Spectral Clustering based on the graph p-Laplacian

Thomas Bühler TB@CS.UNI-SB.DE
Matthias Hein HEIN@CS.UNI-SB.DE
Saarland University, Computer Science Department, Campus E1 1, 66123 Saarbrücken, Germany

Abstract

We present a generalized version of spectral clustering using the graph p-Laplacian, a nonlinear generalization of the standard graph Laplacian. We show that the second eigenvector of the graph p-Laplacian interpolates between a relaxation of the normalized and the Cheeger cut. Moreover, we prove that in the limit as $p \to 1$ the cut found by thresholding the second eigenvector of the graph p-Laplacian converges to the optimal Cheeger cut. Furthermore, we provide an efficient numerical scheme to compute the second eigenvector of the graph p-Laplacian. The experiments show that the clustering found by p-spectral clustering is at least as good as normal spectral clustering, but often leads to significantly better results.

1. Introduction

In recent years, spectral clustering has become one of the major clustering methods. The reasons are its generality, efficiency and its rich theoretical foundation. Spectral clustering can be applied to any kind of data with a suitable similarity measure and the clustering can be computed for millions of points. The theoretical background includes motivations based on balanced graph cuts, random walks and perturbation theory. We refer to (von Luxburg, 2007) and references therein for a detailed introduction to various aspects of spectral clustering.

In this paper our focus lies on the motivation of spectral clustering as a relaxation of balanced graph cut criteria. It is well known that the second eigenvectors of the unnormalized and normalized graph Laplacians correspond to relaxations of the ratio cut (Hagen & Kahng, 1991) and normalized cut (Shi & Malik, 2000). There are also relaxations of balanced graph cut criteria based on semi-definite programming (De Bie & Cristianini, 2006), which turn out to be better than the standard spectral ones but are computationally more expensive.

In this paper we establish a connection between the Cheeger cut and the second eigenvector of the graph p-Laplacian, a nonlinear generalization of the graph Laplacian. A p-Laplacian which differs slightly from the one used in this paper has been used for semi-supervised learning by Zhou and Schölkopf (2005). Our main motivation for the use of eigenvectors of the graph p-Laplacian was the generalized isoperimetric inequality of Amghibech (2003) which relates the second eigenvalue of the graph p-Laplacian to the optimal Cheeger cut. The isoperimetric inequality becomes tight as $p \to 1$, so that the second eigenvalue converges to the optimal Cheeger cut value. In this article we extend the isoperimetric inequality of Amghibech to the unnormalized graph p-Laplacian. However, our key result is to show that the cut obtained by thresholding the second eigenvector of the p-Laplacian converges to the optimal Cheeger cut as $p \to 1$, which provides theoretical evidence that p-spectral clustering is superior to the standard case. Moreover, we provide an efficient algorithmic scheme for the (approximate) computation of the second eigenvector of the p-Laplacian and the resulting clustering. This allows us to do p-spectral clustering also for large scale problems. Our experimental results show that as one varies p from 2 (standard spectral clustering) to 1 the value of the Cheeger cut obtained by thresholding the second eigenvector of the graph p-Laplacian is always decreasing.

In Section 2, we review balanced graph cut criteria. In Section 3, we introduce the graph p-Laplacian followed by the definition of eigenvectors of nonlinear operators. In Section 4, we provide the theoretical key result relating the cut found by thresholding the second eigenvector of the graph p-Laplacian to the optimal Cheeger cut. The algorithmic scheme is presented in Section 5

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

and extensive experiments on various datasets, including large scale ones, are given in Section 6.

2. Balanced graph cut criteria

Given a set of points in a feature space and a similarity measure, the data can be transformed into a weighted, undirected graph G, where the vertices V represent the points in the feature space and the positive edge weights W encode the similarity of pairs of points. A clustering of the points is then equivalent to a partition of V into subsets $C_1, \ldots, C_k$ (which will be called clusters in the following). The usual objective for such a partitioning is to have high within-cluster similarity and low inter-cluster similarity. Additionally, the clusters should be balanced in the sense that the "size" of the clusters should not differ too much. All the graph cut criteria presented in this section implement these objectives with slightly different emphasis on the individual properties.

Before the definition of the balanced graph cut criteria, we have to introduce some notation. The number of points is denoted by $n = |V|$ and the complement of a set $A \subset V$ is written as $\overline{A} = V \backslash A$. The degree function $d : V \to \mathbb{R}$ of the graph is given as $d_i = \sum_{j=1}^{n} w_{ij}$ and the cut of $A \subset V$ and $\overline{A}$ is defined as

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

Moreover, we denote by $|A|$ the cardinality of the set A and by $\mathrm{vol}(A) = \sum_{i \in A} d_i$ the volume of A. In the balanced graph cut criteria one either tries to balance the cardinality or the volume of the clusters.

