(f^k) = \lambda^k > \frac{R_1(f^{k+1}) - R_2(f^{k+1})}{S_1(f^{k+1}) - S_2(f^{k+1})} = F(f^{k+1}).$$

Note, that the constraint of the inner problem plays no role in the proof. However, it is necessary as otherwise the problem would be unbounded from below. □

The sequence $F(f^k)$ is not only monotonically decreasing but converges to a generalized nonlinear eigenvector as introduced in [14].

Theorem 4.1 *Each cluster point f^* of the sequence f^k produced by the RatioDCA is a nonlinear eigenvector with eigenvalue $\lambda^* = \frac{R(f^*)}{S(f^*)} \in \big[0, F(f^0)\big]$ in the sense that it fulfills*

$$0 \in \partial R_1(f^*) - \partial R_2(f^*) - \lambda^* \big(\partial S_1(f^*) - \partial S_2(f^*) \big).$$

If $S_1 - S_2$ is continuously differentiable at f^*, then F has a critical point at f^*.

Proof: By Lemma 4.1 the sequence $F(f^k)$ is monotonically decreasing. By assumption $S = S_1 - S_2$ and $R = R_1 - R_2$ are nonnegative and hence F is bounded below by zero. Thus we have convergence towards a limit

$$\lambda^* = \lim_{k \to \infty} F(f^k) \ .$$

Note that $\|f^k\|_2^2 \leq 1$ for every k, thus the sequence f^k is contained in a compact set, which implies that there exists a subsequence f^{k_j} converging to some element f^*. As the sequence $F(f^{k_j})$ is a subsequence of a convergent sequence, it has to converge towards the same limit, hence also

$$\lim_{j \to \infty} F(f^{k_j}) = \lambda^* \ .$$

As shown before, the objective of the inner optimization problem is nonpositive at the optimal point. Assume now that $\min_{\|f\|_2^2 \leq 1} \Phi_{f^*}(f) < 0$. Then using the proof of Lemma 4.1, the vector $f^{**} = \arg\min_{\|f\|_2^2 \leq 1} \Phi_{f^*}(f)$ satisfies

$$R_1(f^{**}) - R_2(f^{**}) < \lambda^* \big(S_1(f^{**}) - S_2(f^{**}) \big) \ ,$$

where we used the definition of the subdifferential and the 1-homogeneity of S. Hence

$$F(f^{**}) < \lambda^* = F(f^*) \ ,$$

which is a contradiction to the fact that the sequence $F(f^k)$ has converged to λ^*. Thus we must have $\min_{\|f\|_2^2 \leq 1} \Phi_{f^*}(f) = 0$, i.e. the function $\Phi_{f^*}(f)$ is nonnegative in the unit ball. Using the fact that for any $\alpha \geq 0$,

$$\Phi_{f^*}(\alpha f) = \alpha \Phi_{f^*}(f) \ ,$$

we can even conclude that the function $\Phi_{f^*}(f)$ is nonnegative everywhere, and thus $\min_f \Phi_{f^*}(f) = 0$. Note that $\Phi_{f^*}(f^*) = 0$, which implies that f^* is a global minimizer of Φ_{f^*}, and hence

$$0 \in \partial \Phi_{f^*}(f^*) = \partial R_1(f^*) - r_2(f^*) + \lambda^* \big(\partial S_2(f^*) - s_1(f^*) \big) \ ,$$

which implies that f^* is an eigenvector with eigenvalue λ^*. Note that this argument was independent of the choice of the subsequence, thus every convergent subsequence converges to an eigenvector with the same eigenvalue λ^*. Clearly we have $\lambda^* \leq F(f^0)$. □

In the balanced graph cut problem (3) we minimize implicitly over non-constant functions. Thus it is important to guarantee that the RatioDCA for this particular problem always converges to a non-constant vector.

