Journal of Machine Learning Research 21 (2020) 1-66 Submitted 2/18; Revised 3/19; Published 1/20

Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms

Anna Little LITTL119@MSU.EDU
Department of Computational Mathematics, Science, and Engineering
Michigan State University, East Lansing, MI 48824, USA

Mauro Maggioni MAUROMAGGIONIJHU@ICLOUD.COM
Department of Applied Mathematics and Statistics, Department of Mathematics
Johns Hopkins University, Baltimore, MD 21218, USA

James M. Murphy JM.MURPHY@TUFTS.EDU
Department of Mathematics
Tufts University, Medford, MA 02139, USA

Editor: Ryota Tomioka

Abstract

We consider the problem of clustering with the longest-leg path distance (LLPD) metric, which is informative for elongated and irregularly shaped clusters. We prove finite-sample guarantees on the performance of clustering with respect to this metric when random samples are drawn from multiple intrinsically low-dimensional clusters in high-dimensional space, in the presence of a large number of high-dimensional outliers. By combining these results with spectral clustering with respect to LLPD, we provide conditions under which the Laplacian eigengap statistic correctly determines the number of clusters for a large class of data sets, and prove guarantees on the labeling accuracy of the proposed algorithm. Our methods are quite general and provide performance guarantees for spectral clustering with any ultrametric. We also introduce an efficient, easy to implement approximation algorithm for the LLPD based on a multiscale analysis of adjacency graphs, which allows for the runtime of LLPD spectral clustering to be quasilinear in the number of data points.

Keywords: unsupervised learning, spectral clustering, manifold learning, fast algorithms, shortest path distance

1. Introduction

Clustering is a fundamental unsupervised problem in machine learning, seeking to detect group structures in data without any references or labeled training data. Determining clusters can become harder as the dimension of the data increases: one of the

manifestations of the curse of dimension is that points drawn from high-dimensional distributions are far from their nearest neighbors, which can make noise and outliers challenging to address (Hughes, 1968; Györfi et al., 2006; Bellman, 2015). However, many clustering problems for real data involve data that exhibit low dimensional structure, which can be exploited to circumvent the curse of dimensionality. Various assumptions are imposed on the data to model low-dimensional structure, including requiring that the clusters be drawn from affine subspaces (Parsons et al., 2004; Chen and Lerman, 2009a,b; Vidal, 2011; Zhang et al., 2012; Elhamifar and Vidal, 2013; Wang et al., 2014; Soltanolkotabi et al., 2014) or more generally from low-dimensional mixture models (McLachlan and Basford, 1988; Arias-Castro, 2011; Arias-Castro et al., 2011, 2017).

When the shape of clusters is unknown or deviates from both linear structures (Vidal, 2011; Soltanolkotabi and Candès, 2012) or well-separated approximately spherical structures (for which K-means performs well (Mixon et al., 2017)), *spectral clustering* (Ng et al., 2002; Von Luxburg, 2007) is a very popular approach, often robust with respect to the geometry of the clusters and of noise and outliers (Arias-Castro, 2011; Arias-Castro et al., 2011). Spectral clustering requires an initial distance or similarity measure, as it operates on a graph constructed between near neighbors measured and weighted based on such distance. In this article, we propose to analyze low-dimensional clusters when spectral clustering is based on the *longest-leg path distance* (LLPD) metric, in which the distance between points x, y is the minimum over all paths between x, y of the longest edge in the path. Distances in this metric exhibit stark phase transitions between within-cluster distances and between-cluster distances. We are interested in performance guarantees with this metric which will explain this phase transition. We prove theoretical guarantees on the performance of LLPD as a discriminatory metric, under the assumption that data is drawn randomly from distributions supported near low-dimensional sets, together with a possibly very large number of outliers sampled from a distribution in the high dimensional ambient space. Moreover, we show that LLPD spectral clustering correctly determines the number of clusters and achieves high classification accuracy for data drawn from certain non-parametric mixture models. The existing state-of-the-art for spectral clustering struggles in the highly noisy setting, in the case when clusters are highly elongated—which leads to large within-cluster variance for traditional distance metrics—and also in the case when clusters have disparate volumes. In contrast, our method can tolerate a large amount of noise, even in its natural non-parametric setting, and it is essentially invariant to geometry of the clusters.

In order to efficiently analyze large data sets, a fast algorithm for computing LLPD is required. Fast nearest neighbor searches have been developed for Euclidean distance on intrinsically low-dimensional sets (and other doubling spaces) using cover trees (Beygelzimer et al., 2006), among other popular algorithms (e.g. k-d trees (Bentley, 1975)), and have been successfully employed in fast clustering algorithms. These algorithms compute the $O(1)$ nearest neighbors for *all points* in $O(n \log(n))$, where

n is the number of points, and are hence crucial to the scalability of many machine learning algorithms. LLPD seems to require the computation of a minimizer over a large set of paths. We introduce here an algorithm for LLPD, efficient and easy to implement, with the same quasilinear computational complexity as the algorithms above: this makes LLPD nearest neighbor searches scalable to large data sets. We moreover present a fast eigensolver for the (dense) LLPD graph Laplacian that allows for the computation of the approximate eigenvectors of this operator in essentially linear time.

1.1. Summary of Results

The major contributions of the present work are threefold.

First, we analyze the *finite sample behavior of LLPD* for points drawn according to a flexible probabilistic data model, with points drawn from low dimensional structures contaminated by a large number of high dimensional outliers. We derive bounds for maximal within-cluster LLPD and minimal between-cluster LLPD that hold with high probability, and also derive a lower bound for the minimal LLPD to a point's k_{nse} nearest neighbor in the LLPD metric. These results rely on a combination of techniques from manifold learning and percolation theory, and may be of independent interest.

Second, we deploy these finite sample results to prove that, under our data model, the *eigengap statistic for LLPD-based Laplacians correctly determines the number of clusters*. While the eigengap heuristic is often used in practice, existing theoretical analyses of spectral clustering fail to provide a rich class of data for which this estimate is provably accurate. Our results regarding the eigengap are quite general and can be applied to give state-of-the-art performance guarantees for spectral clustering with any ultrametric, not just the LLPD. Moreover, we prove that *the LLPD-based spectral embedding learned by our method is clustered correctly by K-means with high probability*, with misclassification rate improving over the existing state-of-the-art for Euclidean spectral clustering.

Finally, we present a *fast and easy to implement approximation algorithm for LLPD*, based on a multiscale decomposition of adjacency graphs. Let $k_{\ell\ell}$ be the number of LLPD nearest neighbors sought. Our approach generates approximate $k_{\ell\ell}$-nearest neighbors in the LLPD at a cost of $O(n(k_{\text{Euc}}C_{\text{NN}} + m(k_{\text{Euc}} \vee \log(n)) + k_{\ell\ell}))$, where n is the number of data points, k_{Euc} is the number of nearest neighbors used to construct an initial adjacency graph on the data, C_{NN} is the cost of a Euclidean nearest neighbor query, m is related to the approximation scheme, and $\vee$ denotes the maximum. Under the realistic assumption $k_{\text{Euc}}, k_{\ell\ell}, m \ll \log(n)$, this algorithm is $O(n\log^2(n))$ for data with low intrinsic dimension. If $k_{\text{Euc}}, k_{\ell\ell}, m = O(1)$ with respect to n, this reduces to $O(n\log(n))$. We quantify the resulting approximation error, which can be uniformly bounded independent of the data. We moreover *develop a fast eigensolver to compute the K principal eigenfunctions of the dense approximate LLPD Laplacian in $O(n(k_{Euc}C_{NN}+m(k_{Euc}\vee\log(n)\vee K^2)))$ time*. If $k_{\text{Euc}}, K, m = O(1)$

with respect to n, this reduces to $O(n\log(n))$. This allows for the fast computation of the eigenvectors without resorting to constructing a sparse Laplacian. *The proposed method is demonstrated on a variety of synthetic and real data sets, with performance consistently with our theoretical results.*

Article outline. In Section 2, we present an overview of clustering methods, with an emphasis on those most closely related to the one we propose. A summary of our data model and main results, together with motivating examples, are in Section 3. In Section 4, we analyze the LLPD for non-parametric mixture models. In Section 5, performance guarantees for spectral clustering with LLPD are derived, including guarantees on when the eigengap is informative and on the accuracy of clustering the spectral embedding obtained from the LLPD graph Laplacian. Section 6 proposes an efficient approximation algorithm for LLPD yielding faster nearest neighbor searches and computation of the eigenvectors of the LLPD Laplacian. Numerical experiments on representative data sets appear in Section 7. We conclude and discuss new research directions in Section 8.

1.2. Notation

In Table 1, we introduce notation we will use throughout the article.

2. Background

The process of determining groupings within data and assigning labels to data points according to these groupings without supervision is called *clustering* (Hastie et al., 2009). It is a fundamental problem in machine learning, with many approaches known to perform well in certain circumstances, but not in others. In order to provide performance guarantees, analytic, geometric, or statistical assumptions are placed on the data. Perhaps the most popular clustering scheme is K-means (Steinhaus, 1957; Friedman et al., 2001; Hastie et al., 2009), together with its variants (Ostrovsky et al., 2006; Arthur and Vassilvitskii, 2007; Park and Jun, 2009), which are used in conjunction with feature extraction methods. This approach partitions the data into a user-specified number K groups, where the partition is chosen to minimize within-cluster dissimilarity: $C^* = \arg\min_{C=\{C_k\}_{k=1}^K} \sum_{k=1}^K \sum_{x\in C_k} \|x-\bar{x}_k\|_2^2$. Here, $\{C_k\}_{k=1}^K$ is a partition of the points, C_k is the set of points in the k^{th} cluster and $\bar{x}_k$ denotes the mean of the k^{th} cluster. Unfortunately, the K-means algorithm and its refinements perform poorly for data sets that are not the union of well-separated, spherical clusters, and are very sensitive to outliers. In general, density-based methods such as density-based spatial clustering of applications with noise (DBSCAN) and variants (Ester et al., 1996; Xu et al., 1998) or spectral methods (Shi and Malik, 2000; Ng et al., 2002) are required to handle irregularly shaped clusters.

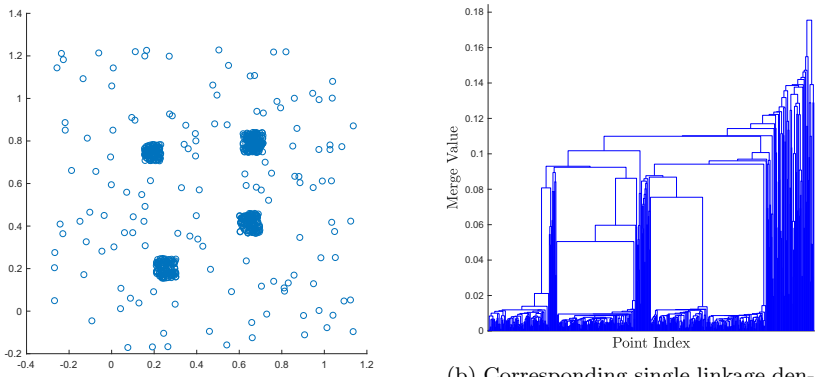
$X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$	Data points to cluster
d	Intrinsic dimension of cluster sets
K	Number of clusters
$\{X_l\}_{l=1}^K$	Discrete data clusters
$\widetilde{X}$	Discrete noise data
X_N	Denoised data; $X_N \subseteq X$
N	Number of points remaining after denoising
$n_{\min}$	Smallest number of points in a cluster
k_{Euc}	Number of nearest neighbors in construction of initial NN-graph
$k_{\ell\ell}$	Number of nearest neighbors for LLPD
k_{nse}	Number of nearest neighbors for LLPD denoising
C_{NN}	Complexity of computing a Euclidean NN
W	Weight matrix
L_{SYM}	Symmetric normalized Laplacian
σ	Scaling parameter in construction of weight matrix
$\{(\phi_i, \lambda_i)\}_{i=1}^n$	Eigenvectors and eigenvalues of an $n \times n$ L_{SYM}
ϵ_{in}	Maximum within cluster LLPD; see (3.7)
ϵ_{nse}	Minimum LLPD of noise points to k_{nse} nearest neighbor; see (3.7)
ϵ_{btw}	Minimum between cluster LLPD; see (3.7)
ϵ_{sep}	Minimum between cluster LLPD after denoising; see (5.3)
δ	Minimum Euclidean distance between clusters; see Definition 3.2
θ	Denoising parameter; see Definition 3.8
ζ_n, ζ_θ	LDLN data cluster balance parameters; see (3.6)
ζ_N	Empirical cluster balance parameter after denoising; see Assumption 1
ρ	Arbitrary metric
$\rho_{\ell\ell}$	LLPD metric; see Definition 2.1
$\mathcal{H}^d$	d-dimensional Hausdorff measure
$B_\epsilon(x)$	D-dimensional ball of radius ϵ centered at x
B_1	Unit ball, with dimension clear from context
$a \vee b$, $a \wedge b$	Maximum, minimum of a and b
$a \lesssim b$, $a \gtrsim b$	$a \leq Cb$, $a \geq Cb$ for some absolute constant $C > 0$

Table 1: Notation used throughout the article.

2.1. Hierarchical Clustering

Hierarchical clustering algorithms build a family of clusters at distinct hierarchical levels. Their results are readily presented as a *dendrogram* (see Figure 1). Hierarchical clustering algorithms can be *agglomerative*, where individual points start as their own clusters and are iteratively merged, or *divisive*, where the full data set is iteratively split until some stopping criterion is reached. It is often challenging to infer a global

partition of the data from hierarchical algorithms, as it is unclear where to cut the dendrogram.



(a) Data to cluster

(b) Corresponding single linkage dendrogram

Figure 1: Four two-dimensional clusters together with noise (Zelnik-Manor and Perona, 2004) appear in (a). In (b) is the corresponding single-linkage dendrogram. Each point begins as its own cluster, and at each level of the dendrogram, the two nearest clusters are merged. It is hard to distinguish between the noise and cluster points from the single linkage dendrogram, as it is not obvious where the four clusters are.

For agglomerative methods, it must be determined which clusters ought to be merged at a given iteration. This is done by a *cluster dissimilarity metric* ρ_c. For two clusters C_i, C_j, $\rho_c(C_i, C_j)$ small means the clusters are candidates for merger. Let ρ_X be a metric defined on all the data points in X. Standard ρ_c, and the corresponding clustering methods, include:

- $\rho_{SL}(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} \rho_X(x_i, x_j)$: *single linkage clustering.*
- $\rho_{CL}(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} \rho_X(x_i, x_j)$: *complete linkage clustering.*
- $\rho_{GA}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x_i \in C_i} \sum_{x_j \in C_j} \rho_X(x_i, x_j)$: *group average clustering.*

In Section 6 we make theoretical and practical connections between the proposed method and single linkage clustering.

2.2. Spectral Clustering

Spectral clustering methods (Shi and Malik, 2000; Meila and Shi, 2001; Ng et al., 2002; Von Luxburg, 2007) use a spectral decomposition of an *adjacency* or *Laplacian matrix* to define an embedding of the data, and then cluster the embedded data using a standard algorithm, commonly K-means. The basic idea is to construct a weighted graph on the data that represents local relationships. The graph has low edge weights for points far apart from each other and high edge weights for points close together. This graph is then partitioned into clusters so that there are large edge weights within each cluster, and small edge weights between each cluster. Spectral clustering in fact relaxes an NP-hard graph partition problem (Chung, 1997; Shi and Malik, 2000).

We now introduce notation related to spectral clustering that will be used throughout this work. Let $f_\sigma : \mathbb{R} \to [0,1]$ denote a kernel function with scale parameter σ. Given a metric $\rho : \mathbb{R}^D \times \mathbb{R}^D \to [0,\infty)$ and some discrete set $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, let $W_{ij} = f_\sigma(\rho(x_i, x_j))$ be the corresponding weight matrix. Let $d_i = \sum_{j=1}^n W_{ij}$ denote the degree of point x_i, and define the diagonal degree matrix $D_{ii} = d_i, D_{ij} = 0$ for $i \neq j$. The graph Laplacian is then defined by $L = D - W$, which is often normalized to obtain the symmetric Laplacian $L_{\text{SYM}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ or random walk Laplacian $L_{\text{RW}} = I - D^{-1}W$. Using the eigenvectors of L to define an embedding leads to *unnormalized* spectral clustering, whereas using the eigenvectors of L_{SYM} or L_{RW} leads to *normalized* spectral clustering. While both normalized and unnormalized spectral clustering minimize between-cluster similarity, only normalized spectral clustering maximizes within-cluster similarity, and is thus preferred in practice (Von Luxburg, 2007).

In this article we consider spectral clustering with L_{SYM} and construct the spectral embedding defined according to the popular algorithm of Ng et al. (2002). When appropriate, we will use $L_{\text{SYM}}(X, \rho, f_\sigma)$ to denote the matrix L_{SYM} computed on the data set X using metric ρ and kernel f_σ. We denote the eigenvalues of L_{SYM} (which are identical to those of L_{RW}) by $\lambda_1 \leq \ldots \leq \lambda_n$, and the corresponding eigenvectors by $\phi_1, \ldots, \phi_n$. To cluster the data into K groups according to Ng et al. (2002), one first forms an $n \times K$ matrix Φ whose columns are given by $\{\phi_i\}_{i=1}^K$; these K eigenvectors are called the K *principal eigenvectors*. The rows of Φ are then normalized to obtain the matrix V, that is $V_{ij} = \Phi_{ij}/(\sum_j \Phi_{ij}^2)^{1/2}$. Let $\{\mathbf{v}_i\}_{i=1}^n \in \mathbb{R}^K$ denote the rows of V. Note that if we let $g : \mathbb{R}^D \to \mathbb{R}^K$ denote the spectral embedding, $\mathbf{v}_i = g(x_i)$. Finally, K-means is applied to cluster the $\{\mathbf{v}_i\}_{i=1}^n$ into K groups, which defines a partition of our data points $\{x_i\}_{i=1}^n$. One can use L_{RW} similarly (Shi and Malik, 2000).

Choosing K is an important aspect of spectral clustering, and various spectral-based mechanisms have been proposed in the literature (Azran and Ghahramani, 2006b,a; Zelnik-Manor and Perona, 2004; Sanguinetti et al., 2005). The eigenvalues of L_{SYM} have often been used to heuristically estimate the number of clusters as the largest empirical eigengap $\hat{K} = \arg\max_i \lambda_{i+1} - \lambda_i$, although there are many data sets for which this heuristic is known to fail (Von Luxburg, 2007); this estimate is called the *eigengap* statistic. We remark that sometimes in the literature it is required that not only should $\lambda_{\hat{K}+1} - \lambda_{\hat{K}}$ be maximal, but also that λ_i should be close to 0 for $i \leq \hat{K}$; we shall not make this additional assumption on $\lambda_i, i \leq \hat{K}$, though we find in practice it is usually satisfied when the eigengap is accurate.

A description of the spectral clustering algorithm of Ng et al. (2002) in the case that K is not known a priori appears in Algorithm 1; the algorithm can be modified in the obvious way if K is known and does not need to be estimated, or when using a sparse Laplacian, for example when W is defined by a sparse nearest neighbors graph.

In addition to determining K, performance guarantees for K-means (or other clustering methods) on the spectral embedding is a topic of active research (Schiebinger

Algorithm 1 Spectral clustering with metric ρ

Input: $\{x_i\}_{i=1}^n$ (data) , $\sigma > 0$ (scaling parameter), f_σ (kernel function)
Output: Y (Labels)

1: Compute the weight matrix $W \in \mathbb{R}^{n\times n}$ with $W_{ij} = f_\sigma(\rho(x_i, x_j))$.
2: Compute the diagonal degree matrix $D \in \mathbb{R}^{n\times n}$ with $D_{ii} = \sum_{j=1}^n W_{ij}$.
3: Form the symmetric normalized Laplacian $L_{\mathrm{SYM}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$.
4: Compute the eigendecomposition $\{(\phi_k, \lambda_k)\}_{k=1}^n$, sorted so that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.
5: Estimate the number of clusters K as $\hat{K} = \arg\max_k \lambda_{k+1} - \lambda_k$.
6: For $1 \leq i \leq n$, let $\mathbf{v}_i = (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_{\hat{K}}(x_i))/||(\phi_1(x_i), \phi_2(x_i), \ldots, \phi_{\hat{K}}(x_i))||_2$ define the (row normalized) spectral embedding.
7: Compute labels Y by running K-means on the data $\{\mathbf{v}_i\}_{i=1}^n$ using $\hat{K}$ as the number of clusters.

et al., 2015; Arias-Castro et al., 2017). However, spectral clustering typically has poor performance in the presence of noise and highly elongated clusters.

2.3. Background on LLPD

Many clustering and machine learning algorithms make use of Euclidean distances to compare points. While universal and popular, this distance is data-independent, not adapted to the geometry of the data. Many data-dependent metrics have been developed, for example diffusion distances (Coifman et al., 2005; Coifman and Lafon, 2006), which are induced by diffusion processes on a data set, and path-based distances (Fischer and Buhmann, 2003; Chang and Yeung, 2008). We shall consider a path-based distance for undirected graphs.

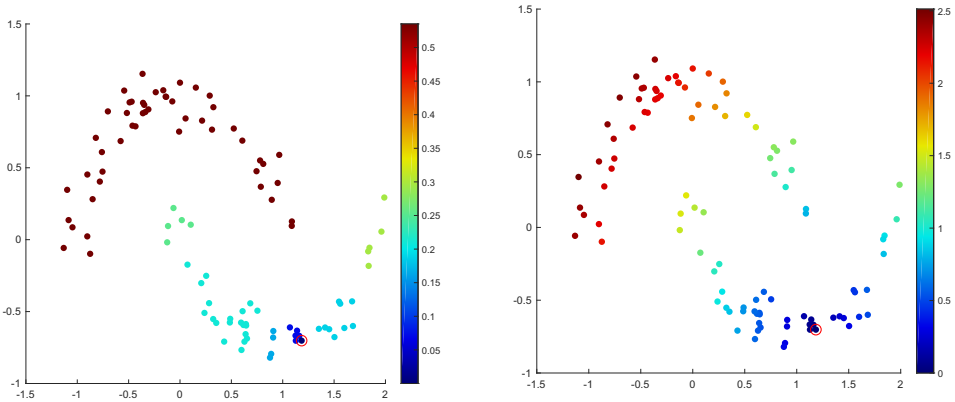
Definition 2.1 *For $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$, let G be the complete graph on X with edges weighted by Euclidean distance between points. For $x_i, x_j \in X$, let $\mathcal{P}(x_i, x_j)$ denote the set of all paths connecting x_i, x_j in G. The* longest-leg path distance (LLPD) *is:*

$$\rho_{\ell\ell}(x_i, x_j) = \min_{\{y_l\}_{l=1}^L \in \mathcal{P}(x_i, x_j)} \max_{l=1,2,\ldots,L-1} \|y_{l+1} - y_l\|_2.$$

In this article we use LLPD with respect to the Euclidean distance, but our results very easily generalize to other base distances. Our goal is to analyze the effects of transforming an original metric through the *min-max* distance along paths in the definition of LLPD above. We note that the LLPD is an *ultrametric*, i.e.

$$\forall x, y, z \in X \quad \rho_{\ell\ell}(x, y) \leq \max\{\rho_{\ell\ell}(x, z), \rho_{\ell\ell}(y, z)\}\,. \tag{2.2}$$

This property is central to the proofs of Sections 4 and 5. Figure 2 illustrates how LLPD successfully differentiates elongated clusters, whereas Euclidean distance does not.



(a) LLPD from the marked point.

(b) Euclidean distances from the marked point.

Figure 2: In this example, LLPD is compared with Euclidean distance. The distance from the red circled source point is shown in each subfigure. Notice that the LLPD has a phase transition that separates the clusters clearly, and that all distances within-cluster are comparable.

2.3.1. Probabilistic Analysis of LLPD

Existing theoretical analysis of LLPD is based on studying the uniform distribution on certain geometric sets. The degree and connectivity properties of near-neighbor graphs defined on points sampled uniformly from $[0,1]^d$ and their connections with percolation have been studied extensively (Appel and Russo, 1997a,b, 2002; Penrose, 1997, 1999). Related results in the case of points drawn from low-dimensional structures were studied by Arias-Castro (2011). These results motivate some of the ideas in this article; detailed references are given below when appropriate.

2.3.2. Spectral Clustering with LLPD

Spectral clustering with LLPD has been shown to enjoy good empirical performance (Fischer et al., 2001; Fischer and Buhmann, 2003; Fischer et al., 2004) and is made more robust by incorporating outlier removal (Chang and Yeung, 2008). The method and its variants generally perform well for non-convex and highly elongated clusters, even in the presence of noise. However, no theoretical guarantees seem to be available. Moreover, numerical implementation of LLPD spectral clustering appears underdeveloped, and existing methods have been evaluated mainly on small, low-dimensional data data setsets. This article derives theoretical guarantees on performance of LLPD spectral clustering which confirms empirical insights, and also provides a fast implementation of the method suitable for large data sets.

2.3.3. Computing LLPD

The problem of computing this distance is referred to by many names in the literature, including the maximum capacity path problem, the widest path problem, and the bottleneck edge query problem (Pollack, 1960; Hu, 1961; Camerini, 1978; Gabow

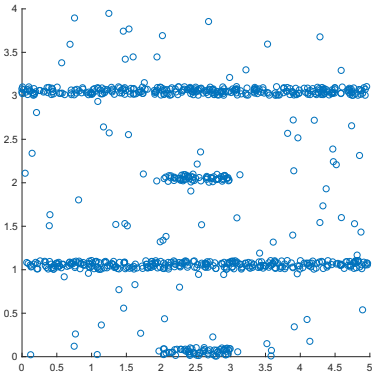
and Tarjan, 1988). A naive computation of LLPD distances is expensive, since the search space $\mathcal{P}(x, y)$ is potentially very large. However, for a fixed pair of points x, y connected in a graph $G = G(V, E)$, $\rho_{\ell\ell}(x, y)$ can be computed in $O(|E|)$ (Punnen, 1991). There has also been significant work on the related problem of finding bottleneck spanning trees. For a fixed root vertex $s \in V$, the *minimal bottleneck spanning tree* rooted at s is the spanning tree whose maximal edge length is minimal. The bottleneck spanning tree can be computed in $O(\min\{n\log(n) + |E|, |E|\log(n)\})$ (Camerini, 1978; Gabow and Tarjan, 1988).
Computing all LLPDs for all points is the *all points path distance (APPD)* problem. Naively applying the bottleneck spanning tree construction to each point gives an APPD runtime of $O(\min\{n^2\log(n)+n|E|, n|E|\log(n)\})$. However the APPD distance matrix can be computed in $O(n^2)$, for example with a modified SLINK algorithm (Sibson, 1973), or with Cartesian trees (Alon and Schieber, 1987; Demaine et al., 2009, 2014). We propose to approximate LLPD and implement LLPD spectral clustering with an algorithm near-linear in n, which enables the analysis of very large data sets (see Section 6).

3. Major Contributions

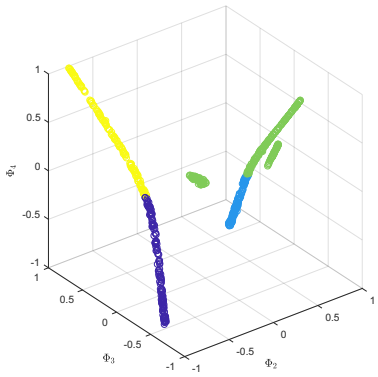
In this section we present a simplified version of our main theoretical result. More general versions of these results, with detailed constants, will follow in Sections 4 and 5. We first discuss a motivating example and outline our data model and assumptions, which will be referred to throughout the article.