The ratio cut $\mathrm{RCut}(C, \overline{C})$ (Hagen & Kahng, 1991) and the normalized cut $\mathrm{NCut}(C, \overline{C})$ (Shi & Malik, 2000) for a partition of V into $C, \overline{C}$ are defined as

$$\begin{aligned} \mathrm{RCut}(C, \overline{C}) &= \frac{\mathrm{cut}(C, \overline{C})}{|C|} + \frac{\mathrm{cut}(C, \overline{C})}{|\overline{C}|}, \\ \mathrm{NCut}(C, \overline{C}) &= \frac{\mathrm{cut}(C, \overline{C})}{\mathrm{vol}(C)} + \frac{\mathrm{cut}(C, \overline{C})}{\mathrm{vol}(\overline{C})}. \end{aligned}$$

A slightly different balancing behavior is induced by the corresponding ratio Cheeger cut $\mathrm{RCC}(C, \overline{C})$ and normalized Cheeger cut $\mathrm{NCC}(C, \overline{C})$ defined as

$$\begin{aligned} \mathrm{RCC}(C, \overline{C}) &= \frac{\mathrm{cut}(C, \overline{C})}{\min\{|C|, |\overline{C}|\}}, \\ \mathrm{NCC}(C, \overline{C}) &= \frac{\mathrm{cut}(C, \overline{C})}{\min\{\mathrm{vol}(C), \mathrm{vol}(\overline{C})\}}. \end{aligned}$$

One has the following simple relation between the normalized cut $\mathrm{NCut}(C, \overline{C})$ and the normalized Cheeger cut $\mathrm{NCC}(C, \overline{C})$:

$$\mathrm{NCC}(C, \overline{C}) \;\leq\; \mathrm{NCut}(C, \overline{C}) \;\leq\; 2\,\mathrm{NCC}(C, \overline{C}).$$

The analogous result holds for the ratio cut $\mathrm{RCut}(C, \overline{C})$ and the ratio Cheeger cut $\mathrm{RCC}(C, \overline{C})$. It is known that finding the global optimum of all these balanced graph cut criteria is NP-hard, see (von Luxburg, 2007). In Section 4, we will show how spectral relaxations of these criteria are related to the eigenproblem of the graph p-Laplacian.

Up to now the cuts are just defined for a partition of V into two sets. For a partition of V into k sets $C_1, \ldots, C_k$ the ratio and normalized cut can be generalized (von Luxburg, 2007) as

$$\mathrm{RCut}(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, \overline{C_i})}{|C_i|}, \qquad (1)$$

$$\mathrm{NCut}(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, \overline{C_i})}{\mathrm{vol}(C_i)}. \qquad (2)$$

There seems to exist no generally accepted multi-partition version of the Cheeger cuts. We come back to this issue in Section 5, when we discuss how to get multiple clusters using the second eigenvector of the graph p-Laplacian.

3. The graph p-Laplacian

It is well known, see e.g. (Hein et al., 2007), that the standard graph Laplacian Δ_2 can be defined as the operator which induces the following quadratic form for a function $f : V \to R$:

$$\langle f, \Delta_2 f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2.$$

For the standard inner product one gets the unnormalized graph Laplacian $\Delta_2^{(\mathrm{u})}$ which in matrix notation is given as $\Delta_2^{(\mathrm{u})} = D - W$, and for the weighted inner product, $\langle f, g \rangle = \sum_{i=1}^{n} d_i f_i g_i$, one obtains the normalized[1] graph Laplacian $\Delta_2^{(\mathrm{n})}$ given as $\Delta_2^{(\mathrm{n})} = \mathbb{I} - D^{-1}W$. One can ask now if there exists an operator Δ_p which induces the general form (for $p > 1$),

$$\langle f, \Delta_p f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} |f_i - f_j|^p.$$

It turns out that this question can be answered positive, see (Amghibech, 2003). The resulting operator

[1] Note that our notation differs from the one in (Hein et al., 2007) where they denote our normalized graph Laplacian as "random walk graph Laplacian".

Δ_p is the graph p-Laplacian (which we abbreviate as p-Laplacian if no confusion is possible). Similar to the graph Laplacian we obtain, dependent on the choice of the inner product, the unnormalized and normalized p-Laplacian $\Delta_p^{(\mathrm{u})}$ and $\Delta_p^{(\mathrm{n})}$. Let $i \in V$, then

$$(\Delta_p^{(\mathrm{u})} f)_i = \sum_{j \in V} w_{ij}\, \phi_p\,(f_i - f_j),$$
$$(\Delta_p^{(\mathrm{n})} f)_i = \frac{1}{d_i} \sum_{j \in V} w_{ij}\, \phi_p\,(f_i - f_j).$$

where $\phi_p : \mathbb{R} \to \mathbb{R}$ is defined for $x \in \mathbb{R}$ as

$$\phi_p(x) = |x|^{p-1}\, \mathrm{sign}(x).$$

Note that $\phi_2(x) = x$, so that we recover the standard graph Laplacians for $p = 2$. In general, the p-Laplacian is a nonlinear operator: $\Delta_p(\alpha f) \neq \alpha\, \Delta_p f$ for $\alpha \in \mathbb{R}$.

3.1. Eigenvalues and eigenvectors of the graph p-Laplacian

Since our goal is to use the p-Laplacian for spectral clustering, the natural question arises how one can define eigenvectors and eigenvalues for such a nonlinear operator. For notational simplicity we restrict us in this section to the case of the unnormalized p-Laplacian $\Delta_p^{(\mathrm{u})}$ but all definitions and results carry over to the normalized version $\Delta_p^{(\mathrm{n})}$.

Definition 3.1 *The real number λ_p is called an eigenvalue for the p-Laplacian $\Delta_p^{(u)}$ if there exists a function $v : V \to \mathbb{R}$ such that*

$$(\Delta_p^{(u)} v)_i = \lambda_p\, \phi_p\,(v_i)\,, \quad \forall\, i = 1, ..., n.$$

The function v is called a p-eigenfunction of $\Delta_p^{(u)}$ corresponding to the eigenvalue λ_p.

