Local Spectral Clustering of Density Upper Level Sets

Anonymous Authors[1]

Abstract

Spectral clustering methods are a family of popular non-parametric clustering tools. Recent works have proposed and analyzed *local* spectral methods, which extract clusters using locally-biased random walks around a user-specified seed node, and are known to have worst-case guarantees for certain graph-based measures of cluster quality. In contrast to existing works, we analyze the personalized PageRank (PPR) algorithm in the classical statistical learning setup, where we obtain samples from an unknown distribution, and aim to identify connected regions of high-density (density clusters). With respect to a neighborhood graph formed over these samples, we provide guarantees for a pair of cluster quality criteria evaluated on empirical analogues to these density clusters. As a result, the PPR algorithm is shown to extract sufficiently salient density clusters.

1. Introduction

Let $\mathbf{X} := (x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Our statistical learning task is clustering: splitting data into groups which satisfy some notion of within-group similarity and between-group difference.

In particular, spectral clustering methods are a family of powerful non-parametric clustering algorithms. Given a symmetric adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with (i, j)th entry representing the similarity between data points x_i and x_j, we form the random walk transition probability matrix $\mathbf{W}$, and corresponding graph Laplacian matrix $\mathbf{L}$[1]:

$$\mathbf{W} := \mathbf{D}^{-1}\mathbf{A}; \quad \mathbf{L} = \mathbf{I}_n - \mathbf{W} \tag{1}$$

where the degree matrix $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} := \sum_j \mathbf{A}_{ij}$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Often, either of the Laplacian matrices $\mathbf{L}_{sym} := \mathbf{D}^{\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$ or $\mathbf{L}_{unn} := \mathbf{D} - \mathbf{A}$ are used instead.

Roughly speaking, spectral clustering techniques first embed the data $\mathbf{X}$ using the spectrum of the graph Laplacian, and subsequently use this *spectral embedding* to find a clustering of the data. When applied to large graphs (or large point clouds) classical global spectral methods can be computationally cumbersome and insensitive to the local geometry of the distribution of the samples (Mahoney et al., 2012; Leskovec et al., 2010). This in turn has led to the investigation of local spectral algorithms (Spielman & Teng, 2013; Andersen et al., 2006; Leskovec et al., 2010) which leverage locally-biased spectra computed using random walks around a user-specified seed node.

A natural model to consider when analyzing point cloud data such as $\mathbf{X}$ is the following:

$$x_i \sim \mathbb{P}, \quad \text{independently, for } i = 1, \ldots, n, \tag{2}$$

with f the density function of $\mathbb{P}$ with respect to the uniform measure over $\mathbb{R}^d$. In this case, we are interested in understanding what the output of a clustering algorithm on this finite sample reveals about the unknown density f. For $\lambda > 0$ and the upper level set $\{x : f(x) \geq \lambda\}$, it is intuitive (Hartigan, 1981; Chaudhuri & Dasgupta, 2010) to define clusters as the connected components $\mathbb{C}_f(\lambda)$ of the upper level set; we call these connected regions of high density *density clusters*, and study the ability of spectral methods to identify such clusters.

Graph connectivity criteria. A somewhat more standard mode of understanding spectral clustering methods is to view them as approximating some graph connectivity criteria.

For $\mathbf{A}$ as before, consider the graph $G = (V, E)$, with vertices $V = \{v_1, \ldots, v_n\}$ corresponding to the n rows of $\mathbf{A}$, and (possibly weighted) edges $E = \{(v_i, v_j, \mathbf{A}_{ij}) : 1 \leq i < j \leq n, \mathbf{A}_{ij} > 0\}$ (Since $\mathbf{A}$ is symmetric, G is an undirected graph; also, by convention, we preclude self-loops, hence $i \neq j$). There are many (Yang & Leskovec, 2015; Fortunato, 2010) graph-based measures which assess the cluster quality of a subset $S \subseteq V$ (or more generally the quality of a partition $S_1 \cup \ldots \cup S_m = V$, for $m \geq 2$.)

Arguably a natural way to assess cluster quality is via a pair of criteria capturing the *external* and *internal connectivity* of S, respectively. As the names suggest, external connectivity

should relate to the number of edges between S and its complement V/S (hereafter denoted S^c), while internal connectivity in turn measures the number of edges between subsets within S. The graph clustering task then becomes to find a subset S (or, for global algorithms, a partition $S_1 \cup \ldots \cup S_m = V$), which has both small external and large internal connectivity.

We will assess the external connectivity of a subset $S \subseteq V$ through its normalized cut. The cut of S is

$$\mathrm{cut}(S; G) := \sum_{u \in S} \sum_{v \in S^c} \mathbf{1}\left((u, v) \in E\right)$$

– where $S^c = V/S$ is the complement of S in V – and the volume of S is

$$\mathrm{vol}(S; G) := \sum_{u \in S} \sum_{v \in V} \mathbf{1}\left((u, v) \in E\right).$$

Then, the *normalized cut* of S is given by

$$\Phi(S; G) := \frac{\mathrm{cut}(S; G)}{\min\left\{\mathrm{vol}(S; G), \mathrm{vol}(S^c; G)\right\}} \tag{3}$$

Intuitively, a set with low *normalized cut* has many more edges which do not cross the cut than edges which do cross the cut.