3.1. Motivating Examples

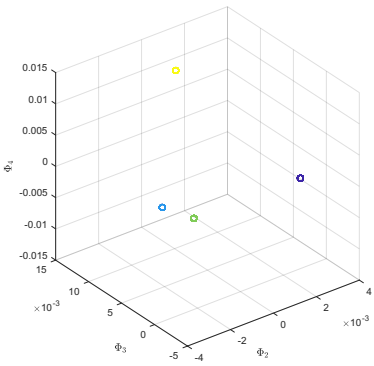
In this subsection we illustrate in which regimes LLPD spectral clustering advances the state-of-art for clustering. As will be explicitly described in Subsection 3.2, we model clusters as connected, high-density regions, and we model noise as a low-density region separating the clusters. Our method easily handles highly elongated and irregularly shaped clusters, where traditional K-means and even spectral clustering fail. For example, consider the four elongated clusters in $\mathbb{R}^2$ illustrated in Figure 3. After denoising the data with nearest neighbor thresholding (see Section 5.2.1 for a description of the denoising procedure and Section 5.2.4 for a discussion of how to tune the thresholding parameter), both K-means and Euclidean spectral clustering split one or more of the most elongated clusters, whereas the LLPD spectral embedding perfectly separates them. Moreover, the eigenvalues of the LLPD Laplacian correctly infer there are 4 clusters, unlike the Euclidean Laplacian.
There are naturally situations where LLPD spectral clustering will not perform well, such as for certain types of structured noise. For example, consider the dumbbell shown in Figure 4. When there is a high-density bridge connecting the dumbbell, LLPD will not be able to distinguish the two balls. However, it is worth noting that this property is precisely what allows for robust performance with elongated clusters,



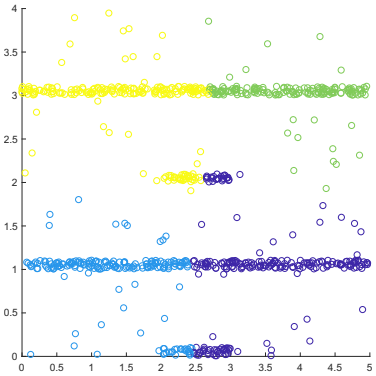
(a) Original data set.



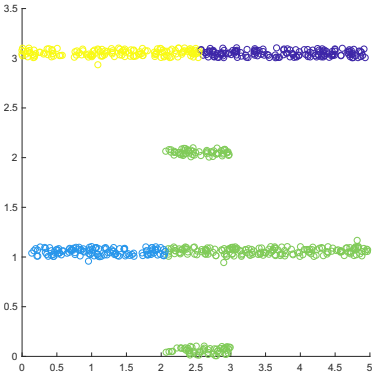
(b) 3 dimensional spectral embedding with Euclidean distances, labeled with K-means. The data has been denoised based on thresholding with Euclidean distances.



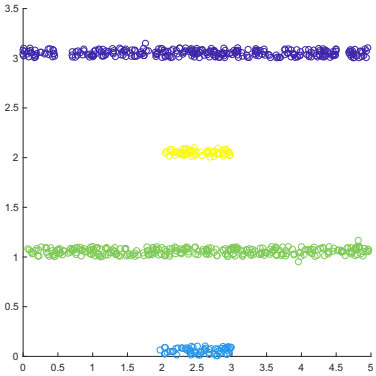
(c) 3 dimensional spectral embedding with LLPD, labeled with K-means. The data has been denoised based on thresholding with LLPD.



(d) K-means labels.



(e) Spectral clustering results with Euclidean distances.



(f) Spectral clustering results with LLPD.

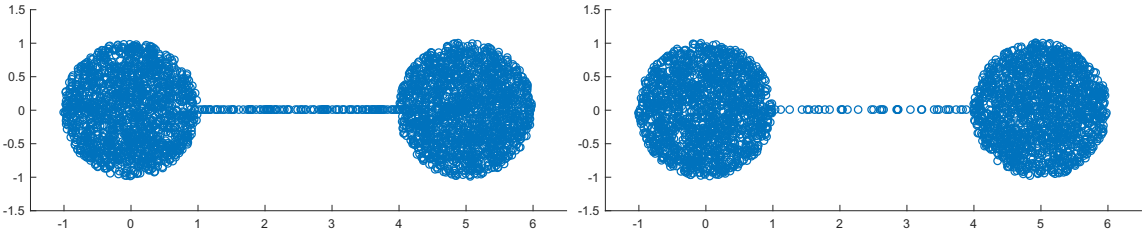
Figure 3: The data set consists of four elongated clusters in $\mathbb{R}^2$, together with ambient noise. The labels given by K-means are quite inaccurate, as are those given by regular spectral clustering. The labels given by LLPD spectral clustering are perfect. Note that Φ_i denotes the i^{th} principal eigenvector of L_{SYM}. For both variants of spectral clustering, the K-means algorithm was run in the 4 dimensional embedding space given by the first 4 principal eigenvectors of L_{SYM}.

and that if the bridge has a lower density than the clusters, LLPD spectral clustering performs very well.

3.2. Low Dimensional Large Noise (LDLN) Data Model and Assumptions

We first define the *low dimensional, large noise (LDLN)* data model, and then establish notation and assumptions for the LLPD metric and denoising procedure on data drawn from this model.

We consider a collection of K disjoint, connected, approximately d-dimensional sets $\mathcal{X}_1, \ldots, \mathcal{X}_K$ embedded in a measurable, D-dimensional ambient set $\mathcal{X} \subset \mathbb{R}^D$. We recall the definition of *d-dimensional Hausdorff measure* as follows (Benedetto and Czaja, 2010). For $A \subset \mathbb{R}^D$, let $\text{diam}(A) = \sup_{x,y \in A} \|x - y\|_2$. Fix $\delta > 0$ and for any



(a) Two clusters connected by a bridge of roughly the same empirical density.

(b) Two clusters connected by a bridge of lower empirical density.

Figure 4: In (a), two spherical clusters are connected with a bridge of approximately the same density; LLPD spectral clustering fails to distinguish between theses two clusters. Despite the fact that the bridge consists of a very small number of points relative to the entire data set, it is quite adversarial for the purposes of LLPD separation. This is a limitation of the proposed method: it is robust to large amounts of diffuse noise, but not to a potentially small amount of concentrated, adversarial noise. Conversely, if the bridge is of lower density, as in (b), then the proposed method will succeed.

$A \subset \mathbb{R}^D$, let

$$\mathcal{H}^d_\delta(A) = \inf\left\{\sum_{i=1}^{\infty} \mathrm{diam}(U_i)^d \mid A \subset \bigcup_{i=1}^{\infty} U_i,\ \mathrm{diam}(U_i) < \delta\right\}.$$

The d-dimensional Hausdorff measure of A is $\mathcal{H}^d(A) = \lim_{\delta \to 0^+} \mathcal{H}^d_\delta(A)$. Note that $\mathcal{H}^D(A)$ is simply a rescaling of the Lebesgue measure in $\mathbb{R}^D$.

Definition 3.1 *A set $S \subset \mathbb{R}^D$ is an element of $\mathcal{S}_d(\kappa, \epsilon_0)$ for some $\kappa \geq 1$ and $\epsilon_0 > 0$ if it has finite d-dimensional Hausdorff measure, is connected, and:*

$$\forall x \in S, \quad \forall \epsilon \in (0, \epsilon_0), \qquad \kappa^{-1}\epsilon^d \leq \frac{\mathcal{H}^d(S \cap B_\epsilon(x))}{\mathcal{H}^d(B_1)} \leq \kappa \epsilon^d\,.$$

Note that $\mathcal{S}_d(\kappa, \epsilon_0)$ includes d-dimensional smooth compact manifolds (which have finite positive reach (Federer, 1959)). With some abuse of notation, we denote by $\mathrm{Unif}(S)$ the probability measure $\mathcal{H}^d/\mathcal{H}^d(S)$. For a set A and $\tau \geq 0$, we define

$$B(A, \tau) := \{x \in \mathbb{R}^D : \exists y \in A \text{ with } \|x - y\|_2 \leq \tau\}.$$

Clearly $B(A, 0) = A$.

Definition 3.2 (LDLN model) *The* Low-Dimensional Large Noise (LDLN) model *consists of a D-dimensional ambient set $\mathcal{X} \subset \mathbb{R}^D$ and K cluster regions $\mathcal{X}_1, \ldots, \mathcal{X}_K \subset \mathcal{X}$ and noise set $\tilde{\mathcal{X}} \subset \mathbb{R}^D$ such that:*

(i) $0 < \mathcal{H}^D(\mathcal{X}) < \infty$;

(ii) $\mathcal{X}_l = B(S_l, \tau)$ *for* $S_l \in \mathcal{S}_d(\kappa, \epsilon_0)$, $l = 1, \ldots, K$, $\tau \geq 0$ *fixed;*

(iii) $\tilde{\mathcal{X}} = \mathcal{X} \setminus (\mathcal{X}_1 \cup \ldots \cup \mathcal{X}_K)$;

(iv) the minimal Euclidean distance δ *between two cluster regions satisfies*

$$\delta := \min_{l \neq s} dist(\mathcal{X}_l, \mathcal{X}_s) = \min_{l \neq s} \min_{x \in \mathcal{X}_l, y \in \mathcal{X}_s} \|x - y\|_2 > 0.$$

Condition (i) says that the ambient set $\mathcal{X}$ is nontrivial and has bounded D-dimensional volume; condition (ii) says that the cluster regions behave like tubes of radius τ around well-behaved d-dimensional sets; condition (iii) defines the noise as consisting of the high-dimensional ambient region minus any cluster region; condition (iv) states that the cluster regions are well-separated.

Definition 3.3 (LDLN data) *Given a LDLN model, LDLN data consists of sets* X_l*, each consisting of* n_l *i.i.d. draws from* $\mathrm{Unif}(\mathcal{X}_l)$*, for* $1 \leq l \leq K$*, and* $\tilde{X}$ *consisting of* $\tilde{n}$ *i.i.d. draws from* $\mathrm{Unif}(\tilde{\mathcal{X}})$*. We let* $X = X_1 \cup \cdots \cup X_K \cup \tilde{X}$*,* $n := n_1 + \ldots + n_K + \tilde{n}, n_{min} := \min_{1 \leq l \leq K} n_l$*.*

Remark 3.4 *Although our model assumes sampling from a uniform distribution on the cluster regions, our results easily extend to any probability measure* μ_l *on* $\mathcal{X}_l$ *such that there exist constants* $0 < C_1 \leq C_2 < \infty$ *so that* $C_1 \mathcal{H}^d(S)/\mathcal{H}^d(\mathcal{X}_l) \leq \mu_l(S) \leq C_2 \mathcal{H}^d(S)/\mathcal{H}^d(\mathcal{X}_l)$ *for any measurable subset* $S \subset \mathcal{X}_l$*, and the same generalization holds for sampling from the noise set* $\tilde{\mathcal{X}}$*. The constants in our results change but nothing else; thus for ease of exposition we assume uniform sampling.*

Remark 3.5 *We could also consider a fully probabilistic model with the data consisting of* n *i.i.d samples from a mixture model* $\sum_{l=1}^{K} q_l \,\mathrm{Unif}(\mathcal{X}_l) + \tilde{q}\,\mathrm{Unif}(\tilde{\mathcal{X}})$*, with suitable mixture weights* $q_1, \ldots, q_K, \tilde{q}$ *summing to* 1*. Then with high probability we would have* n_i *(now a random variable) close to* $q_i n$ *and* $\tilde{n}$ *close to* $\tilde{q}n$*, falling back to the above case. We will use the model above in order to keep the notation simple.*

We define two cluster balance parameters for the LDLN data model:

$$\zeta_n := \frac{\sum_{l=1}^{K} n_l}{n_{\min}} \quad , \quad \zeta_\theta := \frac{\sum_{l=1}^{K}, p_{l,\theta}}{p_{\min,\theta}} \tag{3.6}$$

where $p_{l,\theta} := \mathcal{H}^D(B(\mathcal{X}_l, \theta) \setminus \mathcal{X}_l)/\mathcal{H}^D(\tilde{\mathcal{X}})$, $p_{\min,\theta} := \min_{1 \leq l \leq K} p_{l,\theta}$, and θ is related to the denoising procedure (see Definition 3.8). The parameter ζ_n measures the balance of cluster sample size and the parameter ζ_θ depends on the balance in surface area of the cluster sets $\mathcal{X}_l$. When all n_i are equal and the cluster sets have the same geometry, $\zeta_n = \zeta_\theta = K$.

Let $\rho_{\ell\ell}$ refer to LLPD in the full set X. For $A \subset X$, let $\rho_{\ell\ell}^A$ refer to LLPD when paths are restricted to being contained in the set A. For $x \in X$, let $\beta_{k_{\mathrm{nse}}}(x, A)$ denote the LLPD from x to its $k_{\mathrm{nse}}^{\mathrm{th}}$ LLPD-nearest neighbor when paths are restricted to the set A:

$$\beta_{k_{\text{nse}}}(x, A) := \min_{B \subset A \setminus \{x\}, |B| = k_{\text{nse}}} \max_{y \in B} \rho^A_{\ell\ell}(x, y).$$

Let ϵ_{in} be the maximal within-cluster LLPD, ϵ_{nse} the minimal distance of noise points to their $k_{\text{nse}}^{\text{th}}$ LLPD-nearest neighbor in the absence of cluster points, and ϵ_{btw} the minimal between-cluster LLPD:

$$\epsilon_{\text{in}} := \max_{1 \le l \le K} \max_{x \neq y \in X_l} \rho_{\ell\ell}(x, y), \quad \epsilon_{\text{nse}} := \min_{x \in \tilde{X}} \beta_{k_{\text{nse}}}(x, \tilde{X}), \quad \epsilon_{\text{btw}} := \min_{l \neq l'} \min_{x \in \mathcal{X}_l, y \in \mathcal{X}_{l'}} \rho_{\ell\ell}(x, y)\,. \tag{3.7}$$

Definition 3.8 (Denoised LDLN data) *We preprocess LDLN data (denoising) by removing any points that have a large LLPD to their k_{nse}^{th} LLPD-nearest neighbor, i.e. by removing all points $x \in X$ which satisfy $\beta_{k_{nse}}(x, X) > \theta$ for some thresholding parameter θ. We let $N \le n$ denote the number of points which survive thresholding, and $X_N \subset X$ be the corresponding subset of points.*

A discussion of how to tune θ in practice appears in Section 5.2.4.

3.3. Overview of Main Results

This article investigates geometric conditions implying $\epsilon_{\text{in}} \ll \epsilon_{\text{nse}}$ with high probability. In this context higher density sets are separated by lower density regions; the points in these lower density regions will be referred to as noise and outliers interchangeably. In this regime, noise points are identified and removed with high probability, leading to well-separated clusters that are internally coherent in the sense of having uniformly small within-cluster distances. The proposed clustering method is shown to be highly robust to the choice of scale parameter in the kernel function, and to produce accurate clustering results even in the context of very large amounts of noise and highly nonlinear or elongated clusters. Theorem 3.10 simplifies two major results of the present article, Theorem 5.12 and Corollary 5.14, which establish conditions guaranteeing two desirable properties of LLPD spectral clustering. First, that the K^{th} eigengap of L_{SYM} is the largest gap with high probability, so that the eigengap statistic correctly estimates the number of clusters. Second, that embedding the data according to the principal eigenvectors of the LLPD Laplacian L_{SYM} followed by a simple clustering algorithm correctly labels all points. Throughout the theoretical portions of this article, we will define the accuracy of a clustering algorithm as follows. Let $\{y_i\}_{i=1}^n$ be ground truth labels taking values in $[K] = \{1, \ldots, K\}$, and let $\{\hat{y}_i\}_{i=1}^n \in [K]$ be the labels learned from running a clustering algorithm. Following Abbe (2018), we define the agreement function between y and $\hat{y}$ as

$$A(y, \hat{y}) = \max_{\pi \in \Pi^K} \frac{1}{n} \sum_{i=1}^n \mathbb{1}(\pi(\hat{y}_i) = y_i), \tag{3.9}$$

where the maximum is taken over all permutations of the label set Π^K, and the *accuracy* of a clustering algorithm as the value of the resulting agreement function. The agreement function can be computed numerically using the Hungarian algorithm (Munkres, 1957). If ground truth labels are only available on a data subset (as for LDLN data where noise points are unlabeled), then the accuracy is computed by restricting to the labeled data points. In Section 7, additional notions of accuracy will be introduced for the empirical evaluation of LLPD spectral clustering.

Theorem 3.10 *Under the LDLN data model and assumptions, suppose that the cardinality $\tilde{n}$ of the noise set and the tube radius τ are such that*

$$\tilde{n} \leq \left(\frac{C_2}{C_1}\right)^{\frac{k_{nse}D}{k_{nse}+1}} n_{min}^{\frac{D}{d+1}\left(\frac{k_{nse}}{k_{nse}+1}\right)} \quad , \quad \tau < \frac{C_1}{8} n_{min}^{-(d+1)} \wedge \frac{\epsilon_0}{5} \, .$$

Let $f_\sigma(x) = e^{-x^2/\sigma^2}$ be the Gaussian kernel and assume $k_{nse} = O(1)$. If n_{min} is large enough and θ, σ satisfy

$$C_1 n_{min}^{-\frac{1}{d+1}} \leq \theta \leq C_2 \tilde{n}^{-\left(\frac{k_{nse}+1}{k_{nse}}\right)\frac{1}{D}} \tag{3.11}$$

$$C_3(\zeta_n + \zeta_\theta)\theta \leq \sigma \leq C_4 \delta (\log(\zeta_n + \zeta_\theta))^{-1/2} \tag{3.12}$$

then with high probability the denoised LDLN data X_N satisfies:

(i) the largest gap in the eigenvalues of $L_{SYM}(X_N, \rho_{\ell\ell}^{X_N}, f_\sigma)$ is $\lambda_{K+1} - \lambda_K$.

(ii) spectral clustering with LLPD with K principal eigenvectors achieves perfect accuracy on X_N.

The constants $\{C_i\}_{i=1}^4$ depend on the geometric quantities $K, d, D, \kappa, \tau, \{\mathcal{H}^d(S_l)\}_{l=1}^K, \mathcal{H}^D(\tilde{\mathcal{X}})$, but do not depend on $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

Section 4 verifies that with high probability a point's distance to its $k_{\text{nse}}^{\text{th}}$ nearest neighbor (in LLPD) scales like $n_{\min}^{-(d+1)}$ for cluster points and $\tilde{n}^{-\left(\frac{k_{\text{nse}}+1}{k_{\text{nse}}}\right)\frac{1}{D}}$ for noise points; thus when the denoising parameter θ satisfies (3.11), we successfully distinguish the cluster points from the noise points, and this range is large when the number of noise points $\tilde{n}$ is small relative to $n_{\min}^{\frac{D}{d+1}\left(\frac{k_{\text{nse}}}{k_{\text{nse}}+1}\right)}$. Thus, Theorem 3.10 illustrates that when clusters are (intrinsically) low-dimensional, a number of noise points exponentially (in D/d) larger than $n_{\min}$ may be tolerated. If the data is denoised at an appropriate threshold level, the maximal eigengap heuristic correctly identifies the number of clusters and spectral clustering achieves high accuracy for any kernel scale σ satisfying (3.12). This range for σ is large whenever the cluster separation δ is large relative to the denoising parameter θ. Section 7 discusses how to empirically select σ; (7.1) in particular suggests an automated procedure for doing so. We note that the case when k_{nse} is not $O(1)$ is discussed in Section 5.2.4.

In the noiseless case ($\tilde{n} = 0$) when clusters are approximately balanced ($\zeta_n, \zeta_\theta = O(1)$), Theorem 3.10 can be further simplified as stated in the following corollary. Note that no denoising is necessary in this case; one simply needs the kernel scale σ to be not small relative to the maximal within cluster distance (which is upper bounded by $n_{\min}^{-(d+1)}$) and not large relative to the distance between clusters δ.

Corollary 3.13 (Noiseless, Balanced Case) *Under the LDLN data model and assumptions, further assume the cardinality of the noise set $\tilde{n} = 0$ and the tube radius τ satisfies $\tau < \frac{C_1}{8} n_{min}^{-(d+1)} \wedge \frac{\epsilon_0}{5}$. Let $f_\sigma(x) = e^{-x^2/\sigma^2}$ be the Gaussian kernel and assume $k_{nse}, K, \zeta_n = O(1)$. If n_{min} is large enough and σ satisfies*

$$C_1 n_{min}^{-\frac{1}{d+1}} \leq \sigma \leq C_4 \delta$$

for constants C_1, C_4 not depending on $n_1, \ldots, n_K, \sigma$, then with high probability the LDLN data X satisfies:

(i) the largest gap in the eigenvalues of $L_{SYM}(X, \rho_{\ell\ell}^X, f_\sigma)$ is $\lambda_{K+1} - \lambda_K$.

(ii) spectral clustering with LLPD with K principal eigenvectors achieves perfect accuracy on X.

Remark 3.14 *If one extends the LDLN model to allow the S_l sets to have different dimensions d_l and $\mathcal{X}_l$ to have different tube widths τ_l, that is, $S_l \in \mathcal{S}_{d_l}(\kappa, \epsilon_0)$ and $\mathcal{X}_l = B(S_l, \tau_l)$, Theorem 3.10 still holds with $\max_l \tau_l$ replacing τ and $\max_l n_l^{-1/(d_l+1)}$ replacing $n_{min}^{-1/(d+1)}$. Alternatively, σ can be set in a manner that adapts to local density (Zelnik-Manor and Perona, 2004).*

Remark 3.15 *The constants in Theorem 3.10 and Corollary 3.13 have the following dimensional dependencies.*

1. *$C_1 \lesssim \min_l (\kappa \mathcal{H}^d(S_l)/\mathcal{H}^d(B_1))^{\frac{1}{d}}$ for $\tau = 0$. Letting $rad(\mathcal{M})$ denote the geodesic radius of a manifold $\mathcal{M}$, if S_l is a complete Riemannian manifold with non-negative Ricci curvature, then by the Bishop-Gromov inequality (Bishop and Crittenden, 2011), $(\mathcal{H}^d(S_l)/\mathcal{H}^d(B_1))^{\frac{1}{d}} \leq rad(S_l)$; noting that κ is at worst exponential in d, it follows that C_1 is then dimension independent for $\tau = 0$. For $\tau > 0$, C_1 is upper bounded by an exponential in D/d.*

2. *$C_2 \lesssim (\mathcal{H}^D(\tilde{\mathcal{X}})/\mathcal{H}^D(B_1))^{\frac{1}{D}}$. Assume $\mathcal{H}^D(\tilde{\mathcal{X}}) \gtrsim \mathcal{H}^D(\mathcal{X})$: if $\mathcal{X}$ is the unit D-dimensional ball, then C_2 is dimension independent; if $\mathcal{X}$ is the unit cube, then C_2 scales like $\sqrt{D}$. This illustrates that when $\mathcal{X}$ is not elongated in any direction, we expect C_2 to scale like $rad(\mathcal{X})$.*

3. *C_3, C_4 are independent of d and D.*

4. Finite Sample Analysis of LLPD

In this section we derive high probability bounds for the maximal within-cluster LLPD and the minimal between-cluster LLPD, and also derive a bound for the minimal $k_{\text{nse}}^{\text{th}}$ LLPD-nearest neighbor distance. From these results we infer a sampling regime where LLPD is able to effectively differentiate between clusters and noise.

4.1. Upper-Bounding Within-Cluster LLPD

For bounding the within-cluster LLPD, we seek a uniform upper bound on $\rho_{\ell\ell}$ that holds with high probability. The following two results are essentially Lemma 1 and Theorem 1 in Arias-Castro (2011) with all constants explicitly computed; the proofs are in Appendix A.

Lemma 4.1 *Let $S \in \mathcal{S}_d(\kappa, \epsilon_0)$, and let $\epsilon, \tau > 0$ with $\epsilon < \frac{2\epsilon_0}{5}$. Then $\forall x \in B(S, \tau)$,*

$$C_1 \epsilon^d (\tau \wedge \epsilon)^{D-d} \leq \mathcal{H}^D(B(S,\tau) \cap B_\epsilon(x))/\mathcal{H}^D(B_1) \leq C_2 \epsilon^d (\tau \wedge \epsilon)^{D-d}, \tag{4.2}$$

for constants $C_1 = \kappa^{-2} 2^{-2D-d}, C_2 = \kappa^2 2^{2D+2d}$ independent of ϵ.

Theorem 4.3 *Let $S \in \mathcal{S}_d(\kappa, \epsilon_0)$ and let $\tau > 0$, $\epsilon < \epsilon_0$. Let $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \text{Unif}(B(S,\tau))$ and $C = \kappa^2 2^{2D+d}$. Then*

$$n \geq \frac{C\mathcal{H}^D(B(S,\tau))}{\left(\frac{\epsilon}{4}\right)^d (\tau \wedge \frac{\epsilon}{4})^{D-d} \mathcal{H}^D(B_1)} \log \frac{C\mathcal{H}^D(B(S,\tau))}{\left(\frac{\epsilon}{8}\right)^d (\tau \wedge \frac{\epsilon}{8})^{D-d} \mathcal{H}^D(B_1) t} \implies \mathbb{P}(\max_{i,j} \rho_{\ell\ell}(x_i, x_j) < \epsilon) \geq 1 - t\,.$$

When τ is sufficiently small and ignoring constants, the sampling complexity suggested in Theorem 4.3 depends only on d. The following corollary uses the above result to bound ϵ_{in} in the LDLN data model; the proof also is given in Appendix A.

Corollary 4.4 *Assume the LDLN data model and assumptions, and let $0 < \tau < \frac{\epsilon}{8} \wedge \frac{\epsilon_0}{5}$, $\epsilon < \epsilon_0$, and $C = \kappa^5 2^{4D+5d}$. Then*

$$n_l \geq \frac{C\mathcal{H}^d(S_l)}{\left(\frac{\epsilon}{4}\right)^d \mathcal{H}^d(B_1)} \log \frac{C\mathcal{H}^d(S_l) K}{\left(\frac{\epsilon}{8}\right)^d \mathcal{H}^d(B_1) t} \quad \forall l = 1, \ldots, K \implies \mathbb{P}(\epsilon_{\text{in}} < \epsilon) \geq 1 - t\,. \tag{4.5}$$

The case $\tau = 0$ corresponds to cluster regions being elements of $\mathcal{S}_d(\kappa, \epsilon_0)$, and is proved similarly to Theorem 4.3 (the proof is omitted):

Theorem 4.6 *Let $S \in \mathcal{S}_d(\kappa, \epsilon_0)$, $\tau = 0$, and let $\epsilon \in (0, \epsilon_0)$. Suppose $x_1, \ldots, x_n \overset{i.i.d.}{\sim}$ Unif(S). Then*

$$n \geq \frac{\kappa\mathcal{H}^d(S)}{\left(\frac{\epsilon}{4}\right)^d \mathcal{H}^d(B_1)} \log \frac{\kappa\mathcal{H}^d(S)}{\left(\frac{\epsilon}{8}\right)^d \mathcal{H}^d(B_1) t} \implies \mathbb{P}(\max_{i,j} \rho_{\ell\ell}(x_i, x_j) < \epsilon) \geq 1 - t.$$

Thus, up to geometric constants, for $\tau = 0$ the uniform bound on LLPD depends only on the intrinsic dimension, d, not the ambient dimension, D. When $d \ll D$, this leads to a huge gain in sampling complexity, compared to sampling in the ambient dimension.

4.1.1. Comparison with Existing Asymptotic Estimates

To put Theorem 4.6 in context, we remark on known asymptotic results for LLPD in the case $S = [0,1]^d$ (Appel and Russo, 1997a,b, 2002; Penrose, 1997, 1999). Note that this assumes $\tau = 0$, that is, the cluster is truly intrinsically d-dimensional. Let G_n^d denote a random graph with n vertices, with edge weights $W_{ij} = \|x_i - x_j\|_\infty$, where $x_1, \dots, x_n \overset{i.i.d.}{\sim} \mathrm{Unif}([0,1]^d)$. For $\epsilon > 0$, let $G_n^d(\epsilon)$ be the thresholded version of G_n^d, where edges with W_{ij} greater than ϵ are deleted. Define the random variable $c_{n,d} = \inf\{\epsilon > 0 : \ G_n^d(\epsilon) \text{ is connected}\}$. It is known (Penrose, 1999) that $\max_{i,j} \rho_{\ell\ell}(x_i, x_j) = c_{n,d}$ for a fixed realization of the points $\{x_i\}_{i=1}^n$. Moreover, Appel and Russo (2002) showed $c_{n,d}$ has an almost sure limit in n:

$$\lim_{n\to\infty} (c_{n,d})^d \frac{n}{\log(n)} = \begin{cases} 1 & d = 1, \\ \frac{1}{2d} & d \geq 2. \end{cases}$$

Therefore $\max_{i,j} \rho_{\ell\ell}(x_i, x_j) \sim (\log(n)/n)^{\frac{1}{d}}$, almost surely as $n \to \infty$. Since the ℓ^2 and ℓ^∞ norms are equivalent up to a $\sqrt{d}$ factor, a similar result holds in the case of ℓ^2 norm being used for edge weights. To compare this asymptotic limit with our results, let $\epsilon_* = \max_{i,j} \rho_{\ell\ell}(x_i, x_j)$. By Theorem 4.3, $\epsilon_*^{-d} \log(\epsilon_*^{-d}) \gtrsim n$. Since $\epsilon_* \sim (\log n/n)^{\frac{1}{d}}$, $\epsilon_*^{-d} \log(\epsilon_*^{-d}) \sim (n/\log n) \log(n/\log n) \sim n$ as $n \to \infty$. This shows that our lower bound for $\epsilon_*^{-d} \log(\epsilon_*^{-d})$ matches the one given by the asymptotic limit and is thus sharp.

