Journal of Machine Learning Research 18 (2017) 1-57 Submitted 8/14; Revised 7/16; Published 3/17

Spectral Clustering Based on Local PCA

Ery Arias-Castro EARIASCA@UCSD.EDU
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093, USA

Gilad Lerman LERMAN@UMN.EDU
Department of Mathematics
University of Minnesota, Twin Cities
Minneapolis, MN 55455, USA

Teng Zhang TENG.ZHANG@UCF.EDU
Department of Mathematics
University of Central Florida
Orlando, FL 32816, USA

Editor: Mikhail Belkin

Abstract

We propose a spectral clustering method based on local principal components analysis (PCA). After performing local PCA in selected neighborhoods, the algorithm builds a nearest neighbor graph weighted according to a discrepancy between the principal subspaces in the neighborhoods, and then applies spectral clustering. As opposed to standard spectral methods based solely on pairwise distances between points, our algorithm is able to resolve intersections. We establish theoretical guarantees for simpler variants within a prototypical mathematical framework for multi-manifold clustering, and evaluate our algorithm on various simulated data sets.

Keywords: multi-manifold clustering, spectral clustering, local principal component analysis, intersecting clusters

1. Introduction

The task of multi-manifold clustering, where the data are assumed to be located near surfaces embedded in Euclidean space, is relevant in a variety of applications. In cosmology, it arises as the extraction of galaxy clusters in the form of filaments (curves) and walls (surfaces) (Valdarnini, 2001; Martínez and Saar, 2002); in motion segmentation, moving objects tracked along different views form affine or algebraic surfaces (Ma et al., 2008; Fu et al., 2005; Vidal and Ma, 2006; Chen et al., 2009); this is also true in face recognition, in the context of images of faces in fixed pose under varying illumination conditions (Ho et al., 2003; Basri and Jacobs, 2003; Epstein et al., 1995).

We consider a stylized setting where the underlying surfaces are nonparametric in nature, with a particular emphasis on situations where the surfaces intersect. Specifically, we assume the surfaces are smooth, for otherwise the notion of continuation is potentially ill-posed. For example, without smoothness assumptions, an L-shaped cluster is indistinguishable from the union of two line-segments meeting at right angle.

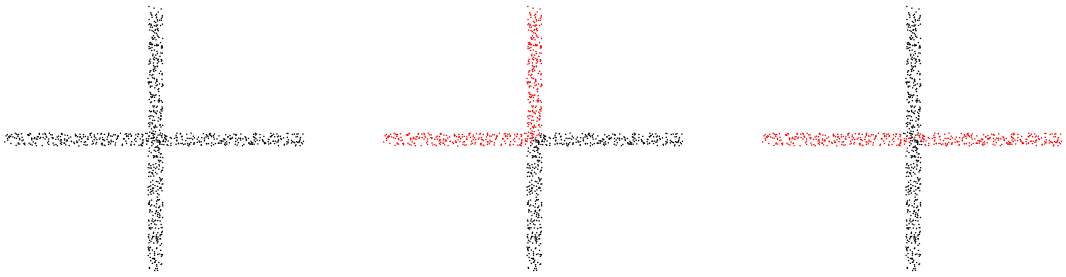

Figure 1: Two rectangular clusters intersecting at right angle. Left: the original data. Center: a typical output of the standard spectral clustering method of Ng et al. (2002), which is generally unable to resolve intersections. Right: a typical output of our method.

Spectral methods (Luxburg, 2007) are particularly suited for nonparametric settings, where the underlying clusters are usually far from convex, making standard methods like K-means irrelevant. However, a drawback of standard spectral approaches such as the well-known variant of Ng, Jordan, and Weiss (2002) is their inability to separate intersecting clusters. Indeed, consider the simplest situation where two straight clusters intersect at right angle, pictured in Figure 1 below. The algorithm of Ng et al. (2002) is based on pairwise affinities that are decreasing in the distances between data points, making it insensitive to smoothness and, therefore, intersections. And indeed, this algorithm typically fails to separate intersecting clusters, even in the easiest setting of Figure 1.