Lemma 4.1 *For every balanced graph cut problem, the RatioDCA converges to a non-constant f^* given that the initial vector f^0 is non-constant.*

Proof: For every convex function $S : \mathbb{R}^n \to \mathbb{R}$ satisfying the conditions of Lemma 3.1, the subdifferential is given as

$$\partial S(f) = \{ u \in \mathbb{R}^n \,|\, u \in U, \quad \langle u, f \rangle = S(f) \},$$

where $\langle u, \mathbf{1} \rangle = 0$ for all $u \in U$. Thus it always holds $\langle s_1(f^k), \mathbf{1} \rangle = 0$. Moreover, $R(f) = \frac{1}{2} \sum_{i,j=1}^n w_{ij} |f_i - f_j|$ is invariant under addition of a constant. Together with $R_2 \equiv 0$, we therefore derive that the objective is invariant under addition of the constant vector. Moreover, the constant vector always attains objective value zero. As derived in Proposition 4.1 the optimal objective value is always non-negative as long as descent is possible and thus RatioDCA converges towards a non-constant vector. Note, moreover that if the method terminates and the objective value is zero, the previous non-constant iterate also is a minimizer. □

Now we are ready to state the following key property of our balanced graph clustering algorithm.

Theorem 4.2 *Let $(A, \overline{A})$ be a given partition of V and let $S : V \to \mathbb{R}_+$ satisfy one of the conditions stated in Theorem 3.1. If one uses as initialization of RatioDCA, $f^0 = \mathbf{1}_A$, then either RatioDCA terminates after one step or it yields an f^1 which after optimal thresholding as in Theorem 3.1 gives a partition $(B, \overline{B})$ which satisfies*

$$\frac{\mathrm{cut}(B, \overline{B})}{\hat{S}(B)} < \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)}.$$

Proof: As $f^0 = \mathbf{1}_A$ we get with

$$R(\mathbf{1}_A) = \frac{1}{2} \sum_{i,j=1}^n w_{ij} |\mathbf{1}_A(i) - \mathbf{1}_A(j)| = \mathrm{cut}(A, \overline{A}), \text{ and } S(\mathbf{1}_A) = \hat{S}(A),$$

for the initial value of the ratio $F(f^0) = \mathrm{cut}(A, \overline{A}) / \hat{S}(A)$. Proposition 4.1 now states that either RatioDCA produces a f^1 with $F(f^1) < F(f^0)$ or it terminates. If the second case happens we are done, let us consider the first case. By Theorem 3.1 there exists a set C^* in the set of sets

$C_i = \{j \in V \,|\, f_j^1 > f_i^1\}$, $i = 1, \ldots, n-1$ such that $F(f^1) \geq F(\mathbf{1}_{C^*})$. In total we get that there exists a partition $(C^*, \overline{C^*})$ such that

$$\mathrm{cut}(C^*, \overline{C^*})/\hat{S}(C^*) = F(\mathbf{1}_{C^*}) \leq F(f^1) < F(f^0) = F(\mathbf{1}_A) = \mathrm{cut}(A, \overline{A})/\hat{S}(A).$$

□

The above "improvement theorem" implies that we can use the result of any other graph partitioning method as initialization. In particular, we can always improve the result of spectral clustering.

4.2 Solution of the Convex Inner Optimization Problems

The performance of RatioDCA depends heavily on how fast we can solve the corresponding inner problem. We propose to use a primal-dual algorithm for the inner problem and show experimentally that this approach yields faster convergence than the FISTA method of [5] which was applied in [14]. Let us restrict our attention to the case where $R(f) = R_1(f) = \frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j|$ and $S_2 = 0$. In other words, we apply the RatioDCA algorithm to (3) with $S = S_1$ which is what we need, e.g., for the tight relaxations of the Cheeger cut, normalized cut and truncated Cheeger cut families. Hence, the inner problem of the RatioDCA algorithm (line 4) has the form

$$f^{k+1} = \underset{\|u\|_2 \leq 1}{\arg\min} \{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j| - \lambda^k \langle u, s_1(f^k)\rangle\}. \tag{4}$$

Recently, Arrow-Hurwicz-type primal-dual algorithms have become popular, e.g., in image processing, to solve problems whose objective function consists of the sum of convex terms, cf., e.g., [10, 7]. We propose to use the following primal-dual algorithm of [7] where it is referred to as Algorithm 2. We call this method a *primal-dual hybrid gradient algorithm* (PDHG) here since this term is used for similar algorithms in the literature. Note that the operator $P_{\|\cdot\|_\infty \leq 1}$ in the first step is the componentwise projection onto the interval $[-1, 1]$. For the sake of readability, we define the linear operator $B : \mathbb{R}^V \to \mathbb{R}^E$ by $Bu = (w_{ij}(u_i - u_j))_{i,j=1}^n$ and its transpose is then $B^{\mathrm{T}}\beta = \left(\sum_{j=1}^n w_{ij}(\beta_{i,j} - \beta_{j,i})\right)_{i=1}^n$.