4.2. Lower-Bounding Between-Cluster Distances and kNN LLPD

Having shown conditions guaranteeing that all points within a cluster are close together in the LLPD, we now derive conditions guaranteeing that points in different clusters are far apart in LLPD. Points in the noise region may generate short paths between the clusters: we will upper-bound the number of between-clusters noise points that can be tolerated. Our approach is related to percolation theory (Gilbert, 1961; Roberts and Storey, 1968; Stauffer and Aharony, 1994) and analysis of single linkage clustering (Hartigan, 1981). The following theorem is in fact inspired by Lemma 2 in Hartigan (1981).

Theorem 4.7 *Under the LDLN data model and assumptions, with* ϵ_{btw} *as in* (3.7)*, for* $\epsilon > 0$

$$\tilde{n} \leq \frac{t^{\lfloor \frac{\delta}{\epsilon} \rfloor^{-1}} \mathcal{H}^D(\tilde{\mathcal{X}})}{\epsilon^D \mathcal{H}^D(B_1)} \implies \mathbb{P}\left(\epsilon_{\mathrm{btw}} > \epsilon\right) \geq 1 - t\,.$$

Proof We say that the ordered set of points $x_{i_1}, \dots, x_{i_{k_{\mathrm{nse}}}}$ forms an ϵ-chain of length k_{nse} if $\|x_{i_j} - x_{i_{j+1}}\|_2 \leq \epsilon$ for $1 \leq j \leq k_{\mathrm{nse}} - 1$. The probability that an ordered set of k_{nse} points forms an ϵ-chain is bounded above by $\left(\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right)^{k_{\mathrm{nse}}-1}$. There are $\frac{\tilde{n}!}{(\tilde{n}-k_{\mathrm{nse}})!}$

ordered sets of k_{nse} points. Letting $A_{k_{\text{nse}}}$ be the event that there exist k_{nse} points forming an ϵ-chain of length k_{nse}, we have

$$\mathbb{P}(A_{k_{\text{nse}}}) \leq \frac{\tilde{n}!}{(\tilde{n}-k_{\text{nse}})!}\left(\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right)^{k_{\text{nse}}-1} \leq \tilde{n}\left(\frac{\mathcal{H}^D(B_1)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\tilde{n}\epsilon^D\right)^{k_{\text{nse}}-1}.$$

Note that $A_{k_{\text{nse}}+1} \subset A_{k_{\text{nse}}}$. In order for there to be a path between $\mathcal{X}_i$ and $\mathcal{X}_j$ (for some $i \neq j$) with all legs bounded by ϵ, there must be at least $\lfloor \delta/\epsilon \rfloor - 1$ points in $\tilde{\mathcal{X}}$ forming an ϵ-chain. Thus recalling $\epsilon_{\text{btw}} = \min_{l\neq s} \min_{x\in\mathcal{X}_l, y\in\mathcal{X}_s} \rho_{\ell\ell}(x,y)$, we have:

$$\mathbb{P}\left(\epsilon_{\text{btw}} \leq \epsilon\right) \leq \mathbb{P}\left(\bigcup_{k_{\text{nse}}=\lfloor\frac{\delta}{\epsilon}\rfloor-1}^{\infty} A_{k_{\text{nse}}}\right) = \mathbb{P}\left(A_{\lfloor\frac{\delta}{\epsilon}\rfloor-1}\right) \leq \tilde{n}\left(\frac{\mathcal{H}^D(B_1)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\tilde{n}\epsilon^D\right)^{\lfloor\frac{\delta}{\epsilon}\rfloor-2} \leq t$$

as long as $\log t \geq \log\tilde{n} + (\lfloor\delta/\epsilon\rfloor - 2)(\log\tilde{n} + \log\epsilon^D + \log\mathcal{H}^D(B_1)/\mathcal{H}^D(\tilde{\mathcal{X}}))$. A simple calculation proves the claim. ■

Remark 4.8 *The above bound is independent of the number of clusters K, as the argument is completely based on the minimal distance that must be crossed between-clusters.*

Combining Theorem 4.7 with Theorem 4.3 or 4.6 allows one to derive conditions guaranteeing the maximal within cluster LLPD is smaller than the minimal between cluster LLPD with high probability, which in turn can be used to derive performance guarantees for spectral clustering on the cluster points. Since however it is not known a priori which points are cluster points, one must robustly distinguish the clusters from the noise. We propose removing any point whose LLPD to its $k_{\text{nse}}^{\text{th}}$ LLPD-nearest neighbor is sufficiently large (denoised LDLN data). The following theorem guarantees that, under certain conditions, all noise points that are not close to a cluster region will be removed by this procedure. The argument is similar to that in Theorem 4.7, although we replace the notion of an ϵ-chain of length k_{nse} with that of an ϵ-group of size k_{nse}.

Theorem 4.9 *Under the LDLN data model and assumptions, with ϵ_{nse} as in* (3.7)*, for $\epsilon > 0$*

$$\tilde{n} \leq \frac{2t^{\frac{1}{k_{nse}+1}}}{(k_{nse}+1)}\left(\frac{\mathcal{H}^D(\tilde{\mathcal{X}})}{\mathcal{H}^D(B_1)}\right)^{\frac{k_{nse}}{k_{nse}+1}} \epsilon^{-D\frac{k_{nse}}{k_{nse}+1}} \implies \mathbb{P}\left(\epsilon_{\text{nse}} > \epsilon\right) \geq 1-t\,.$$

Proof Let $\{x_i\}_{i=1}^{\tilde{n}}$ denote the points in $\tilde{X}$. Let $A_{k_{\text{nse}},\epsilon}$ be the event that there exists an ϵ-group of size k_{nse}, that is, there exist k_{nse} points such that the LLPD between all pairs is at most ϵ. Note that $A_{k_{\text{nse}},\epsilon}$ can also be described as the event that there

exists an ordered set of k_{nse} points $x_{\pi_1}, \ldots, x_{\pi_{k_{\text{nse}}}}$ such that $x_{\pi_i} \in \bigcup_{j=1}^{i-1} B_\epsilon(x_{\pi_j})$ for all $2 \le i \le k_{\text{nse}}$. Let $C_{\pi,i}$ denote the event that $x_{\pi_i} \in \bigcup_{j=1}^{i-1} B_\epsilon(x_{\pi_j})$. For a fixed ordered set of points associated with the ordered index set π, we have

$$\mathbb{P}\left(x_{\pi_i} \in \bigcup_{j=1}^{i-1} B_\epsilon(x_{\pi_j}) \text{ for } 2 \le i \le k_{\text{nse}}\right) = \mathbb{P}(C_{\pi,2})\mathbb{P}(C_{\pi,3}|C_{\pi,2}) \ldots \mathbb{P}\left(C_{\pi,k_{\text{nse}}} | \bigcap_{j=2}^{k_{\text{nse}}-1} C_{\pi,j}\right)$$
$$\le \frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\left(2\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right) \ldots \left((k_{\text{nse}}-1)\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right) = (k_{\text{nse}}-1)!\left(\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right)^{k_{\text{nse}}-1}.$$

There are $\frac{\tilde{n}!}{(\tilde{n}-k_{\text{nse}})!}$ ordered sets of k_{nse} points, so that

$$\mathbb{P}(A_{k_{\text{nse}},\epsilon}) \le \frac{\tilde{n}!}{(\tilde{n}-k_{\text{nse}})!}(k_{\text{nse}}-1)!\left(\frac{\mathcal{H}^D(B_\epsilon)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\right)^{k_{\text{nse}}-1} \le \tilde{n}(k_{\text{nse}}-1)!\left(\frac{\mathcal{H}^D(B_1)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\tilde{n}\epsilon^D\right)^{k_{\text{nse}}-1} \le t$$

as long as $\tilde{n}(k_{\text{nse}}-1)!\left(\frac{\mathcal{H}^D(B_1)}{\mathcal{H}^D(\tilde{\mathcal{X}})}\tilde{n}\epsilon^D\right)^{k_{\text{nse}}-1} \le t$, which occurs if $\tilde{n} \le \frac{2t^{\frac{1}{k_{\text{nse}}}}\mathcal{H}^D(\tilde{\mathcal{X}})^{\frac{k_{\text{nse}}-1}{k_{\text{nse}}}}}{k_{\text{nse}}\mathcal{H}^D(B_1)^{\frac{k_{\text{nse}}-1}{k_{\text{nse}}}}\epsilon^{\frac{D(k_{\text{nse}}-1)}{k_{\text{nse}}}}}$ for $k_{\text{nse}} \ge 2$. Since $\mathbb{P}(\epsilon_{\text{nse}} > \epsilon) = \mathbb{P}(\min_{x\in\tilde{X}} \beta_{k_{\text{nse}}}(x, \tilde{X}) > \epsilon) = 1 - \mathbb{P}(A_{k_{\text{nse}}+1,\epsilon})$, the theorem holds for $k_{\text{nse}} \ge 1$. ■

Remark 4.10 *The theorem guarantees* $\epsilon_{\text{nse}} \ge \left(\frac{2\mathcal{H}^D(\tilde{\mathcal{X}})(2t)^{\frac{1}{k_{nse}}}}{\mathcal{H}^D(B_1)((k_{nse}+1)\tilde{n})^{\frac{k_{nse}+1}{k_{nse}}}}\right)^{\frac{1}{D}}$ *with probability at least* $1-t$. *The lower bound for* ϵ_{nse} *is maximized at the unique maximizer in* $k_{nse} > 0$ *of* $f(k_{nse}) = (2t)^{\frac{1}{k_{nse}}}((k_{nse}+1)\tilde{n})^{-\frac{k_{nse}+1}{k_{nse}}}$, *which occurs at the positive root* k_{nse*} *of* $k_{nse} - \log(k_{nse}+1) = \log\tilde{n} - \log(2t)$. *Notice that* $k_{nse*} = O(\log\tilde{n})$, *so we may, and will, restrict our attention to* $k_{nse} \le k_{nse*} = O(\log\tilde{n})$.

4.3. Robust Denoising with LLPD

Combining Corollary 4.4 ($\tau > 0$ but small) or Theorem 4.6 ($\tau = 0$) with Theorem 4.9 determines how many noise points can be tolerated while within-cluster LLPD remain small relative to $k_{\text{nse}}^{\text{th}}$ nearest neighbor LLPD of noise points. Any $C \ge 1$ in the following theorem guarantees $\epsilon_{\text{in}} < \epsilon_{\text{nse}}$; when $C \gg 1$, $\epsilon_{\text{in}} \ll \epsilon_{\text{nse}}$, and LLPD easily differentiates the clusters from the noise. The proof is given in Appendix A. A similar result for the set-up of Theorem 4.3 is omitted for brevity.

Theorem 4.11 *Assume the LDLN data model and assumptions, and define*

$$\tau_* := \max_{l=1,\ldots,k}\left(\frac{\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{n_l\mathcal{H}^d(B_1)}\log\left(2^d n_l\frac{2K}{t}\right)\right)^{\frac{1}{d}}.$$

Let $0 \le \tau < \frac{\tau_*}{8} \wedge \frac{\epsilon_0}{5}$ *and let* $\epsilon_{\text{in}}, \epsilon_{\text{nse}}$ *as in* (3.7). *For any* $C > 0$,

$$\tilde{n} < \frac{\left(\frac{t}{2}\right)^{\frac{1}{k_{nse}+1}}}{k_{nse}+1}\left(\frac{\mathcal{H}^D(\tilde{\mathcal{X}})}{\mathcal{H}^D(B_1)}\right)^{\frac{k_{nse}}{k_{nse}+1}}\left(\frac{1}{C\tau_*}\right)^{D\frac{k_{nse}}{k_{nse}+1}} \implies \mathbb{P}(C\epsilon_{\text{in}} < \epsilon_{\text{nse}}) \ge 1-t\,.$$

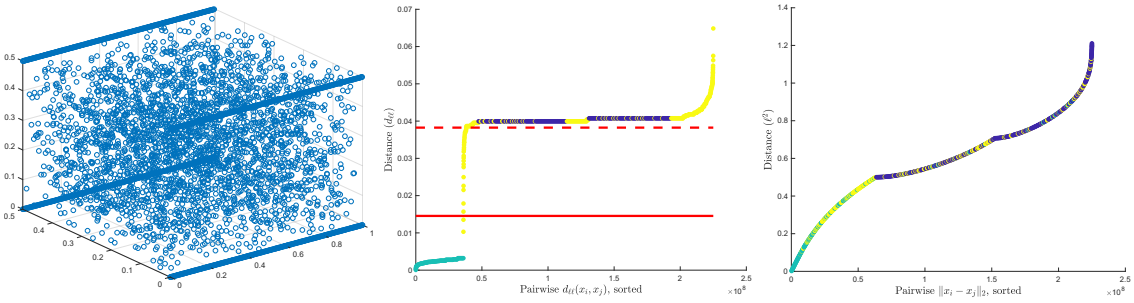
Ignoring log terms and geometric constants, the number of noise points $\tilde{n}$ can be taken as large as $\min_l n_l^{\frac{D}{d}\left(\frac{k_{\text{nse}}}{k_{\text{nse}}+1}\right)}$. Hence if $d \ll D$, an enormous amount of noise points are tolerated while ϵ_{in} is still small relative to ϵ_{nse}. This result is deployed to prove LLPD spectral clustering is robust to large amounts of noise in Theorem 5.12 and Corollary 5.14, and is in particular relevant to condition (5.15), which articulates the range of denoising parameters for which LLPD spectral clustering will perform well.

4.4. Phase Transition in LLPD

In this section, we numerically validate our denoising scheme on simple data. The data is a mixture of five uniform distributions: four from non-adjacent edges of $[0,1] \times [0,\frac{1}{2}] \times [0,\frac{1}{2}]$, and one from the interior of $[0,1] \times [0,\frac{1}{2}] \times [0,\frac{1}{2}]$. Each distribution contributed 3000 sample points. Figure 5a shows the data and Figure 5b all sorted LLPDs. The sharp phase transition is explained mathematically by Theorems 4.6 and 4.7. Indeed, $d = 1, D = 3$ in this example, so Theorem 4.6 guarantees that with high probability, the maximum within cluster LLPD, call it ϵ_{in}, scales as $\epsilon_{\text{in}}^{-1} \log(\epsilon_{\text{in}}^{-1}) \gtrsim n$ while Theorem 4.7 guarantees that with high probability, the minimum between cluster LLPD, call it ϵ_{btw}, scales as $\epsilon_{\text{btw}} \gtrsim n^{-\frac{1}{3}}$. The empirical estimates can be compared with the theoretical guarantees, which are shown on the plot. The guarantees require a confidence level, parametrized by t; this parameter was chosen to be $t = .01$ for this example. The solid red line denotes the maximum within cluster LLPD guaranteed with probability exceeding $1 - t = .99$, and the dashed red line denotes the minimum between cluster LLPD, guaranteed with probability exceeding $1 - t$. It is clear from Figure 5 that the theoretical lower lower bound on ϵ_{btw} is rather sharp, while the theoretical upper bound on ϵ_{in} is looser. Despite the lack of sharpness in estimating ϵ_{in}, the theoretical bounds are quite sufficient to separate the within-cluster and between cluster LLPD. When $d \ll D$, the difference between these theoretical bounds becomes much larger.

5. Performance Guarantees for Ultrametric and LLPD Spectral Clustering

In this section we first derive performance guarantees for spectral clustering with any ultrametric. We show that when the data consists of cluster cores surrounded by noise, the weight matrix W used in spectral clustering is, for a certain range of scales σ, approximately block diagonal with constant blocks. In this range of σ, the number of clusters can be inferred from the maximal eigengap of L_{SYM}, and spectral clustering achieves high labeling accuracy. On the other hand, for Euclidean spectral clustering it is hard to choose a scale parameter that is simultaneously large enough to guarantee a strong connection between every pair of points in the same cluster and small enough to produce weak connections between clusters (and even



(a) Four clusters in $[0,1] \times [0,\frac{1}{2}] \times [0,\frac{1}{2}]$.

(b) Corresponding pairwise LLPD, sorted.

(c) Corresponding pairwise ℓ^2 distances, sorted.

Figure 5: (a) The clusters are on edges of the rectangular prism so that the pairwise LLPDs between the clusters is at least 1. The interior is filled with noise points. Each cluster has 3000 points, as does the interior noise region. (b) The sorted $\rho_{\ell\ell}$ plot shows within-cluster LLPDs in green, between-cluster LLPDs in blue, and LLPDs involving noise points in yellow. There is a clear phase transition between the within-cluster and between-cluster LLPDs. This empirical observation can be compared with the theoretical guarantees of Theorems 4.6 and 4.7. Setting $t = .01$ in those theorems yield corresponding maximum within-cluster LLPD (shown with the solid red line) and minimum between-cluster distance (shown with the dashed red line). The empirical results confirm our theoretical guarantees. Notice moreover that there is no clear separation between the Euclidean distances, which are shown in (c). This illustrates the challenges faced by classical spectral clustering, compared to LLPD spectral clustering, for this data set.

when possible the shape (e.g. elongation) of clusters affects the ability to identify the correct clusters). The resulting Euclidean weight matrix is not approximately block diagonal for any choice of σ, and the eigengap of L_{SYM} becomes uninformative and the labeling accuracy potentially poor. Moreover, using an ultrametric for spectral clustering leads to direct lower bounds on the degree of noise points, since if a noise point is close to any cluster point, it is close to all points in the given cluster. It is well-known that spectral clustering is unreliable for points of low degree and in this case L_{SYM} may have arbitrarily many small eigenvalues (Von Luxburg, 2007).

After proving results for general ultrametrics, we derive specific performance guarantees for LLPD spectral clustering on the LDLN data model. We remove low density points by considering each point's LLPD-nearest neighbor distances, then derive bounds on the eigengap and labeling accuracy which hold even in the presence of noise points with weak connections to the clusters. We prove there is a large range of values of both the thresholding and scale parameter for which we correctly recover the clusters, illustrating that LLPD spectral clustering is robust to the choice of parameters and presence of noise. In particular, when the clusters have a very low-dimensional structure and the noise is very high-dimensional, that is, when $d \ll D$, an enormous amount of noise points can be tolerated. Throughout this section, we use the notation established in Subsection 2.2.

5.1. Ultrametric Spectral Clustering

Let $\rho : \mathbb{R}^D \times \mathbb{R}^D \to [0, \infty)$ be an ultrametric; see (2.2). We analyze L_{SYM} under the assumptions of the following cluster model. As will be seen in Subsection 5.2, this cluster model holds for data drawn from the LDLN data model with the LLPD ultrametric, but it may be of interest in other regimes and for other ultrametrics. The model assumes there are K sets forming cluster cores and each cluster core has a halo of noise points surrounding it; for $1 \leq l \leq K$, A_l denotes the cluster core and C_l the associated halo of noise points. For LDLN data, the parameter ϵ_{sep} corresponds to the minimal between cluster distance after denoising.

Assumption 1 (Ultrametric Cluster Model) *For $1 \leq l \leq K$, assume A_l and C_l are disjoint finite sets, and let $\tilde{A}_l = A_l \cup C_l$. Let $N = |\cup_l \tilde{A}_l|$. Assume that for some $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{sep}}$:*

$$\rho(x_i^l, x_j^l) \leq \epsilon_{\text{in}} \qquad \forall x_i^l, x_j^l \in A_l, 1 \leq l \leq K, \tag{5.1}$$

$$\epsilon_{\text{in}} < \rho(x_i^l, x_j^l) \leq \theta \qquad \forall x_i^l \in A_l, x_j^l \in C_l, 1 \leq l \leq K, \tag{5.2}$$

$$\rho(x_i^l, x_j^s) \geq \epsilon_{\text{sep}} \qquad \forall x_i^l \in \tilde{A}_l, x_j^s \in \tilde{A}_s, 1 \leq l \neq s \leq K. \tag{5.3}$$

Moreover, let $\zeta_N = \max_{1 \leq l \leq K} \frac{N}{|\tilde{A}_l|}$.

Theorem 5.5 shows that under Assumption 1, the maximal eigengap of L_{SYM} corresponds to the number of clusters K and spectral clustering with K principal eigenvectors achieves perfect labeling accuracy. The label accuracy result is obtained by showing the spectral embedding with K principal eigenvectors is a *perfect representation* of the sets $\tilde{A}_l$, as defined in Vu (2018).

Definition 5.4 (Perfect Representation) *A clustering representation is perfect if there exists an $r > 0$ such that*

- *Vertices in the same cluster have distance at most r.*
- *Vertices from different clusters have distance at least $4r$ from each other.*

There are multiple clustering algorithms which are guaranteed to perfectly recover the labels of $\tilde{A}_l$ from a perfect representation, including K-means with furthest point initialization and single linkage clustering. Again following the terminology in Vu (2018), we will refer to all such clustering algorithms as *clustering by distances*. The proof of Theorem 5.5 is in Appendix B.

Theorem 5.5 *Assume the ultrametric cluster model. Then $\lambda_{K+1} - \lambda_K$ is the largest gap in the eigenvalues of $L_{SYM}(\cup_l \tilde{A}_l, \rho, f_\sigma)$ provided*

$$\frac{1}{2} \geq \underbrace{5(1 - f_\sigma(\epsilon_{\text{in}}))}_{\textit{Cluster Coherence}} + \underbrace{6\zeta_N f_\sigma(\epsilon_{\text{sep}})}_{\textit{Cluster Separation}} + \underbrace{4(1 - f_\sigma(\theta))}_{\textit{Noise}} + \beta \,, \tag{5.6}$$

where $\beta = O((1 - f_\sigma(\epsilon_{\text{in}}))^2 + \zeta_N^2 f_\sigma(\epsilon_{\text{sep}})^2 + (1 - f_\sigma(\theta))^2)$ *denotes higher-order terms. Moreover, if*

$$\frac{C}{K^3 \zeta_N^2} \geq (1 - f_\sigma(\epsilon_{\text{in}})) + \zeta_N f_\sigma(\epsilon_{\text{sep}}) + (1 - f_\sigma(\theta)) + \beta \ , \tag{5.7}$$

where C *is an absolute constant, then clustering by distances on the* K *principal eigenvectors of* $L_{SYM}(\cup_l \tilde{A}_l, \rho, f_\sigma)$ *perfectly recovers the cluster labels.*

For condition (5.6) to hold, the following three terms must all be small:

- **Cluster Coherence**: This term is minimized by choosing σ large so that $f_\sigma(\epsilon_{\text{in}}) \approx 1$; the larger the scale parameter, the stronger the within-cluster connections.
- **Cluster Separation**: This term is minimized by choosing σ small so that $f_\sigma(\epsilon_{\text{sep}}) \approx 0$; the smaller the scale parameter, the weaker the between-cluster connections. Note ζ_N is minimized when clusters are balanced, in which case $\zeta_N = K$.
- **Noise**: This term is minimized by choosing σ large, so that once again $f_\sigma(\theta) \approx 1$. When the scale parameter is large, noise points around the cluster will be well connected to their designated cluster.
- **Higher Order Terms**: This term consists of terms that are quadratic in $(1 - f_\sigma(\epsilon_{\text{in}})), f_\sigma(\epsilon_{\text{sep}}), (1 - f_\sigma(\theta))$, which are small in our regime of interest.

Solving for the scale parameter σ will yield a range of σ values where the eigengap statistic is informative; this is done in Corollary 5.14 for LLPD spectral clustering on the LDLN data model.

Condition (5.7) guarantees that clustering the LLPD spectral embedding results in perfect label accuracy, and requires a stronger scaling with respect to K than Condition (5.6), as when $\zeta_N = O(K)$ there is an additional factor of K^{-5} on the left hand side of the inequality. Determining whether this scaling in K is optimal is a topic of ongoing research, though the present article is more concerned with scaling in n, d, and D. Condition (5.7) in fact guarantees perfect accuracy for clustering by distances on the spectral embedding regardless of whether the K principal eigenvectors of L_{SYM} are row-normalized or not. Row normalization is proposed in Ng et al. (2002) and generally results in better clustering results when some points have small degree (Von Luxburg, 2007); however it is not needed here because the properties of LLPD cause all points to have similar degree.

Remark 5.8 *One can also derive a label accuracy result by applying Theorem 2 from Ng et al. (2002), restated in Arias-Castro (2011), to show the spectral embedding satisfies the so-called* orthogonal cone property (OCP) *(Schiebinger et al., 2015). Indeed, let* $\{\phi_k\}_{k=1}^K$ *be the principal eigenvectors of* L_{SYM}*. The OCP guarantees that in the representation* $x \mapsto \{\phi_k(x)\}_{k=1}^K$*, distinct clusters localize in nearly orthogonal directions, not too far from the origin. Proposition 1 from Schiebinger et al. (2015) can then be applied to conclude* K*-means on the spectral embedding achieves high accuracy. Specifically, if (5.6) is satisfied, then with probability at least* $1-t$*,* K*-means on the* K *principal eigenvectors of* $L_{SYM}(\cup_l \tilde{A}_l, \rho, f_\sigma)$ *achieves accuracy at least* $1-\frac{cK^9\zeta_N^3(f_\sigma(\epsilon_{\text{sep}})^2+\beta)}{t}$ *where* c *is an absolute constant and* β *denotes higher order terms. This approach results in a less restrictive scaling for* $1-f_\sigma(\epsilon_{\text{in}}), f_\sigma(\epsilon_{\text{sep}})$ *in terms of* K, ζ_N *than given in Condition (5.7), but does not guarantee perfect accuracy, and also requires row normalization of the spectral embedding as proposed in Ng et al. (2002). The argument using this approach to proving cluster accuracy is not discussed in this article, for reasons of space and as to not introduce additional excessive notation.*

5.2. LLPD Spectral Clustering with kNN LLPD Thresholding

We now return to the LDLN data model defined in Subsection 3.2 and show that it gives rise to the ultrametric cluster model described in Assumption 1 when combined with the LLPD metric. Theorem 5.5 can thus be applied to derive performance guarantees for LLPD spectral clustering on the LDLN data model. All of the notation and assumptions established in Subsection 3.2 hold throughout Subsection 5.2.

5.2.1. Thresholding

Before applying spectral clustering, we denoise the data by removing any points having sufficiently large LLPD to their $k_{\text{nse}}^{\text{th}}$ LLPD-nearest neighbor. Motivated by the sharp phase transition illustrated in Subsection 4.4, we choose a threshold θ and discard a point $x \in X$ if $\beta_{k_{\text{nse}}}(x, X) > \theta$. Note that the definition of ϵ_{nse} guarantees that we can never have a group of more than k_{nse} noise points where all pairwise LLPD are smaller than ϵ_{nse}, because if we did there would be a point $x \in \tilde{X}$ with $\beta_{k_{\text{nse}}}(x, \tilde{X}) < \epsilon_{\text{nse}}$. Thus if $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{nse}}$ then, after thresholding, the data will consist of the cluster cores X_l with θ-groups of at most k_{nse} noise points emanating from the cluster cores, where a θ-group denotes a set of points where the LLPD between all pairs of points in the group is at most θ.

We assume LLPD is re-computed on the denoised data set X_N, whose cardinality we define to be N, and let $\rho_{\ell\ell}^{X_N}$ denote the corresponding LLPD metric. The points remaining after thresholding consist of the sets $\tilde{A}_l$, where

$$\begin{aligned} A_l &= \{x_i \in X_l\} \cup \{x_i \in \tilde{X} \mid \rho_{\ell\ell}(x_i, x_j) \leq \epsilon_{\text{in}} \text{ for some } x_j \in X_l\}, \\ C_l &= \{x_i \in \tilde{X} \mid \epsilon_{\text{in}} < \rho_{\ell\ell}(x_i, x_j) \leq \theta \text{ for some } x_j \in X_l\}, \\ \tilde{A}_l &= A_l \cup C_l = \{x_i \in X_l\} \cup \{x_i \in \tilde{X} \mid \rho_{\ell\ell}(x_i, x_j) \leq \theta \text{ for some } x_j \in X_l\}. \end{aligned} \tag{5.9}$$

The cluster core A_l consists of the points X_l plus any noise points in $\tilde{X}$ that are indistinguishable from X_l, being within the maximal within-cluster LLPD of X_l. The set C_l consists of the noise points in $\tilde{X}$ that are θ-close to X_l in LLPD.