As argued in (Agarwal et al., 2005, 2006; Shashua et al., 2006), a multiway affinity is needed to capture complex structure in data (here, smoothness) beyond proximity attributes. For example, Chen and Lerman (2009a) use a flatness affinity in the context of *hybrid linear modeling*, where the surfaces are assumed to be affine subspaces, and subsequently extended to algebraic surfaces via the 'kernel trick' (Chen, Atev, and Lerman, 2009). Moving beyond parametric models, Arias-Castro, Chen, and Lerman (2011) introduce a localized measure of flatness.

Continuing this line of work, we suggest a spectral clustering method based on the estimation of the local linear structure (tangent bundle) via local principal component analysis (PCA). The idea of using local PCA combined with spectral clustering has precedents in the literature. In particular, our method is inspired by the work of Goldberg, Zhu, Singh, Xu, and Nowak (2009), where the authors develop a spectral clustering method within a semi-supervised learning framework. This approach is in the zeitgeist. While writing this paper, we became aware of two concurrent publications, by Wang, Jiang, Wu, and Zhou (2011) and by Gong, Zhao, and Medioni (2012), both proposing approaches very similar to ours.[1] We also mention the multiscale, spectral-flavored algorithm of Kushnir, Galun, and Brandt (2006), which is also based on local PCA. We comment on these spectral methods

1. The present paper was indeed developed in parallel with these two, first as a short version submitted to ICML in 2011.

in more detail later on. In fact, an early proposal also based on local PCA appears in the literature on subspace clustering (Fan et al., 2006)—although the need for localization is perhaps less intuitive in this setting.

The basic proposition of local PCA combined with spectral clustering has two main stages. The first one forms an affinity between a pair of data points that takes into account both their Euclidean distance and a measure of discrepancy between their tangent spaces. Each tangent space is estimated by PCA in a local neighborhood around each point. The second stage applies standard spectral clustering with this affinity. As a reality check, this relatively simple algorithm succeeds at separating the straight clusters in Figure 1. We tested our algorithm in more elaborate settings, some of them described in Section 4.

Other methods with a spectral component include those of Polito and Perona (2001) and Goh and Vidal (2007), which (roughly speaking) embed the points by a variant of LLE (Saul and Roweis, 2003) and then group the points by K-means clustering. There is also the method of Elhamifar and Vidal (2011), which chooses the neighborhood of each point by computing a sparse linear combination of the remaining points followed by an application of the spectral graph partitioning algorithm of Ng et al. (2002). Note that these methods work under the assumption that the surfaces do not intersect.

Besides spectral-type approaches to multi-manifold clustering, other methods appear in the literature. Among these, Gionis et al. (2005) and Haro et al. (2007) allow for intersecting surfaces but assume that they have different intrinsic dimension or density—and the proposed methodology is entirely based on such assumptions. We also mention the K-manifold method of Souvenir and Pless (2005), which propose an EM-type algorithm; and that of Guo et al. (2007), which propose to minimize an energy functional based on pairwise distances and local curvatures, leading to a combinatorial optimization.

Our contribution is the design and detailed study of a prototypical spectral clustering algorithm based on local PCA, tailored to settings where the underlying clusters come from sampling in the vicinity of smooth surfaces that may intersect. We endeavored to simplify the algorithm as much as possible without sacrificing performance. We provide theoretical results for simpler variants within a standard mathematical framework for multi-manifold clustering. To our knowledge, these are the first mathematically backed successes at the task of resolving intersections in the context of multi-manifold clustering, with the exception of (Arias-Castro et al., 2011), where the corresponding algorithm is shown to succeed at separating intersecting curves. The salient features of our algorithm are illustrated via numerical experiments.

The rest of the paper is organized as follows. In Section 2, we introduce our method in various variants. In Section 3, we analyze the simpler variants in a standard mathematical framework for multi-manifold learning. In Section 4, we perform some numerical experiments illustrating several features of our approach. In Section 5, we discuss possible extensions.

2. The Methodology

We introduce our algorithm and simpler variants that are later analyzed in a mathematical framework. We start with some review of the literature, zooming in on the most closely related publications.