Given $S \subseteq V$, the subgraph induced by S is given by $G[S] = (S, E_S)$, where $(u, v) \in E_S$ if both u and v are in S and $(u, v) \in E_S$. Letting $|S| = m$, $\mathbf{A}_S$ then denotes the $m \times m$ adjacency matrix representation of $G[S]$; similarly, $\mathbf{D}_S$ is the diagonal degree matrix with entries $(\mathbf{D}_S)_{ii} = \sum_{j : v_j \in S} \mathbf{A}_{ij}$, and $\mathbf{W} = \mathbf{D}_S^{-1}\mathbf{A}_S$ is the corresponding random walk matrix (again, defined only over $G[S]$.)

Our internal connectivity parameter $\Psi(S)$ will capture the time it takes for the random walk governed by $\mathbf{W}_S$ to mix (that is, approach a stationary distribution) uniformly over S. Denoting the stationary distribution of this random walk by $\widetilde{\boldsymbol{\pi}} = (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_m)$, the *relative pointwise mixing time* $\tau_\infty(S; G[S])$ is defined to be: the smallest integer $t_0 > 0$ such that for all $v = v_i, v' = v_j \in S$:

$$\left| \frac{e_v \mathbf{W}_S^t - \widetilde{\pi}_j}{\widetilde{\pi}_j} \right| \leq \frac{1}{4}$$

for all $t > t_0$ [2] (Of course, given the definition of $\mathbf{W}_S$ it is well known that the stationary distribution $\boldsymbol{\pi}_S$ will be defined by $\widetilde{\pi}_i = (\mathbf{D}_S)_{ii}/\mathrm{vol}(S; G[S])$.

Intuitively, the smaller the pointwise mixing time, the more connected every pair of points v and v' are in the graph $G[S]$. Therefore, the internal connectivity parameter $\Psi(S; G)$ is simply one over the mixing time:

$$\Psi(S; G) = \frac{1}{\tau_\infty(S; G[S])} \tag{4}$$

If S has normalized cut no greater than Φ, and inverse mixing time no less than Ψ, we will refer to it as a (Φ, Ψ)-cluster. Both local (Zhu et al., 2013) and global (Kannan et al., 2004) spectral algorithms have been shown to output clusters (or partitions) which provably satisfy approximations to the optimal (Φ, Ψ)-cluster (or partition) for a given graph G. [3]

Personalized PageRank. As mentioned previously, global algorithms which find spectral cuts may be computationally infeasible for large graphs; in this setting, local algorithms may be preferred or even required. We will restrict our attention in particular to one such popular algorithm: *personalized PageRank* (PPR). The personalized PageRank algorithm was first introduced by (Haveliwala, 2003) and variants of this algorithm have been studied further in several recent works (Spielman & Teng, 2011; 2014; Zhu et al., 2013; Andersen et al., 2006; Mahoney et al., 2012).

The random walk matrix $\mathbf{W}$ over the graph $G = (V, E)$ with associated adjacency matrix $\mathbf{A}$ is given by (1). PPR is then defined with respect to the following inputs: a user-specified seed node $v_i \in V$, and $\alpha \in [0, 1]$ a teleportation parameter. Letting $v = v_i$ for notational simplicity, and e_v be the indicator vector for v (meaning e_v has a 1 in the ith location and 0 everywhere else), the *PPR vector* is given by the recursive formulation

$$\mathbf{p}(v, \alpha; G) := \alpha \mathbf{e}_v + (1 - \alpha)\mathbf{p}(v, \alpha; G)\mathbf{W} \tag{5}$$

We note in passing that, for $\alpha > 0$, the vector $\mathbf{p}(v, \alpha; G)$ can be well-approximated by a simple local computation (of a random walk with restarts at the node v.) We also point out that, from a density clustering standpoint, since density clusters are inherently local, using the PPR algorithm eases the analysis, and as we will observe in the sequel our analysis requires fewer global regularity conditions relative to more classical global spectral algorithms.

To compute a cluster $\widehat{C} \subset V$ using the PPR vector, we will take sweep cuts of $\mathbf{p}(v, \alpha; G)$. Denote the entries of $\mathbf{p}(v, \alpha; G)$ by $\mathbf{p}(v, \alpha; G) = (p_1, \ldots, p_n)$, and let the *stationary distribution* $\boldsymbol{\pi}$ of the random walk defined by $\mathbf{W}$ be

[2] Given a starting node v and and a random walk defined by transition probability matrix $\mathbf{P}$, the rotation $e_v\mathbf{P}^t$ is used to denote the distribution of the random walk after t steps.