5.2.2. Supporting Lemmata

The following two lemmata are needed to prove Theorem 5.12, the main result of this subsection. The first one guarantees that the sets defined in (5.9) describe exactly the points which survive thresholding, that is $X_N = \cup_l \tilde{A}_l$.

Lemma 5.10 *Assume the LDLN data model and assumptions, and let $\tilde{A}_l$ be as in (5.9). If $k_{nse} < n_{min}$, $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{nse}}$, then $\beta_{k_{nse}}(x, X) \leq \theta$ if and only if $x \in \tilde{A}_l$ for some $1 \leq l \leq K$.*

Proof Assume $\beta_{k_{\text{nse}}}(x, X) \leq \theta$. If $x \in \cup_l X_l$, then clearly $x \in \cup_l \tilde{A}_l$, so assume $x \in \tilde{X}$. We claim there exists some $y \in \cup_l X_l$ such that $\rho_{\ell\ell}(x, y) \leq \theta$. Suppose not; then there exist k_{nse} points $\{x_i\}_{i=1}^{k_{\text{nse}}}$ in $\tilde{X}$ distinct from x with $\rho_{\ell\ell}(x, x_i) \leq \theta$; thus $\epsilon_{\text{nse}} \leq \theta$, a contradiction. Hence, there exists $y \in X_l$ such that $\rho_{\ell\ell}(x, y) \leq \theta$ and $x \in \tilde{A}_l$.
Now assume $x \in \tilde{A}_l$ for some $1 \leq l \leq K$. Then clearly there exists $y \in \cup_l X_l$ with $\rho_{\ell\ell}(x, y) \leq \theta$. Since $\epsilon_{\text{in}} < \theta$, x is within LLPD θ of all points in X_l, and since $k_{\text{nse}} < n_{\min}, \beta_{k_{\text{nse}}}(x, X) \leq \beta_{n_{\min}}(x, X) \leq \theta$. ■

Next we show that when there is sufficient separation between the cluster cores, the LLPD between any two points in distinct clusters is bounded by $\delta/2$, and thus the assumptions of Theorem 5.5 will be satisfied with $\epsilon_{\text{sep}} = \delta/2$.

Lemma 5.11 *Assume the LDLN data model and assumptions, and assume $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{nse}} \wedge \delta/(4k_{nse})$, $A_l, C_l, \tilde{A}_l$ as defined in (5.9), and $k_{nse} < n_{min}$. Then Assumption 1 is satisfied with $\rho = \rho_{\ell\ell}^{X_N}$, $\epsilon_{\text{sep}} = \delta/2$.*

Proof First note that if $x \in A_l$, then $\rho_{\ell\ell}(x, y) \leq \epsilon_{\text{in}}$ for all $y \in X_l$, and thus $x \notin C_l$, so A_l and C_l are disjoint.
Let $x_i^l, x_j^l \in A_l$. Then there exists $y_i, y_j \in X_l$ with $\rho_{\ell\ell}(x_i^l, y_i) \leq \epsilon_{\text{in}}$ and $\rho_{\ell\ell}(x_j^l, y_j) \leq \epsilon_{\text{in}}$, so $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \rho_{\ell\ell}(x_i^l, y_i) \vee \rho_{\ell\ell}(y_i, y_j) \vee \rho_{\ell\ell}(y_j, x_j^l) \leq \epsilon_{\text{in}}$. Since x_i^l, x_j^l were arbitrary, $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \epsilon_{\text{in}}$ for all $x_i^l, x_j^l \in A_l$. We now show that in fact $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^l) \leq \epsilon_{\text{in}}$. Suppose not. Since $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \epsilon_{\text{in}}$, there exists a path in X from x_i^l to x_j^l with all legs bounded by ϵ_{in}. Since $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^l) > \epsilon_{\text{in}}$, one of the points along this path must have been removed by thresholding, i.e. there exists y on the path with $\beta_{k_{\text{nse}}}(y, X) > \theta$. But then for all $x^l \in A_l$, $\rho_{\ell\ell}(y, x^l) \leq \rho_{\ell\ell}(y, x_i^l) \vee \rho_{\ell\ell}(x_i^l, x^l) \leq \epsilon_{\text{in}}$, so $\beta_{k_{\text{nse}}}(y, X) \leq \epsilon_{\text{in}}$ since $k_{\text{nse}} < n_{\min}$; contradiction.
Let $x_i^l \in A_l, x_j^l \in C_l$. Then there exist points $y_i, y_j \in X_l$ such that $\rho_{\ell\ell}(x_i^l, y_i) \leq \epsilon_{\text{in}}$ and $\epsilon_{\text{in}} < \rho_{\ell\ell}(x_j^l, y_j) \leq \theta$. Thus $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \rho_{\ell\ell}(x_i^l, y_i) \vee \rho_{\ell\ell}(y_i, y_j) \vee \rho_{\ell\ell}(y_j, x_j^l) \leq \epsilon_{\text{in}} \vee \epsilon_{\text{in}} \vee \theta = \theta$.

Now suppose $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \epsilon_{\text{in}}$. Then $\rho_{\ell\ell}(x_j^l, y_i) \leq \rho_{\ell\ell}(x_j^l, x_i^l) \vee \rho_{\ell\ell}(x_i^l, y_i) \leq \epsilon_{\text{in}}$ so that $x_j^l \in A_l$ since $y_i \in X_l$; this is a contradiction since $x_j^l \in C_l$ and A_l and C_l are disjoint. We thus conclude $\epsilon_{\text{in}} < \rho_{\ell\ell}(x_i^l, x_j^l) \leq \theta$. Since x_i^l, x_j^l were arbitrary, $\epsilon_{\text{in}} < \rho_{\ell\ell}(x_i^l, x_j^l) \leq \theta$ for all $x_i^l \in A_l, x_j^l \in C_l$. We now show in fact $\epsilon_{\text{in}} < \rho_{\ell\ell}^{X_N}(x_i^l, x_j^l) \leq \theta$. Clearly, $\epsilon_{\text{in}} < \rho_{\ell\ell}(x_i^l, x_j^l) \leq \rho_{\ell\ell}^{X_N}(x_i^l, x_j^l)$. Now suppose $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^l) > \theta$. Since $\rho_{\ell\ell}(x_i^l, x_j^l) \leq \theta$, there exists a path in X from x_i^l to x_j^l with all legs bounded by θ. Since $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^l) > \theta$, one of the points along this path must have been removed by thresholding, i.e. there exists y on the path with $\beta_{k_{\text{nse}}}(y, X) > \theta$. But then for all $x^l \in \tilde{A}_l$, $\rho_{\ell\ell}(y, x^l) \leq \rho_{\ell\ell}(y, x_i^l) \vee \rho_{\ell\ell}(x_i^l, x^l) \leq \theta$, so $\beta_{k_{\text{nse}}}(y, X) \leq \theta$ since $k_{\text{nse}} < n_{\min}$, which is a contradiction.

Finally, we show we can choose $\epsilon_{\text{sep}} = \delta/2$, that is, $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^s) \geq \delta/2$ for all $x_i^l \in \tilde{A}_l, x_j^s \in \tilde{A}_s, l \neq s$. We first verify that every point in $\tilde{A}_l$ is within Euclidean distance θk_{nse} of a point in X_l. Let $x \in \tilde{A}_l$ and assume $x \in \tilde{X}$ (otherwise there is nothing to show). Then there exists a point $y \in X_l$ with $\rho_{\ell\ell}(x, y) \leq \theta$, i.e. there exists a path of points from x to y with the length of all legs bounded by θ. Note there can be at most k_{nse} consecutive noise points along this path, since otherwise we would have a $z \in \tilde{X}$ with $\beta_{k_{\text{nse}}}(z, \tilde{X}) \leq \theta$ which contradicts $\epsilon_{\text{nse}} > \theta$. Let y^* be the last point in X_l on this path. Since $\theta < \delta/(4k_{\text{nse}}) < \delta/(2k_{\text{nse}} + 1)$, $\text{dist}(X_l, X_s) \geq \delta > 2\theta k_{\text{nse}} + \theta$, and the path cannot contain any points in $X_s, l \neq s$; thus the path from y^* to x consists of at most k_{nse} points in $\tilde{X}$, so $\|x - y^*\|_2 \leq k_{\text{nse}}\theta$. Thus: $\min_{1 \leq l \neq s \leq K} \text{dist}(\tilde{A}_l, \tilde{A}_s) \geq \min_{1 \leq l \neq s \leq K} \text{dist}(X_l, X_s) - 2\theta k_{\text{nse}} \geq \delta - 2\theta k_{\text{nse}} > \delta/2$ since $\theta < \delta/(4k_{\text{nse}})$. Now by Lemma 5.10, there are no points outside of $\cup_l \tilde{A}_l$ which survive thresholding, so we conclude $\rho_{\ell\ell}^{X_N}(x_i^l, x_j^s) \geq \delta/2$ for all $x_i^l \in \tilde{A}_l, x_j^s \in \tilde{A}_s$. ■

5.2.3. Main Result

We now state our main result for LLPD spectral clustering with kNN LLPD thresholding.

Theorem 5.12 *Assume the LDLN data model and assumptions. For a chosen θ and k_{nse}, perform thresholding at level θ as above to obtain X_N, and assume $k_{nse} < n_{min}$, $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{nse}} \wedge \delta/(4k_{nse})$. Then $\lambda_{K+1} - \lambda_K$ is the largest gap in the eigenvalues of $L_{SYM}(X_N, \rho_{\ell\ell}^{X_N}, f_\sigma)$ provided that*

$$\frac{1}{2} \geq 5(1 - f_\sigma(\epsilon_{\text{in}})) + 6\zeta_N f_\sigma(\delta/2) + 4(1 - f_\sigma(\theta)) + \beta \ , \tag{5.13}$$

and clustering by distances on the K principal eigenvectors of $L_{SYM}(X_N, \rho_{\ell\ell}^{X_N}, f_\sigma)$ perfectly recovers the cluster labels provided that

$$\frac{C}{K^3 \zeta_N^2} \geq (1 - f_\sigma(\epsilon_{\text{in}})) + \zeta_N f_\sigma(\delta/2) + (1 - f_\sigma(\theta)) + \beta \ ,$$

where $\beta = O\left((1-f_\sigma(\epsilon_{\text{in}})^2 + \zeta_N^2 f_\sigma(\delta/2)^2 + (1-f_\sigma(\theta))^2\right)$ *denotes higher-order terms and* C *is an absolute constant. In addition, for* $n_{\min}$ *large enough with probability at least* $1 - O(n_{\min}^{-1})$, $\zeta_N \leq 2\zeta_n + 3k_{nse}\zeta_\theta$ *for the LDLN data model balance parameters* ζ_n, ζ_θ.

Proof Define the sets $A_l, C_l, \tilde{A}_l$ as in (5.9). By Lemma 5.10, removing all points satisfying $\beta_{k_{\text{nse}}}(x, X) > \theta$ leaves us with exactly $X_N = \cup_l \tilde{A}_l$. By Lemma 5.11, all assumptions of Theorem 5.5 are satisfied for ultrametric $\rho_{\ell\ell}^{X_N}$ and we can apply Theorem 5.5 with $\epsilon_{\text{in}}, \epsilon_{\text{nse}}$ as defined in Subsection 3.2 and $\epsilon_{\text{sep}} = \delta/2$. All that remains is to verify the bound on ζ_N.

Recall $\tilde{A}_l = X_l \cup \{x_i \in \tilde{X} \mid \rho_{\ell\ell}(x_i, x_j) \leq \theta \text{ for some } x_j \in X_l\}$; let m_l denote the cardinality of $\{x_i \in \tilde{X} \mid \rho_{\ell\ell}(x_i, x_j) \leq \theta \text{ for some } x_j \in X_l\}$ so that $\zeta_N = \max_{1 \leq l \leq K} \frac{\sum_{i=1}^K n_i + m_i}{n_l + m_l}$. For $1 \leq l \leq K$, let $\omega_l = \sum_{x \in \tilde{X}} \mathbb{1}_{x \in B(\mathcal{X}_l, \theta) \setminus \mathcal{X}_l}$ denote the number of noise points that fall within a tube of width θ around the cluster region $\mathcal{X}_l$. Note that $\omega_l \sim \text{Bin}(\tilde{n}, p_{l,\theta})$ where $p_{l,\theta} = \mathcal{H}^D(B(\mathcal{X}_l, \theta) \setminus \mathcal{X}_l)/\mathcal{H}^D(\tilde{\mathcal{X}})$ is as defined in Section 3.2. The assumptions of Theorem 5.12 guarantee that $m_l \leq k_{\text{nse}}\omega_l$, since ω_l is the number of groups attaching to $\mathcal{X}_l$, and each group consists of at most k_{nse} noise points. To obtain a lower bound for m_l, note that $m_l \geq \sum_{x \in \tilde{X}} \mathbb{1}_{x \in B(X_l, \theta) \setminus \mathcal{X}_l}$, where $B(X_l, \theta) \subset B(\mathcal{X}_l, \theta)$ is formed from the discrete sample points X_l. Since $B(X_l, \theta) \to B(\mathcal{X}_l, \theta)$ as $n_l \to \infty$, for $n_{\min}$ large enough $\mathcal{H}^D(B(X_l, \theta) \setminus \mathcal{X}_l) \geq \frac{1}{2}\mathcal{H}^D(B(\mathcal{X}_l, \theta) \setminus \mathcal{X}_l)$, and $m_l \geq \omega_{l,2}$ where $\omega_{l,2} \sim \text{Bin}(\tilde{n}, p_{l,\theta}/2)$.

We first consider the high noise case $\tilde{n}p_{\min,\theta} \geq n_{\min}$, and define $\zeta_l = \frac{\sum_{i=1}^K n_i + m_i}{n_l + m_l}$. We have

$$\zeta_l \leq \frac{\sum_{i=1}^K n_i}{n_l} + \frac{\sum_{i=1}^K m_i}{m_l} \leq \frac{\sum_{i=1}^K n_i}{n_l} + k_{\text{nse}} \frac{\sum_{i=1}^K \omega_i}{\omega_{l,2}}.$$

A multiplicative Chernoff bound (Hagerup and Rüb, 1990) gives $\mathbb{P}(\omega_i \geq (1+\delta_1)\tilde{n}p_{i,\theta}) \leq \exp(-\delta_1^2 \tilde{n}p_{i,\theta}/3) \leq \exp(-\delta_1^2 n_{\min}/3)$ for any $0 \leq \delta_1 \leq 1$. Choosing $\delta_1 = \sqrt{3\log(Kn_{\min})/n_{\min}}$ and taking a union bound gives $\omega_i \leq (1+\delta_1)\tilde{n}p_{i,\theta}$ for all $1 \leq i \leq K$ with probability at least $1 - n_{\min}^{-1}$. A lower Chernoff bound also gives $\mathbb{P}(\omega_{i,2} \leq (1-\delta_2)\tilde{n}p_{i,\theta}/2) \leq \exp(-\delta_2^2 \tilde{n}p_{i,\theta}/4) \leq \exp(-\delta_2^2 n_{\min}/4)$ for any $0 \leq \delta_2 \leq 1$ and choosing $\delta_2 = \sqrt{4\log(Kn_{\min})/n_{\min}}$ gives $\omega_{i,2} \geq (1-\delta_2)\tilde{n}p_{i,\theta}/2$ for all $1 \leq i \leq K$ with probability at least $1 - n_{\min}^{-1}$. Thus with probability at least $1 - O(n_{\min}^{-1})$, one has

$$\frac{\sum_{i=1}^K \omega_i}{\omega_{2,l}} \leq \frac{\sum_{i=1}^K 2(1+\delta_1)\tilde{n}p_{i,\theta}}{(1-\delta_2)\tilde{n}p_{l,\theta}} \leq 3\frac{\sum_{i=1}^K p_{i,\theta}}{p_{l,\theta}}$$

for all $1 \leq l \leq K$ for $n_{\min}$ large enough, giving $\zeta_N = \max_{1 \leq l \leq K} \zeta_l \leq \zeta_n + 3\zeta_\theta$.

We next consider the small noise case $\tilde{n}p_{\min,\theta} \leq n_{\min}$. A Chernoff bound gives $\mathbb{P}(\omega_i \geq (1 + (\delta_i \vee \sqrt{\delta_i})\tilde{n}p_{i,\theta}) \leq \exp(-\delta_i^2 \tilde{n}p_{i,\theta}/3)$ for any $\delta_i \geq 0$. We choose $\delta_i = 3\log(Kn_{\min})/(\tilde{n}p_{i,\theta})$, so that with probability at least $1 - n_{\min}^{-1}$ we have

$$\omega_i \leq (1 + (\delta_i \vee \sqrt{\delta_i})\tilde{n}p_{i,\theta}) \leq 2\tilde{n}p_{i,\theta} + 6\log(Kn_{\min})$$

for all $1 \leq i \leq K$ and we obtain for $n_{\min}$ large enough

$$\begin{aligned}
\zeta_N &\leq \frac{\sum_{i=1}^K n_i + k_{\text{nse}}\omega_i}{n_{\min}} \\
&\leq \frac{\sum_{i=1}^K n_i + k_{\text{nse}}(2\tilde{n}p_{i,\theta} + 6\log(Kn_{\min}))}{n_{\min}} \\
&\leq \frac{\sum_{i=1}^K 2n_i + 2k_{\text{nse}}\tilde{n}p_{i,\theta}}{n_{\min}} \\
&\leq \frac{\sum_{i=1}^K 2n_i + 2k_{\text{nse}}n_{\min}p_{i,\theta}/p_{\min,\theta}}{n_{\min}} \\
&= 2\zeta_n + 2k_{\text{nse}}\zeta_\theta.
\end{aligned}$$

Combining the two cases $\zeta_N \leq 2\zeta_n + 3k_{\text{nse}}\zeta_\theta$ that with probability at least $1 - O(n_{\min}^{-1})$. ■

Theorem 5.13 illustrates that after thresholding, the number of clusters can be reliably estimated by the maximal eigengap for the range of σ values where (5.13) holds. The following corollary combines what we know about the behavior of ϵ_{in} and ϵ_{nse} for the LDLN data model (as analyzed in Section 4) with the derived performance guarantees for spectral clustering to give the range of θ, σ values where $\lambda_{K+1} - \lambda_K$ is the largest gap with high probability. We remind the reader that although the LDLN data model assumes uniform sampling, Theorem 5.13 and Corollary 5.14 can easily be extended to a more general sampling model.

Corollary 5.14 *Assume the notation of Theorem 5.12 holds. Then for n_{min} large enough, for any $\tau < \frac{C_1}{8} n_{min}^{-(d+1)} \wedge \frac{\epsilon_0}{5}$ and any*

$$C_1 n_{min}^{-\frac{1}{d+1}} \leq \theta \leq \left[C_2 \tilde{n}^{-\left(\frac{k_{nse}+1}{k_{nse}}\right)\frac{1}{D}} \right] \wedge \delta(4k_{nse})^{-1}, \tag{5.15}$$

we have that $\lambda_{K+1} - \lambda_K$ is the largest gap in the eigenvalues of L_{SYM} with high probability, provided that

$$C_3\theta \leq \sigma \leq \frac{C_4\delta}{f_1^{-1}(C_5(\zeta_n + k_{nse}\zeta_\theta)^{-1})} \tag{5.16}$$

where all C_i are constants independent of $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$.

Proof
By Corollary 4.4, for $n_{\min}$ large enough, ϵ_{in} satisfies $n_{\min} \lesssim \epsilon_{\text{in}}^{-d} \log(\epsilon_{\text{in}}^{-d}) \leq \epsilon_{\text{in}}^{-(d+1)}$, i.e. $\epsilon_{\text{in}} \leq C_1 n_{\min}^{-\frac{1}{d+1}}$ with high probability, as long as $\tau < \frac{C_1}{8} n_{\min}^{-\frac{1}{d+1}} \wedge \frac{\epsilon_0}{5}$. By Theorem 4.9, with high probability $\epsilon_{\text{nse}} \geq C_2 \tilde{n}^{-\frac{k_{\text{nse}}+1}{k_{\text{nse}}D}}$. We now apply Theorem 5.12. Note that for

an appropriate choice of constants, the assumptions of Corollary 5.14 guarantee $\epsilon_{\text{in}} \leq \theta < \epsilon_{\text{nse}} \wedge \delta/(4k_{\text{nse}})$ with high probability. There exist constants C_6, C_7, C_8 independent of $n_1, \ldots, n_K, \tilde{n}, \theta, \sigma$, such that inequality (5.13) is guaranteed as long as $f_\sigma(\epsilon_{\text{in}}) \geq C_6$, $\zeta_N f_\sigma(\delta/2) \leq C_7$, and $f_\sigma(\theta) \geq C_8$. Solving for σ, we obtain $\left(\frac{\epsilon_{\text{in}}}{f_1^{-1}(C_6)} \vee \frac{\theta}{f_1^{-1}(C_8)}\right) \leq \sigma \leq \frac{\delta}{2f_1^{-1}(C_7\zeta_N^{-1})}$. Combining with our bound for ϵ_{in} and recalling $\zeta_N \leq 2\zeta_n + 3k_{\text{nse}}\zeta_\theta$ with high probability by Theorem 5.12, this is implied by (5.15) and (5.16) and for an appropriate choice of relabelled constants.

■

This corollary illustrates that when $\tilde{n}$ is small relative to $n_{\min}^{\frac{D}{d+1}\left(\frac{k_{\text{nse}}}{k_{\text{nse}}+1}\right)}$, we obtain a large range of values of both the thresholding parameter θ and scale parameter σ where the maximal eigengap heuristic correctly identifies the number of clusters, i.e. LLPD spectral clustering is robust with respect to both of these parameters.

5.2.4. Parameter Selection

In terms of implementation, only the parameters k_{nse} and θ must be chosen, and then L_{SYM} can be computed for a range of σ values. Ideally k_{nse} is chosen to maximize the upper bound in (5.15), since $\tilde{n}^{-\left(\frac{k_{\text{nse}}+1}{k_{\text{nse}}}\right)\frac{1}{D}}$ is increasing in k_{nse} while δ/k_{nse} is decreasing in k_{nse}. Numerical experiments indicate robustness with respect to this parameter choice, and $k_{\text{nse}} = 20$ was used for all experiments reported in Section 7.
Regarding the thresholding parameter θ, ideally $\theta = \epsilon_{\text{in}}$, since this guarantees that all cluster points will be kept and the maximal number of noise points will be removed, i.e. we have perfectly denoised the data. However, ϵ_{in} is not known explicitly and must be estimated from the data. In practice the thresholding can be done by computing $\beta_{k_{\text{nse}}}(x, X)$ for all data points and clustering the data points into groups based on these values, or by choosing θ to correspond to the elbow in a graph of the sorted nearest neighbor distances as illustrated in Section 7. This latter approach for estimating θ is very similar to the proposal in Ester et al. (1996) for estimating the scale parameter in DBSCAN, although we use LLPD instead of Euclidean nearest neighbor distances. Note the thresholding procedure precedes the application of the spectral clustering kernel; it can be done once and then L_{SYM} computed for various σ values.

5.3. Comparison with Related Methods

We now theoretically compare the proposed method with related methods. LLPD spectral clustering combines spectral graph methods with a notion of distance that incorporates density, so we naturally focus on comparisons to spectral clustering and density-based methods.

5.3.1. Comparison with Theoretical Guarantees for Spectral Clustering

Our results on the eigengap and misclassification rate for LLPD spectral clustering are naturally comparable to existing results for Euclidean spectral clustering. Arias-Castro (2011); Arias-Castro et al. (2011, 2017) made a series of contributions to the theory of spectral clustering performance guarantees. We focus on the results in Arias-Castro (2011), where the author proves performance guarantees on spectral clustering by considering the same data model as the one proposed in the present article, and proceeds by analyzing the corresponding Euclidean weight matrix.

Our primary result, Theorem 5.12, is most comparable to Proposition 4 in Arias-Castro (2011), which estimates $\lambda_K \leq Cn^{-3}, \lambda_{K+1} \geq Cn^{-2}$ for some constant C. From the theoretical point of view, this does not necessarily mean $\lambda_{K+1} - \lambda_K \geq \lambda_{l+1} - \lambda_l, l \neq K$. Compared to that result, Theorem 5.12 enjoys a much stronger conclusion for guaranteeing the significance of the eigengap. From a practical point of view, it is noted in Arias-Castro (2011) that Proposition 4 is not a useful condition for actual data. Our method is shown to correctly estimate the eigengap in both high-dimensional and noisy settings, where the eigengap with Euclidean distance is uninformative; see Section 7.

Theorem 5.12 also provides conditions guaranteeing LLPD spectral clustering achieves perfect labeling accuracy. The proposed conditions are sufficient to guarantee the representation of the data in the coordinates of the principal eigenvectors of the LLPD Laplacian is a perfect representation. An alternative approach to ensuring spectral clustering accuracy is presented in Schiebinger et al. (2015), which develops the notion of the *orthogonal cone property (OCP)*. The OCP characterizes low-dimensional embeddings that represent points in distinct clusters in nearly orthogonal directions. Such embeddings are then easily clustered with, for example, K-means. The two crucial parameters in the approach of Schiebinger et al. (2015) measure how well-separated each cluster is from the others, and how internally well-connected each distinct cluster is. The results of Section 4 prove that under the LDLN data model, points in the same cluster are very close together in LLPD, while points in distinct clusters are far apart in LLPD. In this sense, the results of Schiebinger et al. (2015) suggest that LLPD spectral clustering ought to perform well in the LDLN regime. Indeed, the LLPD is nearly invariant to cluster geometry, unlike Euclidean distance. As clusters become more anisotropic, the within-cluster distances stay almost the same when using LLPD, but increase when using Euclidean distances. In particular, the framework of Schiebinger et al. (2015) implies that performance of LLPD spectral clustering will degrade slowly as clusters are stretched, while performance of Euclidean spectral clustering will degrade rapidly. We remark that the OCP framework has been generalized to continuum setting for the analysis of mixture models (Garcia Trillos et al., 2019b).

This observation may also be formulated in terms of the spectrum of the Laplacian. For the (continuous) Laplacian Δ on a domain $\mathcal{M} \subset \mathbb{R}^{\mathbb{D}}$, Szegö (1954); Weinberger

(1956) prove that among unit-volume domains, the second Neumann eigenvalue $\lambda_2(\Delta)$ is minimal when the underlying $\mathcal{M}$ is the ball. One can show that as the ball becomes more elliptical in an area-preserving way, the second eigenvalue of the Laplacian decreases. Passing to the discrete setting (Garcia Trillos et al., 2019a), this implies that as clusters become more elongated and less compact, the second eigenvalues on the individual clusters (ignoring between-cluster interactions, as proposed in Maggioni and Murphy (2019)) decreases. Spectral clustering performance results are highly dependent on these second eigenvalues of the Laplacian when localized on individual clusters (Arias-Castro, 2011; Schiebinger et al., 2015), and in particular performance guarantees weaken dramatically as they become closer to 0. In this sense, Euclidean spectral clustering is not robust to elongating clusters. LLPD spectral clustering, however, uses a distance that is nearly invariant to this kind of geometric distortion, so that the second eigenvalues of the LLPD Laplacian localized on distinct clusters stay far from 0 even in the case of highly elongated clusters. In this sense, LLPD spectral clustering is more robust than Euclidean spectral clustering for elongated clusters.

The same phenomenon is observed from the perspective of graph-cuts. It is well-known (Shi and Malik, 2000) that spectral clustering on the graph with weight matrix W approximates the minimization of the multiway normalized cut functional

$$\mathrm{Ncut}(C_1, C_2, \ldots, C_K) = \underset{(C_1,C_2,\ldots,C_K)}{\arg\min} \sum_{k=1}^{K} \frac{W(C_k, X \setminus C_k)}{\mathrm{vol}(C_k)},$$

where

$$W(C_k, X \setminus C_k) = \sum_{x_i \in C_k} \sum_{x_j \notin C_k} W_{ij}, \ \mathrm{vol}(C_k) = \sum_{x_i \in C_k} \sum_{x_j \in X} W_{ij}.$$

As clusters become more elongated, cluster-splitting cuts measured in Euclidean distance become cheaper and the optimal graph cut shifts from one that separates the clusters to one that splits them. On the other hand, when using the LLPD a cluster-splitting cut only becomes marginally cheaper as the cluster stretches, so that the optimal graph cut preserves the clusters rather than splits them.