2.1 Some Precedents

Using local PCA within a spectral clustering algorithm was implemented in four other publications we know of (Goldberg et al., 2009; Kushnir et al., 2006; Gong et al., 2012; Wang et al., 2011). As a first stage in their semi-supervised learning method, Goldberg, Zhu, Singh, Xu, and Nowak (2009) design a spectral clustering algorithm. The method starts by subsampling the data points, obtaining 'centers' in the following way. Draw $\boldsymbol{y}_1$ at random from the data and remove its ℓ-nearest neighbors from the data. Then repeat with the remaining data, obtaining centers $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots$. Let $\boldsymbol{C}_i$ denote the sample covariance in the neighborhood of $\boldsymbol{y}_i$ made of its ℓ-nearest neighbors. An m-nearest-neighbor graph is then defined on the centers in terms of the Mahalanobis distances. Explicitly, the centers $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are connected in the graph if $\boldsymbol{y}_j$ is among the m nearest neighbors of $\boldsymbol{y}_i$ in Mahalanobis distance

$$\|\boldsymbol{C}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{y}_j)\|, \tag{1}$$

or vice-versa. The parameters ℓ and m are both chosen of order $\log n$. An existing edge between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ is then weighted by $\exp(-H_{ij}^2/\eta^2)$, where H_{ij} denotes the Hellinger distance between the probability distributions $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_i)$ and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_j)$. The spectral graph partitioning algorithm of Ng, Jordan, and Weiss (2002)—detailed in Algorithm 1—is then applied to the resulting affinity matrix, with some form of constrained K-means. We note that Goldberg et al. (2009) evaluate their method in the context of semi-supervised learning where the clustering routine is only required to return subclusters of actual clusters. In particular, the data points other than the centers are discarded. Note also that their evaluation is empirical.

Algorithm 1 Spectral Graph Partitioning (Ng, Jordan, and Weiss, 2002)

Input:

Affinity matrix $\boldsymbol{W} = (W_{ij})$, size of the partition K

Steps:

1: Compute $\boldsymbol{Z} = (Z_{ij})$ according to $Z_{ij} = W_{ij}/\sqrt{\Delta_i \Delta_j}$, with $\Delta_i := \sum_{j=1}^n W_{ij}$.
2: Extract the top K eigenvectors of $\boldsymbol{Z}$.
3: Renormalize each row of the resulting $n \times K$ matrix.
4: Apply K-means to the row vectors.

The algorithm proposed by Kushnir, Galun, and Brandt (2006) is multiscale and works by coarsening the neighborhood graph and computing sampling density and geometric information inferred along the way such as obtained via PCA in local neighborhoods. This bottom-up flow is then followed by a top-down pass, and the two are iterated a few times. The algorithm is too complex to be described in detail here, and probably too complex to be analyzed mathematically. The clustering methods of Goldberg et al. (2009) and ours can be seen as simpler variants that only go bottom up and coarsen the graph only once.

In the last stages of writing this paper, we learned of the works of Wang, Jiang, Wu, and Zhou (2011) and Gong, Zhao, and Medioni (2012), who propose algorithms very similar to our Algorithm 3 detailed below. Note that these publications do not provide any theoretical guarantees for their methods, which is one of our main contributions here.

2.2 Our Algorithms

We now describe our method and propose several variants. Our setting is standard: we observe data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^D$ that we assume were sampled in the vicinity of K smooth surfaces embedded in $\mathbb{R}^D$. The setting is formalized later in Section 3.1.

2.2.1 Connected Component Extraction: Comparing Local Covariances

We start with our simplest variant, which is also the most natural. The method depends on a neighborhood radius $r > 0$, a spatial scale parameter $\varepsilon > 0$ and a covariance (relative) scale $\eta > 0$. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ denotes its Euclidean norm, and for a (square) matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes its spectral norm. For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \ldots, n\}$. Given a data set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, for any point $\boldsymbol{x} \in \mathbb{R}^D$ and $r > 0$, define the neighborhood

$$N_r(\boldsymbol{x}) = \{\boldsymbol{x}_j : \|\boldsymbol{x} - \boldsymbol{x}_j\| \leq r\}.$$

Algorithm 2 Connected Component Extraction: Comparing Covariances

Input:

Data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$; neighborhood radius $r > 0$; spatial scale $\varepsilon > 0$, covariance scale $\eta > 0$.

Steps:

1: For each $i \in [n]$, compute the sample covariance matrix $\boldsymbol{C}_i$ of $N_r(\boldsymbol{x}_i)$.