[3] In the case of (Kannan et al., 2004), the internal connectivity parameter ϕ is actually the conductance, i.e. the minimum normalized cut within the subgraph $G[S]$. See Theorem 3.1 for details; however, note that $\phi^2/\log(\mathrm{vol}(S)) \leq O(\Psi)$, and so the lower bound on ϕ translates to a lower bound on Ψ.

given by

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n), \quad \pi_j := \frac{\mathbf{D}_{jj}}{\mathrm{vol}(V; G)}.$$

Then, for a number $\beta \geq 0$, the sweep cut S_β is

$$S_\beta = \{u_j \in V : p_j > \beta \pi_j\}. \tag{6}$$

We will choose one such sweep cut to be our cluster estimate $\widehat{C}$. (We delay formal introduction of the local clustering algorithm we analyze until 2, after we have given a method for forming a graph over the data $\mathbf{X}$.)

Large sample behavior. Let (r_n) be a sequence of positive numbers. Given a sequence of kernel functions $k_n : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ of the form $k_n(x, x') = k(\|x - x'\| / r_n)$ for k a non-increasing function, and data $\mathbf{X} = \{x_1, \ldots, x_n\}$ sampled from $\mathbb{P}$ as before, form the (weighted, complete) similarity graph $G_n = (\mathbf{X}, E_n)$ with $E_n = \{k(x_i, x_j) : 1 \leq i < j \leq n\}$. (Here, $\|\cdot\|$ is used to denote Euclidean norm).

In this context, continuous analogues to (for instance) normalized cut have been defined, over the data-manifold rather than the graph, and convergence of finite sample graph-theoretic functionals to their continuous counterparts has been shown (Trillos et al., 2016; Ery et al., 2012; Maier et al., 2011). However these continuous analogues are not always easily interpretable – and their corresponding minimizers not always easily identifiable – for the particular density function under consideration. Of course, relating these partitions to the arguably more simply defined high density clusters can be also challenging in general. Intuitively, however, under the right conditions such high-density clusters should have more edges within themselves than to the remainder of the graph. We formalize this intuition next.

1.1. Summary of results

Hereafter, we consider the uniform kernel function for a fixed $r > 0$,

$$k(x, x') = \mathbf{1}(\|x - x'\| \leq r) \tag{7}$$

and the associated neighborhood graph

$$G_{n,r} = (\mathbf{X}, E_{n,r}),\ (x_i, x_j) \in E_{n,r} \text{ if } k(x_i, x_j) = 1 \tag{8}$$

For a given high density cluster $\mathcal{C} \subseteq \mathbb{C}_f(\lambda)$, we call $\mathcal{C}[\mathbf{X}] = \mathcal{C} \cap \mathbf{X}$ the *empirical density cluster*. We now introduce a notion of consistency for the task of density cluster estimation:

Definition 1 (Consistent density cluster estimation)**.** *For an estimator $\widehat{\mathcal{C}}_n \subset \mathbf{X}$, and any $\mathcal{C}, \mathcal{C}' \in \mathbb{C}_f(\lambda)$, we say $\widehat{\mathcal{C}}_n$ is a consistent estimator of $\mathcal{C}$ if the following statement holds: as the sample size $n \to \infty$, each of the following*

$$\mathcal{C}[\mathbf{X}] \subseteq \widehat{\mathcal{C}}_n, \text{ and } \widehat{\mathcal{C}}_n \cap \mathcal{C}'[\mathbf{X}] = \emptyset \tag{9}$$

occur with probability tending to 1.

Our results can now be summarized by the following two points:

1. Under a natural set of geometric conditions[4], the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$ can be bounded. Theorems 1 and Theorem 2 provide an upper and lower bound, respectively
2. We show these bounds on the graph connectivity criteria have algorithmic consequences for personalized PageRank. An immediate consequence of Theorems 1 and 2, along with the previous work of (Zhu et al., 2013), is to yield an upper bound on the normalized cut of the set $\widehat{C}$ output by Algorithm 1, as well as upper bounding the symmetric set difference between $\widehat{C}$ and $\mathcal{C}[\mathbf{X}]$. Furthermore, in 4 we show that a careful analysis of the form typical to local clustering algorithms yields Theorem 3, which states that Algorithm 1, properly initialized, performs consistent density cluster estimation in the sense of (9).

Organization. In Section 5, we provide some example density functions, to clarify the relevance of our results, and empirically demonstrate that violations of the geometric conditions we set out in Section 2 manifestly impact density cluster recovery (i.e. the conditions are not superfluous), before concluding in **??**. First, however, we summarize some related work.

1.2. Related Work

In addition to the background given above, a few related lines of work are worth highlighting.