The origin of this definition of an eigenvector for nonlinear operators lies in the Rayleigh-Ritz principle, a variational characterization of eigenvalues and eigenvectors for linear operators. For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, it is well-known that one can obtain the smallest eigenvalue $\lambda^{(1)}$ and the corresponding eigenvector $v^{(1)}$ satisfying $Av^{(1)} = \lambda^{(1)}\, v^{(1)}$ via the variational characterization

$$v^{(1)} = \arg\min_{f \in \mathbb{R}^n} \frac{\langle f, A\, f \rangle_{\mathbb{R}^n}}{\|f\|_2^2},$$

where the p-norm is defined as $\|f\|_p^p := \sum_{i=1}^n |f_i|^p$. Note that this characterization implies that (up to rescaling) $v^{(1)}$ is the global minimizer of $\langle f, Af \rangle$ subject to $\|f\|_2 = 1$. This variational characterization can now be carried over to nonlinear operators. We define for the unnormalized p-Laplacian $\Delta_p^{(\mathrm{u})}$,

$$Q_p(f) := \left\langle f, \Delta_p^{(\mathrm{u})} f \right\rangle = \frac{1}{2} \sum_{i,j=1}^n w_{ij} |f_i - f_j|^p,$$

and define similarly the functional $F_p : \mathbb{R}^V \to \mathbb{R}$,

$$F_p(f) := \frac{Q_p(f)}{\|f\|_p^p}.$$

Theorem 3.1 *The functional F_p has a critical point at $v \in \mathbb{R}^V$ if and only if v is a p-eigenfunction of $\Delta_p^{(u)}$. The corresponding eigenvalue λ_p is given as $\lambda_p = F_p(v)$. Moreover, we have $F_p(\alpha f) = F_p(f)$ for all $f \in \mathbb{R}^V$ and $\alpha \in \mathbb{R}$.*

Proof: One can check that the condition for a critical point of F_p at v can be rewritten as

$$\Delta_p v - \frac{Q_p(v)}{\|v\|_p^p} \phi_p(v) = 0.$$

Thus, with Definition 3.1 v is an eigenvector of Δ_p. Moreover, the equation implies that a given eigenvector v to the eigenvalue λ_p is a critical point of F_p if $\lambda_p = F_p(v)$. Summing up the eigenvector equation of Definition 3.1 shows this equality. The last statement follows directly from the definition. □

This theorem shows that in order to get all eigenvectors and eigenvalues of $\Delta_p^{(\mathrm{u})}$ we have to find all critical points of the functional F_p. Moreover, with $F_p(\alpha f) = F_p(f)$, we observe that the usual property for linear operators that eigenvectors are invariant under scaling carries over to the nonlinear case. The following proposition is a generalization of a result by Fiedler (1973) to the graph p-Laplacian. It relates the connectivity of the graph to properties of the first eigenvalue $\lambda_p^{(1)}$ of the p-Laplacian. We denote by $\mathbf{1}_A \in \mathbb{R}^V$ the function which is one on A and zero else.

Proposition 3.1 *The multiplicity of the first eigenvalue $\lambda_p^{(1)} = 0$ of the p-Laplacian $\Delta_p^{(u)}$ is equal to the number K of connected components $C_1, \ldots, C_K$ of the graph. The corresponding eigenspace for $\lambda_p^{(1)} = 0$ is given as $\{\sum_{i=1}^K \alpha_i \mathbf{1}_{j \in C_i} \,|\, \alpha_i \in \mathbb{R},\, i = 1, \ldots, K\}$.*

Proof: We have $Q_p(f) \geq 0$, so that all eigenvalues λ_p of $\Delta_p^{(\mathrm{u})}$ are non-negative. Similar to the case $p = 2$, one can check that $\sum_{i,j=1}^n w_{ij} |f_i - f_j|^p = 0$, if and only if f is constant on each connected component. □

In spectral clustering the graph is usually assumed to be connected, so that $v_p^{(1)} = c\mathbf{1}$ for $c \in \mathbb{R}$, otherwise

spectral clustering is trivial. For the following we assume that the graph is connected. The previous proposition suggests that similar to the standard case $p = 2$ we need at least the second eigenvector to construct a partitioning of the graph. For $p = 2$, we get the second eigenvector again by the variational Rayleigh-Ritz principle,

$$v^{(2)} = \arg\min_{f\in\mathbb{R}^n} \left\{ \frac{\langle f, \Delta_2^{(\mathrm{u})} f\rangle}{\|f\|_2^2} \mid \langle f, \mathbf{1}\rangle = 0. \right\}.$$

This form is not suited for the p-Laplacian since its eigenvectors are not necessarily orthogonal. However, for a function with $\langle f, \mathbf{1}\rangle = 0$ one has

$$\|f\|_2^2 = \left\| f - \frac{1}{n}\langle f, \mathbf{1}\rangle\ \mathbf{1} \right\|_2^2 = \min_{c\in\mathbb{R}} \|f - c\,\mathbf{1}\|_2^2 .$$

Thus, we can write equivalently,

$$v^{(2)} = \arg\min_{f\in\mathbb{R}^n} \frac{\langle f, \Delta_2^{(\mathrm{u})} f\rangle}{\min_{c\in\mathbb{R}} \|f - c\,\mathbf{1}\|_2^2}.$$