Algorithm PDHG – Solution of the inner problem of RatioDCA for (3) and S convex

1: **Initialization:** $u^0, \bar{u}^0, \beta^0 = 0, \gamma, \sigma_0, \tau_0 > 0$ with $\sigma_0\tau_0 \leq 1/\|B\|_2^2$
2: **repeat**
3: $\quad \beta^{l+1} = P_{\|\cdot\|_\infty \leq 1}(\beta^l + \sigma_l B\bar{u}^l)$
4: $\quad u^{l+1} = \frac{1}{1+\tau_l}\big(u^l - \tau_l(B^{\mathrm{T}}\beta^{l+1} - 2\lambda^k s_1(f^k))\big)$
5: $\quad \theta_l = 1/\sqrt{1 + 2\gamma\tau_l}, \quad \tau_{l+1} = \theta_l\tau_l, \quad \sigma_{l+1} = \sigma_l/\theta_l$
6: $\quad \bar{u}^{l+1} = u^{l+1} + \theta_l(u^{l+1} - u^l)$
7: **until** duality gap $< \epsilon$
8: **Output:** $f^{k+1} \approx u^{l+1}/\|u^{l+1}\|_2$

Although PDHG and FISTA have the same guaranteed converges rates of $\mathcal{O}(1/l^2)$, our experiments show that for clustering applications, PDHG can outperform FISTA substantially. In Fig.1, we illustrate this difference on a toy problem. Note that a single step takes about the same computation time for both algorithms so that the number of iterations is a valid criterion for comparison. Let us now turn to the inner problem of RatioDCA for the case where $S_2 \neq 0$. As an example, we consider our tight relaxation of the hard balanced cut. According to Table 1, S now has the form

$$S(f) = \underbrace{g_{\max,\frac{K}{|V|}}(f) - g_{\min,\frac{K}{|V|}}(f)}_{=S_1} - \underbrace{\left(g_{\max,\frac{K-1}{|V|}}(f) - g_{\min,\frac{K-1}{|V|}}(f)\right)}_{=S_2}.$$

Let us introduce the shorter notation $g := g_{\max,\frac{K-1}{|V|}}$ for the function that yields the sum of the $K-1$ largest components of the input vector. Observe that in this notation,

$$S_2(f) = g(f) + g(-f).$$

Since both g and $g(-\cdot)$ are convex and thus S_2 is convex. So, the convex inner problem we have to solve in RatioDCA is given by

$$f^{k+1} = \underset{\|u\|_2 \le 1}{\arg\min} \{ \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} |f_i - f_j| + \lambda^k \left(g(u) + g(-u) - \left\langle u, s_1(f^k) \right\rangle \right) \}. \tag{5}$$

In the algorithm given below, the following versions of the continuous knapsack problem arise as subproblems

$$\tilde{u} = \underset{u \in \mathbb{R}^n}{\arg\min} \frac{1}{2} \|u - v\|_2^2 \quad \text{subject to} \quad 0 \le u \le 1,\ \mathbf{1}^{\mathrm{T}} u = K - 1,$$

where $v \in \mathbb{R}^n$ is some given vector. Clearly, $\tilde{u}$ can be written as the orthogonal projection onto the set $C := \{u \,:\, 0 \le u \le 1,\ \mathbf{1}^{\mathrm{T}} u = K - 1\}$, i.e.,

$$\tilde{u} = P_C(v).$$

Although there is no closed-form solution, P_C can be computed in linear time. In our implementation, we use the algorithm proposed in [19]. We obtain the following PDHG algorithm:

Algorithm PDHG – Solution of the inner problem of RatioDCA for the tight relaxation of the hard balanced cut