Global spectral clustering methods were first developed in the context of graph partitioning (Fiedler, 1973; Donath & Hoffman, 1973) and their performance is well-understood in this context (see, for instance, (Tolliver & Miller, 2006; von Luxburg, 2007)). In a similar vein, several recent works (McSherry, 2001; Lei & Rinaldo, 2015; Rohe et al., 2011; Abbe, 2018; Chaudhuri et al., 2012; Balakrishnan et al., 2011) have studied the efficacy of spectral methods in successfully recovering the community structure in various variants of the stochastic block model.

[4]We formally introduce the geometric conditions in Section 2. They preclude clusters which are too thin and long, or those for which the gap in density between the high density area and the outside is not sufficiently large

Building on the work of Koltchinskii & Gine (2000) the works (von Luxburg et al., 2008; Hein et al., 2005) for instance, have studied the limiting behaviour of spectral clustering algorithms. These works show that when samples are obtained from a distribution, following appropriate graph construction, in certain cases the spectrum of the Laplacian converges to that of the Laplace-Beltrami operator on the data-manifold. However, relating the partition obtained using the Laplace-Beltrami operator, to the more intuitively defined high-density clusters, can be challenging in general.

Perhaps most similar to our results are (Vempala & Wang, 2004; Shi et al., 2009; Schiebinger et al., 2015), which study the consistency of spectral algorithms in recovering the latent labels in certain parametric and non-parametric mixture models. These results focus on global rather than local algorithms, and as such impose global rather than local conditions on the nature of the density. Moreover, they do not in general ensure recovery of density clusters which is a focus of our work.

2. Background and Assumptions.

We begin by introducing an defining well-conditioned density clusters before turning to formally define the PPR algorithm under consideration, Algorithm 1, and discussing the choice of tuning-parameters.

2.1. Well-conditioned density clusters.

In order to provide meaningful bounds on the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$, we must introduce some assumptions on the density f.

Let $B(x,r) = \{y \in \mathbb{R}^d : \|y-x\| \leq r\}$ be a closed ball of radius r around the point x. Given a set $\mathcal{A} \subset \mathbb{R}^d$, and a number $\sigma > 0$, define the σ-expansion of $\mathcal{A}$ to be $\mathcal{A}_\sigma = \mathcal{A} + B(0,\sigma) = \{y \in \mathbb{R}^d : \inf_{x\in\mathcal{A}} \|y-x\| \leq \sigma\}$. We are now ready to give the assumptions, which we state with respect to a density cluster $\mathcal{C} \in \mathbb{C}_f(\lambda)$ for some $\lambda > 0$, and expansion parameter $\sigma > 0$:

(A1) *Bounded density within cluster:* There exist numbers $0 < \lambda_\sigma < \Lambda_\sigma < \infty$ such that:

$$\lambda_\sigma = \inf_{x\in\mathcal{C}_\sigma} f(x) \leq \sup_{x\in\mathcal{C}_\sigma} f(x) \leq \Lambda_\sigma \qquad (10)$$

(A2) *Low noise density:* For some $\gamma > 0$, there exists a constant $c_1 > 0$ such that for all $x \in \mathbb{R}^d$ with $0 < \rho(x,\mathcal{C}_\sigma) \leq \sigma$,

$$\inf_{x'\in\mathcal{C}_\sigma} f(x') - f(x) \geq c_1 \rho(x,\mathcal{C}_\sigma)^\gamma,$$

where $\rho(x,\mathcal{A}) = \min_{x_0\in\mathcal{A}} \|x - x_0\|$ for $\mathcal{A} \subset \mathbb{R}^d$.

(A3) *Cluster separation:* For all $\mathcal{C}' \in \mathbb{C}_f(\lambda)$,

$$\rho(\mathcal{C}_\sigma, \mathcal{C}'_\sigma) > \sigma,$$

where $\rho(\mathcal{A},\mathcal{A}') = \min_{x\in\mathcal{A}} \rho(x,\mathcal{A}')$ for $\mathcal{A}' \subset \mathbb{R}^d$.

(A4) *Cluster diameter:* There exists $D < \infty$ such that for all $x, x' \in \mathcal{C}_\sigma$:

$$\|x - x'\| \leq D,$$

and $\mathcal{C}_\sigma$ is a convex set.

We note that σ plays several roles here, precluding arbitrarily narrow clusters and long clusters in (A1) and (A4), flat densities around the level set in (A2), and poorly separated clusters in (A3).

Assumptions (A1)-(A3) are used to upper bound $\Phi(\mathcal{C}[\mathbf{X}]; G_{n,r})$, whereas (A1) and (A4) are necessary to lower bound $\Psi(\mathcal{C}[\mathbf{X}]; G_{n,r})$. We note that the lower bound on minimum density in (10) and (A3) combined are similar to the (σ,ϵ)-saliency of (Chaudhuri & Dasgupta, 2010), a standard density clustering assumption, while (A2) is seen in, for instance, (Singh et al., 2009), (as well as many other works on density clustering and level set estimation.) It is worth highlighting that these assumptions are all local in nature, a benefit of studying a local algorithm such as PPR.