2: Remove $\boldsymbol{x}_i$ when there is $\boldsymbol{x}_j$ such that $\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \leq r$ and $\|\boldsymbol{C}_j - \boldsymbol{C}_i\| > \eta r^2$.

3: Compute the following affinities between data points:

$$W_{ij} = \mathbb{1}_{\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \varepsilon\}} \cdot \mathbb{1}_{\{\|\boldsymbol{C}_i - \boldsymbol{C}_j\| \leq \eta r^2\}}. \tag{2}$$

4: Extract the connected components of the resulting graph.

5: Each point removed in Step 2 is grouped with the closest point that survived Step 2.

In words, the algorithm first computes local covariances (Step 1). It removes points that are believed to be very near an intersection (Step 2)—we elaborate on this below. With the remaining points, it creates an unweighted graph: the nodes of this graph are the data points and edges are formed between two nodes if both the distance between these nodes and the distance between the local covariance structures at these nodes are sufficiently small (Step 3). The connected components of the resulting graph are extracted (Step 4) and the points that survived Step 2 are labeled accordingly. Each point removed in Step 2 receives the label of its closest labeled point (Step 5).

In principle, the neighborhood size r is chosen just large enough that performing PCA in each neighborhood yields a reliable estimate of the local covariance structure. For this, the number of points inside the neighborhood needs to be large enough, which depends on the sample size n, the sampling density, intrinsic dimension of the surfaces and their surface area (Hausdorff measure), how far the points are from the surfaces (i.e., noise level), and the regularity of the surfaces. The spatial scale parameter ε depends on the sampling density and r. It is meant to be larger than r and needs to be large enough that a point has plenty

of points within distance ε, including some across an intersection, so each cluster is strongly connected. At the same time, ε needs to be small enough that a local linear approximation to the surfaces is a relevant feature of proximity. Its choice is rather similar to the choice of the scale parameter in standard spectral clustering (Ng et al., 2002; Zelnik-Manor and Perona, 2005). The covariance scale η needs to be large enough that centers from the same cluster and within distance ε of each other have local covariance matrices within distance ηr^2, but small enough that points from different clusters near their intersection have local covariance matrices separated by a distance substantially larger than ηr^2. This depends on the curvature of the surfaces and the incidence angle at the intersection of two (or more) surfaces. Note that a typical covariance matrix over a ball of radius r has norm of order r^2, which justifies using our choice of parametrization. In the mathematical framework we introduce later on, these parameters can be chosen automatically as done in (Arias-Castro et al., 2011), at least when the points are sampled exactly on the surfaces. We will not elaborate on that since in practice this does not inform our choice of parameters.

The rationale behind Step 2 is as follows. As we just discussed, the parameters need to be tuned so that points from the same cluster and within distance ε have local covariance matrices within distance ηr^2. Strictly speaking, this is true of points away from the boundary of the underlying surface. Hence, although $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in Step 2 are meant to be from different clusters, they could be from the same surface near its boundary. In the former situation, since they are near each other, in our model this will imply that they are close to an intersection. Therefore, roughly speaking, Step 2 removes points near an intersection, but also points near the boundaries of the underlying surfaces. The reason that we require $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to be within distance r, as opposed to ε, is because otherwise removing the points in Step 2 would create a "gap" of ε near an intersection which then cannot be bridged in Steps 3-4. (Alternatively, one could replace r with $\xi \in (r, \varepsilon)$ in Step 2, but this would add this additional parameter ξ to the algorithm.) This step is in fact crucial as the local covariance varies smoothly along the intersection of two smooth surfaces.

Although this method works in simple situations like that of two intersecting segments (Figure 1), it is not meant to be practical. Indeed, extracting connected components is known to be sensitive to spurious points and therefore unstable. Furthermore, we found that comparing local covariance matrices as in affinity (2) tends to be less stable than comparing local projections as in affinity (3), which brings us to our next variant.

2.2.2 Connected Component Extraction: Comparing Local Projections

We present another variant also based on extracting the connected components of a neighborhood graph that compares orthogonal projections onto the largest principal directions. See Algorithm 3.

We note that the local intrinsic dimension is determined by thresholding the eigenvalues of the local covariance matrix, keeping the directions with eigenvalues within some range of the largest eigenvalue. The same strategy is used by Kushnir et al. (2006), but with a different threshold. The method is a hard version of what we implemented, which we describe in Algorithm 4.

