Convergence of a Steepest Descent Algorithm for Ratio Cut Clustering

Xavier Bresson[*], Thomas Laurent[†], David Uminsky[‡] and James H. von Brecht[§]

May 1, 2012

Abstract

Unsupervised clustering of scattered, noisy and high-dimensional data points is an important and difficult problem. Tight continuous relaxations of balanced cut problems have recently been shown to provide excellent clustering results. In this paper, we present an explicit-implicit gradient flow scheme for the relaxed ratio cut problem, and prove that the algorithm converges to a critical point of the energy. We also show the efficiency of the proposed algorithm on the two moons dataset.

1 Introduction

Partitioning data points into sensible groups is a fundamental problem in machine learning and has a wide range of applications. An efficient approach to deal with this problem is to cast the data partitioning problem as a graph clustering problem. Given a set of data points $V = \{x_1, \ldots, x_n\}$ and similarity weights $\{w_{i,j}\}_{1 \leq i,j \leq n}$, the clustering problem aims at finding a balanced cut of the graph of the data. In this work, we consider the balanced cut of Hagen and Kahng [5] known as ratio cut. The ratio cut problem is

$$\text{Minimize} \quad \text{RatioCut}(S) = \frac{\sum_{x_i \in S} \sum_{x_j \in S^c} w_{i,j}}{|S|} + \frac{\sum_{x_i \in S} \sum_{x_j \in S^c} w_{i,j}}{|S^c|} \tag{1}$$

$$\text{over all subsets } S \subsetneq V.$$

Here $|S|$ denotes the number of data points in S. While the problem, as stated above, is NP-hard, it has the following *tight continuous* relaxation:

$$\text{Minimize} \quad E(f) = \frac{\frac{1}{2}\sum_{i,j} w_{i,j}|f_i - f_j|}{\sum_i |f_i - m(f)|} \tag{2}$$

$$\text{over all non-constant functions } f : V \to \mathbb{R}.$$

Here $m(f)$ stands for the average of $f \in \mathbb{R}^n$ and f_i stands for $f(x_i)$. Recently, various algorithms have been proposed [12, 6, 7, 1, 10] to minimize relaxations of balance cut problem similar to (2). In this work, we present an explicit-implicit gradient flow algorithm, then prove that the iterates converge to critical points of the energy. We also present numerical experiments to show the robustness and efficiency of the algorithm.

[*] Department of Computer Science, City University of Hong Kong, Hong Kong (xbresson@cityu.edu.hk).

[†] Department of Mathematics, University of California Riverside, Riverside CA 92521 (laurent@math.ucr.edu)

[‡] Department of Mathematics, University of California Los Angeles, Los Angeles CA 90095 (duminsky@math.ucla.edu)

[§] Department of Mathematics, University of California Los Angeles, Los Angeles CA 90095 (jub@math.ucla.edu)

1.1 The Tight Continuous Relaxation

We begin by first explaining the meaning of the term *tight relaxation*. Since E is invariant under the addition of a constant, problem (2) is equivalent to

$$\text{Minimize} \quad \frac{\frac{1}{2}\sum_{i,j} w_{i,j}|f_i - f_j|}{\sum_i |f_i|} \qquad (3)$$
$$\text{over all } f : V \to \mathbb{R} \text{ s.t. } m(f) = 0 \text{ and } f \neq 0.$$

If the graph is connected then the total variation functional $\frac{1}{2}\sum_{i,j} w_{i,j}|f_i - f_j|$ defines a norm on the space of mean zero functions; we denote it by $\|f\|_{TV}$. The denominator of (3) is simply the ℓ^1-norm, and we denote it by $\|f\|_1$.

The continuous problem (3) is a tight relaxation of (1) in the following sense— if S^* is a solution of (1), then any nonzero, binary function of mean zero

$$f^*(x_i) = \begin{cases} a & \text{if } x_i \in S^* \\ b & \text{if } x_i \in (S^*)^c \end{cases} \qquad (4)$$

is a solution of problem (3). This is a consequence of the fact that the the extreme points of the TV-unit ball

$$\{f \in \mathbb{R}^n : \|f\|_{TV} \leq 1, m(f) = 0\}$$

are binary functions (see [12] for a proof of this fact). Therefore, if we fix $\|f\|_{TV} = 1$ and maximize the convex functional in the denominator of (3), the minimum of the ratio is attained at an extreme point. That is, at a binary function of mean zero. Binary functions of mean zero are always of the form

$$f = \lambda\left(|S^c|\chi_S - |S|\chi_{S^c}\right), \quad S \subsetneq V, \quad \lambda \neq 0,$$

where χ_S is the characteristic function of the set S. For such a function, we easily check that $E(f) = \text{RatioCut}(S)/2$. From this observation we can see that if S^* is a solution of the ratio cut problem (1), then $f^* = \lambda\left(|(S^*)^c|\chi_{S^*} + |S^*|\chi_{(S^*)^c}\right)$ is a solution of the continuous relaxation (3) for any $\lambda \neq 0$. A different proof of the fact that problem (2) is a tight relaxation of problem (1) can be found in [10].