Now we make the crucial distinction between the assumptions made in density clustering, and the stronger conditions we will require here. First, we introduce

$$\mathbf{\Phi}(\sigma,\lambda,\lambda_\sigma,\gamma) := \frac{\lambda}{\lambda_\sigma}\frac{(\lambda_\sigma - \frac{\sigma^\gamma}{\gamma+1})}{\lambda_\sigma}$$
$$\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D) := \frac{\sigma^d}{D^d}\frac{\lambda_\sigma^7}{\Lambda_\sigma^7}\frac{1}{\log(\Lambda_\sigma^2/(\lambda_\sigma\sigma^d\nu_d))}.$$

Well-conditioned density clusters satisfy all of the given assumptions, for parameters which results in 'good' values of $\mathbf{\Phi}$ and $\mathbf{\Psi}$.

Definition 2 (Well-conditioned density clusters)**.** *For $\lambda > 0$ and $\mathcal{C} \in \mathbb{C}_f(\lambda)$, let $\mathcal{C}$ satisfy (A1) - (A4) with respect to parameters $\sigma, \lambda_\sigma, \gamma > 0$ and $\Lambda_\sigma, D < \infty$. Letting $\kappa_1(\mathcal{C})$ and $\kappa_2(\mathcal{C})$ be given by*

$$\kappa_1(\mathcal{C}) := \frac{\mathbf{\Phi}(\sigma,\lambda,\lambda_\sigma,\gamma)}{\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)}$$
$$\kappa_2(\mathcal{C}) := \kappa_1(\mathcal{C}) \cdot \sqrt{\mathbf{\Psi}(\sigma,\lambda_\sigma,\Lambda_\sigma,D)},$$

we call $\mathcal{C}$ a (κ_1,κ_2)-well-conditioned density cluster (with respect to $\sigma,\lambda_\sigma,\gamma,\Lambda_\sigma$ and D).

As the notation suggests, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are related to the graph connectivity criteria. In fact, as we will see in the next section, for well-conditioned density clusters they are exactly (up to constants) the bounds we derive.

2.2. Algorithm under consideration.

Algorithm 1 will be the simple variant of PPR we analyze. It will take as input the data $\mathbf{X}$ along with user-specified parameters $r, \alpha, \mathrm{vol}_0$, and $v \in \mathbf{X}$.

Algorithm 1 PPR on a neighborhood graph

Input: data $\mathbf{X}$, radius r, teleportation parameter $\alpha \in [0,1]$, seed node $v \in \mathbf{X}$, target volume vol_0.
Output: $\widehat{C} \subset V$.

1: Form the neighborhood graph $G_{n,r}$ as given in (8)
2: Compute PPR vector $\mathbf{p}(v, \alpha; G_{n,r})$ as defined by (5).
3: For $\beta \in [\frac{1}{8}, \frac{1}{2}]$ compute sweep cuts S_β as defined by (6).
4: Return

$$\widehat{C} = \underset{\beta \in [\frac{1}{8}, \frac{1}{2}]}{\arg\min}\, \Phi(S_\beta; G_{n,r})$$

As is typical in the local clustering literature, our results will be stated with respect to specific choices or ranges of each of the user-specified parameters, which in this case may depend on the underlying (unknown) density.

In particular, for a well conditioned density cluster $\mathcal{C}$ (with respect to some $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ and D), we require

$$\begin{aligned} \alpha \in [1/2, 9/10] \cdot \mathbf{\Psi}(\sigma, \lambda_\sigma, \Lambda_\sigma, D), r \in [1/2, 1] \cdot \sigma \\ \mathrm{vol}_0 \in [1/2, 1] \cdot \mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}]), v \in \mathcal{C}_\sigma[\mathbf{X}]^g \end{aligned} \quad (11)$$

where $\mathcal{C}_\sigma[\mathbf{X}]^g \subseteq \mathcal{C}_\sigma[\mathbf{X}]$ is some 'good' subset of $\mathcal{C}_\sigma[\mathbf{X}]$ which, as we will see, satisfies $\mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}]^g) \geq \mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}])/2$. (Intuitively one can think of $\mathcal{C}_\sigma[\mathbf{X}]^g$ as being the nodes sufficiently close to the center of $\mathcal{C}_\sigma[\mathbf{X}]$, although we provide no formal justification to this effect.) We call a PPR algorithm run with hyperparameters satisfying (11) *well-initialized*.

3. Local Clustering on Density Level Sets

In this section we provide bounds for the normalized cut and inverse mixing time of an empirical density cluster $\mathcal{C}[\mathbf{X}]$. As a result we can lower bound $\Phi(\widehat{C}; G_{n,r})$ for $\widehat{C}$ the output of a well-initialized PPR algorithm, and upper bound the symmetric set difference between $\widehat{C}$ and $\mathcal{C}[\mathbf{X}]$.