---

Input: accuracy ϵ
Initialization: $u^0, \bar{u}^0, \beta^0 = 0, \gamma, \sigma_0, \tau_0 > 0$ with $\sigma_0 \tau_0 \le 1/\|(I, B^{\mathrm{T}})^{\mathrm{T}}\|_2^2$
repeat
  $\beta_1^{l+1} = -2\lambda^k P_C(-(\beta_1^l + \sigma_l \bar{u}^l)/(2\lambda^k))$
  $\beta_2^{l+1} = P_{\|\cdot\|_\infty \le 1}(\beta_2^l + \sigma_l B \bar{u}^l)$
  $y = \frac{1}{1+\tau_l}(u^l - \tau_l(\beta_1^{l+1} + B^{\mathrm{T}} \beta_2^{l+1}) + 2\tau_l \lambda^k s_1(f^k))$
  $u^{l+1} = y - \frac{2\tau_l \lambda^k}{1+\tau_l} P_C(\frac{1+\tau_l}{2\tau_l \lambda^k} y)$
  $\theta_l = 1/\sqrt{1 + 2\gamma\tau_l}, \quad \tau_{l+1} = \theta_l \tau_l, \quad \sigma_{l+1} = \sigma_l / \theta_l$
  $\bar{u}^{l+1} = u^{l+1} + \theta_l (u^{l+1} - u^l)$
until duality gap $< \epsilon$
Output: $f^{k+1} \approx u^{l+1} / \|u^{l+1}\|_2$

General structure of PDHG. We have described PDHG algorithms for two different inner problems which arise in RatioDCA. The aim of this paragraph is to show the general form of the PDHG algorithm and to explain how the preceding two important examples can be derived. Let us consider a general problem of the form

$$\min_{u \in \mathbb{R}^n} \{ G(u) + \sum_{i=1}^{L} \Phi_i(A_i u) \}, \tag{6}$$

where $G : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\Phi_i : \mathbb{R}^{m_i} \to \mathbb{R} \cup \{+\infty\}$, for $i = 1, \ldots, L$ are proper, convex and lower-semicontinuous functions and $A_i \in \mathbb{R}^{m_i \times n}$, $i = 1, \ldots, L$ are linear mappings. The dual problem can be written in terms of the corresponding conjugate functions as

$$-\min_{\beta_i \in \mathbb{R}^{m_i}} \{ G^*(-\sum_{i=1}^{L} A_i^{\mathrm{T}} \beta_i) + \sum_{i=1}^{L} \Phi_i^*(\beta_i) \}. \tag{7}$$

For a proper, convex and lower-semicontinuous function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and a vector $v \in \mathbb{R}$, the problem

$$\operatorname{prox}_h(v) = \underset{u \in \mathbb{R}^n}{\arg\min} \{ \frac{1}{2} \|u - v\|_2^2 + h(u) \}.$$

is called a proximal problem. Moreover, we refer to $\operatorname{prox}_h(v)$ as the proximum of h at v. The general PDHG algorithm has the following form, cf., [7].

Algorithm PDHG

1: **Input:** accuracy ϵ
2: **Initialization:** $u^0, \bar{u}^0, \beta^0 = 0, \gamma, \sigma_0, \tau_0 > 0$ with $\sigma_0\tau_0 \leq 1/\|(A_1^{\mathrm{T}}, \ldots, A_L^{\mathrm{T}})^{\mathrm{T}}\|_2^2$
3: **repeat**
4: $\quad\beta_i^{l+1} = \mathrm{prox}_{\sigma_l \Phi_i^*}(\beta_i^l + \sigma_l A_i \bar{u}^l) \quad$ for $i = 1, \ldots, L$
5: $\quad u^{l+1} = \mathrm{prox}_{\tau_l G}(u^l - \tau_l \sum_{i=1}^L A_i^{\mathrm{T}} \beta_i^{l+1})$
6: $\quad\theta_l = 1/\sqrt{1+2\gamma\tau_l}, \quad \tau_{l+1} = \theta_l \tau_l, \quad \sigma_{l+1} = \sigma_l/\theta_l$
7: $\quad\bar{u}^{l+1} = u^{l+1} + \theta_l(u^{l+1} - u^l)$
8: **until** duality gap $< \epsilon$
9: **Output:** u^{l+1}

Notice that this algorithm exploits the additive structure of the objective function. It decouples G, $\Phi_1, \ldots, \Phi_L$ in the sense that in the subproblems of lines 4 and 5 we have to solve proximal problems with respect to only one of these functions at a time. Consequently, the PDHG algorithm performs very well for many applications and is relatively simple to implement. The following convergence result was shown in [7].