1.2 Explicit-implicit gradient Flow

Let

$$T(f) = \|f\|_{TV} \text{ and } B(f) = \sum_i |f_i - m(f)|. \qquad (5)$$

Note that both T and B are convex. If T and B were differentiable, the explicit-implicit gradient flow of $E = T/B$ would be

$$\frac{f^{k+1} - f^k}{\tau^k} = -\frac{\nabla T(f^{k+1}) - E(f^k)\nabla B(f^k)}{B(f^k)} \qquad (6)$$

where τ^k is the time step. Since T and B are not differentiable, we replace (6) with its non-smooth equivalent:

$$g^k = f^k + \frac{\tau^k}{B(f^k)}E(f^k)v^k \quad \text{for some } v^k \in \partial B(f^k) \qquad (7)$$

$$f^{k+1} = \arg\min_f \left\{T(f) + B(f^k)\frac{\|f - g^k\|^2}{2\tau^k}\right\}. \qquad (8)$$

The minimization problem (8) is a standard ROF problem [11] that can be solved efficiently using approaches such as augmented Lagrangian method [4] or primal-dual method [3]. The scheme (7)–(8), as will be shown in the next section, decreases the energy and preserve the zero mean properties of the successive iterates. In order to remain away from the origin, where the energy is not defined, we project each iterate onto the sphere $\mathcal{S}^{n-1} = \{u \in \mathbb{R}^n : \|u\|_2 = 1\}$ at the end of each step. In numerical experiments we observe faster convergence when the time step is chosen to be

$$\tau^k = c\frac{B(f^k)}{E(f^k)}, \quad c > 0. \qquad (9)$$

With these choices, we arrive at our proposed algorithm to find critical points of the ratio cut functional (2):

$$g^k = f^k + cv^k \quad \text{for some } v^k \in \partial B(f^k) \tag{10}$$

$$h^k = \arg\min_f \left\{ T(f) + E(f^k) \frac{\|f - g^k\|^2}{2c} \right\} \tag{11}$$

$$f^{k+1} = \frac{h^k}{\|h^k\|_2}, \tag{12}$$

which we formalize in Algorithm 1.

Algorithm 1 Steepest descent of the RatioCut functional (2)

```
$f^{k=0}$ nonzero function with mean zero.
$c$ positive constant.
while loop not converged do
    $w^k \in \text{sign}(f^k), \quad v^k = w^k - m(w^k), \quad \lambda^k = \frac{\|f^k\|_{TV}}{\|f^k\|_1}$
    $g^k = f^k + c\, v^k$
    $h^k = \arg\min_f \ \{ \|f\|_{TV} + \frac{\lambda^k}{2c} ||f - g^k||_2^2 \ \}$
    $f^{k+1} = \frac{h^k}{\|h^k\|_2}$
end while
```

Let $\{f^k\}$ denote a sequence of iterates generated by Algorithm 1, starting from a non-zero function f^0 with $m(f^0) = 0$. In section 2, we show that any accumulation point of this sequence is a critical point of the the ratio cut functional (2). Moreover we show that $\|f^{k+1} - f^k\|_2 \to 0$ as $k \to \infty$, so that either the sequence converges or the set of accumulation points is a connected subset of the sphere $\mathcal{S}^{n-1}$. In section 3 we demonstrate the efficiency of Algorithm 1 on the two moons example.

2 Convergence

Given a connected graph, we want to minimize

$$E(f) = \frac{\sum_{i,j=1}^n w_{i,j} |f_i - f_j|}{\sum_{i=1}^n |f_i - m(f)|} = \frac{T(f)}{B(f)}$$

over the space of non-constant functions $f \in \mathbb{R}^n$. (Note that E is not defined for constant functions). This is equivalent to minimizing E over the set of non-constant functions with mean zero, which we write as

$$\mathcal{F} = \{f \in \mathbb{R}^n : m(f) = 0 \text{ and } f \neq 0\}.$$

We define $\mathbf{1} := (1, \dots, 1)^T \in \mathbb{R}^n$, so that $m(f) = \langle \mathbf{1}, f \rangle / n$ and $\mathbf{1}^\perp$ gives the space of functions with mean zero. Clearly $\mathcal{F}$ is an open subset of $\mathbf{1}^\perp$. As we assume a connected graph, T and B define norms on $\mathbf{1}^\perp$. Since all norms are equivalent in finite dimensions, there exist constants $\beta > \alpha > 0$ such that

$$\alpha B(f) \leq T(f) \leq \beta B(f) \quad \text{for all } f \in \mathbf{1}^\perp.$$