For notational simplicity, hereafter for $S \subseteq \mathbf{X}$ we will refer to $\Phi(S; G_{n,r})$ as $\Phi_{n,r}(S)$, and likewise with $\Psi(S; G_{n,r})$ and $\Psi_{n,r}(S)$. We will also use $\nu(\cdot)$ to denote the uniform measure over $\mathbb{R}^d$, and $\nu_d = \nu\big(B(0,d)\big)$ as the measure of the unit ball.

We begin with a lower bound on the normalized cut in Theorem 1. We will require Assumptions (A1)-(A3) to hold; however, the upper bound on density in (10) will not be needed and so we omit the parameter Λ_σ from the statement of the theorem.

Theorem 1. *For some $\lambda > 0$, let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1)-(A3) for some $\sigma, \lambda_\sigma, c_1, \gamma > 0$. Then, for any $r < \sigma$ and $\delta \in (0,1]$, the following statements hold with probability at least $1 - \delta$: Fix $\epsilon > 0$. Then, for*

$$n \geq \frac{9 \log(2/\delta)}{\epsilon^2} \left(\frac{1}{\lambda_\sigma^2 \nu(\mathcal{C}_\sigma) \nu_d r^d} \right)^2 \quad (12)$$

we have

$$\frac{\Phi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}])}{r} \leq c_\sigma \frac{\lambda}{\lambda_\sigma} \frac{(\lambda_\sigma - \frac{r^\gamma}{\gamma+1})}{\lambda_\sigma} + \epsilon \quad (13)$$

where $c_0 = 4d$ and $c_\sigma = c_0/\sigma$.

The proof of Theorem 1, along with all other proofs, can be found in the supplementary document. A few remarks are in order.

Remark 1. Note that the bound of (13) depends exponentially on d; precisely, on $C_0 = d2^{2d+1}$. It is possible to improve this dependency to the order of $(1 + \frac{r}{\sigma})^{2d}$. However, as indicated by (11), we will think of r as being a constant radius of order σ (rather than $r = r_n \to 0$ as $n \to \infty$) and therefore even this improvement still retains an exponential dependency on the dimension d.

Remark 2. Aside from the looseness implied by Remark 1, the error bound of (13) is almost tight. Specifically, choosing

$$\mathcal{A}_\sigma = B(0, \sigma),$$
$$f(x) = \begin{cases} \lambda \text{ for } x \in \mathcal{A}_\sigma, \\ \lambda - \rho(x, \mathcal{A}_\sigma)^\gamma \text{ for } 0 < \rho(x, \mathcal{A}_\sigma) < r \end{cases}$$

we have that for n within constant order of the lower bound in (12), with probability at least $1 - \delta$

$$\frac{\Phi_{n,r}(\mathcal{A}_\sigma[\mathbf{X}])}{r} \geq c \frac{(\lambda - \frac{r^{\epsilon+1}}{\epsilon+1})}{\lambda} - \epsilon$$

for some constant c which depends only on dimension. (Note that a factor of $1/\sigma$ in c_σ is not replicated in this lower bound.) For justification see the supplementary materials.

We now provide an upper bound for $\Psi_{n,r}(\mathcal{C}[\mathbf{X}])$.

Theorem 2. *Fix $\lambda > 0$, and let $\mathcal{C} \in \mathbb{C}_f(\lambda)$ satisfy Assumptions (A1) and (A4) for some $\sigma, \lambda_\sigma, \Lambda_\sigma, D > 0$. Then, for any $r < \sigma/4d$, the following statement holds: with proba-*

bility one

$$\liminf_{n\to\infty}\Psi_{n,r}(\mathcal{C}_\sigma[\mathbf{X}]) \geq c_1\frac{\Lambda_\sigma^8}{\lambda_\sigma^8}(\log\mu + c_\lambda)\Big(c_\lambda(d^3\log\mu + c_2) + \frac{d^2D^2}{r^2}\big(\log d + c_2\frac{d^2D^2}{r^2}\frac{\Lambda_\sigma^2}{\lambda_\sigma^2}c_\lambda + c_3 + \log\log\mu\big)\Big), \tag{14}$$

where $\mu = \log(\frac{2D}{r})$, $c_1 = 3852800/3$, $c_2 = 331776$, *and* $c_3 = \log 2$ *are all dimension free, and* $c_\lambda = \log(\Lambda_\sigma^2/\lambda_\sigma^2)$.

Approximate density cluster recovery with PPR. In (Zhu et al., 2013), building on the work of (Andersen et al., 2006) and others, theory is developed which links algorithmic performance of PPR to the normalized cut and mixing time parameters. We collect some of the main results of (Zhu et al., 2013) in Lemma 1.

For $G = (V, E)$ consider some $A \subseteq V$, and let $\Phi(A; G)$ and $\Psi(A; G)$ be defined as in (3) and (4), respectively.