This motivates the definition of $F_p^{(2)} : \mathbb{R}^V \to \mathbb{R}$,

$$F_p^{(2)}(f) = \frac{Q_p(f)}{\min_{c\in\mathbb{R}} \|f - c\mathbf{1}\|_p^p}.$$

Theorem 3.2 *The second eigenvalue $\lambda_p^{(2)}$ of the graph p-Laplacian $\Delta_p^{(u)}$ is equal to the global minimum of the functional $F_p^{(2)}$. The corresponding eigenvector $v_p^{(2)}$ of $\Delta_p^{(u)}$ is then given as $v_p^{(2)} = u^* - c^*\mathbf{1}$ for any global minimizer u^* of $F_p^{(2)}$, where $c^* = \arg\min_{c\in\mathbb{R}} \sum_{i=1}^n |u_i^* - c|^p$. Furthermore, the functional $F_p^{(2)}$ satisfies $F_p^{(2)}(tu + c\mathbf{1}) = F_p^{(2)}(u)$, for all t, $c \in \mathbb{R}$.*

Proof: Can be found in (Bühler & Hein, 2009). □

Thus, instead of solving the complicated nonlinear equation of Definition 3.1 to obtain the second eigenvector of the graph p-Laplacian, we just have to find the global minimum of the functional $F_p^{(2)}$. In the next section, we discuss the relation between the second eigenvalue $\lambda_p^{(2)}$ of the graph p-Laplacian and the balanced graph cuts of Section 2. In Section 5, we provide an algorithmic framework to compute the second eigenvector of the p-Laplacian efficiently.

4. Spectral properties of the graph p-Laplacian and the Cheeger cut

Now that we have discussed the variational characterization of the second eigenvector of the p-Laplacian, we will provide the relation to the relaxation of balanced graph cut criteria as it can be done for the standard graph Laplacian.

4.1. Spectral relaxation of balanced graph cuts

It is well known that the second eigenvector of the unnormalized and normalized standard graph Laplacians ($p = 2$) is the solution of a relaxation of the ratio cut $\mathrm{RCut}(C, \overline{C})$ and normalized cut $\mathrm{NCut}(C, \overline{C})$, see e.g. (von Luxburg, 2007). We will show now that the second eigenvector $v_p^{(2)}$ of the p-Laplacian can also be seen as a relaxation of balanced graph cuts.

Theorem 4.1 *For $p > 1$ and every partition of V into $C, \overline{C}$ there exists a function $f_{p,C} \in \mathbb{R}^V$ such that the functional $F_p^{(2)}$ associated to the unnormalized p-Laplacian satisfies*

$$F_p^{(2)}(f_{p,C}) = \mathrm{cut}(C, \overline{C}) \left| \frac{1}{|C|^{\frac{1}{p-1}}} + \frac{1}{\left|\overline{C}\right|^{\frac{1}{p-1}}} \right|^{p-1},$$

with the special cases,

$$F_2^{(2)}(f_{2,C}) = \mathrm{RCut}(C, \overline{C}),$$
$$\lim_{p\to 1} F_p^{(2)}(f_{p,C}) = \mathrm{RCC}(C, \overline{C}).$$

Moreover, one has $F_p^{(2)}(f_{p,C}) \leq 2^{p-1}\,\mathrm{RCC}(C, \overline{C})$. Equivalent statements hold for a function $g_{p,C}$ for the normalized cut and the normalized p-Laplacian $\Delta_p^{(n)}$.

Proof: Let $p > 1$, then we define for a partition $C, \overline{C}$ of V the function $f_{p,C} : V \to \mathbb{R}$ as

$$(f_{p,C})_i = \begin{cases} 1/\left|C\right|^{\frac{1}{p-1}} & , \ i \in C, \\ -1/\left|\overline{C}\right|^{\frac{1}{p-1}} & , \ i \in \overline{C}. \end{cases}$$

One has $Q_p(f_{p,C}) = \sum_{i\in C,\, j\in\overline{C}} \left| \frac{1}{|C|^{\frac{1}{p-1}}} + \frac{1}{\left|\overline{C}\right|^{\frac{1}{p-1}}} \right|^p$.
Moreover, one has
$\min_{c\in\mathbb{R}} \|f_{p,C} - c\mathbf{1}\|_p^p = \|f_{p,C}\|_p^p = \frac{1}{|C|^{\frac{1}{p-1}}} + \frac{1}{\left|\overline{C}\right|^{\frac{1}{p-1}}}$.
With $F_p^{(2)}(f_{p,C}) = Q_p(f_{p,C}) / \min_{c\in\mathbb{R}} \|f - c\mathbf{1}\|_p^p$, we get

$$F_p^{(2)}(f_{p,C}) = \sum_{i\in C, y\in\overline{C}} w_{ij} \left| \frac{1}{|C|^{\frac{1}{p-1}}} + \frac{1}{\left|\overline{C}\right|^{\frac{1}{p-1}}} \right|^{p-1}$$
$$\leq \sum_{i\in C, y\in\overline{C}} w_{ij} \left| \frac{2}{\min\{|C|, \left|\overline{C}\right|\}^{\frac{1}{p-1}}} \right|^{p-1} = 2^{p-1}\,\mathrm{RCC}(C, \overline{C}).$$

The first equality shows the general result and simplifies to the ratio cut for $p = 2$. The limit $p \to 1$ follows with $\lim_{\alpha\to\infty}(a^\alpha + b^\alpha)^{1/\alpha} = \max\{a, b\}$. □