A somewhat different approach to analyzing the performance of spectral clustering is developed in Balakrishnan et al. (2011), which proposes as a model for spectral clustering *noisy hierarchical block matrices (noisy HBM)* of the form $W = A + R$ for ideal A and a noisy perturbation R. The ideal A is characterized by on and off-diagonal block values that are constrained to fall in certain ranges, which models concentration of within-cluster and between-cluster distances. The noisy perturbation R is a random, mean 0 matrix having rows with independent, subgaussian entries, characterized by a variance parameter σ_{noise}. The authors propose a modified spectral clustering algorithm (using the K-centering algorithm) which, under certain assumptions on the idealized within-cluster and between-cluster distances, learns all clusters above a certain size for a range of σ_{noise} levels. The proposed theoretical analysis of

LLPD in Section 4 shows that under the LDLN data model and for n sufficiently large, the Laplacian matrix (and weight matrix) is nearly block constant with large separation between clusters. Our Theorems 4.3, 4.7, 4.9, 4.11 may be interpreted as showing that the (LLPD-denoised) weight matrix associated to data generated from the LDLN model may fit the idealized model suggested by Balakrishnan et al. (2011). In particular, when $R = 0$, the results for this noisy HBM are comparable with, for example, Theorem 5.12. However, the proposed method does not consider hierarchical clustering, but instead shows localization properties of the eigenvectors of L_{SYM}. In particular, the proposed method is shown to correctly learn the number of clusters K through the eigengap, assuming the LDLN model, which is not considered in Balakrishnan et al. (2011).

5.3.2. Comparison with Density-Based Methods

The *DBSCAN* algorithm labels points as cluster or noise based on density to nearby points, then creates clusters as maximally connected regions of cluster points. While popular, DBSCAN is extremely sensitive to the selection of parameters to distinguish between cluster and noise points, and often performs poorly in practice. The DBSCAN parameter for noise detection is comparable to the denoising parameter θ used in LLPD spectral clustering, though LLPD spectral clustering is quite robust to θ in theory and practice. Moreover, DBSCAN does not enjoy the robust theoretical guarantees provided in this article for LLPD spectral clustering on the LDLN data model, although some results are known for techniques related to DBSCAN (Rinaldo and Wasserman, 2010; Sriperumbudur and Steinwart, 2012).

In order to address the shortcomings of DBSCAN, the *fast search and find of density peaks clustering (FSFDPC)* algorithm was proposed (Rodriguez and Laio, 2014). This method first learns modes in the data as points of high density far (in Euclidean distance) from other points of high density, then associates each point to a nearby mode in an iterative fashion. While more robust than DBSCAN, FSFDPC cannot learn clusters that are highly nonlinear or elongated. Maggioni and Murphy (2019) proposed a modification to FSFDPC called *learning by unsupervised nonlinear diffusion (LUND)* which uses diffusion distances instead of Euclidean distances to learn the modes and make label assignments, allowing for the clustering of a wide range of data, both theoretically and empirically. While LUND enjoys theoretical guarantees and strong empirical performance (Murphy and Maggioni, 2018, 2019), it does not perform as robustly on the proposed LDLN model for estimation of K or for labeling accuracy. In particular, the eigenvalues of the diffusion process which underlies diffusion distances (and thus LUND) do not exhibit the same sharp cutoff phenomenon as those of the LLPD Laplacian under the LDLN data model.

Cluster trees are a related density-based method that produces a multiscale hierarchy of clusterings, in a manner related to single linkage clustering. Indeed, for data sampled from some density μ on a Euclidean domain X, a cluster tree is the family of clusterings $\mathcal{T} = \{C_r\}_{r=0}^{\infty}$, where C_r are the connected regions of the set $\{x \mid \mu(x) \geq r\}$.

Chaudhuri and Dasgupta (2010) studied cluster trees where $X \subset \mathbb{R}^D$ is a subset of Euclidean space, showing that if sufficiently many samples are drawn from μ, depending on D, then the clusters in the empirical cluster tree closely match the population level clusters given by thresholding μ. Balakrishnan et al. (2013) generalized this work to the case when the underlying distribution is supported on an intrinsically d-dimensional set, showing that the performance guarantees depend only on d, not D.

The cluster tree itself is related to the LLPD as $\rho_{\ell\ell}(x_i, x_j)$ is equal to the smallest r such that x_i, x_j are in the same connected component of a complete Euclidean distance graph with edges of length $\geq r$ removed. Furthermore, the model of Balakrishnan et al. (2013) assumes the support of the density is near a low-dimensional manifold, which is comparable to the LDLN model of assuming the clusters are τ-close to low-dimensional elements of $\mathcal{S}_d(\kappa, \epsilon_0)$. The notion of separation in Chaudhuri and Dasgupta (2010); Balakrishnan et al. (2013) is also comparable to the notion of between-cluster separation in the present manuscript. On the other hand, the proposed method considers a more narrow data model (LDLN versus arbitrary density μ), and proves strong results for LLPD spectral clustering including inference of K, labeling accuracy, and robustness to noise and choice of parameters. The proof techniques are also rather different for the two methods, as the LDLN data model provides simplifying assumptions which do not hold for a general probability density function. Indeed, the approach presented in this article achieves precise finite sample estimates for the LLPD using percolation theory and the chain arguments of Theorems 4.7 and 4.9, in contrast to the general consistency results on cluster trees derived from a wide class of probability distributions (Chaudhuri and Dasgupta, 2010; Balakrishnan et al., 2013).

6. Numerical Implementations of LLPD

We first demonstrate how LLPD can be accurately approximated from a sequence of m multiscale graphs in Section 6.1. Section 6.2 discusses how this approach can be used for fast LLPD-nearest neighbor queries. When the data has low intrinsic dimension d and $m = O(1)$, the LLPD nearest neighbors of all points can be computed in $O(DC^d n \log(n))$ for a constant C independent of n, d, D. Although there are theoretical methods for obtaining the exact LLPD (Demaine et al., 2009, 2014; McKenzie and Damelin, 2019) with the same computational complexity, they are not practical for real data. The method proposed here is an efficient alternative, whose accuracy can be controlled by choice of m.

The proposed LLPD approximation procedure can also be leveraged to define a fast eigensolver for LLPD spectral clustering, which is discussed in Section 6.3. The ultrametric structure of the weight matrix allows for fast matrix-vector multiplication, so that $L_{\text{SYM}}x$ (and thus the eigenvectors of L_{SYM}) can be computed with complexity $O(mn)$ for a dense L_{SYM} defined using LLPD. When the number of scales $m \ll n$,

this is a vast improvement over the typical $O(n^2)$ needed for a dense Laplacian. Once again when the data has low intrinsic dimension and $K, m = O(1)$, LLPD spectral clustering can be implemented in $O(DC^d n \log(n))$.
Connections with single linkage clustering are discussed in Section 6.4, as the LLPD approximation procedure gives a pruned single linkage dendrogram. Matlab code implementing both the fast LLPD nearest neighbor searches and LLPD spectral clustering is publicly available at `https://bitbucket.org/annavlittle/llpd_code/branch/v2.1`. The software auto-selects both the number of clusters K and kernel scale σ.

6.1. Approximate LLPD from a Sequence of Thresholded Graphs

The notion of *nearest neighbor graph* is important for the formal analysis which follows.

Definition 6.1 *Let (X, ρ) be a metric space. The* (symmetric) k-nearest neighbors graph on X with respect to ρ *is the graph with nodes X and an edge between x_i, x_j of weight $\rho(x_i, x_j)$ if x_j is among the k points with smallest ρ-distance to x_i or if x_i is among the k points with smallest ρ-distance to x_j.*

Let $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ and G be some graph defined on X. Let $D_G^{\ell\ell}$ denote the matrix of exact LLPDs obtained from all paths in the graph G; note this is a generalization of Definition 2.1, which considers G to be a complete graph. We define an approximation $\hat{D}_G^{\ell\ell}$ of $D_G^{\ell\ell}$ based on a sequence of thresholded graphs. Let E-nearest neighbor denote a nearest neighbor in the Euclidean metric, and LLPD-nearest neighbor denote a nearest neighbor in the LLPD metric.

Definition 6.2 *Let X be given and let k_{Euc} be a positive integer. Let $G(\infty)$ denote the complete graph on X, with edge weights defined by Euclidean distance, and $G_{k_{Euc}}(\infty)$ the k_{Euc} E-nearest neighbors graph on X as in Definition 6.1. For a threshold $t > 0$, let $G(t), G_{k_{Euc}}(t)$ be the graphs obtained from $G(\infty), G_{k_{Euc}}(\infty)$, respectively, by discarding all edges of magnitude greater than t.*

We approximate $\rho_{\ell\ell}(x_i, x_j) = (D^{\ell\ell}_{G(\infty)})_{ij}$ as follows. Given a sequence of thresholds $t_1 < t_2 < \cdots < t_m$, compute $G_{k_{\text{Euc}}}(\infty)$ and $\{G_{k_{\text{Euc}}}(t_s)\}_{s=1}^m$. Then this sequence of graphs may be used to approximate $\rho_{\ell\ell}$ by finding the smallest threshold t_s for which two path-connected components C_1, C_2 merge: for $x \in C_1, y \in C_2$, we have $\rho_{\ell\ell}(x, y) \approx t_s$. We thus approximate $\rho_{\ell\ell}(x_i, x_j)$ by $(\hat{D}^{\ell\ell}_{G_{ij}})_{ij} = \inf_s\{t_s \mid x_i \sim x_j \text{ in } G_{ij}(t_s)\}$, where $x_i \sim x_j$ denotes that the two points are path connected. We let $\mathbb{D} = \{\mathbf{C}_{t_s}\}_{s=1}^m$ denote the dendrogram which arises from this procedure. More specifically, $\mathbf{C}_{t_s} = \{C^1_{t_s}, \ldots C^{\nu_s}_{t_s}\}$ are the connected components of $G_{k_{\text{Euc}}}(t_s)$, so that ν_s is the number of connected components at scale t_s.
The error incurred in this estimation of $\rho_{\ell\ell}$ is a result of two approximations: (a) approximating LLPD in $G(\infty)$ by LLPD in $G_{k_{\text{Euc}}}(\infty)$; (b) approximating LLPD in

$G_{k_{\text{Euc}}}(\infty)$ from the sequence of thresholded graphs $\{G_{k_{\text{Euc}}}(t_s)\}_{s=1}^m$. Since the optimal paths which determine $\rho_{\ell\ell}$ are always paths in a *minimal spanning tree* (MST) of $G(\infty)$ (Hu, 1961), we do not incur any error from (a) whenever an MST of $G(\infty)$ is a subgraph of $G_{k_{\text{Euc}}}(\infty)$. González-Barrios and Quiroz (2003) show that when sampling a compact, connected manifold with sufficiently smooth boundary, the MST is a subgraph of $G_{k_{\text{Euc}}}(\infty)$ with high probability for $k_{\text{Euc}} = O(\log(n))$. Thus for $k_{\text{Euc}} = O(\log(n))$, we do not incur any error from (a) in within-cluster LLPD, as the nearest neighbor graph for each cluster will contain the MST for the given cluster. When the clusters are well-separated, we generally will incur some error from (a) in the between-cluster LLPD, but this is precisely the regime where a high amount of error can be tolerated. The following proposition controls the error incurred by (b).

Proposition 6.3 *Let G be a graph on X and $x_i, x_j \in X$ such that $(\hat{D}_G^{\ell\ell})_{ij} = t_s$. Then $(D_G^{\ell\ell})_{ij} \leq (\hat{D}_G^{\ell\ell})_{ij} \leq t_s/(t_{s-1})(D_G^{\ell\ell})_{ij}$.*

Proof There is a path in G connecting x_i, x_j with every leg of length $\leq t_s$, since $(\hat{D}_G^{\ell\ell})_{ij} = t_s$. Hence, $(D_G^{\ell\ell})_{ij} \leq t_s = (\hat{D}_G^{\ell\ell})_{ij}$. Moreover, $t_{s-1} \leq (D_G^{\ell\ell})_{ij}$, since no path in G with all legs $\leq t_{s-1}$ connects x_i, x_j. It follows that $t_s \leq \frac{t_s}{t_{s-1}}(D_G^{\ell\ell})_{ij}$, hence $(\hat{D}_G^{\ell\ell})_{ij} \leq \frac{t_s}{t_{s-1}}(D_G^{\ell\ell})_{ij}$. ■

Thus if $\{t_s\}_{s=1}^m$ grows exponentially at rate $(1+\epsilon)$, the ratio $\frac{t_s}{t_{s-1}}$ is bounded uniformly by $(1+\epsilon)$, and a uniform bound on the relative error is: $(D_G^{\ell\ell})_{ij} \leq (\hat{D}_G^{\ell\ell})_{ij} \leq (1+\epsilon)(D_G^{\ell\ell})_{ij}$. Alternatively, one can choose the $\{t_s\}_{s=1}^m$ to be fixed percentiles in the distribution of edge magnitudes of G.

Algorithm 2 summarizes the multiscale construction which is used to approximate LLPD. At each scale t_s, the connected components of $G_{k_{\text{Euc}}}(t_s)$ are computed; the component identities are then stored in an $n \times m$ matrix, and the rows of the matrix are then sorted to obtain a hierarchical clustering structure. This sorted matrix of connected components (denoted CC_{sorted} in Algorithm 2) can be used to quickly obtain the LLPD-nearest neighbors of each point, as discussed in Section 6.2. Note that if $G_{k_{\text{Euc}}}(\infty)$ is disconnected, one can add additional edges to obtain a connected graph.

6.2. LLPD Nearest Neighbor Algorithm

We next describe how to perform fast LLPD nearest neighbor queries using the multiscale graphs introduced in Section 6.1. Algorithm 3 gives pseudocode for the approximation of each point's $k_{\ell\ell}$ LLPD-nearest neighbors, with the approximation based on the multiscale construction in Algorithm 2.

Figure 6a illustrates how Algorithm 3 works on a data set consisting of 11 points and 4 scales. Letting π denote the ordering of the points in CC_{sorted} as produced by Algorithm 2, CC_{sorted} is queried to find $x_{\pi(6)}$'s 8 LLPD-nearest neighbors (nearest

Algorithm 2 Approximate LLPD

Input: $X, \{t_s\}_{s=1}^m, k_{\text{Euc}}$
Output: $\mathbb{D} = \{\mathbf{C}_{t_s}\}_{s=1}^m$, point order $\pi(i)$, CC_{sorted}

Form a k_{Euc} E-nearest neighbors graph on X; call it $G_{k_{\text{Euc}}}(\infty)$.
Sort the edges of $G_{k_{\text{Euc}}}(\infty)$ into the bins defined by the thresholds $\{t_s\}_{s=1}^m$.
for $s = 1 : m$ **do**
    Form $G_{k_{\text{Euc}}}(t_s)$ and compute its connected components $\mathbf{C}_{t_s} = \{C_{t_s}^1, \dots C_{t_s}^{\nu_s}\}$.
end for
Create an $n \times m$ matrix CC storing each point's connected component at each scale.
Sort the rows of CC based on $\mathbf{C}_{t_m}$ (the last column).
for $s = m : 2$ **do**
    for $i = 1 : \nu_s$ **do**
        Sort the rows of CC corresponding to $C_{t_s}^i$ according to $\mathbf{C}_{t_{s-1}}$.
    end for
end for
Let CC_{sorted} denote the $n \times m$ matrix containing the final sorted version of CC.
Let $\pi(i)$ denote the point order encoded by CC_{sorted}.

neighbors are shown in bold). Starting in the first column of CC_{sorted} which corresponds to the finest scale ($s = 1$), points in the same connected component as the base point are added to the nearest neighbor set, and the LLPD to these points is recorded as t_1. Assuming the nearest neighbors set does not yet contain $k_{\ell\ell}$ points, one then adds to it any points not yet in the nearest neighbor set which are in the same connected component as the base point at the second finest scale, and records the LLPD to these neighbors as t_2 (see the second column of Figure 6a which illustrates $s = 2$ in the pseudocode). One continues in this manner until $k_{\ell\ell}$ neighbors are found.

Remark 6.4 *For a fixed x, there might be many points of equal LLPD to x. This is in contrast to the case for Euclidean distance, where such phenomena typically occur only for highly structured data, for example, for data consisting of points lying on a sphere and x the center of the sphere. In the case that $k_{\ell\ell}$ LLPD nearest neighbors for x are sought and there are more than $k_{\ell\ell}$ points at the same LLPD from x, Algorithm 3 returns a sample of these LLPD-equidistant points in $O(m+k_{\ell\ell})$ by simply returning the first $k_{\ell\ell}$ neighbors encountered in a fixed ordering of the data; a random sample could be returned for an additional cost.*

Figure 6b shows a plot of the empirical runtime of the proposed algorithm against number of points in log scale, suggesting nearly linear runtime. This is confirmed theoretically as follows.

Theorem 6.5 *Algorithm 3 has complexity $O(n(k_{Euc}C_{NN} + m(k_{Euc} \vee \log(n)) + k_{\ell\ell}))$.*

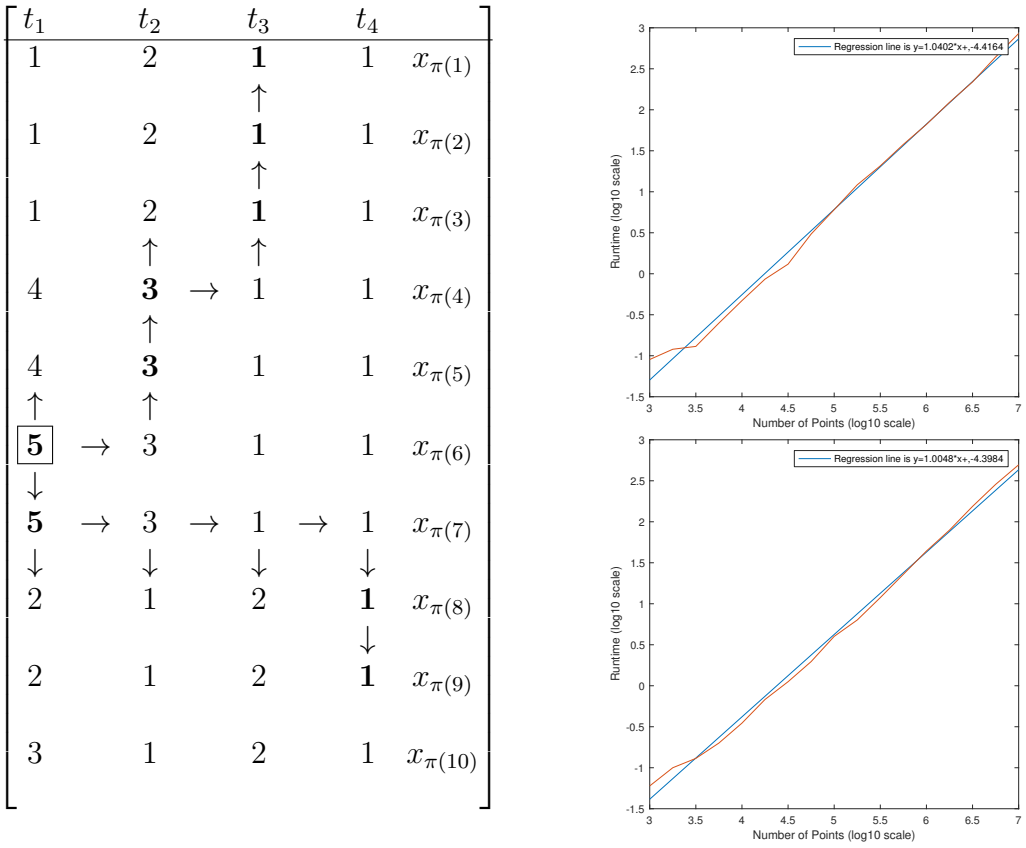
Algorithm 3 Fast LLPD nearest neighbor queries

Input: $X, \{t_s\}_{s=1}^m, k_{\text{Euc}}, k_{\ell\ell}$
Output: $n \times n$ sparse matrix $\hat{D}^{\ell\ell}_{G_{k_{\text{Euc}}}}$ giving approximate $k_{\ell\ell}$ LLPD-nearest neighbors

Use Algorithm 2 to obtain $\pi(i)$ and CC_{sorted}.
for $i = 1 : n$ **do**
  $\hat{D}^{\ell\ell}_{\pi(i),\pi(i)} = t_1$
  $NN = 1$ % Number of nearest neighbors found
  $i_{\text{up}} = 1$
  $i_{\text{down}} = 1$
  for $s = 1 : m$ **do**
    while $CC_{\text{sorted}}(i_{\text{up}}, s) = CC_{\text{sorted}}(i_{\text{up}} - 1, s)$ and $NN < k_{\ell\ell}$ and $i_{\text{up}} > 1$ **do**
      $i_{\text{up}} = i_{\text{up}} - 1$
      $\hat{D}^{\ell\ell}_{\pi(i),\pi(i_{\text{up}})} = t_s$
      $NN = NN + 1$
    end while
    while $CC_{\text{sorted}}(i_{\text{down}}, s) = CC_{\text{sorted}}(i_{\text{down}} + 1, s)$ and $NN < k_{\ell\ell}$ and $i_{\text{down}} < n$ **do**
      $i_{\text{down}} = i_{\text{down}} + 1$
      $\hat{D}^{\ell\ell}_{\pi(i),\pi(i_{\text{down}})} = t_s$
      $NN = NN + 1$
    end while
  end for
end for
return $\hat{D}^{\ell\ell}_{G_{k_{\text{Euc}}}}$

Proof
The major steps of Algorithm 3 (which includes running Algorithm 2) are:

- Generating the k_{Euc} E-nearest neighbors graph $G_{k_{\text{Euc}}}(\infty)$: $O(k_{\text{Euc}} n C_{\text{NN}})$, where C_{NN} is the cost of an E-nearest neighbor query. For high-dimensional data $C_{\text{NN}} = O(nD)$. When the data has low intrinsic dimension $d < D$ cover trees (Beygelzimer et al., 2006) allows $C_{\text{NN}} = O(DC^d \log(n))$, after a pre-processing step with cost $O(C^d Dn \log(n))$.
- Binning the edges of $G_{k_{\text{Euc}}}(\infty)$: $O(k_{\text{Euc}} n(m \wedge \log(k_{\text{Euc}} n)))$. Binning without sorting is $O(k_{\text{Euc}} nm)$; if the edges are sorted first, the cost is $O(k_{\text{Euc}} n \log(k_{\text{Euc}} n))$.
- Forming $G_{k_{\text{Euc}}}(t_s)$, for $s = 1, \ldots, m$, and computing its connected components: $O(k_{\text{Euc}} mn)$.
- Sorting the connected components matrix to create CC_{sorted}: $O(mn \log(n))$.
- Finding each point's $k_{\ell\ell}$ LLPD-nearest neighbors by querying CC_{sorted}: $O(n(m + k_{\ell\ell}))$.



(a) Illustration of Algorithm 3

(b) Complexity plots for Algorithm 3

Figure 6: Algorithm 3 is demonstrated on a simple example in (a). The figure illustrates how CC_{sorted} is queried to return $x_{\pi(6)}$'s 8 LLPD-nearest neighbors. Nearest neighbors are shown in bold, and $\hat{\rho}_{\ell\ell}(x_{\pi(6)}, x_{\pi(7)}) = t_1$, $\hat{\rho}_{\ell\ell}(x_{\pi(6)}, x_{\pi(5)}) = t_2$, etc. Note each upward or downward arrow represents a comparison which checks whether two points are in the same connected component at the given scale. In (b), the runtime of Algorithm 3 on uniform data in $[0,1]^2$ is plotted against number of points in log scale. The slope of the line is approximately 1, indicating that the algorithm is essentially quasilinear in the number of points. Here, $k_{\text{Euc}} = 20, k_{\ell\ell} = 10, D = 2$, and the thresholds $\{t_s\}_{s=1}^m$ correspond to fixed percentiles of edge magnitudes in $G_{k_{\text{Euc}}}(\infty)$. The top plot has $m = 10$ and the bottom plot $m = 100$.

Observe that $O(C_{\text{NN}})$ always dominates $O(m \wedge \log(k_{\text{Euc}} n))$. Hence, the overall complexity is $O(n(k_{\text{Euc}} C_{\text{NN}} + m(k_{\text{Euc}} \vee \log(n)) + k_{\ell\ell}))$. ■

Corollary 6.6 *If $k_{Euc}, k_{\ell\ell}, m = O(1)$ with respect to n and the data has low intrinsic dimension so that $C_{NN} = O(DC^d \log(n))$, Algorithm 3 has complexity $O(DC^d n \log(n))$.*

If $k_{\ell\ell} = O(n)$ or the data has high intrinsic dimension, the complexity is $O(n^2)$. Hence, d, m, k_{Euc}, and $k_{\ell\ell}$ are all important parameters affecting the computational complexity.

Remark 6.7 *One can also incorporate a minimal spanning tree (MST) into the construction, i.e. replace $G_{k_{Euc}}(\infty)$ with its MST. This will reduce the number of edges which must be binned to give a total computational complexity of $O(n(k_{Euc}C_{NN} + m\log(n) + k_{\ell\ell}))$. Computing the LLPD with and without the MST has the same complexity when $k_{Euc} \leq O(\log(n))$, so for simplicity we do not incorporate MSTs in our implementation.*

6.3. A Fast Eigensolver for LLPD Laplacian

In this section we describe an algorithm for computing the eigenvectors of a dense Laplacian defined using approximate LLPD with complexity $O(mn)$. The ultrametric property of the LLPD makes L_{SYM} highly compressible, which can be exploited for fast eigenvector computations. Assume LLPD is approximated using m scales $\{t_s\}_{s=1}^m$ and the corresponding thresholded graphs $G_{k_{\text{Euc}}}(t_s)$ as described in Algorithm 2. Let $n_i = |C_{t_1}^i|$ for $1 \leq i \leq \nu_1$ denote the cardinalities of the connected components of $G_{k_{\text{Euc}}}(t_1)$, and $V = \sum_{k=1}^m \nu_k$ the total number of connected components across all scales.

In order to develop a fast algorithm for computing the eigenvectors of the LLPD Laplacian $L_{\text{SYM}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, it suffices to describe a fast method for computing the matrix-vector multiplication $x \mapsto L_{\text{SYM}}x$, where L_{SYM} is defined using $W_{ij} = e^{-\rho_{\ell\ell}(x_i,x_j)^2/\sigma^2}$ (Trefethen and Bau, 1997). Assume without loss of generality that we order the entries of both x and W according to the point order π defined in Algorithm 2. Note because L_{SYM} is block constant with ν_1^2 blocks, any eigenvector will also be block constant with ν_1 blocks, and it suffices to develop a fast multiplier for $x \mapsto L_{\text{SYM}}x$ when $x \in \mathbb{R}^n$ has the form: $x = [x_1 1_{n_1}\ x_2 1_{n_2} \dots x_{\nu_1} 1_{\nu_1}]$ where $1_{n_i} \in \mathbb{R}^{n_i}$ is the all one's vector. Assuming LLPD's have been precomputed using Algorithm 2, Algorithm 4 gives pseudocode for computing Wx with complexity $O(mn)$. Since $L_{\text{SYM}}x = x - D^{-1/2}WD^{-1/2}x$, $W(D^{-1/2}x)$ is computable via Algorithm 4, and $D^{-1/2}$ is diagonal, a straight forward generalization of Algorithm 4 gives $L_{\text{SYM}}x$ in $O(mn)$. Since the matrix-vector multiplication has reduced complexity $O(mn)$, the decomposition of the principal K eigenvectors can be done with complexity $O(K^2mn)$ (Trefethen and Bau, 1997), which in the practical case $K, m = O(1)$, is essentially linear in n. Thus the total complexity of implementing LLPD spectral clustering including the LLPD approximation discussed in Section 6.1 becomes $O(n(k_{\text{Euc}}C_{\text{NN}} + m(k_{\text{Euc}} \vee \log(n) \vee K^2)))$. We defer timing studies and theoretical analysis of the fast eigensolver algorithm to a subsequent article, in the interest of space. We remark that the strategy proposed for computing the eigenvectors of the LLPD Laplacian could in principal be used for Laplacians derived from other distances. However, without the compressible ultrametric structure, the approximation using only $m \ll n$ scales is likely to be poor, leading to inaccurate approximate eigenvectors.