Lemma 1 (PPR clustering)**.** *There exists a set* $A^g \subset A$ *with* $\mathrm{vol}(A^g; G) \geq \mathrm{vol}(A; G)/2$ *such that the following statement holds: Choose any* $v \in A^g$, *fix* $\alpha = 9/10\Psi(A; G)$, *and compute the page rank vector* $\mathbf{p}(v, \alpha; G)$. *Letting*

$$\widehat{C} = \arg\min_{\beta\in[\frac{1}{8},\frac{1}{2}]} \Phi(S_\beta; G)$$

the following guarantees hold:

$$\mathrm{vol}(\widehat{C} \setminus A) \leq \frac{24\Phi(A; G)}{\Psi(A; G)}\mathrm{vol}(A)$$
$$\mathrm{vol}(A \setminus \widehat{C}) \leq \frac{30\Phi(A; G)}{\Psi(A; G)}\mathrm{vol}(A)$$
$$\Phi(\widehat{C}; G) = O\left(\frac{\Phi(A; G)}{\sqrt{\Psi(A; G)}}\right)$$

Corollary 1 immediately follows.

Corollary 1. *Fix* $\lambda > 0$, *and let* $\mathcal{C} \in \mathbb{C}_f(\lambda)$ *be a* (κ_1, κ_2)*-well conditioned cluster (with respect to some* $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ *and* D*). Then, if Algorithm 1 is well-initialized (in the sense that the choices of input parameters satisfy (11)), the following guarantees hold for output set* $\widehat{C} \subset \mathbf{X}$*:*

$$\mathrm{vol}(\widehat{C} \setminus \mathcal{C}_\sigma[\mathbf{X}]), \mathrm{vol}(\mathcal{C}_\sigma[\mathbf{X}] \setminus \widehat{C}) \leq 30\kappa_1(\mathcal{C})\mathrm{vol}(\mathcal{C}_\sigma[X])$$
$$\Phi_{n,r}(\widehat{C}) = O\left(\kappa_2(\mathcal{C})\right)$$

4. Consistent cluster estimation with PPR.

While Corollary 1 is nice, it does not imply consistency of the type laid out in Definition 1.

However, a slightly more subtle argument regarding the way a random walk on $G_{n,r}$ distributes mass over $\mathcal{C}[\mathbf{X}]$ and $\mathcal{C}'[\mathbf{X}]$ will do the trick, yielding Theorem 3.

Theorem 3. *Fix* $\lambda > 0$, *and let* $\mathcal{C} \in \mathbb{C}_f(\lambda)$ *be a* (κ_1, κ_2)*-well conditioned cluster (with respect to some* $\sigma, \lambda_\sigma, \gamma, \Lambda_\sigma$ *and* D*), with*

$$\kappa_2 \leq \left(\frac{1}{5} - \frac{\epsilon}{4}\right)\frac{\lambda_\sigma^2}{\Lambda_\sigma^2}\frac{\sigma^d\nu_d}{\nu(\mathcal{C}_\sigma)} \tag{15}$$

for some $\epsilon > 0$.

Then, if Algorithm 1 is well-initialized (in the sense that the choices of input parameters satisfy (11)), the output set $\widehat{C} \subset \mathbf{X}$ *is a consistent estimator for* $\mathcal{C}$, *in the sense of Definition 1.*

Remark 3. When $\lambda_\sigma = \Lambda_\sigma$ and $\nu(\mathcal{C}_\sigma) = \sigma^d\nu_d$ – in some sense, the 'best' case given our assumptions – the upper bound of (15) reduces to $\kappa_2 \leq 1/5 - \epsilon/4$. That $\kappa_2 = O(1)$ is a minimal requirement for the bounds of Corollary 1 not to be facile; therefore, in this case the assumed upper bound of (15) is not really a stricter requirement.

5. Examples

Example 1 is intended to show how the machinery developed above translates in a specific, common mixture model, and the extent to which bounds are (or are not) tight. Example 2 will try to delve into some of the details of how PPR interpolates the conductance and density cut, and will show a case where a poorly conditioned density cluster is not recovered by PPR . Example 3 will emphasize finite sample cluster recovery, for a well-conditioned but non-convex mixture model

Examples 1 and 2 should be thought of as shedding light on the population performance of PPR, whereas Example 3 shows performance on a finite sample.

1. *Gaussian Mixture Model:* We will compute optimal $\mathbf{\Phi}$ and $\mathbf{\Psi}$ for given λ, and show the following
   - A graph comparing $\mathbf{\Phi}$ to $\Phi_{n,r}$ as the value of λ changes.
   - A graph comparing $\mathbf{\Psi}$ to $\Psi_{n,r}$ as the value of λ changes.
   - That for some values of λ, the conditions required for Theorem 3 hold.
2. *Thin and long parallel clusters, with* ϵ*-uniform noise:* We will (try to) show that the set outputted by PPR interpolates between the minimum normalized cut solution (fatter) and the density cluster (thinner). The *conductance* is

$$\Phi^\star(G_{n,r}) := \min_{C\subset\mathbf{X}} \Phi_{n,r}(C)$$

and the conductance cut is $C^\star \subset \mathbf{X}$ which achieves the minimum.