We note that neither algorithm includes an intersection-removal step as Step 2 in Algorithm 2. The reason is that the algorithms work without such a step. Indeed, we show in

Algorithm 3 Connected Component Extraction: Comparing Projections

Input:

Data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$; neighborhood radius $r > 0$, spatial scale $\varepsilon > 0$, projection scale $\eta > 0$.

Steps:

1: For each $i \in [n]$, compute the sample covariance matrix $\boldsymbol{C}_i$ of $N_r(\boldsymbol{x}_i)$.

2: Compute the projection $\boldsymbol{Q}_i$ onto the eigenvectors of $\boldsymbol{C}_i$ with corresponding eigenvalue exceeding $\sqrt{\eta}\,\|\boldsymbol{C}_i\|$.

3: Compute the following affinities between data points:

$$W_{ij} = \mathbb{1}_{\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \le \varepsilon\}} \cdot \mathbb{1}_{\{\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| \le \eta\}}. \tag{3}$$

4: Extract the connected components of the resulting graph.

Theorem 1 that Algorithm 3 is able to separate the clusters—the only drawback is that it may possibly treat the intersection of two surfaces as a cluster. And we show via numerical experiments in Section 4 that Algorithm 4 performs well in a number of situations.

2.2.3 Covariances or Projections?

In our numerical experiments, we tried working both directly with covariance matrices as in (2) and with projections as in (3). Note that in our experiments we used spectral graph partitioning with soft versions of these affinities, as described in Section 2.2.4. We found working with projections to be more reliable. The problem comes, in part, from boundaries. When a surface has a boundary, local covariances over neighborhoods that overlap with the boundary are quite different from local covariances over nearby neighborhoods that do not touch the boundary. Consider the example of two segments, S_1 and S_2, intersecting at an angle of $\theta \in (0, \pi/2)$ at their middle point, specifically

$$S_1 = [-1, 1] \times \{0\}, \qquad S_2 = \{(x, x\tan\theta) : x \in [-\cos\theta, \cos\theta]\}.$$

Assume there is no noise and that the sampling is uniform. Assume $r \in (0, \frac{1}{2}\sin\theta)$ so that the disc centered at $\boldsymbol{x}_1 := (1/2, 0)$ does not intersect S_2, and the disc centered at $\boldsymbol{x}_2 := (\frac{1}{2}\cos\theta, \frac{1}{2}\sin\theta)$ does not intersect S_1. Let $\boldsymbol{x}_0 = (1, 0)$. For $\boldsymbol{x} \in S_1 \cup S_2$, let $\boldsymbol{C}_{\boldsymbol{x}}$ denote the local covariance at $\boldsymbol{x}$ over a ball of radius $r < \frac{1}{2}\sin\theta$, so that the neighborhoods of $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$ are "pure". The situation is described in Figure 2.

Simple calculations yield

$$\boldsymbol{C}_{\boldsymbol{x}_0} = \frac{r^2}{12}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}_{\boldsymbol{x}_1} = \frac{r^2}{3}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}_{\boldsymbol{x}_2} = \frac{r^2}{3}\begin{pmatrix} \cos^2\theta & \sin(\theta)\cos(\theta) \\ \sin(\theta)\cos(\theta) & \sin^2\theta \end{pmatrix},$$

so that

$$\|\boldsymbol{C}_{\boldsymbol{x}_0} - \boldsymbol{C}_{\boldsymbol{x}_1}\| = \frac{r^2}{4}, \quad \|\boldsymbol{C}_{\boldsymbol{x}_1} - \boldsymbol{C}_{\boldsymbol{x}_2}\| = \frac{r^2}{3}\sin\theta.$$

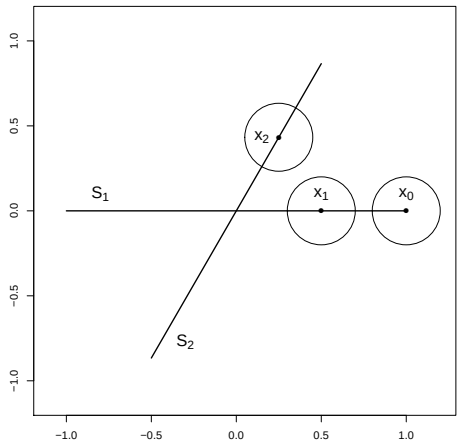

Figure 2: Two segments intersecting. The local covariances (within the disc neighborhoods drawn) at $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are closer than the local covariances at $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$, even though $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are on the same segment.