Therefore

$$\alpha \leq E(f) \leq \beta \quad \text{for all } f \in \mathcal{F}.$$

If we let

$$L(f) = \|f\|_1 = \sum_{i=1}^n |f_i|, \ \ P_0 f = f - m(f)\mathbf{1}, \tag{13}$$

then we see that $B(f) = L(P_0 f)$. Note that $P_0 = \text{Id} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, so that the matrix P_0 simply gives the orthogonal projection onto $\mathbf{1}^\perp$. As $L(f)$ is convex, so is $B(f) = L(P_0 f)$, and we also have

$$\partial B(f) = P_0 \ \text{sign}(P_0 f).$$

It is then easy to see that $\langle \partial B(f), \mathbf{1} \rangle = 0$ for all f. If $f \in \mathbf{1}^\perp$, then $B(f) = L(f)$ and $\partial B(f)$ is simply the projection of $\partial L(f)$ on $\mathbf{1}^\perp$, i.e. $\partial B(f) = P_0 \operatorname{sign}(f)$.

Starting from a non-constant function f, we define g and h according to Algorithm 1

$$g = f + cv, \qquad \text{where} \qquad v \in \partial B(f) \tag{14}$$

$$h = \arg\min_u \left\{ T(u) + E(f) \frac{\|u-g\|_2^2}{2c} \right\}, \tag{15}$$

which we write succinctly as

$$h \in \mathcal{H}^c(f).$$

Since g is not uniquely defined when $B(f)$ is non-differentiable, in general $\mathcal{H}^c(f)$ may have more than one element. Therefore the map $\mathcal{H}^c$ is a *set-valued map* defined over the space of non-constant functions (see Definition 2 in the following subsection).

2.1 Estimates

Lemma 1 (Elementary properties of $\mathcal{H}^c$)**.** *Let g and h be defined by* (14)–(15)*.*

1. *If f is not constant, then h is not constant. Moreover, the energy inequality*

$$E(f) \geq E(h) + \frac{E(f)}{B(h)} \frac{\|h-f\|_2^2}{c} \tag{16}$$

holds. As a consequence, $E(h) < E(f)$ unless $h = f$.

2. *If f is not constant, then*

$$\|h\|_2 \leq \|g\|_2 \leq \|f\|_2 + 2c\sqrt{n}. \tag{17}$$

3. *If $f \in \mathbb{R}^n$, then $\|g\|_2 > \|f\|_2$, or, to be more precise:*

$$\|g\|_2^2 = ||f||_2^2 + 2cB(f) + c^2||\partial B(f)||_2^2.$$

4. *If $f \in \mathcal{F}$, then $g, h \in \mathcal{F}$.*

Proof. (1.) The definition (15) of h implies that $E(f)\frac{h-g}{c} \in -\partial T(h)$, and therefore, since T is convex,

$$T(f) \geq T(h) + \left\langle -E(f)\frac{h-g}{c}, f-h \right\rangle \tag{18}$$

$$= T(h) - E(f) \left\langle \frac{h-(f+cv)}{c}, f-h \right\rangle \tag{19}$$

$$= T(h) + \frac{E(f)}{c}\|h-f\|_2^2 - E(f) \langle v, h-f \rangle. \tag{20}$$

Since B is also convex, we have $B(h) \geq B(f) + \langle v, h-f \rangle$, and therefore adding these two last inequalities,

$$T(f) + E(f)B(h) \geq T(h) + E(f)B(f) + \frac{E(f)}{c}\|h-f\|_2^2.$$

In other words,

$$E(f)B(h) \geq T(h) + \frac{E(f)}{c}\|h-f\|_2^2.$$

Since f is not constant, we have $E(f) > 0$. Note that if h were constant, then $B(h) = 0$ which would imply $h = f$. This is a contradiction since f is not constant. Thus $B(h) > 0$, so we may divide in the last expression to obtain (16).

(2.) To prove that $\|h\|_2 \leq \|g\|_2$, note

$$h = \operatorname{prox}_\Phi(g) := \arg\min_u \left\{ \Phi(u) + \frac{\|u-g\|_2^2}{2} \right\} \qquad \text{where } \Phi(u) = \frac{c}{E(f)} T(u).$$

Since proximal mappings are Lipshitz continuous with constant one, and since $\text{prox}_\Phi(0) = 0$, we have

$$\|h\|_2 = \|\text{prox}_\Phi(g) - \text{prox}_\Phi(0)\|_2 \leq \|g\|_2. \tag{21}$$

To establish the inequality $\|g\|_2 \leq \|f\|_2 + 2c\sqrt{n}$, note that $\|\text{sign}(P_0 f)\|_\infty \leq 1$ and therefore

$$\|\partial B(f)\|_2 \leq \sqrt{n}\|\partial B(f)\|_\infty = \sqrt{n}\|\text{sign}(P_0 f) - m(\text{sign}(P_0 f))\mathbf{1}\|_\infty \leq 2\sqrt{n} \quad \text{ for all } f \in \mathbb{R}^n. \tag{22}$$

The upper bound then follows from the definition of g and the triangle inequality.