Algorithm 4 Fast LLPD matrix-vector multiplication

Input: $\{t_s\}_{s=1}^m$, $\mathbb{D} = \{\mathbf{C}_{t_s}\}_{s=1}^m$, $f_\sigma(t)$, x
Output: Wx

Enumerate all connected components at all scales: $\mathbf{C} = [C_{t_1}^1 \dots C_{t_1}^{\nu_1} \dots C_{t_m}^1 \dots C_{t_m}^{\nu_m}]$.
Let $\mathcal{V}$ be the collection of V nodes corresponding to the elements of $\mathbf{C}$.
For $i = 1, \dots, V$, let $\mathcal{C}(i)$ be the set of direct children of node i in dendrogram $\mathbb{D}$.
For $i = 1, \dots, V$, let $\mathcal{P}(i)$ be the direct parent of node i in dendrogram $\mathbb{D}$.

for $i = 1 : \nu_1$ **do**
  $\Sigma(i) = n_i x_i$
end for
for $i = (\nu_1 + 1) : V$ **do**
  $\Sigma(i) = \sum_{j \in \mathcal{C}(i)} \Sigma(j)$
end for

for $i = 1 : \nu_1$ **do**
  $\alpha_i(1) = i$
  for $j = 2 : m$ **do**
    $\alpha_i(j) = \mathcal{P}(\alpha_i(j-1))$
  end for
end for

Let $K = [f_\sigma(t_1)\ f_\sigma(t_2) \cdots f_\sigma(t_m)]$ be a vector of kernel evaluations at each scale.

for $i = 1 : \nu_1$ **do**
  $\xi_i(j) = \Sigma(\alpha_i(j))$
  $d\xi_i(1) = \xi_i(1)$
  for $j = 2 : m$ **do**
    $d\xi_i(j) = \xi_i(j+1) - \xi_i(j)$
  end for
  Let I_i be the index set corresponding to $C_{t_1}^i$.
  $(Wx)_{I_i} = \sum_{s=1}^m d\xi_i(s) K(s)$
end for

6.4. LLPD as Approximate Single Linkage Clustering

The algorithmic implementation giving $\hat{D}^{\ell\ell}_{G_{k_{\text{Euc}}}}$ approximates the true LLPD $\rho_{\ell\ell}$ by merging path connected components at various scales. In this sense, our approach is reminiscent of single linkage clustering (Hastie et al., 2009). Indeed, the connected component structure defined in Algorithm 2 can be viewed as an approximate single linkage dendrogram.

Single linkage clustering generates, from $X = \{x_i\}_{i=1}^n$, a dendrogram $\mathbb{D}_{\text{SL}} = \{\mathbf{C}_k\}_{k=0}^{n-1}$, where $\mathbf{C}_k : \{1, 2, \dots, n\} \to \{C_k^1, C_k^2, \dots, C_k^{n-k}\}$ assigns x_i to its cluster at level k of the dendrogram ($\mathbf{C}_0$ assigns each point to a singleton cluster). Let d_k be the Euclidean distance between the clusters merged at level k: $d_k = \min_{i \neq j} \min_{x \in C_i^{k-1}, y \in C_j^{k-1}} \|x - y\|_2$. Note that $\{d_k\}_{k=1}^{n-1}$ is non-decreasing, and when strictly increasing, the clusters pro-

duced by single linkage clustering at the k^{th} level are the path connected components of $G(d_k)$. In the more general case, the path connected components of $G(d_k)$ may correspond to multiple levels of the single linkage hierarchy. Let $\{t_s\}_{s=1}^m$ be the thresholds used in Algorithm 2, and assume that $G_{k_{\text{Euc}}}(\infty)$ contains an MST of $G(\infty)$ as a subgraph. Let $\mathbb{D} = \{\mathbf{C}_{t_s}\}_{s=1}^m$ be the path-connected components with edges $\leq t_s$. $\mathbb{D}$ is a compressed dendrogram, obtained from the full dendrogram $\mathbb{D}_{\text{SL}}$ by pruning at certain levels. Let $\tau_s = \inf\{k \mid d_k \geq t_s,\ d_k < d_{k+1}\}$, and define the *pruned dendrogram* as $P(\mathbb{D}_{\text{SL}}) = \{\mathbf{C}_{\tau_s}\}_{s=1}^m$. In this case, the dendrogram obtained from the approximate LLPD is a pruning of an exact single linkage dendrogram. We omit the proof of the following in the interest of space.

Proposition 6.8 *If $G_{k_{Euc}}(\infty)$ contains an MST of $G(\infty)$ as a subgraph, $P(\mathbb{D}_{SL}) = \mathbb{D}$.*

Note that the approximate LLPD algorithm also offers an inexpensive approximation of single linkage clustering. A naive implementation of single linkage clustering is $O(n^3)$, while the SLINK algorithm (Sibson, 1973) improves this to $O(n^2)$. Thus to generate $\mathbb{D}$ by first performing exact single linkage clustering, then pruning, is $O(n^2)$, whereas to approximate $\mathbb{D}$ directly via approximate LLPD is $O(n\log(n))$; see Figure 7.

Figure 7: The cost of constructing the full single linkage dendrogram with SLINK is $O(n^2)$, and the cost of pruning is $O(mn)$, where m is the number of pruning cuts, so that acquiring $\mathbb{D}$ in this manner has overall complexity $O(n^2)$. The proposed method, in contrast, computes $\mathbb{D}$ with complexity $O(n\log(n))$.

7. Numerical Experiments

In this section we illustrate LLPD spectral clustering on four synthetic data sets and five real data sets. LLPD was approximated using Algorithm 2, and data sets were denoised by removing all points whose $k_{\text{nse}}^{\text{th}}$ nearest neighbor LLPD exceeded θ. Algorithm 4 was then used to compute approximate eigenpairs of the LLPD Laplacian for a range of σ values. The parameters $\hat{K}, \hat{\sigma}$ were then estimated from the multiscale spectral decompositions via

$$\hat{K} = \arg\max_i \max_\sigma (\lambda_{i+1}(\sigma) - \lambda_i(\sigma)) \quad , \quad \hat{\sigma} = \arg\max_\sigma \left(\lambda_{\hat{K}+1}(\sigma) - \lambda_{\hat{K}}(\sigma)\right) , \qquad (7.1)$$

and a final clustering was obtained by running K-means on the spectral embedding defined by the principal K eigenvectors of $L_{\text{SYM}}(\hat{\sigma})$. For each data set, we investigate (1) whether $\hat{K} = K$ and (2) the labeling accuracy of LLPD spectral clustering given

K. We compare the results of (1) and (2) with those obtained from Euclidean spectral clustering, where $\hat{K}, \hat{\sigma}$ are estimated using an identical procedure, and also compare the results of (2) with the labeling accuracy obtained by applying K-means directly. To make results as comparable as possible, Euclidean spectral clustering and K-means were run on the LLPD denoised data sets. All results are reported in Table 2.
Labeling accuracy was evaluated using three statistics: *overall accuracy (OA)*, *average accuracy (AA)*, and *Cohen's* κ. OA is the metric used in the theoretical analysis, namely the proportion of correctly labeled points after clusters are aligned, as defined by the agreement function (3.9). AA computes the overall accuracy on each cluster separately, then averages the results, in order to give small clusters equal weight to large ones. Cohen's κ measures agreement between two labelings, corrected for random agreement (Banerjee et al., 1999). Note that AA and κ are computed using the alignment that is optimal for OA. We note that accuracy is computed only on the points with ground truth labels, and in particular, any noise points remaining after denoising are ignored in the accuracy computations. For the synthetic data, where it is known which points are noise and which are from the clusters, one can assign labels to noise points according to Euclidean distance to the nearest cluster. For all synthetic data sets considered, the empirical results observed changed only trivially, and we do not report these results.
Parameters were set consistently across all examples, unless otherwise noted. The initial E-nearest neighbor graph was constructed using $k_{\text{Euc}} = 20$. The scales $\{t_s\}_{s=1}^{m}$ for approximation were chosen to increase exponentially while requiring $m = 20$. Nearest neighbor denoising was performed using $k_{\text{nse}} = 20$. The denoising threshold θ was chosen by estimating the elbow in a graph of sorted nearest neighbor distances. For each data set, L_{SYM} was computed for 20 σ values equally spaced in an interval. All code and scripts to reproduce the results in this article are publicly available[1].

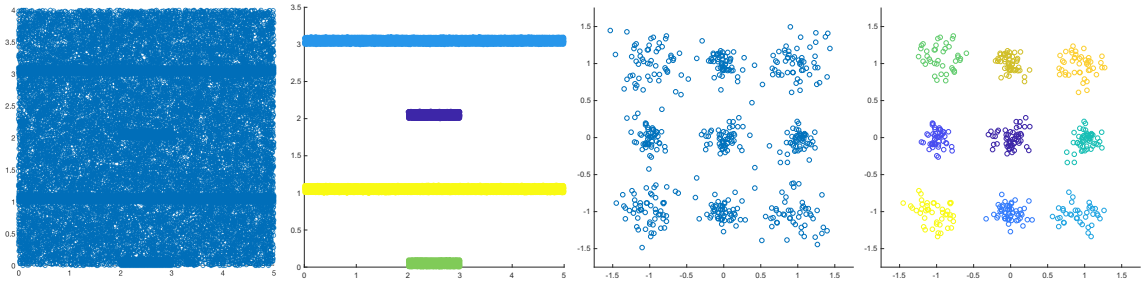
7.1. Synthetic Data

The four synthetic data sets considered are:

- **Four Lines** This data set consists of four highly elongated clusters in $\mathbb{R}^2$ with uniform two-dimensional noise added; see Figure 8a. The longer clusters have $n_i = 40000$ points, the smaller ones $n_i = 8000$, with $\tilde{n} = 20000$ noise points. This data set is too large to cluster with a dense Euclidean Laplacian.
- **Nine Gaussians** Each of the nine clusters consist of $n_i = 50$ random samples from a two-dimensional Gaussian distribution; see Figure 8c. All of the Gaussians have distinct means. Five have covariance matrix $0.01I$ while four have covariance matrix $0.04I$, resulting in clusters of unequal density. The noise consists of $\tilde{n} = 50$ uniformly sampled points.

1. `https://bitbucket.org/annavlittle/llpd_code/branch/v2.1`

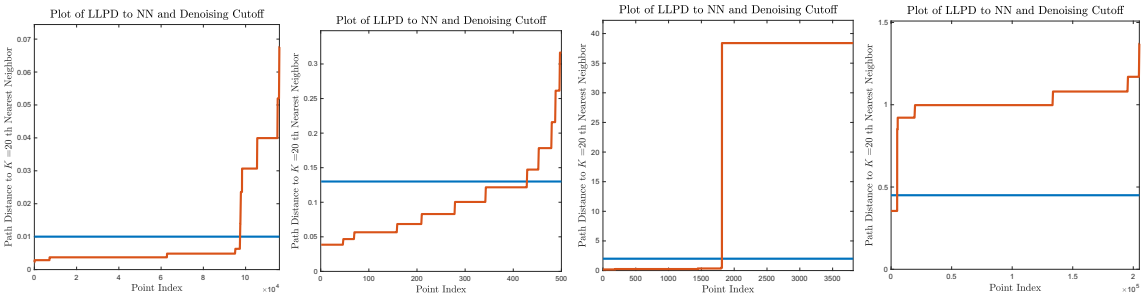
- **Concentric Spheres** Letting $\mathbb{S}_r^d \subset \mathbb{R}^{d+1}$ denote the d-dimensional sphere of radius r centered at the origin, the clusters consist of points uniformly sampled from three concentric 2-dimensional spheres embedded in $\mathbb{R}^{1000}$: $n_1 = 250$ points from $\mathbb{S}_1^2$, $n_2 = 563$ points from $\mathbb{S}_{1.5}^2$, and $n_3 = 1000$ points from $\mathbb{S}_2^2$, so that the cluster densities are constant. The noise consists of an additional $\tilde{n} = 2000$ points uniformly sampled from $[-2, 2]^{1000}$.
- **Parallel Planes** Five $d = 5$ dimensional planes are embedded in $[0, 1]^{25}$ by setting the last $D - d = 20$ coordinates to a distinct constant for each plane; we sample uniformly $n_i = 1000$, $1 \leq i \leq 5$ points from each plane and add $\tilde{n} = 200000$ noise points uniformly sampled from $[0, 1]^{25}$. Only 2 of the last 20 coordinates contribute to the separability of the planes, so that the Euclidean distance between consecutive parallel planes is approximately 0.35. We note that for this data set, it is possible to run Euclidean spectral clustering after denoising with the LLPD.



(a) Four Lines (b) LLPD spectral clustering on denoised Four Lines (c) Nine Gaussians (d) LLPD spectral clustering on denoised Nine Gaussians

Figure 8: Two dimensional synthetic data sets and LLPD spectral clustering results for the denoised data sets. In Figures 8b and 8d, color corresponds to the label returned by LLPD spectral clustering.

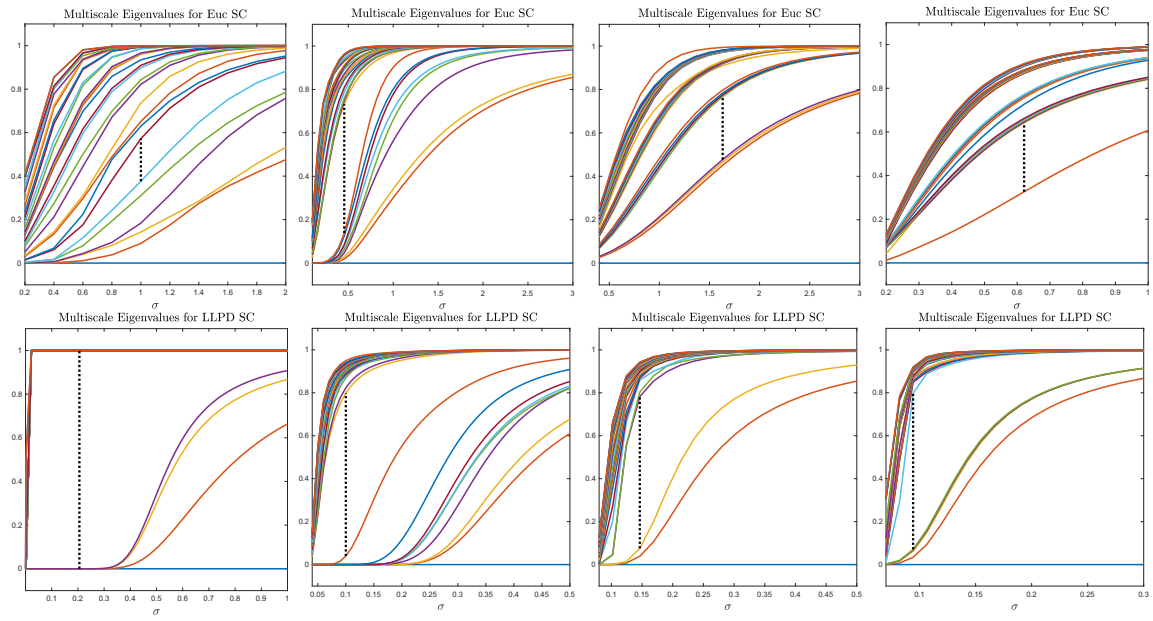
Figure 9 illustrates the denoising procedure. Sorted LLPD-nearest neighbor distances are shown in blue, and the denoising threshold θ (selected by choosing the graph elbow) is shown in red. All plots exhibit an elbow pattern, which is shallow when D/d is small (Figure 9b) and sharp when D/d is large (Figure 9c; the sharpness is due to the drastic difference in nearest neighbor distances for cluster and noise points). Figure 10 shows the multiscale eigenvalue plots for the synthetic data sets. For the four lines data, Euclidean spectral clustering is run with $n = 1160$ since it is prohibitively slow for $n = 116000$; however all relevant proportions such as $\tilde{n}/n_i$ are the same. LLPD spectral clustering correctly infers K for all synthetic data sets; Euclidean spectral clustering fails to correctly infers K except for the nine Gaussians example. See Table 2 for all $\hat{K}$ values and empirical accuracies. Although accuracy is reported on the cluster points only, we remark that labels can be extended to any noise points which survive denoising by considering the label of the closest cluster set, and the empirical accuracies reported in Table 2 remain essentially unchanged.



(a) Four Lines (b) Nine Gaussians (c) Concentric Spheres (d) Parallel Planes

Figure 9: LLPD to $k_{\text{nse}}^{\text{th}}$ LLPD-nearest neighbor (blue) and threshold θ used for denoising the data (red).

In addition to learning the number of clusters K, the multiscale eigenvalue plots can also be used to infer a good scale σ for LLPD spectral clustering as $\hat{\sigma} = \arg\max_{\sigma} \left(\lambda_{\hat{K}+1}(\sigma) - \lambda_{\hat{K}}(\sigma)\right)$. For the two dimensional examples, the right panel of Figure 8 shows the results of LLPD spectral clustering with $\hat{K}$, $\hat{\sigma}$ inferred from the maximal eigengap with LLPD. Robustly estimating K and σ makes LLPD spectral clustering essentially parameter free, and thus highly desirable for the analysis of real data.



(a) Four Lines (b) Nine Gaussians (c) Concentric Spheres (d) Parallel Planes

Figure 10: Multiscale eigenvalues of L_{SYM} for synthetic data sets using Euclidean distance (top) and LLPD (bottom).

7.2. Real Data

We apply our method on the following real data sets:

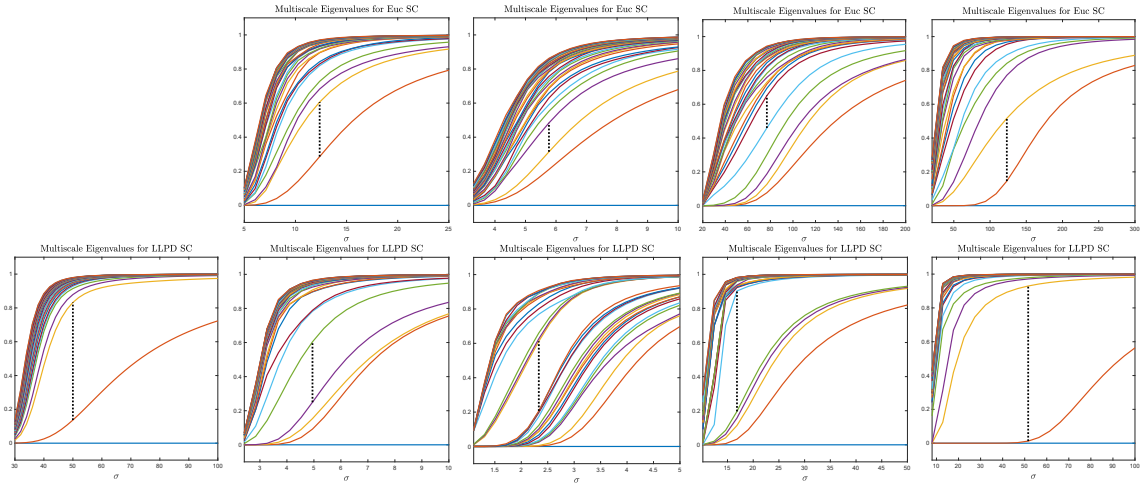
(a) Skins data (b) DrivFace Representative Faces (c) COIL objects (d) COIL 16 objects

Figure 11: Representative objects from (a) Skins, (b) DrivFace, (c) COIL, and (d) COIL 16 data sets.

- **Skins** This large data set consists of RGB values corresponding to pixels sampled from two classes: human skin and other[2]. The human skin samples are widely sampled with respect to age, gender, and skin color; see Bhatt et al. (2009) for details on the construction of the data set. This data set consists of 245057 data points in $D = 3$ dimensions, corresponding to the RGB values. Note LLPD was approximated from scales $\{t_s\}_{s=1}^m$ defined by 10 percentiles, as opposed to the default exponential scaling. See Figure 11a.
- **DrivFace** The DrivFace data set is publicly available[3] from the UCI Machine Learning Repository (Lichman, 2013). This data set consists of 606 80×80 pixel images of the faces of four drivers, 2 male and 2 female. See Figure 11b
- **COIL** The COIL (Columbia University Image Library) data set[4] consists of images of 20 different objects captured at varying angles (Nene et al., 1996). There are 1440 different data points, each of which is a 32×32 image, thought of as a $D = 1024$ dimensional point cloud. See Figure 11c.
- **COIL 16** To ease the problem slightly, we consider a 16 class subset of the full COIL data, shown in Figure 11d.
- **Pen Digits** This data set[5] consists of 3779 spatially resampled digital signals of hand-written digits in 16 dimensions (Alimoglu and Alpaydin, 1996). We consider a subset consisting of five digits: $\{0, 2, 3, 4, 6\}$.
- **Landsat** The landsat satellite data we consider consists of pixels in 3×3 neighborhoods in a multispectral camera with four spectral bands[6]. This leads to a total ambient dimension of $D = 36$. The data considered consists of $K = 4$

2. https://archive.ics.uci.edu/ml/datasets/skin+segmentation
3. https://archive.ics.uci.edu/ml/datasets/DrivFace
4. http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
5. https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
6. https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)

classes, consisting of pixels of different physical materials: red soil, cotton, damp soil, and soil with vegetable stubble.



(a) Skins: LLPD eigenvalues (b) DrivFace (c) COIL 16 (d) Pen Digits (e) Landsat

Figure 12: Multiscale eigenvalues of L_{SYM} for real data sets using Euclidean distance (top, (b)-(e), does not appear for (a)) and LLPD (bottom (a)-(e)).

Labeling accuracy results as well as the $\hat{K}$ values returned by our algorithm are given in Table 2. LLPD spectral clustering correctly estimates K for all data sets except the full COIL data set and Landsat. Euclidean spectral clustering fails to correctly detect K on all real data sets. Figure 12 shows both the Euclidean and LLPD eigenvalues for Skins, DrivFace, COIL 16, Pen Digits, and Landsat. Euclidean spectral clustering results for Skins are omitted because Euclidean spectral clustering with a dense Laplacian is computationally intractable with such a large sample size. At least 90% of data points were retained during the denoising procedure with the exception of Skins (88.0% retained) and Landsat (67.2% retained). After denoising, LLPD spectral clustering achieved an overall accuracy exceeding 98.6% on all real data sets except COIL 20 (90.5%). Euclidean spectral clustering performed well on DrivFaces (OA 94.1%) and Pen Digits (OA 98.1%), but poorly on the remaining data sets, where the overall accuracy ranged from $68.9\% - 76.8\%$. K-means also performed well on DrivFaces (OA 87.5%) and Pen Digits (OA 97.6%) but poorly on the remaining data sets, where the overall accuracy ranged from $54.7\% - 78.5\%$.

8. Conclusions and Future Directions

This article developed finite sample estimates on the behavior of the LLPD metric, derived theoretical guarantees for spectral clustering with the LLPD, and introduced fast approximate algorithms for computing the LLPD and LLPD spectral clustering. The theoretical guarantees on the eigengap provide mathematical rigor for the

Data Set	Accuracy Statistic	K-means	Euclidean SC	LLPD SC
Four Lines ($n = 116000, \tilde{n} = 20000, N = 97361, D = 2, d = 1, K = 4, \zeta_N = 12.0239, \theta = .01, \hat{\sigma} = 0.2057, \delta = .9001$)	OA	.4951	.6838	**1.000**
	AA	.4944	.6995	**1.000**
	κ	.3275	.5821	**1.000**
	$\hat{K}$	-	6	**4**
Nine Gaussians ($n = 500, \tilde{n} = 0, N = 428, D = 2, d = 2, K = 9, \zeta_N = 10.7, \theta = .13, \hat{\sigma} = 0.1000, \delta = .1094$)	OA	**.9930**	**.9930**	**.9930**
	AA	**.9920**	**.9920**	**.9920**
	κ	**.9921**	**.9921**	**.9921**
	$\hat{K}$	-	**9**	**9**
Concentric Spheres ($n = 3813, \tilde{n} = 2000, N = 1813, D = 1000, d = 2, K = 3, \zeta_N = 7.2520, \theta = 2, \hat{\sigma} = 0.1463, \delta = .5$)	OA	.3464	.3519	**.9989**
	AA	.3463	.3438	**.9988**
	κ	.0094	.0155	**.9981**
	$\hat{K}$	-	4	**3**
Parallel Planes ($n = 205000, \tilde{n} = 200000, N = 5000, D = 30, d = 10, K = 5, \zeta_N = 5, \theta = .45, \hat{\sigma} = 0.0942, \delta = .3553$)	OA	.5594	.3964	**.9990**
	AA	.5594	.3964	**.9990**
	κ	.4493	.2455	**.9987**
	$\hat{K}$	-	2	**5**
Skins ($n = 245057, N = 215694, D = 3, K = 2, \zeta_N = 4.5343, \theta = 2, \hat{\sigma} = 50, \hat{\delta} = 0$)	OA	.5473	-	**.9962**
	AA	.4051	-	**.9970**
	κ	-.1683	-	**.9890**
	$\hat{K}$	-	-	**2**
DrivFaces ($n = 612, N = 574, D = 6400, K = 4, \zeta_N = 6.4494, \theta = 10, \hat{\sigma} = 4.9474, \hat{\delta} = 9.5976$)	OA	.8746	.9408	**1.000**
	AA	.8882	.9476	**1.000**
	κ	.9198	.9198	**1.000**
	$\hat{K}$	-	2	**4**
COIL 20 ($n = 1440, N = 1351, D = 1024, K = 20, \zeta_N = 27.5714, \theta = 4.5, \hat{\sigma} = 1.9211, \hat{\delta} = 3.3706$)	OA	.6555	.6890	**.9055**
	AA	.6290	.6726	**.8833**
	κ	.6368	.6724	**.9004**
	$\hat{K}$	-	3	17
COIL 16 ($n = 1152, N = 1088, D = 1024, K = 16, \zeta_N = 22.1837, \theta = 3.9, \hat{\sigma} = 2.3316, \hat{\delta} = 5.4350$)	OA	.7500	.7330	**1.000**
	AA	.7311	.6782	**1.000**
	κ	.7330	.6864	**1.000**
	$\hat{K}$	-	3	**16**
Pen Digits ($n = 3779, N = 3750, D = 16, K = 5, \zeta_N = 5.2228, \theta = 60, \hat{\sigma} = 16.8421, \hat{\delta} = 11.3137$)	OA	.9760	.9813	**.9949**
	AA	.9764	.9816	**.9949**
	κ	.9700	.9767	**.9937**
	$\hat{K}$	-	6	**5**
Landsat ($n = 1136, N = 763, D = 36, K = 4, \zeta_N = 8.4778, \theta = 32, \hat{\sigma} = 51.5789, \hat{\delta} = 28.2489$)	OA	.7851	.7680	**.9869**
	AA	.8619	.8532	**.9722**
	κ	.6953	.6722	**.9802**
	$\hat{K}$	-	2	2

Table 2: In all examples, LLPD spectral clustering performs at least as well as K-means and Euclidean spectral clustering, and it typically outperforms both. Best results for each method and performance metric are bolded. For each data set, we include parameters that determine the theoretical results. For both real and synthetic data sets, n (the total number of data points), N (the number of data points after denoising), D (the ambient dimension of the data), K (the number of clusters in the data), ζ_N (cluster balance parameter on the denoised data), θ (LLPD denoising threshold), and $\hat{\sigma}$ (learned scaling parameter in LLPD weight matrix) are given. For the synthetic data, $\tilde{n}$ (number of noise points), d (intrinsic dimension of the data), and δ (minimal Euclidean distance between clusters) are given, since these are known or can be computed exactly. For the real data, $\hat{\delta}$ (the minimal Euclidean distanced between clusters, after denoising) is provided. We remark that for the Skins data set, a very small number of points (which are integer triples in $\mathbb{R}^3$) appear in both classes, so that $\hat{\delta} = 0$. Naturally these points are not classified correctly, which leads to a slightly imperfect accuracy for LLPD spectral clustering.

heuristic claim that the eigengap determines the number of clusters, and theoretical guarantees on labeling accuracy improve on the state of the art in the LDLN data model. Moreover, the proposed approximation scheme enables efficient LLPD spectral clustering on large, high-dimensional data sets. Our theoretical results are verified numerically, and it is shown that LLPD spectral clustering determines the number of clusters and labels points with high accuracy in many cases where Euclidean spectral clustering fails. In a sense, the method proposed in this article combines two different clustering techniques: density techniques like DBSCAN and single linkage clustering, and spectral clustering. The combination allows for improved robustness and performance guarantees compared to either set of techniques alone.