Thus, since one minimizes over all functions in the eigenproblem for the second eigenvector of the p-Laplacian $\Delta_p^{(\mathrm{u})}$ and $\Delta_p^{(\mathrm{n})}$ it is a relaxation of the

ratio/normalized cut for $p = 2$ and for the ratio/normalized Cheeger cut in the limit of $p \to 1$. In the interval $1 < p < 2$ the eigenproblem can be seen as as a relaxation of the interpolation between ratio/normalized cut and the ratio/normalized Cheeger cut, for the functional $F_p^{(2)}$ of $\Delta_p^{(\mathrm{u})}$ we get,

$$F_p^{(2)}(f_{p,C}) = \mathrm{cut}(C,\overline{C})\left|\frac{1}{|C|^{\frac{1}{p-1}}} + \frac{1}{\left|\overline{C}\right|^{\frac{1}{p-1}}}\right|^{p-1},$$

which can be understood using the inequalities between l_p-norms, for $\alpha \geq \beta \geq 1$ one has $\|x\|_\beta \geq \|x\|_\alpha$,

$$\frac{1}{|C|} + \frac{1}{|\overline{C}|} \geq \left(\frac{1}{|C|^\alpha} + \frac{1}{|\overline{C}|^\alpha}\right)^{\frac{1}{\alpha}} \geq \max\left\{\frac{1}{|C|}, \frac{1}{|\overline{C}|}\right\},$$

with $\alpha = 1/(p-1)$ and thus for $1 < p < 2$, one has $\infty > \alpha > 1$.

The spectral relaxation of ratio (Hagen & Kahng, 1991) and normalized cut (Shi & Malik, 2000) was one of the main motivations for standard spectral clustering. There exist other possibilities to relax the ratio and normalized cut problem, see (De Bie & Cristianini, 2006), which lead to a semi-definite program. These relaxations give better bounds on the true cut than the standard spectral relaxation ($p = 2$), though they are computationally expensive. However, up to our knowledge the bounds which can be achieved by semidefinite programming are not as tight as the ones which we provide in the next section for the p-Laplacian as $p \to 1$.

4.2. Isoperimetric Inequality - the second eigenvalue $\lambda_p^{(2)}$ and the Cheeger cut

The isoperimetric inequality (Chung, 1997) for the graph Laplacian ($p = 2$) provides additional theoretical backup for the spectral relaxation. It provides upper and lower bounds on the ratio/normalized Cheeger cut in terms of the second eigenvalue of the graph p-Laplacian. We define the optimal ratio and normalized Cheeger cut values h_{RCC} and h_{NCC} as

$$h_{\mathrm{RCC}} = \inf_C \mathrm{RCC}(C,\overline{C}) \text{ and } h_{\mathrm{NCC}} = \inf_C \mathrm{NCC}(C,\overline{C}).$$

The standard isoperimetric inequality for $p = 2$ (see Chung, 1997) is given as

$$\frac{h_{\mathrm{NCC}}^2}{2} \leq \lambda_2^{(2)} \leq 2\, h_{\mathrm{NCC}},$$

where $\lambda_2^{(2)}$ is the second eigenvalue of the standard normalized graph Laplacian ($p = 2$). The isoperimetric inequality for the normalized p-Laplacian has been proven by Amghibech (2003).

Theorem 4.2 (Amghibech, 2003) *Denote by $\lambda_p^{(2)}$ the second eigenvalue of the* ***normalized*** *p-Laplacian $\Delta_p^{(n)}$. Then for any $p > 1$,*

$$2^{p-1}\left(\frac{h_{\mathrm{NCC}}}{p}\right)^p \leq \lambda_p^{(2)} \leq 2^{p-1}\, h_{\mathrm{NCC}}\ .$$

We extend the result of Amghibech to the unnormalized p-Laplacian.

Theorem 4.3 *Denote by $\lambda_p^{(2)}$ the second eigenvalue of the* ***unnormalized*** *p-Laplacian $\Delta_p^{(u)}$. For $p > 1$,*

$$\left(\frac{2}{\max_i d_i}\right)^{p-1}\left(\frac{h_{\mathrm{RCC}}}{p}\right)^p \leq \lambda_p^{(2)} \leq 2^{p-1} h_{\mathrm{RCC}}\ .$$

Proof: Can be found in (Bühler & Hein, 2009). □

Note that $h_{\mathrm{NCC}} < 1$ and $\frac{h_{\mathrm{RCC}}}{\max_i d_i} < 1$, so that in both cases the left hand side of the bound is smaller than h_{NCC} resp. h_{RCC}. When considering the limit $p \to 1$, one observes that the bounds on λ_p become tight as $p \to 1$. Thus in the limit of $p \to 1$, the second eigenvalue of the unnormalized/normalized p-Laplacian approximates the optimal ratio/normalized Cheeger cut arbitrarily well.

Still the problem remains how to transform the real-valued second eigenvector of the p-Laplacian into a partitioning of the graph. We use the standard procedure and threshold the second eigenvector $v_p^{(2)}$ to obtain the partitioning. The optimal threshold is determined by minimizing the corresponding Cheeger cut. For the second eigenvector $v_p^{(2)}$ of the unnormalized graph p-Laplacian $\Delta_p^{(\mathrm{u})}$ we determine,

$$\operatorname*{arg\,min}_{C_t=\{i\in V \mid v_p^{(2)}(i)>t\}} \mathrm{RCC}(C_t,\overline{C_t}), \tag{3}$$

and similarly for the second eigenvector $v_p^{(2)}$ of the normalized graph p-Laplacian $\Delta_p^{(\mathrm{n})}$ we compute,

$$\operatorname*{arg\,min}_{C_t=\{i\in V \mid v_p^{(2)}(i)>t\}} \mathrm{NCC}(C_t,\overline{C_t}). \tag{4}$$

The obvious question is how good the cut values obtained by thresholding the second eigenvector of the p-Laplacian are compared to optimal Cheeger cut values. The following Theorem answers this question and provides the key motivation for p-spectral clustering.