(3.) Since B is homogeneous of degree one, we have

$$||g||_2^2 = ||f + c\partial B(f)||_2^2 = ||f||_2^2 + 2c\,\langle f, \partial B(f)\rangle + c^2||\partial B(f)||_2^2 = ||f||_2^2 + 2cB(f) + c^2||\partial B(f)||_2^2. \tag{23}$$

(4.) Since $\partial B(f) \subset \mathbf{1}^\perp$, it is clear that $f \in \mathbf{1}^\perp$ implies $g \in \mathbf{1}^\perp$. Equation (23) shows that $||g||_2 > ||f||_2 > 0$ so that g cannot be constant (the only constant function of mean zero is the zero function). Thus $g \in \mathcal{F}$. Suppose that $h \notin \mathbf{1}^\perp$. Since P_0 projects onto $\mathbf{1}^\perp$ and since $T(P_0 u) = T(u)$ for all $u \in \mathbb{R}^n$ (because T is invariant under addition of a constant), we have

$$T(h) + \frac{E(f)}{2c}\|h - g\|_2^2 = T(P_0 h) + \frac{E(f)}{2c}\left(\|P_0 h - g\|_2^2 + \|(\text{Id} - P_0)h\|_2^2\right).$$

This contradicts the definition of h as the global minimizer unless $(\text{Id} - P_0)h = 0$. Thus h has mean zero. By property (1.) we know h is not constant, so $h \in \mathcal{F}$ as well. □

Definiton 1. *Let $f^0 \in \mathcal{F}$. We say that f^k, g^k, h^k is a sequence generated by the algorithm if*

$$f^{k+1} \in P_2(\mathcal{H}^c(f^k)) \qquad \textit{where } P_2 \textit{ is the projection onto the sphere } \mathcal{S}^{n-1}$$

and where g^k and h^k are defined from f^k by (14) and (15).

Lemma 2 (Properties of the iterates)**.** *If f^k, g^k, h^k is a sequence generated by the algorithm, then $E(f^{k+1}) \leq E(f^k)$ with equality if and only if $f^k = f^{k+1}$. Moreover,*

$$\|f^k - h^k\|_2 \to 0 \quad \textit{and} \quad \|f^k - f^{k+1}\|_2 \to 0. \tag{24}$$

Therefore $\mathcal{S}^{n-1}$ is an attractor for the sequence $\{h^k\}$.

Proof. The fact that the energy decreases is a consequence of (16) from Lemma 1 together with the fact that $E(f^{k+1}) = E(h^k)$ due to the invariance of E under scaling. As $f^k \in \mathbf{1}^\perp$ and $||f^k||_2 = 1$ it follows that $E(f^k) \geq \alpha > 0$. From (16) we then have

$$\|h^k - f^k\|_2^2 \leq \frac{c}{\alpha}B(h^k)(E(f^k) - E(f^{k+1})). \tag{25}$$

Now from (17) we have

$$B(h^k) = \|h^k\|_1 \leq \sqrt{n}\|h^k\|_2 \leq \sqrt{n} + 2nc,$$

and therefore

$$\|h^k - f^k\|_2^2 \leq \frac{c}{\alpha}(\sqrt{n} + 2nc)(E(f^k) - E(f^{k+1})) \to 0,$$

where we have used that $E(f^k)$ is a converging sequence since it is decreasing and bounded from below.

We now show $\|f^k - f^{k+1}\|_2 \to 0$. Note that the projection P_2 is smooth on the annulus $\mathcal{A} := \{u \in \mathbb{R}^n : 1/2 \leq \|u\| \leq 3/2\}$ and therefore it is Lipschitz continuous on $\mathcal{A}$ with constant, say, C. Since eventually $h^k \in \mathcal{A}$, we have

$$\|f^k - f^{k+1}\|_2 = \|P_2(f^k) - P_2(h^k)\|_2 \leq C\|f^k - h^k\|_2 \to 0.$$

□

2.2 Proof of convergence

Definiton 2 (Set-valued map). *Let X and Y be two subsets of $\mathbb{R}^n$. If for each $x \in X$ there is a corresponding set $F(x) \subset Y$ then F is called a set-valued map from X to Y . We denote this by $F : X \rightrightarrows Y$. The graph of F, denoted Graph(F) is defined by*

$$\mathit{Graph}(F) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in F(x), x \in X\}.$$

A set-valued map F is called closed if Graph(F) is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$.