It is of interest to generalize and improve the results in this article. Our theoretical results involved two components. First, we proved estimates on distances between points under the LLPD metric, under the assumption that data fits the LDLN model. Second, we proved that the weight matrix corresponding to these distances enjoys a structure which guarantees that the eigengap in the normalized graph Laplacian is informative. The first part of this program is generalizable to other distance metrics and data drawn from different distributions. Indeed, one can interpret the LLPD as a minimum over the ℓ^∞ norm of paths between points. Norms other than the ℓ^∞ norm may correspond to interesting metrics for data drawn from some class of distributions, for example, the geodesic distance with respect to some metric on a manifold. Moreover, introducing a comparison of tangent-planes into the spectral clustering distance metric has been shown to be effective in the Euclidean setting (Arias-Castro et al., 2017), and allows one to distinguish between intersecting clusters in many cases. Introducing tangent plane comparisons into the LLPD construction would perhaps allow the results in this article to generalize to data drawn from intersecting distributions.

An additional problem not addressed in the present article is the consistency of LLPD spectral clustering. It is of interest to consider the behavior as $n \to \infty$ and determine if LLPD spectral clustering converges in the large sample limit to a continuum partial differential equation. This line of work has been fruitfully developed in recent years for spectral clustering with Euclidean distances (Garcia Trillos et al., 2016; Garcia Trillos and Slepcev, 2016a,b).

Acknowledgments

The authors are grateful to two anonymous reviewers, whose comments and suggestions significantly improved the presentation of the manuscript. MM and JMM were partially supported by NSF-IIS-1708553, NSF-DMS-1724979, NSF-CHE-1708353 and AFOSR FA9550-17-1-0280.

Appendix A. Proofs from Section 4

A.1. Proof of Lemma 4.1

Let $y \in S$ satisfy $\|x-y\|_2 \le \tau$. Suppose $\tau \ge \epsilon/4$. For the upper bound, we have:

$$\mathcal{H}^D(B(S,\tau)\cap B_\epsilon(x)) \le \mathcal{H}^D(B_\epsilon(x)) = \mathcal{H}^D(B_1)\epsilon^D \le \mathcal{H}^D(B_1)\epsilon^d 4^{D-d}\left(\tau\wedge\epsilon\right)^{D-d}.$$

For the lower bound, set $z=(1-\alpha)x+\alpha y,\ \alpha=\frac{\epsilon}{4\tau}$. Then $\|z-x\|_2 \le \epsilon/4$ and $\|z-y\|_2 \le \tau-\epsilon/4$, so $B_{\epsilon/4}(z)\subset B(S,\tau)\cap B_\epsilon(x)$, and

$$4^{-D}\mathcal{H}^D(B_1)\epsilon^d(\epsilon\wedge\tau)^{D-d} \le 4^{-D}\mathcal{H}^D(B_1)\epsilon^D = \mathcal{H}^D(B_{\epsilon/4}(z)) \le \mathcal{H}^D(B(S,\tau)\cap B_\epsilon(x)).$$

This shows that (4.2) holds in the case $\tau \ge \epsilon/4$.
Now suppose $\tau < \epsilon/4$. We consider two cases, with the second to be reduced to the first one.
Case 1: $x=y\in S$. Let $\{y_i\}_{i=1}^n$ be a τ-packing of $S\cap B_{\epsilon-\tau}(y)$, i.e.: $S\cap B_{\epsilon-\tau}(y)\subset \bigcup_{i=1}^n B_\tau(y_i)$, and $\|y_i-y_j\|_2>\tau, i\neq j$. We show this implies that $B(S,\tau)\cap B_{\frac{\epsilon}{2}}(y)\subset \bigcup_{i=1}^n B_{2\tau}(y_i)$. Indeed, let $x_0\in B(S,\tau)\cap B_{\frac{\epsilon}{2}}(y)$. Then there is some $x^*\in S$ such that $\|x_0-x^*\|_2\le\tau$, and so $\|x^*-y\|_2\le\|x^*-x_0\|_2+\|x_0-y\|_2\le\tau+\epsilon/2<\epsilon-\tau$ (since $\tau<\epsilon/4$), and hence $x^*\in S\cap B_{\epsilon-\tau}(y)$. Thus there exists y_i^* in the τ-packing of $S\cap B_{\epsilon-\tau}(y)$ such that $x^*\in B_\tau(y_i^*)$, so that $\|x_0-y_i^*\|_2\le\|x_0-x^*\|_2+\|x^*-y_i^*\|_2\le 2\tau$, and $x_0\in B_{2\tau}(y_i^*)$. Hence,

$$\mathcal{H}^D(B(S,\tau)\cap B_{\frac{\epsilon}{2}}(y)) \le \sum_{i=1}^n \mathcal{H}^D(B_{2\tau}(y_i)) = n\mathcal{H}^D(B_1)2^D\tau^D. \tag{A.1}$$

Similarly, it is straight-forward to verify that $\bigcup_{i=1}^n B_{\frac{\tau}{2}}(y_i)\subset B(S,\tau)\cap B_\epsilon(y)$, and since the $\{B_{\frac{\tau}{2}}(y_i)\}_{i=1}^n$ are pairwise disjoint, it follows that

$$n\mathcal{H}^D(B_1)2^{-D}\tau^D = \mathcal{H}^D\left(\cup_{i=1}^n B_{\frac{\tau}{2}}(y_i)\right) \le \mathcal{H}^D(B(S,\tau)\cap B_\epsilon(y)). \tag{A.2}$$

We now estimate n. Indeed, $S\cap B_{\frac{\epsilon}{2}}(y)\subset S\cap B_{\epsilon-\tau}(y)\subset\bigcup_{i=1}^n S\cap B_\tau(y_i)$, so that by assumption $S\in\mathcal{S}_d(\kappa,\epsilon_0)$ and $\epsilon\in(0,\frac{2\epsilon_0}{5})\subset(0,\epsilon_0)$,

$$2^{-d}\epsilon^d\kappa^{-1}\mathcal{H}^d(B_1) \le \mathcal{H}^d(S\cap B_{\frac{\epsilon}{2}}(y)) \le \sum_{i=1}^n \mathcal{H}^d(S\cap B_\tau(y_i)) \le \kappa\tau^d n\mathcal{H}^d(B_1).$$

It follows that

$$2^{-d}\left(\epsilon/\tau\right)^d\kappa^{-2} \le n. \tag{A.3}$$

Similarly, $\bigcup_{i=1}^n S\cap B_{\frac{\tau}{2}}(y_i)\subset S\cap B_\epsilon(y)$ yields

$$n\kappa^{-1}\tau^d 2^{-d}\mathcal{H}^d(B_1) \le \sum_{i=1}^n \mathcal{H}^d(S\cap B_{\frac{\tau}{2}}(y_i)) \le \mathcal{H}^d(S\cap B_\epsilon(y)) \le \kappa\epsilon^d\mathcal{H}^d(B_1),$$

so that

$$n \leq 2^d \kappa^2 (\epsilon/\tau)^d . \tag{A.4}$$

By combining (A.2) and (A.3), we obtain

$$\mathcal{H}^D(B_1) 2^{-(d+D)} \kappa^{-2} \tau^D (\epsilon/\tau)^d \leq \mathcal{H}^D(B(S,\tau) \cap B_\epsilon(y)) \tag{A.5}$$

and by combining (A.1) and (A.4), we obtain

$$\mathcal{H}^D(B(S,\tau) \cap B_{\frac{\epsilon}{2}}(y)) \leq \mathcal{H}^D(B_1) 2^{d+D} \kappa^2 \tau^D (\epsilon/\tau)^d , \tag{A.6}$$

which are valid for any $\epsilon < \epsilon_0, \tau < \epsilon/4$. Replacing $\epsilon/2$ and τ with ϵ and 2τ, respectively, in (A.6), and combining with (A.5), we obtain, for $\epsilon < \epsilon_0/2, \tau < \epsilon/4$,

$$\mathcal{H}^D(B_1) 2^{-(d+D)} \kappa^{-2} \tau^D (\epsilon/\tau)^d \leq \mathcal{H}^D(B(S,\tau) \cap B_\epsilon(y)) \leq \mathcal{H}^D(B_1) 2^{d+2D} \kappa^2 \tau^D (\epsilon/\tau)^d .$$

Case 2: $x \notin S$. Notice that $\|x - y\|_2 \leq \tau \leq \epsilon/4$, so $B_{\frac{3\epsilon}{4}}(y) \subset B_\epsilon(x) \subset B_{\frac{5\epsilon}{4}}(y)$. Thus: $\mathcal{H}^D(B(S,\tau) \cap B_\epsilon(x)) \leq \mathcal{H}^D(B(S,\tau) \cap B_{\frac{5\epsilon}{4}}(y)) \leq \mathcal{H}^D(B_1) 2^{2D+2d} \kappa^2 \tau^D (\epsilon/\tau)^d$, so as long as $\epsilon < \frac{2\epsilon_0}{5}$ we have

$$\mathcal{H}^D(B(S,\tau) \cap B_\epsilon(x)) \geq \mathcal{H}^D(B(S, 3\tau/4) \cap B_{3\epsilon/4}(y)) \geq \mathcal{H}^D(B_1) 2^{-(2D+d)} \kappa^{-2} \tau^D (\epsilon/\tau)^d .$$

We thus obtain the statement in Lemma 4.1.

A.2. Proof of Theorem 4.3

Cover $B(S,\tau)$ with an $\epsilon/4$-packing $\{y_i\}_{i=1}^N$, such that $B(S,\tau) \subset \bigcup_{i=1}^N B_{\epsilon/4}(y_i)$, and $\|y_i - y_j\|_2 > \epsilon/4, \forall i \neq j$. $\{B_{\epsilon/8}(y_i)\}_{i=1}^N$ are thus pairwise disjoint, so that $\sum_{i=1}^N \mathcal{H}^D(B_{\epsilon/8}(y_i) \cap B(S,\tau)) \leq \mathcal{H}^D(B(S,\tau))$. By Lemma 4.1, we may bound $C_1 (\epsilon/8)^d (\epsilon/8 \wedge \tau)^{D-d} \leq \mathcal{H}^D(B_{\epsilon/8}(y_i) \cap B(S,\tau))$, where $C_1 = \kappa^{-2} 2^{-(2D+d)} \mathcal{H}^D(B_1)$. Hence, $N C_1 (\epsilon/8)^d (\epsilon/8 \wedge \tau)^{D-d} \leq \mathcal{H}^D(B(S,\tau))$, so that

$$N \leq C \mathcal{H}^D(B(S,\tau)) (\epsilon/8)^{-d} (\epsilon/8 \wedge \tau)^{-(D-d)} , \; C = \kappa^2 2^{2D+d} (\mathcal{H}^D(B_1))^{-1}.$$

So, $C \mathcal{H}^D(B(S,\tau)) (\epsilon/8)^{-d} (\epsilon/8 \wedge \tau)^{-(D-d)}$ balls of radius $\epsilon/4$ are needed to cover $B(S,\tau)$. We now determine how many samples n must be taken so that each ball contains at least one sample with probability exceeding $1 - t$. If this occurs, then each pair of points is connected by a path with all edges of length at most ϵ. Notice that the distribution of the number of points ω_i in the set $B_{\epsilon/4}(y_i) \cap B(S,\tau)$ is $\omega_i \sim \text{Bin}(n, p_i)$, where

$$p_i = \frac{\mathcal{H}^D(B(S,\tau) \cap B_{\epsilon/4}(y_i))}{\mathcal{H}^D(B(S,\tau))} \geq \frac{C^{-1} (\epsilon/4)^d (\epsilon/4 \wedge \tau)^{D-d}}{\mathcal{H}^D(B(S,\tau))} := p.$$

Since $\mathbb{P}(\exists i : \omega_i = 0) \leq N(1-p)^n \leq N e^{-pn} \leq t$ as long as $n \geq 1/p \log N/t$, it suffices for n to satisfy $n \geq \frac{C\mathcal{H}^D(B(S,\tau))}{(\epsilon/4)^d (\tau \wedge \epsilon/4)^{D-d}} \log \frac{C\mathcal{H}^D(B(S,\tau))}{(\epsilon/8)^d (\tau \wedge \epsilon/8)^{D-d} t}$.

A.3. Proof of Corollary 4.4

For a fixed l, choose a τ packing of S_l, i.e. let $y_1, \ldots, y_m \in S_l$ such that $S_l \subset \cup_i B_\tau(y_i)$ and $\|y_i - y_j\|_2 > \tau$ for $i \neq j$. Then $B(S_l, \tau) \subset B_{2\tau}(y_i)$. Now we control the size of m. Since the $B_{\frac{\tau}{2}}(y_i)$ are disjoint and $S_l \in S(\kappa, \epsilon_0)$, $\mathcal{H}^d(S_l) \geq \sum_{i=1}^m \mathcal{H}^d(S_l \cap B_{\frac{\tau}{2}}(y_i)) \geq m\kappa^{-1}\mathcal{H}^d(B_1)\left(\frac{\tau}{2}\right)^d$, so that $m \leq \kappa \frac{\mathcal{H}^d(S_l)}{\mathcal{H}^d(B_1)}\left(\frac{\tau}{2}\right)^{-d}$. Furthermore, since $\frac{2}{5}\epsilon_0 > 2\tau$ we have by Lemma 4.1:

$$\frac{\mathcal{H}^D(B(S_l,\tau))}{\mathcal{H}^D(B_1)} \leq \sum_{i=1}^m \frac{\mathcal{H}^D(B(S_l,\tau)\cap B_{2\tau}(y_i))}{\mathcal{H}^D(B_1)} \leq m\kappa^2 2^{(2D+2d)}(2\tau)^d \tau^{D-d} \leq \kappa^3 2^{(2D+4d)} \frac{\mathcal{H}^d(S_l)}{\mathcal{H}^d(B_1)} \tau^{D-d}.$$

Combining the above with (4.5) implies

$$n_l \geq \frac{C\mathcal{H}^D(B(S_l,\tau))}{(\epsilon/4)^d \tau^{D-d}\mathcal{H}^D(B_1)} \log \frac{C\mathcal{H}^D(B(S_l,\tau))}{(\epsilon/8)^d \tau^{D-d}\mathcal{H}^D(B_1)t}$$

for $C = \kappa^2 2^{2D+d}$. Thus by Theorem 4.3, $\mathbb{P}(\max_{x\neq y\in X_l} \rho_{\ell\ell}(x,y) < \epsilon) \geq 1 - \frac{t}{K}$. Repeating the above argument for each S_l and letting E_l denote the event that $\max_{x\neq y\in X_l} \rho_{\ell\ell}(x,y) \geq \epsilon$, we obtain $\mathbb{P}(\epsilon_{\text{in}} \geq \epsilon) = \mathbb{P}(\max_{1\leq l\leq K} \max_{x\neq y\in X_l} \rho_{\ell\ell}(x,y) \geq \epsilon) = \mathbb{P}(\cup_l E_l) \leq \sum_l \mathbb{P}(E_l) \leq K(\frac{t}{K}) = t$.

A.4. Proof of Theorem 4.11

Re-writing the inequality assumed in the theorem, we are guaranteed that

$$C\left(\max_{l=1,\ldots,K} \frac{4^d\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{n_l\mathcal{H}^d(B_1)} \log\left(2^d n_l \frac{2K}{t}\right)\right)^{\frac{1}{d}} < \left(\frac{\left(\frac{t}{2}\right)^{\frac{1}{k_{\text{nse}}}}}{((k_{\text{nse}}+1)\tilde{n})^{\frac{k_{\text{nse}}+1}{k_{\text{nse}}}}} \frac{\mathcal{H}^D(\tilde{X})}{\mathcal{H}^D(B_1)}\right)^{\frac{1}{D}}.$$

Let $C\epsilon_1^*$ denote the left hand side of the above inequality and ϵ_2^* the right hand side. Then for all $1 \leq l \leq K$, $n_l \geq \frac{\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{\left(\epsilon_1^*/4\right)^d \mathcal{H}^d(B_1)} \left(\log\left(2^d n_l \frac{2K}{t}\right)\right)$, and since $\log\left(2^d n_l \frac{2K}{t}\right) \geq 1$, clearly $n_l \geq \frac{\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{\left(\epsilon_1^*/4\right)^d \mathcal{H}^d(B_1)}$, and we obtain $n_l \geq \left(\frac{\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{\left(\epsilon_1^*/4\right)^d \mathcal{H}^d(B_1)} \log\left(\frac{\kappa^5 2^{4D+5d}\mathcal{H}^d(S_l)}{\left(\epsilon_1^*/8\right)^d \mathcal{H}^d(B_1)} \frac{2K}{t}\right)\right)$. Since $\tau < \frac{\epsilon_1^*}{8} \wedge \frac{\epsilon_0}{5}$ by assumption, Corollary 4.4 yields $\mathbb{P}\left(\epsilon_{\text{in}} < \epsilon_1^*\right) \geq 1 - \frac{t}{2}$. Also by Theorem 4.9, $\mathbb{P}(\epsilon_{\text{nse}} > \epsilon_2^*) \geq 1 - \frac{t}{2}$. Since we are assuming $C\epsilon_1^* < \epsilon_2^*$, $\mathbb{P}(C\epsilon_{\text{in}} < \epsilon_{\text{nse}}) \geq \mathbb{P}((\epsilon_{\text{in}} < \epsilon_1^*) \cap (\epsilon_{\text{nse}} > \epsilon_2^*)) \geq 1 - t$.

Appendix B. Proof of Theorem 5.5

Let $n_l = |A_l|$ and $m_l = |C_l|$ denote the cardinality of the sets in Assumption 1, and let $\eta_1 := 1 - f_\sigma(\epsilon_{\text{in}}), \eta_\theta := 1 - f_\sigma(\theta), \eta_2 := f_\sigma(\epsilon_{\text{nse}})$.

B.1. Bounding Entries of Weight Matrix W and Degrees

The following Lemma guarantees that the weight matrix will have a convenient structure.

Lemma B.1 *For $1 \leq l \leq K$, let $A_l, C_l, \tilde{A}_l$ be as in Assumption 1. Then:*

1. *For each fixed $x_i^l \in C_l$, x_i^l is equidistant from all points in A_l; more precisely:*

$$\rho(x_i^l, x_j^l) = \min_{x^l \in A_l} \rho(x_i^l, x^l) =: \rho_i^l, \qquad \forall x_i^l \in C_l, x_j^l \in A_l, 1 \leq l \leq K.$$

2. *The distance between any point in $\tilde{A}_l$ and $\tilde{A}_s$ is constant for $l \neq s$, that is:*

$$\rho(x_i^l, x_j^s) = \min_{x^l \in \tilde{A}_l, x^s \in \tilde{A}_s} \rho(x^l, x^s) =: \rho^{l,s} \qquad \forall x_i^l \in \tilde{A}_l, x_j^s \in \tilde{A}_s, 1 \leq l \neq s \leq K.$$

Proof To prove (1), let $x_i^l \in C_l$ and $x_j^l \in A_l$. Since ρ_i^l is the minimum distance between x_i^l and a point in A_l, clearly $\rho(x_i^l, x_j^l) \geq \rho_i^l$. Let x_*^l denote the point in A_l such that $\rho_i^l = \rho(x_i^l, x_*^l)$. Then $\rho(x_i^l, x_j^l) \leq \rho(x_i^l, x_*^l) \vee \rho(x_*^l, x_j^l) \leq \rho(x_i^l, x_*^l) \vee \epsilon_{\text{in}} = \rho_i^l \vee \epsilon_{\text{in}} = \rho_i^l$. Thus $\rho(x_i^l, x_j^l) = \rho_i^l$.
To prove (2), let $x_i^l \in \tilde{A}_l$ and $x_j^s \in \tilde{A}_s$ for $l \neq s$. Clearly, $\rho(x_i^l, x_j^s) \geq \rho^{l,s}$. Now let $x_*^l \in \tilde{A}_l, x_*^s \in \tilde{A}_s$ be the points that achieve the minimum, i.e. $\rho^{l,s} = \rho(x_*^l, x_*^s)$. Note that $\rho(x_i^l, x_*^l) \leq \theta$ and similarly for $\rho(x_j^s, x_*^s)$ (if x_i^l, x_*^l are both in C_l, pick any point $z \in A_l$ to obtain $\rho(x_i^l, x_*^l) \leq \rho(x_i^l, z) \vee \rho(z, x_*^l) \leq \theta$). Thus:

$$\rho(x_i^l, x_j^s) \leq \rho(x_i^l, x_*^l) \vee \rho(x_*^l, x_*^s) \vee \rho(x_*^s, x_j^s) \leq \theta \vee \rho^{l,s} \vee \theta = \rho^{l,s},$$

since $\rho^{l,s} \geq \epsilon_{\text{nse}} > \theta$. Thus $\rho(x_i^l, x_j^s) = \rho^{l,s}$. ■

We now proceed with the proof of Theorem 5.5. By Lemma B.1, the off-diagonal blocks of W are constant, and so letting $w^{l,s} = W_{\tilde{A}_l, \tilde{A}_s}$ denote this constant, W has form

$$W = \begin{bmatrix} W_{\tilde{A}_1, \tilde{A}_1} & w^{1,2} & \dots & w^{1,K} \\ w^{2,1} & W_{\tilde{A}_2, \tilde{A}_2} & \dots & w^{2,K} \\ \vdots & & & \vdots \\ w^{K,1} & w^{K,2} & \dots & W_{\tilde{A_K}, \tilde{A_K}} \end{bmatrix},$$

and $w^{l,s} \leq f_\sigma(\epsilon_{\text{nse}})$ for $1 \leq l \neq s \leq K$ by (5.3).
We now consider an arbitrary diagonal block $W_{\tilde{A}_l, \tilde{A}_l}$. For convenience let x_i^l, $1 \leq i \leq n_l + m_l$ denote the points in $\tilde{A}_l$, ordered so that $x_i^l \in A_l$ for $i = 1, \dots, n_l$ and $x_i^l \in C_l$ for $i = n_l + 1, \dots, n_l + m_l$. For every $x_{i+n_l}^l \in C_l$, let $\rho_{i+n_l}^l$ denote the minimal distance to A_l, i.e. $\rho_{i+n_l}^l = \min_{x^l \in A_l} \rho(x_{i+n_l}^l, x^l)$, $w_i^l = f_\sigma(\rho_{i+n_l}^l)$ for all $1 \leq l \leq K, 1 \leq i \leq m_l$. Then by Lemma B.1, any point in C_l is equidistant from all points in A_l, so that $(W_{\tilde{A}_l, \tilde{A}_l})_{ij} = w_{i-n_l}^l$ for all $x_i^l \in C_l, x_j^l \in A_l$, and by (5.2), $f_\sigma(\epsilon_{\text{in}}) > w_{i-n_l}^l \geq f_\sigma(\theta)$ for $n_l + 1 \leq i \leq n_l + m_l$. Note if $x_i^l, x_j^l \in C_l$, then pick any $x_*^l \in A_l$, and one has $\rho(x_i^l, x_j^l) \leq \rho(x_i^l, x_*^l) \vee \rho(x_*^l, x_j^l) \leq \theta$ by (5.2).

Thus the diagonal blocks of W have the following form:

$$W_{\tilde{A}_l,\tilde{A}_l} = \begin{bmatrix} [f_\sigma(\epsilon_{\text{in}}),1] & [f_\sigma(\epsilon_{\text{in}}),1] & \dots & [f_\sigma(\epsilon_{\text{in}}),1] & w_1^l & w_2^l & \dots & w_{m_l}^l \\ [f_\sigma(\epsilon_{\text{in}}),1] & [f_\sigma(\epsilon_{\text{in}}),1] & \dots & [f_\sigma(\epsilon_{\text{in}}),1] & w_1^l & w_2^l & \dots & w_{m_l}^l \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ [f_\sigma(\epsilon_{\text{in}}),1] & [f_\sigma(\epsilon_{\text{in}}),1] & \dots & [f_\sigma(\epsilon_{\text{in}}),1] & w_1^l & w_2^l & & w_{m_l}^l \\ w_1^l & w_1^l & \dots & w_1^l & [f_\sigma(\theta),1] & [f_\sigma(\theta),1] & \dots & [f_\sigma(\theta),1] \\ w_2^l & w_2^l & \dots & w_2^l & [f_\sigma(\theta),1] & [f_\sigma(\theta),1] & \dots & [f_\sigma(\theta),1] \\ \vdots & & & \vdots & \vdots & & & \vdots \\ w_{m_l}^l & w_{m_l}^l & \dots & w_{m_l}^l & [f_\sigma(\theta),1] & [f_\sigma(\theta),1] & \dots & [f_\sigma(\theta),1] \end{bmatrix}.$$

The entries labeled $[f_\sigma(\epsilon_{\text{in}}),1]$ or $[f_\sigma(\theta),1]$ indicate entries falling in the interval. So we have the following bounds on the entries of W:

$$\begin{aligned} f_\sigma(\epsilon_{\text{in}}) \le (W_{\tilde{A}_l,\tilde{A}_l})_{ij} \le 1 &\qquad \text{for } x_i^l, x_j^l \in A_l, \\ f_\sigma(\theta) \le (W_{\tilde{A}_l,\tilde{A}_l})_{ij} < f_\sigma(\epsilon_{\text{in}}) &\qquad \text{for } x_i^l \in A_l, x_j^l \in C_l, \\ f_\sigma(\theta) \le (W_{\tilde{A}_l,\tilde{A}_l})_{ij} \le 1 &\qquad \text{for } x_i^l, x_j^l \in C_l, \\ 0 \le (W_{\tilde{A}_l,\tilde{A}_s})_{ij} \le f_\sigma(\epsilon_{\text{nse}}) &\qquad \text{for } x_i^l \in \tilde{A}_l, x_j^s \in \tilde{A}_s, l \ne s. \end{aligned}$$

Let $\deg_i^l$ denote the degree of x_i^l, and let $w^l = \sum_{j=1}^{m_l} w_j^l$, $o^l = \sum_{s\ne l}(n_s+m_s)w^{l,s}$. Note that regarding the degrees:

$$\begin{aligned} n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l \le \deg_i^l \le n_l + w^l + o^l &\quad \text{for } x_i^l \in A_l, \\ (n_l+m_l) f_\sigma(\theta) + o^l \le \deg_i^l \le n_l + m_l + o^l &\quad \text{for } x_i^l \in C_l. \end{aligned}$$

where $m_l f_\sigma(\theta) \le w^l \le m_l f_\sigma(\epsilon_{\text{in}}) \le m_l$.