Theorem 4.3 *Assume that the primal problem* (6) *and the dual problem* (7) *have a solution and that the duality gap is zero. Then, the sequences* $(u^l)_{l\in\mathbb{N}}$ *and* $(\beta^l)_{l\in\mathbb{N}}$ *generated by PDHG converge to a solution of the primal and dual problem, respectively.*

We now illustrate how to derive the PDHG algorithms for the two inner problems considered above from the general form of the algorithm. We start with problem (4). It turned out in our experiments that PDHG is slightly faster if applied to the primal problem

$$\min_{u\in\mathbb{R}^n}\{\sum_{i,j=1}^n w_{ij}|f_i - f_j| + \frac{1}{2}\|u - 2\lambda^k s_1(f^k)\|_2^2\}. \tag{8}$$

It is straightforward to see that (8) and (4) are equivalent, i.e., we simply have to normalize the solution of (8) to obtain a solution of (4). The PDHG algorithm then follows from the general version by setting $L = 1$ and

$$\begin{aligned} G &:= \frac{1}{2}\|\cdot - 2\lambda^k s_1(f^k)\|_2^2, \\ A_1 &:= B, \\ \Phi_1 &:= \|\cdot\|_1. \end{aligned}$$

Since $(\|\cdot\|_1)^*$ is equal to the indicator function $\iota_{\|\cdot\|_\infty \leq 1}$, we get $\mathrm{prox}_{\sigma_l \Phi_1^*} = P_{\|\cdot\|_\infty \leq 1}$, i.e., the componentwise projection onto the interval $[-1, 1]$.

Let us now consider the inner problem (5) which appears in RatioDCA for the tight relaxation of the hard balanced cut. Clearly, the objective function is the sum of proper, convex and lower-semicontinuous terms and we can apply PDHG. Instead of (5), we again consider a slightly different but equivalent problem

$$\min_{u\in\mathbb{R}^n}\{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j| + 2\lambda^k\left(g(u) + g(-u)\right) + \frac{1}{2}\|u - \lambda^k s_1(f^k)\|_2^2\}. \tag{9}$$

The PDHG algorithm presented above can be obtained by setting in the general version $L = 2$ and

$$\begin{aligned} G &:= \frac{1}{2}\|\cdot - 2\lambda^k s_1(f^k)\|_2^2 + 2\lambda^k g, \\ A_1 &:= I, \quad A_2 = B, \\ \Phi_1 &:= 2\lambda^k g(-\cdot), \quad \Phi_2 := \|\cdot\|_1. \end{aligned}$$

The subproblem in line 5 of the resulting PDHG algorithm consists in computing the proximum of the conjugate function $(\|\cdot\|_1)^*$, i.e., the orthogonal projection $P_{\|\cdot\|_\infty \leq 1}$. Moreover, we now have to deal with proximal problems with respect to scaled versions of g and g^*. Observe that $g^* = \iota_C$

where ι_C is the indicator function of $C := \{u \in \mathbb{R}^n \,:\, 0 \leq u \leq 1,\; \mathbf{1}^{\mathrm{T}} u = K-1\}$. Consequently, the proximum of g^* is the orthogonal projection onto C. Concerning the proximum of g note that for any constant $\xi > 0$ we have

$$\mathrm{prox}_{\xi g}(u) = u - \xi \mathrm{prox}_{g^*/\xi}(u/\xi) = u - \xi P_C(u/\xi).$$

These results are used in the subproblems of line 4 and 7 of the PDHG algorithm.