Define the compact sets

$$K_1 = \{u \in \mathbb{R}^n : \|u\|_2 = 1 \text{ and } m(u) = 0\} \tag{26}$$
$$K_2 = \{u \in \mathbb{R}^n : 1 \leq \|u\|_2 \leq 1 + 2c\sqrt{n} \text{ and } m(u) = 0\} \tag{27}$$

along with the set-valued map $\mathcal{Y}^c : K_1 \rightrightarrows K_2$

$$\mathcal{Y}^c(f) = f + c\partial B(f).$$

The fact that the range of $\mathcal{Y}^c$ is in K_2 is a consequence of (17).

Lemma 3. *The set-valued map $\mathcal{Y}^c$ is closed.*

Proof. Let us first show that the set-valued map sign : $\mathbb{R}^n \rightrightarrows [-1, 1]^n$ is closed. Let assume that

$$f^k \to f^* \tag{28}$$
$$z^k \in \operatorname{sign}(f^k) \to z^* \tag{29}$$

We want to show that $z^* \in \operatorname{sign}(f^*)$, or equivalently, $z_i^* \in \operatorname{sign}(f_i^*)$ for all $1 \leq i \leq n$. If $f_i^* > 0$ then $f_i^k > 0$ for k large enough. As $z_i^k = 1$ for all such k it follows that $z_i^* = 1 = \operatorname{sign}(f_i^*)$. Similar reasoning applies if $f_i^* < 0$. Lastly, if $f_i^* = 0$ then $\operatorname{sign}(f_i^*) = [-1, 1]$. The entire sequence $\{z_i^k\}_{k=1}^\infty$ therefore lies in $\operatorname{sign}(f_i^*)$, so obviously $z_i^* \in \operatorname{sign}(f_i^*)$ as well.

To show that $\mathcal{Y}^c$ is closed, assume first that

$$f^k \to f^* \tag{30}$$
$$g^k \in \mathcal{Y}^c(f^k) = f^k + c\, P_0 \operatorname{sign}(f^k) \to g^*, \tag{31}$$

where we have used the fact that $\partial B(f) = P_0 \operatorname{sign}(f)$ whenever $f \in K_1$. Thus our goal is to prove that $g^* \in \mathcal{Y}^c(f^*)$. Clearly there exists $z^k \in \operatorname{sign}(f^k)$ such that

$$g^k = f^k + cP_0 z^k. \tag{32}$$

Since z^k lies in a compact set there exists a subsequence $z^{k_i} \to z^*$. So we have

$$f^{k_i} \to f^* \tag{33}$$
$$z^{k_i} \in \operatorname{sign}(f^{k_i}) \to z^* \tag{34}$$

Since sign is closed $z^* \in \operatorname{sign}(f^*)$, which combines with (32) gives

$$g^{k_i} \to f^* + cP_0 z^* \in \mathcal{Y}^c(f^*)$$

where we have used the definition of $\mathcal{Y}^c(f^*)$ and the fact that $f^* \in K_1$. From (31) we then obtain $g^* \in \mathcal{Y}^c(f^*)$ as desired. □

We define the function $\Psi^c : K_1 \times K_2 \to \mathbb{R}^d$

$$\Psi^c(f, g) = \arg\min_u \left\{ T(u) + E(f) \frac{\|u - g\|_2^2}{2c} \right\}$$

Lemma 4. *The function Ψ^c is continuous on $K_1 \times K_2$.*

Proof. Let $h = \Psi^c(f,g)$ and $h' = \Psi^c(f',g')$. Then we have $E(f)\frac{h-g}{c} \in -\partial T(h)$ and $E(f')\frac{h'-g'}{c} \in -\partial T(h')$ so

$$
\begin{aligned}
T(h') &\geq T(h) - \left\langle E(f)\frac{h-g}{c}, h'-h \right\rangle \\
T(h) &\geq T(h') - \left\langle E(f')\frac{h'-g'}{c}, h-h' \right\rangle.
\end{aligned}
$$

By adding these two inequalities,

$$\big\langle E(f)(h-g) - E(f')(h'-g'), h-h' \big\rangle \leq 0.$$

Adding and subtracting we get

$$
\begin{aligned}
&\Big\langle E(f)(h-g) - E(f)(h'-g'), h-h' \Big\rangle + \Big\langle (E(f)-E(f'))(h'-g'), h-h' \Big\rangle \leq 0 \\
&E(f)\Big\langle (h-h') - (g-g'), h-h' \Big\rangle + (E(f)-E(f'))\Big\langle h'-g', h-h' \Big\rangle \leq 0 \\
&E(f)\left(\|h-h'\|_2^2 - \Big\langle g-g', h-h' \Big\rangle \right) + (E(f)-E(f'))\Big\langle h'-g', h-h' \Big\rangle \leq 0 \\
&\|h-h'\|_2^2 \leq \Big\langle g-g', h-h' \Big\rangle - \frac{(E(f)-E(f'))}{E(f)}\Big\langle h'-g', h-h' \Big\rangle
\end{aligned}
$$