Theorem 4.4 *Denote by h_{RCC}^* and h_{NCC}^* the ratio/normalized Cheeger cut values obtained by tresholding the second eigenvector $v_p^{(2)}$ of the unnormalized/normalized p-Laplacian via* (3) *for $\Delta_p^{(u)}$ resp.* (4)

Algorithm 1 p-Laplacian based Spectral Clustering

1: **Input:** weight matrix W, number of desired clusters k, choice of p-Laplacian.
2: **Initialization:** cluster $C_1 = V$, number of clusters $s = 1$
3: **repeat**
4: Minimize $F_p^{(2)} : \mathbb{R}^{C_i} \to \mathbb{R}$ for the chosen p-Laplacian for each cluster C_i, $i = 1, \dots, s$.
5: Compute optimal threshold for dividing each cluster C_i via (3) for $\Delta_p^{(\mathrm{u})}$ or (4) for $\Delta_p^{(\mathrm{n})}$.
6: Choose to split the cluster C_i so that the total multi-partition cut criterion is minimized (ratio cut (1) for $\Delta_p^{(\mathrm{u})}$ and normalized cut (2) for $\Delta_p^{(\mathrm{n})}$).
7: $s \Leftarrow s + 1$
8: **until** number of clusters $s = k$

for $\Delta_p^{(n)}$. Then for $p > 1$,

$$h_{\mathrm{RCC}} \;\leq\; h^*_{\mathrm{RCC}} \;\leq\; p \left(\max_{i \in V} d_i\right)^{\frac{p-1}{p}} \left(h_{\mathrm{RCC}}\right)^{\frac{1}{p}} ,$$
$$h_{\mathrm{NCC}} \;\leq\; h^*_{\mathrm{NCC}} \;\leq\; p \left(h_{\mathrm{NCC}}\right)^{\frac{1}{p}} .$$

Proof: Can be found in (Bühler & Hein, 2009). □

One observes that in the limit of $p \to 1$ both inequalities become tight, which implies that for $p \to 1$ the cut found by thresholding the second eigenvector of the p-Laplacian converges to the optimal Cheeger cut.

5. p-Spectral Clustering

The algorithmic scheme for p-Spectral Clustering is shown in Algorithm 1. More than two clusters are obtained by consecutive splitting of clusters until the desired number of clusters is reached. As multi-partition criterion, we use the established generalized versions of ratio cut (1) and normalized cut (2). However, one could also think about multi-partition versions of the Cheeger cut. The sequential splitting of clusters is the more "traditional" way to do spectral clustering. Alternatively, one uses for the standard graph Laplacian the first k eigenvectors to define a new representation of the data. In this new k-dimensional representation one then applies a standard clustering algorithm like k-means. This alternative is not possible in our case since at the moment we are not able to compute higher-order eigenvectors of the p-Laplacian. However, as Theorem 4.4 shows there is also need for going this way since thresholding will yield the optimal Cheeger cut in the limit $p \to 1$.

The functional $F_p^{(2)} : \mathbb{R}^V \to \mathbb{R}$ is non-convex and thus we cannot guarantee to reach the global minimum. Indeed, a direct minimization for small values of p leads often very fast to convergence to a non-optimal local minimum. Thus we use a different procedure using the fact that for $p = 2$ we can easily compute the global minimizer of $F_2^{(2)}$. It is just the second eigenvector of the standard graph Laplacian, which can be efficiently computed for sparse matrices e.g. using ARPACK. Since the functional $F_p(f)$ is continuous in p, we can hope for close values p_1 and p_2 that the global minimizer of $F_{p_1}^{(2)}$ and $F_{p_2}^{(2)}$ are also close (at least the local minimizer should be close). Moreover, it is well known that Newton-like methods have superlinear convergence close to the local optima (Bertsekas, 1999). These two facts suggest to solve the problem $F_p^{(2)}(u)$ by minimizing a sequence of functionals F_{p_i},

$$F_{p_0}^{(2)}, F_{p_1}^{(2)}, ..., F_p^{(2)}, \quad \text{with } p_0 = 2 > p_1 > ... > p \ ,$$

where each step is initialized with the solution of the previous step and initialization is done with $p_0 = 2$. In the experiments we found that the update rule $p_{t+1} = 0.9\, p_t$ yields a good trade-off between decreasing too fast with the danger that the optimum for $F_{p_t}^{(2)}$ is far away from the optimum of $F_{p_{t+1}}^{(2)}$ and decreasing too slow which yields fast convergence of the Newton method but needs a lot of iterations.

The minimimization of the functionals F_{p_t} is done using a mixture of gradient and Newton steps. However, the Hessian of F_{p_i} is not sparse, which causes problems for large scale problems, but it can be decomposed into

$$H = A + \left(ab^T + ba^T\right) + bb^T \ ,$$

where $a, b \in \mathbb{R}^n$ and the matrix A is sparse. Thus H is a sum of a sparse matrix plus low-rank updates. Thus, we just discard the low-rank updates and use A as a surrogate for the true Hessian. We use the Minimal Residual method (Paige & Saunders, 1975) for solving the linear system of the Newton step as the matrix A is symmetric but not necessarily positive definite. In order to avoid problems with an ill-conditioned matrix A, we add a small ridge. Note that the term $\min_{c \in \mathbb{R}} \|f - c\mathbf{1}\|_p^p$ in the functional $F_p^{(2)}(f)$ is itself a (convex) optimization problem which can be solved very fast using bisection.