5 Experiments

In a first experiment, we study the influence of the different balancing criteria on the obtained clustering. The data is a Gaussian mixture in $\mathbb{R}^{20}$ where the projection onto the first two dimensions is shown in Figure 2 - the remaining 18 dimensions are just noise. The distribution of the 2000 points is [1200,600,200]. A symmetric k-NN-graph with $k = 20$ is built with Gaussian weights $e^{-\frac{2\|x-y\|^2}{\max\{\sigma_{x,k}^2,\sigma_{y,k}^2\}}}$ where $\sigma_{x,k}$ is the k-NN distance of point x. For better interpretation, we report

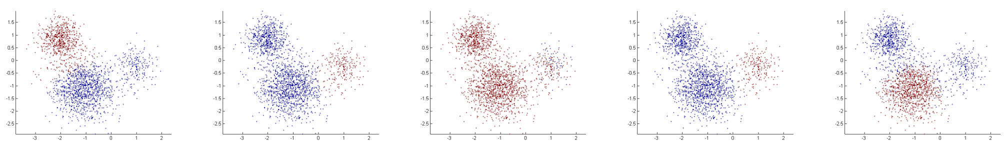

Figure 2: From left to right: Cheeger 1-cut, Normalized 1-cut, truncated Cheeger cut (TCC), hard balanced cut (HBC), hard Cheeger cut (HCC). The criteria are the normalized ones, i.e., the vertex weights are $e_i = d_i$.

all resulting partitions with respect to all balanced graph cut criteria, cut and the size of the largest component in the following table. The parameter for truncated, hard Cheeger cut and hard balanced cut is set to $K = 200$. One observes that the normalized 1-cut results in a less balanced partition but with a much smaller cut than the Cheeger 1-cut, which is itself less balanced than the hard Cheeger cut. The latter is fully balanced but has an even higher cut. The truncated Cheeger cut has a smaller cut than the hard balanced cut but its partition is not feasible. Note that the hard balanced cut is similar to the normalized 1-cut but achieves smaller cut at the prize of a larger maximal component. Thus, the example nicely shows how the different balance criterion influence the final partition.

Criterion \ Obj.	Cut	$\max\{\lvert A\rvert, \lvert\overline{A}\rvert\}$	Ch. 1-cut	N. 1-cut	TCC_{200}	HBC_{200}	HCC_{200}
Cheeger 1-cut	408.4	1301	0.099	0.079	2.042	408.4	0.817
Norm. 1-cut	178.3	1775	0.132	0.075	0.892	178.3	6.858
Trunc. Ch. cut	153.6	1945	0.513	0.263	0.768	∞	∞
Hard bal. cut	175.4	1785	0.134	0.076	0.877	175.4	10.96
Hard Ch. cut	639.2	1000	0.119	0.115	3.196	639.2	0.798

Next we perform unnormalized 1-spectral clustering on the full USPS, normal and extended[2] MNIST-datasets (resp. 9298, 70000 and 630000 points) in the same setting as in [14] with no vertex weights, that is $e_i = 1, \forall i \in V$. As clustering criterion for multi-partitioning we use the multicut version of the normalized 1-cut, given as $\mathrm{RCut}(C_1, \ldots, C_M) = \sum_{i=1}^{M} \frac{\mathrm{cut}(C_i, \overline{C_i})}{|C_i|}$. We successively subdivide clusters until the desired number of clusters ($M = 10$) is reached. This recursive partitioning scheme is used for all methods. In [14] the Cheeger 1-cut has been used which is not compatible with the multi-cut criterion. We expect that using the normalized 1-cut for the bipartitioning steps we should get better results. The results of the other methods for USPS and MNIST (normal) are taken from [14]. Each bipartitioning step is initialized randomly. Out of 100 obtained multi-partitionings we report the results of the best clustering with respect to the multi-cut criterion. The next table shows the obtained RCut and errors.

[2]The extended MNIST dataset is generated by translating each original input image of MNIST by one pixel (8 directions).

Vertices/Edges		N. 1-cut	Ch. 1-cut[14]	S.&B.[29]	1.1-SCl [6]	Standard spectral
USPS	Rcut	0.6629	0.6661	0.6663	0.6676	0.8180
9K/272K	Error	0.1301	0.1349	0.1309	0.1308	0.1686
MNIST (Normal)	Rcut	0.1499	0.1507	0.1545	0.1529	0.2252
70K/1043K	Error	0.1236	0.1244	0.1318	0.1293	0.1883
MNIST (Ext)	Rcut	0.0996	0.0997	–	–	0.1594
630K/9192K	Error	0.1180	0.1223	–	–	0.2297

We see for all datasets improvements in the obtained cut. Also a slight decrease in the obtained error can be observed. The improvements are not so drastic as the clustering is already very good. The problem is that for both datasets one digit is split (0) and two are merged (4 and 9) resulting in seemingly large errors. Similar results hold for the extended MNIST dataset. Note that the resulting error is comparable to recently reported results on semi-supervised learning [20].