From Cauchy-Schwarz we have

$$\|h'-h\|_2 \leq \|g'-g\|_2 + \frac{|E(f')-E(f)|}{E(f)}\,\|h'-g'\|_2 \leq \|g'-g\|_2 + \frac{|E(f')-E(f)|}{E(f)} 2\|g'\|_2$$

The last inequality follows from (21). We then easily conclude that if $(f',g') \to (f,g)$ then $h' \to h$, due to the continuity of E on K_1. □

We next show that the set-valued map $\mathcal{H}^c : K_1 \rightrightarrows \mathcal{F}$

$$\mathcal{H}^c(f) = \Psi^c(f, \mathcal{Y}^c(f))$$

is closed. The fact that the range of $\mathcal{H}$ is in $\mathcal{F}$ is a consequence of Lemma 1.

Lemma 5. *The set-valued map $\mathcal{H}^c$ is closed.*

Proof. Suppose that

$$
\begin{aligned}
&f^k \to f^* &(35)\\
&h^k \in \mathcal{H}^c(f^k) = \Psi^c(f^k, \mathcal{Y}^c(f^k)) \to h^*. &(36)
\end{aligned}
$$

We must show that $h^* \in \mathcal{H}^c(f^*)$. Clearly there exist $g^k \in \mathcal{Y}^c(f^k)$ such that

$$h^k = \Psi^c(f^k, g^k).$$

Since the sequence g^k is in the compact set K_2 there exists $g^* \in K_2$ and a subsequence $g^{k_i} \to g^*$. So we have

$$
\begin{aligned}
&f^{k_i} \to f^* &(37)\\
&g^{k_i} \in \mathcal{Y}^c(f^{k_i}) \to g^*, &(38)
\end{aligned}
$$

from which we conclude that $g^* \in \mathcal{Y}^c(f^*)$ because $\mathcal{Y}^c$ is closed. Now since Ψ^c is continuous we have

$$h^{k_i} = \Psi^c(f^{k_i}, g^{k_i}) \to \Psi^c(f^*, g^*) \in \Psi^c(f^*, \mathcal{Y}^c(f^*)) = \mathcal{H}^c(f^*).$$

But $h^{k_i} \to h^*$, so we may conclude $h^* \in \mathcal{H}^c(f^*)$ as desired. □

Definiton 3 (Critical points)**.** *Let $f \in \mathcal{F}$. We say that f is a critical point of the energy $E(f)$ if there exist $w \in \partial T(f)$ and $v \in \partial B(f)$ so that*

$$0 = w - E(f)v.$$

If both T and B are differentiable at f then the subdifferentials $\partial T(f), \partial B(f)$ are single-valued, so we recover the usual quotient-rule

$$0 = \nabla T(f) - E(f)\nabla B(f).$$

Theorem 1 (Convergence of the algorithm)**.** *Take $f^0 \in \mathcal{F}$ and fix a constant $c > 0$. Let $\{f^k\}_{k=0}^{+\infty} \subset \mathcal{F}$ be a sequence generated by the algorithm. Then*

1. *Any accumulation point f^* of the sequence is a critical point of the energy.*
2. *Either the sequence converges, or the set of accumulation points is a connected subset of $\mathcal{S}^{n-1}$.*

Proof. (1.) The proof is inspired by [8]. Let f^{k_i} denote a subsequence converging to f^*. Since the sequence $\{f^{k_i+1}\}_{i=1}^{\infty}$ lies in a compact set we can extract a further subsequence (still denoted $\{f^{k_i+1}\}$) that converges to some function f'. So we have, as $i \to \infty$

$$f^{k_i} \to f^* \tag{39}$$

$$f^{k_i+1} \to f'. \tag{40}$$

But, because of (24) it must be that $f^* = f'$. Thus we have

$$f^{k_i} \to f^* \tag{41}$$

$$f^{k_i+1} \in P_2(\mathcal{H}^c(f^{k_i})) \to f^*. \tag{42}$$

Clearly there exist $h^{k_i} \in \mathcal{H}^c(f^{k_i})$ such that $f^{k_i+1} = P_2(h^{k_i})$. Since the h^{k_i} eventually lie in the annulus $\mathcal{A} := \{1/2 \leq \|u\|_2 \leq 3/2\}$, we can assume (upon extracting another subsequence) that the $h^{k_i} \in \mathcal{A} \to h^* \in \mathcal{A}$. Therefore we have