6. Experimental evaluation

In all experiments, we used a symmetric K-NN graph with $K = 10$ and weights w_{ij} defined as

$$w_{ij} = \max\{\mathrm{s}_i(j), \mathrm{s}_j(i)\}, \text{ where } \mathrm{s}_i(j) = e^{-\frac{4}{\sigma_i^2}\|x_i - x_j\|^2} ,$$

with σ_i being the Euclidean distance of x_i to its K-nearest neighbor. We evaluate the clustering on

datasets with known number of classes k. We then clustered the data into k clusters and checked the agreement of the found clusters $C_1, \ldots, C_k$ with the class structure using the error measure

$$\text{error}(C_1, .., C_k) = \frac{1}{|V|} \sum_{i=1}^{k} \sum_{j \in C_i} \mathbb{I}_{Y_j \neq Y_i'}, \qquad (5)$$

where Y_j is the true label of j and Y_i' is the dominant label in cluster C_i.

6.1. High-dimensional noisy two moons

The two moons dataset is generated as two half-circles in $\mathbb{R}^2$ which are embedded into a d-dimensional space where Gaussian noise $\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ is added. When varying d, n and σ, we always made the same observation: unnormalized and normalized p-spectral clustering leads for decreasing values of p to cuts with decreasing values of the Cheeger cuts RCC and NCC. In Fig. 1, we illustrate this for the case $d = 100$, $n = 2000$ and $\sigma^2 = 0.02$. Note that this dataset is far from being trivial since the high-dimensional noise has corrupted the graph (see the edge structure in Fig. 1). The histogram of the values of the second eigenvectors for p equal to 2, 1.7, 1.4 and 1.1, show strong differences. For $p = 2$, the values are scattered over the interval, whereas for $p = 1.1$, they are almost concentrated on two peaks. This suggests that for $p = 1.1$, the p-eigenvector is quite close to the function $f_{p,C}$ as defined in Theorem 4.1. The third row in Fig. 1 shows the resulting clusters found by p-spectral clustering with $\Delta_p^{(n)}$. For $p \to 1$, the clustering is almost perfect despite the difficulty of this dataset. In order to illustrate that this result is representative, we have repeated the experiment 100 times. The plot in the bottom left of Fig. 1 shows the mean of the normalized Cheeger cut, the second eigenvalue $\lambda_p^{(2)}$, normalized cut and error as $p \to 1$. One observes that despite there is some variance, the results of p-spectral clustering are significantly better than standard spectral clustering.

6.2. UCI-Datasets

In Table 2 we show results for p-spectral clustering on several UCI datsets both for the unnormalized (right column) and the normalized p-Laplacian (left column). The corresponding Cheeger-cuts (second row) are consistently decreasing as $p \to 1$. For most of the datasets this also implies that the ratio/normalized cut decreases. Note that the error is often constant despite the fact that the cut is still decreasing. Opposite to the other examples, minimizing the cut does not necessarily lead to a smaller error.

Table 1. Top: Results of unnormalized p-spectral clustering with $k = 10$ for USPS and MNIST using the ratio-multi-partition criterion (1). In both cases the RCut and the error significantly decrease as p decreases. *Bottom:* confusion matrix for MNIST of the clusters found by p-spectral clustering for $p = 1.2$. Class 1 has been split into two clusters and class 4 and 9 have been merged. Thus there exists no class 9 in the table. Apart from the merged classes the clustering reflects the class structure quite well.

	USPS		MNIST	
p	RCut	Error	RCut	Error
2.0	0.819	0.233	0.225	0.189
1.9	0.741	0.142	0.209	0.172
1.8	0.718	0.141	0.186	0.170
1.7	0.698	0.139	0.170	0.169
1.6	0.684	0.134	0.164	0.170
1.5	0.676	0.133	0.161	0.133
1.4	0.693	0.141	0.158	0.132
1.3	0.684	0.138	0.155	0.131
1.2	0.679	0.137	0.153	0.129

True/Cluster	0	1	2	3	4	5	6	7	8
0	6845	5	7	0	5	8	26	4	3
1	1	7794	32	8	21	1	2	16	2
2	38	47	6712	25	15	5	8	114	26
3	5	6	31	6939	30	61	2	45	22
4	3	45	2	1	6750	0	14	5	4
5	15	1	4	92	39	6087	61	5	9
6	23	17	6	0	9	23	6797	0	1
7	1	83	22	1	116	2	0	7067	1
8	18	51	13	507	112	122	23	18	5961
9	15	15	3	117	6708	11	4	77	8

6.3. USPS and MNIST

We perform unnormalized p-spectral clustering on the full USPS and MNIST-datasets ($n = 9298$ and $n = 70000$). In Table 1 one observes that for $p \to 1$ the ratio cut as well as the error decreases for both datasets. The error is even misleading since the class separation is quite good but one class has been split which implies that two classes have been merged. This happens for both datasets and in Table 1 we provide the confusion matrix for MNIST for p-spectral clustering with $p = 1.2$. For larger values of number of clusters k we thus expect better results. In the following table we present the runtime behavior (in seconds) for USPS:

p	2.0	1.9	1.8	1.7	1.6	1.5	1.4	1.3	1.2
t	10	81	99	144	224	456	1147	2266	4660

As $p \to 1$, the problem becomes more difficult which is clear since one approximates asymptotically the optimal Cheeger cut. However, there is still room for improvement to speed up our current implementation.

Acknowledgments

This work has been supported by the Excellence Cluster on Multimodal Computing and Interaction at Saarland University.