We will show that

- For sufficiently small ϵ, all three of the conductance cut, PPR cut, and density cut agree.
- For an intermediate value of ϵ, the conductance cut and the density cut disagree. The PPR cut interpolates between the two.
- For a sufficiently large value of ϵ, the PPR cut fails to recover the density cut, and draws closer to the conductance cut.

3. *Non-convex mixture model:* We will show that, for well-conditioned non-convex mixture model, and a finite sample size n, cluster recovery is achieved with high probability over repeated simulations.

6. Discussion

For a clustering algorithm and a given object (such as a graph or set of points), there are an almost limitless number of ways to define what the 'right' clustering is. We have considered a few such ways – density level sets, and the bicriteria of normalized cut, inverse mixing time – and shown that under the right conditions, the latter agree with the former, with resulting algorithmic consequences.

There are still many directions worth pursuing in this area. Concretely, we might wish to generalize our results to hold over a wider range of kernel functions, and hyperparameter inputs to the PPR algorithm. More broadly, we do not provide any sort of theoretical lower bound, although we give empirical evidence in Example 2 that poorly conditioned density clusters are not consistently estimated by PPR . Example 2 also hints at a way of understanding local spectral algorithms – or, at least, PPR– as interpolating between normalized and density cut. Exploring this connection is an avenue for future work.

References

Abbe, E. Community detection and stochastic block models. *Foundations and Trends® in Communications and Information Theory*, 14(1-2):1–162, 2018.

Andersen, R., Chung, F., and Lang, K. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 475–486, 2006.

Balakrishnan, S., Xu, M., Krishnamurthy, A., and Singh, A. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc., 2011.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for the cluster tree. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 343–351. Curran Associates, Inc., 2010.

Chaudhuri, K., Graham, F. C., and Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. In *COLT*, volume 23, pp. 35.1–35.23, 2012.

Donath, W. E. and Hoffman, A. J. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, 17(5):420–425, September 1973.

Ery, A.-C., Pelletier, B., and Pudlo, P. The normalized graph cut and cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44(4):907–937, 12 2012.

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

Fortunato, S. Community detection in graphs. *Physics Reports*, 486(3):75 – 174, 2010. ISSN 0370-1573.

Hartigan, J. A. Consistency of single-linkage for high-density clusters. *Journal of the American Statistical Association*, 1981.

Haveliwala, T. H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

Hein, M., Audibert, J.-Y., and von Luxburg, U. From graphs to manifolds – weak and strong pointwise consistency of graph laplacians. In *Learning Theory*, 2005.

Kannan, R., Vempala, S., and Vetta, A. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004. ISSN 0004-5411.

Koltchinskii, V. and Gine, E. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 02 2000.

Lei, J. and Rinaldo, A. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015.

Leskovec, J., Lang, K. J., and Mahoney, M. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

Mahoney, M. W., Orecchia, L., and Vishnoi, N. K. A local spectral method for graphs: with applications to improving graph partitions and exploring data graphs locally. *Journal of Machine Learning Research*, 13:2339–2365, 2012.

Maier, M., von Luxburg, U., and Hein, M. How the result of graph clustering methods depends on the construction of the graph. *CoRR*, abs/1102.2075, 2011.

McSherry, F. Spectral partitioning of random graphs. In *FOCS*, pp. 529–537, 2001.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915, 08 2011.

Schiebinger, G., Wainwright, M. J., and Yu, B. The geometry of kernelized spectral clustering. *Ann. Statist.*, 43(2): 819–846, 04 2015.

Shi, T., Belkin, M., and Yu, B. Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.*, 37(6B):3960–3984, 12 2009.

Singh, A., Scott, C., and Nowak, R. Adaptive hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B): 2760–2782, 10 2009.

Spielman, D. A. and Teng, S.-H. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Spielman, D. A. and Teng, S.-H. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

Spielman, D. A. and Teng, S.-H. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

Tolliver, D. and Miller, G. L. Graph partitioning by spectral rounding: Applications in image segmentation and clustering. In *Computer Vision and Pattern Recognition,CVPR*, volume 1, pp. 1053–1060, 2006.

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

Vempala, S. and Wang, G. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841 – 860, 2004.

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

von Luxburg, U., Belkin, M., and Bousquet, O. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 04 2008.

Yang, J. and Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, Jan 2015.

Zhu, Z. A., Lattanzi, S., and Mirrokni, V. S. A local algorithm for finding well-connected clusters. In *ICML (3)*, pp. 396–404, 2013.