$$f^{k_i} \to f^* \tag{43}$$

$$h^{k_i} \in \mathcal{H}^c(f^{k_i}) \to h^* \tag{44}$$

and since $\mathcal{H}^c$ is closed $h^* \in \mathcal{H}^c(f^*)$. Since P_2 is continuous in the annulus $\mathcal{A}$ and all limit points of $\{h^k\}$ lie on $\mathcal{S}^{n-1}$, we conclude that

$$f^{k_i+1} = P_2(h^{k_i}) \to P_2(h^*) = h^* \in \mathcal{H}^c(f^*).$$

From (42) we therefore have $f^* \in \mathcal{H}^c(f^*)$. By definition of $\mathcal{H}^c(f^*)$, if $f^* \in \mathcal{H}^c(f^*)$ then there exists $y^* \in \mathcal{Y}^c(f^*)$ so that

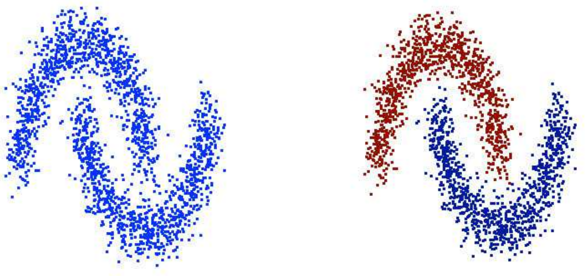
$$f^* = \arg\min_u \left\{ T(u) + E(f^*)\frac{\|u - y^*\|_2^2}{2c} \right\}.$$

Therefore there exists $w^* \in \partial T(f^*)$ so that $0 = cw^* + E(f^*)(f^* - y^*)$. By definition of $\mathcal{Y}^c(f^*)$ there exists $v^* \in \partial B(f^*)$ so that

$$0 = cw^* + E(f^*)(f^* - (f^* + cv^*)) = c(w^* - E(f^*)v^*).$$

Thus f^* is a critical point of the energy according to definition 3.

(2.) For any sequence generated by the algorithm, $||f^{k+1} - f^k||_2 \to 0$ according to lemma 24. Moreover, they lie in the bounded set $\mathcal{S}^{n-1} \subset \mathbb{R}^n$. The hypotheses of Theorem 26.1 of [9] are therefore satisfied, giving the desired conclusion. □



(a) Two moons dataset (b) Desired clustering

Figure 1: Unsupervised clustering of the two moons dataset. Each moon has 1,000 data points in $\mathbb{R}^{100}$.

3 Experiments

We construct the two moons dataset as in [2] (Figure 1). The first moon is a half circle of radius one in $\mathbb{R}^2$, centered at the origin, sampled with a thousand points; the second moon is an upside down half circle also sampled at a thousand points, but centered at $(1, -1/2)$. The dataset is embedded in $\mathbb{R}^{100}$ by adding Gaussian noise with $\sigma = 0.015$. In all experiments we use a 10 nearest neighbors graph with the self-tuning weights as in [13] (the neighbor parameter in the self-tuning is set to 7 and the universal scaling to 1). The constant c in Algorithm 1 is taken to be $c = 1/4$.

Clustering results with different initial conditions are shown in Figure 2. Since the energy is not convex there is no guarantee that the algorithm will converge toward the global minimizer of the ratio cut functional. However, for most initial data, the algorithm indeed finds the correct solution in a very small number of iterative steps.

References

[1] X. Bresson, X.-C. Tai, T.F. Chan, and A. Szlam. Multi-Class Transductive Learning based on ℓ^1 Relaxations of Cheeger Cut and Mumford-Shah-Potts Model. *UCLA CAM Report*, 2012.

[2] T. Bühler and M. Hein. Spectral Clustering Based on the Graph p-Laplacian. In *International Conference on Machine Learning*, pages 81–88, 2009.

[3] 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.

[4] T. Goldstein and S. Osher. The Split Bregman Method for L1-Regularized Problems. *SIAM Journal on Imaging Sciences*, 2(2):323–343, 2009.

[5] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11:1074 –1085, 1992.