Therefore, when $\sin\theta \leq \frac{3}{4}$ (roughly, $\theta \leq 48^o$), the local covariances at $\boldsymbol{x}_0, \boldsymbol{x}_1 \in S_1$ are farther (in operator norm) than those at $\boldsymbol{x}_1 \in S_1$ and $\boldsymbol{x}_2 \in S_2$. As for projections, however,

$$\boldsymbol{Q}_{\boldsymbol{x}_0} = \boldsymbol{Q}_{\boldsymbol{x}_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{Q}_{\boldsymbol{x}_2} = \begin{pmatrix} \cos^2\theta & \sin(\theta)\cos(\theta) \\ \sin(\theta)\cos(\theta) & \sin^2\theta \end{pmatrix},$$

so that

$$\|\boldsymbol{Q}_{\boldsymbol{x}_0} - \boldsymbol{Q}_{\boldsymbol{x}_1}\| = 0, \quad \|\boldsymbol{Q}_{\boldsymbol{x}_1} - \boldsymbol{Q}_{\boldsymbol{x}_2}\| = \sqrt{2}\sin\theta.$$

While in theory points within distance r from the boundary account for a small portion of the sample, in practice this is not the case, at least not with the sample sizes that we are able to rapidly process. In fact, we find that spectral graph partitioning is challenged by having points near the boundary that are far (in affinity) from nearby points from the same cluster. This may explain why the (soft version of) affinity (3) yields better results than the (soft version of) affinity (2) in our experiments.

2.2.4 Spectral Clustering Based on Local PCA

The following variant is more robust in practice and is the algorithm we actually implemented. The method assumes that the surfaces are of same dimension d and that there are K of them, with both parameters K and d known.

We note that $\boldsymbol{y}_1, \dots, \boldsymbol{y}_{n_0}$ forms an r-packing of the data. The underlying rationale for this coarsening is justified in (Goldberg et al., 2009) by the fact that the covariance matrices, and also the top principal directions, change smoothly with the location of the neighborhood, so that without subsampling these characteristics would not help detect the abrupt event of an intersection. The affinity (4) is of course a soft version of (3).

Algorithm 4 Spectral Clustering Based on Local PCA

Input:

Data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$; neighborhood radius $r > 0$; spatial scale $\varepsilon > 0$, projection scale $\eta > 0$; intrinsic dimension d; number of clusters K.

Steps:

0: Pick one point $\boldsymbol{y}_1$ at random from the data. Pick another point $\boldsymbol{y}_2$ among the data points not included in $N_r(\boldsymbol{y}_1)$, and repeat the process, selecting centers $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_0}$.

1: For each $i = 1, \ldots, n_0$, compute the sample covariance matrix $\boldsymbol{C}_i$ of $N_r(\boldsymbol{y}_i)$. Let $\boldsymbol{Q}_i$ denote the orthogonal projection onto the space spanned by the top d eigenvectors of $\boldsymbol{C}_i$.

2: Compute the following affinities between center pairs:

$$W_{ij} = \exp\left(-\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{\varepsilon^2}\right) \cdot \exp\left(-\frac{\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\|^2}{\eta^2}\right). \tag{4}$$

3: Apply spectral graph partitioning (Algorithm 1) to $\boldsymbol{W}$.

4: The data points are clustered according to the closest center in Euclidean distance.

2.2.5 Comparison with Closely Related Methods

We highlight some differences with the other proposals in the literature. We first compare our approach to that of Goldberg et al. (2009), which was our main inspiration.

- *Neighborhoods.* Comparing with Goldberg et al. (2009), we define neighborhoods over r-balls instead of ℓ-nearest neighbors, and connect points over ε-balls instead of m-nearest neighbors. This choice is for convenience, as these ways are in fact essentially equivalent when the sampling density is fairly uniform. This is elaborated at length in (Maier et al., 2009; Brito et al., 1997; Arias-Castro, 2011).