B.2. Bounding Entries of Normalized Weight Matrix $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$

We thus obtain the following entrywise bounds for the diagonal block $D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}$:

$$\begin{aligned} \frac{f_\sigma(\epsilon_{\text{in}})}{n_l + w^l + o^l} \le (D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}})_{ij} \le \frac{1}{n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l} &\qquad \text{for } x_i^l, x_j^l \in A_l, \\ \frac{f_\sigma(\theta)}{n_l + m_l + o^l} \le (D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}})_{ij} \le \frac{1}{(n_l+m_l) f_\sigma(\theta) + o^l} &\qquad \text{for } x_i^l, x_j^l \in C_l. \end{aligned}$$

For $x_i^l \in A_l, x_j^l \in C_l$, we have:

$$\frac{f_\sigma(\theta)}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l + o^l}} \le (D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}})_{ij} < \frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l}\sqrt{(n_l+m_l) f_\sigma(\theta) + o^l}}.$$

Now consider the off-diagonal block $D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_s} D_{\tilde{A}_s}^{-\frac{1}{2}}$ for some $l \ne s$. Since $\deg_i^l \ge f_\sigma(\theta)\min_l(n_l+m_l) = f_\sigma(\theta)\zeta_N^{-1} N$ for all data points, we have:

$$\left| D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_s} D_{\tilde{A}_s}^{-\frac{1}{2}} \right| \le \frac{\zeta_N f_\sigma(\epsilon_{\text{nse}})}{f_\sigma(\theta) N}.$$

B.3. Perturbation to Obtain a Block Diagonal and Block Constant Matrix

Consider the normalized weight matrix $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. This matrix consists of diagonal blocks of the form $D_{\tilde{A}_l}^{-\frac{1}{2}}W_{\tilde{A}_l,\tilde{A}_l}D_{\tilde{A}_l}^{-\frac{1}{2}}$ and off diagonal blocks of the form $D_{\tilde{A}_l}^{-\frac{1}{2}}W_{\tilde{A}_l,\tilde{A}_s}D_{\tilde{A}_s}^{-\frac{1}{2}}$, some $l \neq s$. We will consider the spectral perturbations associated with (1) setting off-diagonal blocks to 0 and (2) making the diagonal blocks essentially block constant. More precisely, we consider the spectral perturbations associated with the matrix perturbations $\mathbf{P_1}, \mathbf{P_2}$ given as:

$$
\begin{aligned}
D^{-\frac{1}{2}}WD^{-\frac{1}{2}} &= \begin{bmatrix}
D_{\tilde{A}_1}^{-\frac{1}{2}}W_{\tilde{A}_1,\tilde{A}_1}D_{\tilde{A}_1}^{-\frac{1}{2}} & D_{\tilde{A}_1}^{-\frac{1}{2}}W_{\tilde{A}_1,\tilde{A}_2}D_{\tilde{A}_2}^{-\frac{1}{2}} & \dots & D_{\tilde{A}_1}^{-\frac{1}{2}}W_{\tilde{A}_1,\tilde{A_K}}D_{\tilde{A_K}}^{-\frac{1}{2}} \\
D_{\tilde{A}_2}^{-\frac{1}{2}}W_{\tilde{A}_2,\tilde{A}_1}D_{\tilde{A}_1}^{-\frac{1}{2}} & D_{\tilde{A}_2}^{-\frac{1}{2}}W_{\tilde{A}_2,\tilde{A}_2}D_{\tilde{A}_2}^{-\frac{1}{2}} & \dots & D_{\tilde{A}_2}^{-\frac{1}{2}}W_{\tilde{A}_2,\tilde{A_K}}D_{\tilde{A_K}}^{-\frac{1}{2}} \\
\vdots & & & \vdots \\
D_{\tilde{A_K}}^{-\frac{1}{2}}W_{\tilde{A_K},\tilde{A}_1}D_{\tilde{A}_1}^{-\frac{1}{2}} & D_{\tilde{A_K}}^{-\frac{1}{2}}W_{\tilde{A_K},\tilde{A}_2}D_{\tilde{A}_2}^{-\frac{1}{2}} & \dots & D_{\tilde{A_K}}^{-\frac{1}{2}}W_{\tilde{A_K},\tilde{A_K}}D_{\tilde{A_K}}^{-\frac{1}{2}}
\end{bmatrix} \\
&\xrightarrow{\mathbf{P_1}} \begin{bmatrix}
D_{\tilde{A}_1}^{-\frac{1}{2}}W_{\tilde{A}_1,\tilde{A}_1}D_{\tilde{A}_1}^{-\frac{1}{2}} & 0 & \dots & 0 \\
0 & D_{\tilde{A}_2}^{-\frac{1}{2}}W_{\tilde{A}_2,\tilde{A}_2}D_{\tilde{A}_2}^{-\frac{1}{2}} & \dots & 0 \\
\vdots & & & \vdots \\
0 & 0 & \dots & D_{\tilde{A_K}}^{-\frac{1}{2}}W_{\tilde{A_K},\tilde{A_K}}D_{\tilde{A_K}}^{-\frac{1}{2}}
\end{bmatrix} := C \\
&\xrightarrow{\mathbf{P_2}} \begin{bmatrix}
B_1 & 0 & \dots & 0 \\
0 & B_2 & \dots & 0 \\
\vdots & & & \vdots \\
0 & 0 & \dots & B_K
\end{bmatrix} := B,
\end{aligned}
$$

where B_l is defined by $B_l = \frac{f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l}\mathbf{1}_{n_l+m_l}$. Throughout the proof $\mathbf{1}_N$ denotes the $N \times N$ matrix of all 1's, I_N the $N \times N$ identity matrix, and $\|\cdot\|_2$ the spectral norm.

B.4. Bounding $\mathbf{P_1}$ (Diagonalization)

We first consider the spectral perturbation due to $\mathbf{P_1}$. Using the bounds from B.2 for an off-diagonal block, the perturbation in the eigenvalues due to $\mathbf{P_1}$ is bounded by:

$$\|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - C\|_2 \leq N\|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - C\|_{\max} \leq \frac{\zeta_N f_\sigma(\epsilon_{\text{nse}})}{f_\sigma(\theta)} = \zeta_N\eta_2 + O(\zeta_N\eta_2\eta_\theta) := P_1.$$

B.5. Bounding $\mathbf{P_2}$ (Constant Blocks)

We now consider the spectral perturbation due to $\mathbf{P_2}$. Because $\mathbf{P_2}$ acts on the blocks of a block diagonal matrix, it is sufficient to bound the perturbation of each block. For the l^{th} block, let

$$D_{\tilde{A}_l}^{-\frac{1}{2}}W_{\tilde{A}_l,\tilde{A}_l}D_{\tilde{A}_l}^{-\frac{1}{2}} - B_l := \begin{bmatrix} Q_l & R_l \\ R_l^T & S_l \end{bmatrix}$$

where Q is $n_l \times n_l$, R is $n_l \times m_l$, and S is $m_l \times m_l$, and R_l^T denotes the transpose of R_l. We will control the magnitude of each entry in Q_l, R_l, S_l using the bounds computed in B.2.

Bounding Q_l: For $x_i^l, x_j^l \in A_l$, we have

$$(B_l)_{ij} = \frac{f_\sigma(\epsilon_{\text{in}})}{n_l + w^l + o^l} \leq \left(D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l, \tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}\right)_{ij} \leq \frac{1}{n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l},$$

so that $(Q_l)_{ij} \leq \frac{1}{n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l + w^l + o^l} \leq \frac{\frac{1}{f_\sigma(\epsilon_{\text{in}})}}{n_l + w^l + o^l} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l + w^l + o^l} = \frac{2\eta_1 + O(\eta_1^2)}{n_l + w^l + o^l}$. Since $(Q_l)_{ij} \geq 0$, the above is in fact a bound for $|(Q_l)_{ij}|$, and we obtain:

$$|(Q_l)_{ij}| \leq \frac{2\eta_1 + O(\eta_1^2)}{(1-\eta_\theta)(n_l + m_l)} = \frac{2\eta_1 + O(\eta_1^2 + \eta_\theta^2)}{(n_l + m_l)}.$$

Bounding R_l: For $x_i^l \in A_l$ and $x_j^l \in C_l$, note that

$$\begin{aligned}
\left(D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l, \tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}\right)_{ij} - (B_l)_{ij} &\leq \frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{n_l f_\sigma(\epsilon_{\text{in}}) + w^l + o^l}\sqrt{(n_l + m_l) f_\sigma(\theta) + o^l}} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l + w^l + o^l} \\
&= \frac{\frac{\sqrt{f_\sigma(\epsilon_{\text{in}})}}{\sqrt{f_\sigma(\theta)}} - f_\sigma(\epsilon_{\text{in}})}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l + o^l}}.
\end{aligned}$$

Similarly:

$$\begin{aligned}
\left(D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l, \tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}\right)_{ij} - (B_l)_{ij} &\geq \frac{f_\sigma(\theta)}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l + o^l}} - \frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l f_\sigma(\theta) + o^l}} \\
&= \frac{f_\sigma(\theta) - \frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{f_\sigma(\theta)}}}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l + o^l}}
\end{aligned}$$

so that $|R_{ij}| \leq \frac{\left(\frac{\sqrt{f_\sigma(\epsilon_{\text{in}})}}{\sqrt{f_\sigma(\theta)}} - f_\sigma(\epsilon_{\text{in}})\right) \vee \left(\frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{f_\sigma(\theta)}} - f_\sigma(\theta)\right)}{\sqrt{n_l + w^l + o^l}\sqrt{n_l + m_l + o^l}}$. Thus we obtain:

$$\begin{aligned}
|R_{ij}| &\leq \left[\left(\frac{\sqrt{f_\sigma(\epsilon_{\text{in}})}}{f_\sigma(\theta)} - \frac{f_\sigma(\epsilon_{\text{in}})}{\sqrt{f_\sigma(\theta)}}\right) \vee \left(\frac{f_\sigma(\epsilon_{\text{in}})}{f_\sigma(\theta)} - \sqrt{f_\sigma(\theta)}\right)\right](n_l + m_l)^{-1} \\
&\leq \left[\left(\frac{\sqrt{1-\eta_1}}{1-\eta_\theta} - \frac{1-\eta_1}{\sqrt{1-\eta_\theta}}\right) \vee \left(\frac{1-\eta_1}{1-\eta_\theta} - \sqrt{1-\eta_\theta}\right)\right](n_l + m_l)^{-1} \\
&\leq \left[\left(\frac{\eta_1}{2} + \frac{\eta_\theta}{2} + O(\eta_1^2 + \eta_\theta^2)\right) \vee \left(\frac{3\eta_\theta}{2} - \eta_1 + O(\eta_1^2 + \eta_\theta^2)\right)\right](n_l + m_l)^{-1} \\
&\leq \left[\frac{3\eta_\theta}{2} + O(\eta_1^2 + \eta_\theta^2)\right](n_l + m_l)^{-1}.
\end{aligned}$$

Bounding S_l: For $x_i^l, x_j^l \in C_l$, note that

$$\left(D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}\right)_{ij} - (B^l)_{ij} \leq \frac{1}{(n_l+m_l)f_\sigma(\theta)+o^l} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} \leq \frac{f_\sigma(\theta)^{-1} - f_\sigma(\epsilon_{\text{in}})}{n_l+m_l+o^l}.$$

Similarly:

$$\begin{aligned}\left(D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}}\right)_{ij} - (B^l)_{ij} &\geq \frac{f_\sigma(\theta)}{n_l f_\sigma(\epsilon_{\text{in}})+m_l+o^l} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} \\ &\geq \frac{f_\sigma(\theta)}{n_l+m_l+o^l} - \frac{f_\sigma(\epsilon_{\text{in}})}{n_l+m_l f_\sigma(\theta)+o^l} = \frac{f_\sigma(\theta) - \frac{f_\sigma(\epsilon_{\text{in}})}{f_\sigma(\theta)}}{n_l+m_l+o^l}.\end{aligned}$$

Thus we have: $|S_{ij}| \leq \frac{\left(\frac{1}{f_\sigma(\theta)} - f_\sigma(\epsilon_{\text{in}})\right) \vee \left(\frac{f_\sigma(\epsilon_{\text{in}})}{f_\sigma(\theta)} - f_\sigma(\theta)\right)}{n_l+m_l+o^l}$, so that

$$\begin{aligned}|S_{ij}| &\leq \left[\left(\frac{1}{f_\sigma(\theta)} - f_\sigma(\epsilon_{\text{in}})\right) \vee \left(\frac{f_\sigma(\epsilon_{\text{in}})}{f_\sigma(\theta)} - f_\sigma(\theta)\right)\right](n_l+m_l)^{-1} \\ &= \left[\left(\eta_1+\eta_\theta+O(\eta_\theta^2)\right) \vee \left(2\eta_\theta-\eta_1+O(\eta_1^2+\eta_\theta^2)\right)\right](n_l+m_l)^{-1} \\ &\leq \left[2\eta_\theta+O(\eta_1^2+\eta_\theta^2)\right](n_l+m_l)^{-1}.\end{aligned}$$

Thus the norm of the spectral perturbation of $D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}} \xrightarrow{\mathbf{P_2}} B_l$ is bounded by

$$\begin{aligned}\|D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}} - B_l\|_2 &\leq (n_l+m_l)\|D_{\tilde{A}_l}^{-\frac{1}{2}} W_{\tilde{A}_l,\tilde{A}_l} D_{\tilde{A}_l}^{-\frac{1}{2}} - B_l\|_{\max} \\ &\leq (n_l+m_l)\left(\|Q\|_{\max} \vee \|R\|_{\max} \vee \|S\|_{\max}\right) \\ &\leq 2\eta_1+2\eta_\theta+O(\eta_1^2+\eta_\theta^2) := P_2^l.\end{aligned}$$

Defining $P_2 := \max_l P_2^l$, the perturbation of all eigenvalues due to $\mathbf{P_2}$ is bounded by P_2.

B.6. Bounding the Eigenvalues of $L_{\mathbf{SYM}}$

Since the eigenvalues of $\mathbf{1}_{n_l+m_l}$ are

$$\lambda_i(\mathbf{1}_{n_l+m_l}) = \begin{cases} 0 & i = 1,\ldots,n_l+m_l-1 \\ n_l+m_l & i = n_l+m_l \end{cases},$$

we have

$$\lambda_i(B_l) = \begin{cases} 0 & i = 1,\ldots,n_l+m_l-1 \\ \frac{(n_l+m_l)f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} & i = n_l+m_l. \end{cases}$$

Note that since the blocks B^l are orthogonal, the eigenvalues of B are simply the union of the eigenvalues of the blocks, and the eigenvalues of $I-B$ are obtained by

subtracting the eigenvalues of B from 1. Thus:

$$\lambda_i^l(I-B) = \begin{cases} 1 - \frac{(n_l+m_l)f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} & i=1,\ 1\le l\le K, \\ 1 & i=2,\dots,n_l+m_l,\ 1\le l\le K. \end{cases}$$

Since $|\lambda_i(L_{\text{SYM}}) - \lambda_i(I-B)| \le \|B - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\|_2 \le P_1+P_2$, and L_{SYM} is positive semi-definite, by the Hoffman-Wielandt Theorem (Stewart, 1990), the eigenvalues of L_{SYM} are:

$$\lambda_i^l(L_{\text{SYM}}) = \left(1 - \frac{(n_l+m_l)f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} \pm (P_1+P_2)\right) \vee 0, \qquad i=1,\ 1\le l\le K, \quad \text{(B.2)}$$
$$\lambda_i^l(L_{\text{SYM}}) = 1 \pm (P_1+P_2), \qquad i=2,\dots,n_l+m_l,\ 1\le l\le K.$$

where:

$$P_1 = \zeta_N\eta_2 + O(\zeta_N^2\eta_2^2+\eta_\theta^2) \quad , \quad P_2 = 2\eta_1+2\eta_\theta+O(\eta_1^2+\eta_\theta^2),$$
$$P_1+P_2 = 2\eta_1+2\eta_\theta+\zeta_N\eta_2+O(\eta_1^2+\eta_\theta^2+\zeta_N^2\eta_2^2).$$

B.7. The Largest Spectral Gap of L_{SYM}

For the remainder of the proof let $\{\lambda_i\}_{i=1}^N$ denote the eigenvalues of L_{SYM} sorted in increasing order, and $\Delta_i = \lambda_{i+1}-\lambda_i$ for $1\le i\le N-1$. Note the condition we will derive to guarantee Δ_K is the largest eigengap also ensures that the K smallest eigenvalues are given by (B.2).
Recalling the definition of ζ_N from Theorem 5.5, for $1\le i\le K$, we have

$$\begin{aligned} 0 \le \lambda_i &\le 1 - \frac{(n_l+m_l)f_\sigma(\epsilon_{\text{in}})}{n_l+w^l+o^l} + (P_1+P_2) \\ &\le 1 - \frac{(n_l+m_l)f_\sigma(\epsilon_{\text{in}})}{(n_l+m_l)+\sum_{s\ne l}(n_s+m_s)f_\sigma(\epsilon_{\text{nse}})} + (P_1+P_2) \\ &\le 1 - \frac{(1-\eta_1)}{1+\zeta_N\eta_2} + (P_1+P_2) \\ &= 1-(1-\eta_1)(1-\zeta_N\eta_2+O(\zeta_N^2\eta_2^2)) + (P_1+P_2) \\ &= \eta_1+\zeta_N\eta_2+(P_1+P_2)+O(\eta_1^2+\zeta_N^2\eta_2^2)). \end{aligned}$$

Thus for $i<K$, the gap is bounded by: $\Delta_i \le \lambda_{i+1} \le \eta_1+\zeta_N\eta_2+(P_1+P_2)+O(\eta_1^2+\zeta_N^2\eta_2^2))$. For $i>K$, we have $\Delta_i \le 1+(P_1+P_2)-(1-(P_1+P_2)) \le 2(P_1+P_2)$. Finally for $i=K$:

$$\begin{aligned} \Delta_K &\ge 1-(P_1+P_2)-(\eta_1+\zeta_N\eta_2+P_1+P_2)+O(\eta_1^2+\zeta_N^2\eta_2^2)) \\ &\ge 1-\eta_1-\zeta_N\eta_2-2(P_1+P_2)+O(\eta_1^2+\zeta_N^2\eta_2^2)). \end{aligned}$$

Thus Δ_K is the largest gap if $\frac{1}{2} \ge \eta_1+\zeta_N\eta_2+2(P_1+P_2)+O(\eta_1^2+\zeta_N^2\eta_2^2) = 5\eta_1+4\eta_\theta+6\zeta_N\eta_2+O(\eta_1^2+\eta_\theta^2+\zeta_N^2\eta_2^2)$, which is the condition of Theorem 5.5.

B.8. Bounding the Spectral Embedding and Labeling Accuracy

We apply Theorem 2 from Fan et al. (2018) to bound the eigenvector perturbation. We let $\Phi = (\phi_1 \dots \phi_K)$ denote the N by K matrix whose columns are the top K eigenvectors of B (defined in B.3), ordered so that ϕ_l corresponds to the block B_l. We let $\tilde{\Phi}$ be the equivalent quantity for $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Defining the coherence of Φ as $\mathrm{coh}(\Phi) = (N/K)\max_i \sum_{j=1}^{K} \Phi_{ij}^2$, we note that

$$\mathrm{coh}(\Phi) \leq \frac{N}{K}\left(\|\phi_1\|_\infty^2 + \dots + \|\phi_K\|_\infty^2\right) \leq \frac{N}{K} \cdot \frac{K\zeta_N}{N} = \zeta_N,$$

i.e. Φ has low coherence since each eigenvector is constant on a cluster. Thus by Theorem 2 from Fan et al. (2018), there exists a rotation R such that

$$\|\tilde{\Phi}R - \Phi\|_{\max} = O\left(\frac{K^{\frac{5}{2}}\zeta_N^2\|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - B\|_\infty}{\lambda_K(B)\sqrt{N}}\right).$$

We recall from Section B.6 that

$$\lambda_K(B) \geq \min_{1\leq l\leq K} \frac{(n_l+m_l)(1-\eta_1)}{n_l+m_l+N\eta_2} = \frac{1-\eta_1}{1+\zeta_N\eta_2} = 1-\eta_1-\zeta_N\eta_2 + O(\eta_1^2+\zeta_N^2\eta_2^2).$$

Letting C_l denote the diagonal blocks of C and using the bounds computed in Sections B.4 and B.5, we have:

$$\begin{aligned}
\|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - B\|_\infty &\leq \|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - C\|_\infty + \max_l \|C_l - B_l\|_\infty \\
&\leq N\|D^{-\frac{1}{2}}WD^{-\frac{1}{2}} - C\|_{\max} + \max_l\ (n_l+m_l)\|C_l - B_l\|_{\max} \\
&\leq 2\eta_1 + 2\eta_\theta + \zeta_N\eta_2 + O(\eta_1^2+\eta_\theta^2+\zeta_N^2\eta_2^2).
\end{aligned}$$

We conclude that

$$\|\tilde{\Phi}R - \Phi\|_{\max} \leq \frac{cK^{\frac{5}{2}}\zeta_N^2}{\sqrt{N}}\left[\eta_1+\eta_\theta+\zeta_N\eta_2+O(\eta_1^2+\eta_\theta^2+\zeta_N^2\eta_2^2)\right] := P_3.$$

for some absolute constant c. Letting $\{r_i\}_{i=1}^N$ denote the rows of Φ and $\{\tilde{r}_i\}_{i=1}^N$ denote the rows of $\tilde{\Phi}R$, we have $\|r_i - \tilde{r}_i\|_2 \leq \sqrt{K}\|\tilde{\Phi}R - \Phi\|_{\max} \leq \sqrt{K}P_3$ for all i. Letting $\pi(i) \in \{1,\dots,K\}$ denote the index of the set $\tilde{A}_l$ which contains the point corresponding to the i^{th} row, we have $r_i = [0\dots(n_{\pi(i)}+m_{\pi(i)})^{-1/2}\dots 0]$ for all i, where the non-zero element occurs in the $\pi(i)^{\text{th}}$ column. Thus the spectral embedding maps all points in $\tilde{A}_l$ inside a sphere in $\mathbb{R}^K$ centered at $z_l = [0\dots(n_l+m_l)^{-1/2}\dots 0]$ with radius $\sqrt{K}P_3$. When $l \neq s$, we have $\|z_l - z_s\|_2 \geq \sqrt{\frac{2}{N}}$. Thus $\sqrt{\frac{2}{N}} > 10\sqrt{K}P_3$ is sufficient to ensure that these spheres are well separated, i.e. the embedding is a perfect representation (see Definition 5.4) of the clusters $\tilde{A}_l$ with $r = 2\sqrt{K}P_3$. Simplifying this condition, we thus obtain perfect label accuracy by clustering by distances on the spectral embedding whenever

$$\frac{1}{K^3\zeta_N^2} \gtrsim \eta_1+\eta_\theta+\zeta_N\eta_2+O(\eta_1^2+\eta_\theta^2+\zeta_N^2\eta_2^2).$$

References

E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

F. Alimoglu and E. Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *TAINN*. Citeseer, 1996.

N. Alon and B. Schieber. *Optimal preprocessing for answering on-line product queries*. Tel-Aviv University. The Moise and Frida Eskenasy Institute of Computer Sciences, 1987.

M. Appel and R. Russo. The maximum vertex degree of a graph on uniform points in $[0,1]^d$. *Advances in Applied Probability*, 29(3):567–581, 1997a.

M. Appel and R. Russo. The minimum vertex degree of a graph on uniform points in $[0,1]^d$. *Advances in Applied Probability*, 29(3):582–594, 1997b.

M. Appel and R. Russo. The connectivity of a graph on uniform points on $[0,1]^d$. *Statistics & Probability Letters*, 60(4):351–357, 2002.

E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transactions on Information Theory*, 57(3):1692–1706, 2011.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587, 2011.

E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local PCA. *Journal of Machine Learning Research*, 18(9):1–57, 2017.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *SODA*, pages 1027–1035. SIAM, 2007.

A. Azran and Z. Ghahramani. A new approach to data driven clustering. In *ICML*, pages 57–64. ACM, 2006a.

A. Azran and Z. Ghahramani. Spectral methods for automatic multiscale data clustering. In *CVPR*, volume 1, pages 190–197. IEEE, 2006b.

S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *NIPS*, pages 954–962, 2011.

S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. In *NIPS*, pages 2679–2687, 2013.

M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23, 1999.

R.E. Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, 2015.

J.J. Benedetto and W. Czaja. *Integration and modern analysis*. Springer Science & Business Media, 2010.

J.L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *ICML*, pages 97–104. ACM, 2006.

R.B. Bhatt, G. Sharma, A. Dhall, and S. Chaudhury. Efficient skin region segmentation using low complexity fuzzy decision tree model. In *INDICON*, pages 1–4. IEEE, 2009.

R.L. Bishop and R.J. Crittenden. *Geometry of manifolds*, volume 15. Academic press, 2011.

P.M. Camerini. The min-max spanning tree problem and some extensions. *Information Processing Letters*, 7(1):10–14, 1978.

H. Chang and D.-Y. Yeung. Robust path-based spectral clustering. *Pattern Recognition*, 41(1):191–203, 2008.

K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *NIPS*, pages 343–351, 2010.

G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9(5):517–558, 2009a.

G. Chen and G. Lerman. Spectral curvature clustering (SCC). *International Journal of Computer Vision*, 81(3):317–330, 2009b.

F. Chung. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.

R.R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

E.D. Demaine, G.M. Landau, and O. Weimann. On Cartesian trees and range minimum queries. In *ICALP*, pages 341–353. Springer, 2009.

E.D. Demaine, G.M. Landau, and O. Weimann. On Cartesian trees and range minimum queries. *Algorithmica*, 68(3):610–625, 2014.

E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (11):2765–2781, 2013.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

J. Fan, W. Wang, and Y. Zhong. An ℓ^∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

B. Fischer and J.M. Buhmann. Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):513–518, 2003.

B. Fischer, T. Zöller, and J. Buhmann. Path based pairwise data clustering with application to texture segmentation. In *Energy minimization methods in computer vision and pattern recognition*, pages 235–250. Springer, 2001.

B. Fischer, V. Roth, and J.M. Buhmann. Clustering with the connectivity kernel. In *NIPS*, pages 89–96, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in Statistics Springer, Berlin, 2001.

H. Gabow and R.E. Tarjan. Algorithms for two bottleneck optimization problems. *Journal of Algorithms*, 9:411–417, 1988.

N. Garcia Trillos and D. Slepcev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220(1):193–241, 2016a.

N. Garcia Trillos and D. Slepcev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 2016b.

N. Garcia Trillos, D. Slepcev, J. Von Brecht T. Laurent, and X. Bresson. Consistency of Cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 17(181): 1–46, 2016.

N. Garcia Trillos, M. Gerlach, M. Hein, and D. Slepčev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, pages 1–61, 2019a.

N. Garcia Trillos, F. Hoffmann, and B. Hosseini. Geometric structure of graph Laplacian embeddings. *arXiv preprint arXiv:1901.10651*, 2019b.

E.N. Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

J.M. González-Barrios and A.J. Quiroz. A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree. *Statistics & Probability Letters*, 62(1):23–34, 2003.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33(6):305–308, 1990.

J.A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Society*, 76(374):388–394, 1981.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2009.

T.C. Hu. Letter to the editor: The maximum capacity route problem. *Operations Research*, 9(6):898–900, 1961.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

M. Maggioni and J.M. Murphy. Learning by unsupervised nonlinear diffusion. *Journal of Machine Learning Research*, 20(160):1–56, 2019.

D. McKenzie and S. Damelin. Power weighted shortest paths for clustering Euclidean data. *Foundations of Data Science*, 1(3):307, 2019.

G.J. McLachlan and K.E. Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

M. Meila and J. Shi. Learning segmentation by random walks. In *NIPS*, pages 873–879, 2001.

D.G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

J.M. Murphy and M. Maggioni. Diffusion geometric methods for fusion of remotely sensed data. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIV*, volume 10644, page 106440I. International Society for Optics and Photonics, 2018.

J.M. Murphy and M. Maggioni. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1829–1845, 2019.

S.A. Nene, S.K. Nayar, and H. Murase. Columbia object image library (COIL-20). 1996.

A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2002.

R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *FOCS*, pages 165–176. IEEE, 2006.

H.S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, 2009.

L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

M. Penrose. The longest edge of the random minimal spanning tree. *Annals of Applied Probability*, 7(2):340–361, 1997.

R. Penrose. A strong law for the longest edge of the minimal spanning tree. *Annals of Probability*, 27(1):246–260, 1999.

M. Pollack. Letter to the editor: The maximum capacity through a network. *Operations Research*, 8(5):733–736, 1960.

A.P. Punnen. A linear time algorithm for the maximum capacity path problem. *European Journal of Operational Research*, 53:402–404, 1991.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

F.D. Roberts and S.H. Storey. A three-dimensional cluster problem. *Biometrika*, 55 (1):258–260, 1968.

A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

G. Sanguinetti, J. Laidler, and N.D. Lawrence. Automatic determination of the number of clusters using spectral algorithms. In *MLSP*, pages 55–60. IEEE, 2005.

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

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

R. Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

M. Soltanolkotabi and E.J. Candès. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

M. Soltanolkotabi, E. Elhamifar, and E.J. Candès. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

B. Sriperumbudur and I. Steinwart. Consistency and rates for clustering with DBSCAN. In *AISTATS*, pages 1090–1098, 2012.

D. Stauffer and A. Aharony. *Introduction to Percolation Theory*. CRC Press, 1994.

H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.*, 4(12):801–804, 1957.

G.W Stewart. Matrix perturbation theory. *Citeseer*, 1990.

G. Szegö. Inequalities for certain eigenvalues of a membrane of given area. *Journal of Rational Mechanics and Analysis*, 3:343–356, 1954.

L.N. Trefethen and D. Bau. *Numerical linear algebra*, volume 50. SIAM, 1997.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

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

V. Vu. A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

X. Wang, K. Slavakis, and G. Lerman. Riemannian multi-manifold modeling. *arXiv preprint arXiv:1410.0095*, 2014.

H.F. Weinberger. An isoperimetric inequality for the N-dimensional free membrane problem. *Journal of Rational Mechanics and Analysis*, 5(4):633–636, 1956.

X. Xu, M. Ester, H.-P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *ICDE*, pages 324–331. IEEE, 1998.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, pages 1601–1608, 2004.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100(3):217–240, 2012.