References

[1] N. Alon and V. D. Milman. λ_1, isoperimetric inequalities for graphs, and superconcentrators. *J. Combin. Theory Ser. B*, 38(1):73–88, 1985.

[2] R. Andersen and K. Lang. An algorithm for improving graph partitions. In *Proc. of the 19th ACM-SIAM Symposium on Discrete Algorithms (SODA 2008)*, pages 651–660, 2008.

[3] S. Arora, J. R. Lee, and A. Naor. Expander flows, geometric embeddings and graph partitioning. In *Proc. 36th Annual ACM Symp. on Theory of Computing (STOC)*, pages 222–231. ACM, 2004.

[4] F. Bach. Convex analysis and optimization with submodular functions, 2010. arXiv:1010.4207v2.

[5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2:183–202, 2009.

[6] T. Bühler and M. Hein. Spectral clustering based on the graph p-Laplacian. In L. Bottou and M. Littman, editors, *Proc. of the 26th Int. Conf. on Machine Learning (ICML)*, pages 81–88. Omnipress, 2009.

[7] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[8] F. Chung. *Spectral Graph Theory*. AMS, Providence, RI, 1997.

[9] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.*, 17:420–425, 1973.

[10] E. Esser, X. Zhang, and T. F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM J. Imaging Sciences*, 3(4):1015–1046, 2010.

[11] S. Fujishige. *Submodular functions and optimization*, volume 58 of *Annals of Discrete Mathematics*. Elsevier B. V., Amsterdam, second edition, 2005.

[12] Stephen Guattery and Gary L. Miller. On the quality of spectral separators. *SIAM Journal on Matrix Analysis and Applications*, 19:701–719, 1998.

[13] L. Hagen and A. B. Kahng. Fast spectral methods for ratio cut partitioning and clustering. *Proc. IEEE Intl. Conf. on Computer-Aided Design*, pages 10–13, November 1991.

[14] M. Hein and T. Bühler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 847–855, 2010.

[15] J.-B. Hiriart-Urruty. Generalized differentiability, duality and optimization for problems dealing with differences of convex functions. In *Convexity and duality in optimization*, pages 37–70. 1985.

[16] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

[17] C. Houdré. Mixed and Isoperimetric Estimates on the Log-Sobolev Constants of Graphs and Markov Chains. *Combinatorica*, 21:489–513, 2001.

[18] Y. Kawahara, K. Nagano, and Y. Okamoto. Submodular fractional programming for balanced clustering. *Pattern Recognition Letters*, 32:235–243, 2011.

[19] K. C. Kiwiel. On linear-time algorithms for the continuous quadratic knapsack problem. *Journal of Optimization Theory and Applications*, 134(3):549–554, 2007.

[20] W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *Proc. of the 27th Int. Conf. on Machine Learning (ICML)*, 2010.

[21] L. Lovász. Submodular functions and convexity. In *Mathematical programming: the state of the art (Bonn, 1982)*, pages 235–257. Springer, Berlin, 1983.

[22] M. Maier, U. von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 1025 – 1032, 2009.

[23] K. Nagano, Y. Kawahara, and S. Iwata. Minimum average cost clustering. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 1759–1767, 2010.

[24] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *21st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[25] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[26] S. B. Patkar and H. Narayanan. Improving graph partitions using submodular functions. *Discrete Appl. Math.*, 131(2):535–553, 2003.

[27] A. Pothen, H. D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11:430 – 452, 1990.

[28] O. S. Rothaus. Analytic inequalities, Isoperimetric Inequalities and Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 64:296–313, 1985.

[29] A. Szlam and X. Bresson. Total variation and Cheeger cuts. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1039–1046. Omnipress, 2010.

[30] J.-P. Tillich. Edge isoperimetric inequalities for product graphs. *Discrete Mathematics*, 213:291–320, 2000.

[31] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.