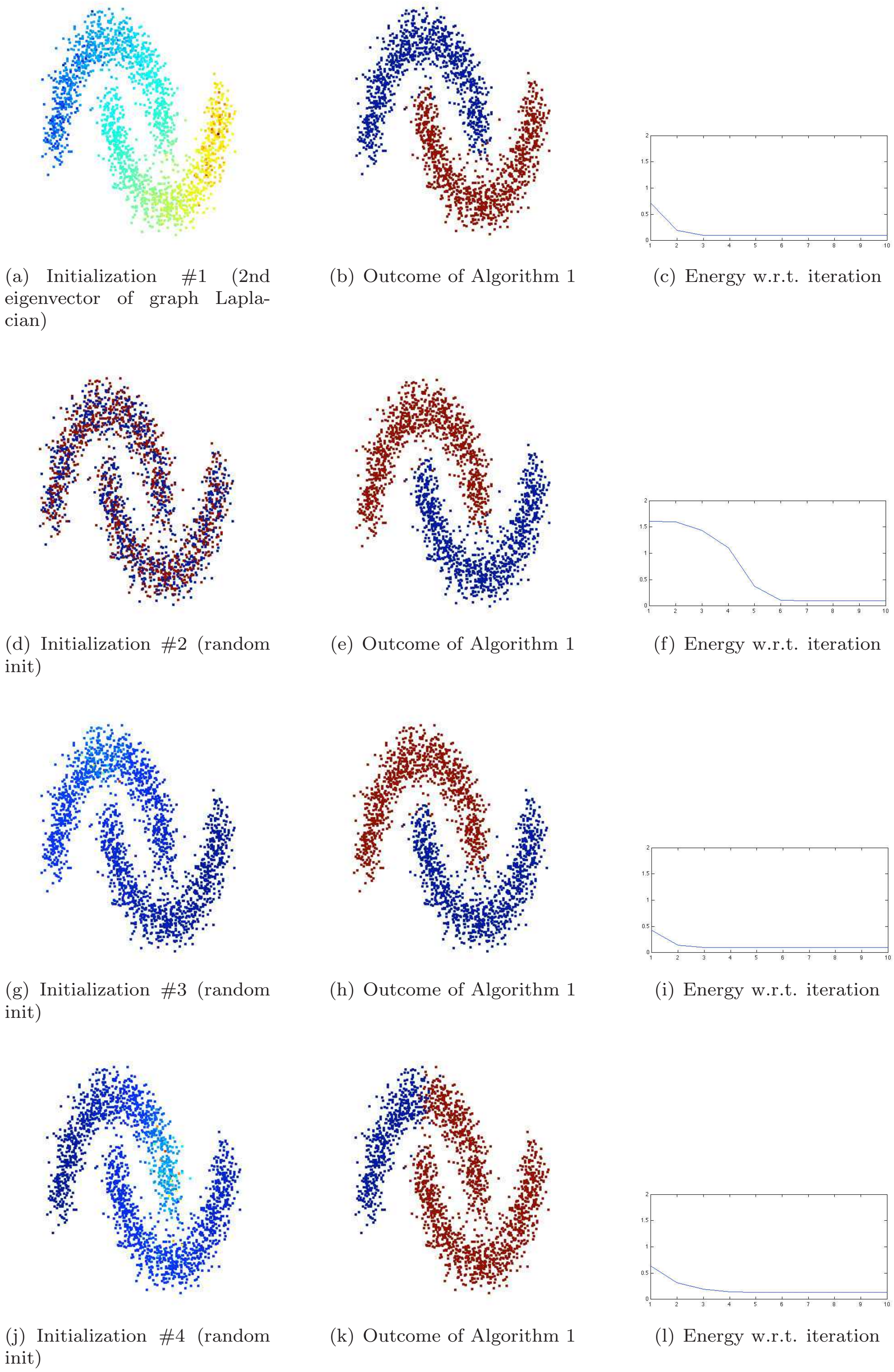
[6] M. Hein and T. Bühler. An Inverse Power Method for Nonlinear Eigenproblems with Applications in 1-Spectral Clustering and Sparse PCA. In *In Advances in Neural Information Processing Systems (NIPS)*, pages 847–855, 2010.

[7] M. Hein and S. Setzer. Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts. In *In Advances in Neural Information Processing Systems (NIPS)*, 2011.

[8] R.R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12(1):108 – 121, 1976.

[9] A. M. Ostrowski. *Solution of Equations in Euclidean and Banach Spaces.* Academic Press, New York, 1973.

[10] S. Rangapuram and M. Hein. Constrained 1-Spectral Clustering. In *International conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1143–1151, 2012.



(a) Initialization #1 (2nd eigenvector of graph Laplacian)

(b) Outcome of Algorithm 1

(c) Energy w.r.t. iteration

(d) Initialization #2 (random init)

(e) Outcome of Algorithm 1

(f) Energy w.r.t. iteration

(g) Initialization #3 (random init)

(h) Outcome of Algorithm 1

(i) Energy w.r.t. iteration

(j) Initialization #4 (random init)

(k) Outcome of Algorithm 1

(l) Energy w.r.t. iteration

Figure 2: Outcomes of Algorithm 1 with different initial data. On the right column the value of the ratio cut functional (2) is plotted versus the number of iterations.

[11] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D*, 60(1-4):259 – 268, 1992.

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

[13] L. Zelnik-Manor and P. Perona. Self-tuning Spectral Clustering. In *In Advances in Neural Information Processing Systems (NIPS)*, 2004.