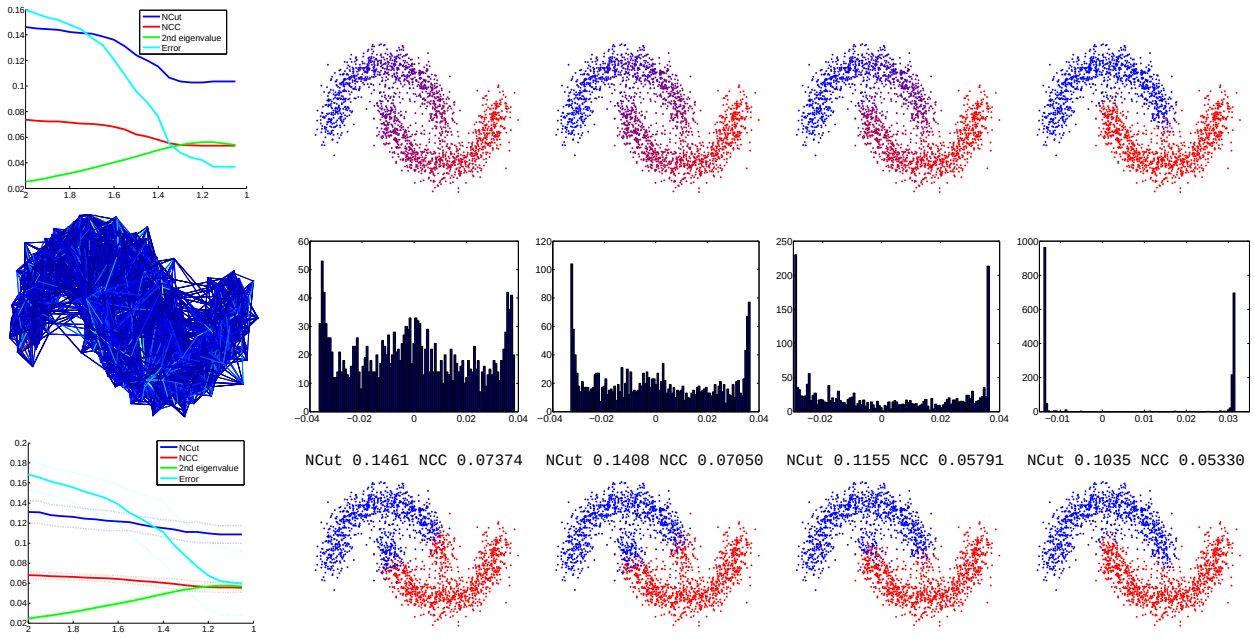

Figure 1. Results for the two moons data set, 2000 points in 100 dimensions, noise variance 0.02. *First row, from left to right:* Second eigenvector of the p-Laplacian for $p = 2.0, 1.7, 1.4, 1.1$. *Second row:* Histogram of the values of the second eigenvector. *Last row:* Resulting clustering after finding optimal threshold according to the NCC criterion. *First column, top:* The values of NCC, the eigenvalue $\lambda_p^{(2)}$, NCut and the error for the example shown on the right. *Middle:* Plot of the edge structure. *Bottom*: Average values plus standard deviation of NCC, NCut, $\lambda_p^{(2)}$ and the error for varying p.

Table 2. Results of unnormalized/normalized p-spectral clustering on UCI-datasets. For each dataset, the rows correspond to NCut, NCC resp. RCut, RCC and error.

	Normalized			Unnormalized		
p	2.0	1.4	1.1	2.0	1.4	1.1
Breast	0.0254	0.0229	0.0289	0.0467	0.0332	0.0332
	0.0209	0.0135	0.0174	0.0300	0.0220	0.0220
	0.293	0.293	0.293	0.293	0.293	0.293
Heart	0.118	0.0796	0.0796	0.108	0.0946	0.0946
	0.0621	0.0579	0.0579	0.0550	0.0473	0.0473
	0.215	0.356	0.356	0.204	0.219	0.219
Ring norm	0.443	0.420	0.420	0.219	0.210	0.210
	0.222	0.210	0.210	0.109	0.105	0.105
	0.281	0.288	0.287	0.290	0.310	0.309
Two norm	0.0821	0.0813	0.0811	0.0392	0.0388	0.0387
	0.0411	0.0407	0.0406	0.0196	0.0194	0.0193
	0.0257	0.0259	0.0261	0.0255	0.0261	0.0269
Wave form	0.101	0.0857	0.0828	0.0485	0.0410	0.0396
	0.0552	0.0460	0.0438	0.0265	0.0221	0.0210
	0.227	0.211	0.201	0.225	0.212	0.201

References

Amghibech, S. (2003). Eigenvalues of the discrete p-Laplacian for graphs. *Ars Combin.*, *67*, 283–302.

Bertsekas, D. (1999). *Nonlinear programming.* Athena Scientific.

Bühler, T., & Hein, M. (2009). Supplementary material. `http://www.ml.uni-saarland.de/Publications/BueHei09tech.pdf`.

Chung, F. (1997). *Spectral graph theory.* AMS.

De Bie, T., & Cristianini, N. (2006). Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems. *J. Mach. Learn. Res.*, *7*, 1409–1436.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math. J.*, *23*, 298–305.

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

Hein, M., Audibert, J.-Y., & von Luxburg, U. (2007). Graph Laplacians and their convergence on random neighborhood graphs. *J. Mach. Learn. Res.*, *8*, 1325–1368.

Paige, C., & Saunders, M. (1975). Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Anal.*, *12*, 617–629.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.*, *22*, 888–905.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*, 395–416.

Zhou, D., & Schölkopf, B. (2005). Regularization on discrete spaces. *Deutsche Arbeitsgemeinschaft für Mustererkennung-Symposium* (pp. 361–368).