- *Mahalanobis distances.* Goldberg et al. (2009) use Mahalanobis distances (1) between centers. In our version, we could for example replace the Euclidean distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ in the affinity (2) with the average Mahalanobis distance

$$\|\boldsymbol{C}_i^{-1/2}(\boldsymbol{x}_i - \boldsymbol{x}_j)\| + \|\boldsymbol{C}_j^{-1/2}(\boldsymbol{x}_j - \boldsymbol{x}_i)\|.$$

  We actually tried this and found that the algorithm was less stable, particularly under low noise. Introducing a regularization in this distance—which requires the introduction of another parameter—solves this problem, at least partially.

  That said, using Mahalanobis distances makes the procedure less sensitive to the choice of ε, in that neighborhoods may include points from different clusters. Think of two parallel line segments separated by a distance of δ, and assume there is no noise, so the points are sampled exactly from these segments. Assuming an infinite sample size, the local covariance is the same everywhere so that points within distance ε are connected by the affinity (2). Hence, Algorithm 2 requires that $\varepsilon < \delta$. In terms of Mahalanobis distances, points on different segments are infinitely separated, so a

version based on these distances would work with any $\varepsilon > 0$. In the case of curved surfaces and/or noise, the situation is similar, though not as evident. Even then, the gain in performance is not obvious since we only require that ε be slightly larger in order of magnitude than r.

- *Hellinger distances.* As we mentioned earlier, Goldberg et al. (2009) use Hellinger distances of the probability distributions $\mathcal{N}(\mathbf{0}, \boldsymbol{C}_i)$ and $\mathcal{N}(\mathbf{0}, \boldsymbol{C}_j)$ to compare covariance matrices, specifically
$$\left(1 - 2^{D/2} \frac{\det(\boldsymbol{C}_i \boldsymbol{C}_j)^{1/4}}{\det(\boldsymbol{C}_i + \boldsymbol{C}_j)^{1/2}}\right)^{1/2}, \tag{5}$$
if $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ are full-rank. While using these distances or the Frobenius distances makes little difference in practice, we find it easier to work with the latter when it comes to proving theoretical guarantees. Moreover, it seems more natural to assume a uniform sampling distribution in each neighborhood rather than a normal distribution, so that using the more sophisticated similarity (5) does not seem justified.

- *K-means.* We use K-means++ for a good initialization. We found that the more sophisticated size-constrained K-means (Bradley et al., 2000) used in (Goldberg et al., 2009) did not improve the clustering results.

As we mentioned above, our work was developed in parallel to that of Wang et al. (2011) and Gong et al. (2012). We highlight some differences. First, there is no subsampling, but rather, the local tangent space is estimated at each data point $\boldsymbol{x}_i$. Wang et al. (2011) fit a mixture of d-dimensional affine subspaces to the data using MPPCA (Tipping and Bishop, 1999), which is then used to estimate the tangent subspaces at each data point. Gong et al. (2012) develop some sort of robust local PCA. While Wang et al. (2011) assume all surfaces are of same dimension known to the user, Gong et al. (2012) estimate that locally by looking at the largest gap in the spectrum of estimated local covariance matrix. This is similar in spirit to what is done in Step 2 of Algorithm 3, but we did not include this step in Algorithm 4 because we did not find it reliable in practice. We also tried estimating the local dimensionality using the method of Little et al. (2009), but this failed in the most complex cases.

Wang et al. (2011) use a nearest-neighbor graph and their affinity is defined as
$$W_{ij} = \Delta_{ij} \cdot \left(\prod_{s=1}^{d} \cos\theta_s(i,j)\right)^{\alpha},$$
where $\Delta_{ij} = 1$ if $\boldsymbol{x}_i$ is among the ℓ-nearest neighbors of $\boldsymbol{x}_j$, or vice versa, while $\Delta_{ij} = 0$ otherwise; $\theta_1(i,j) \geq \cdots \geq \theta_d(i,j)$ are the principal (a.k.a., canonical) angles (Stewart and Sun, 1990) between the estimated tangent subspaces at $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. ℓ and α are parameters of the method. Gong et al. (2012) define an affinity that incorporates the self-tuning method of Zelnik-Manor and Perona (2005); in our notation, their affinity is
$$\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\varepsilon_i \varepsilon_j}\right) \cdot \exp\left(-\frac{(\sin^{-1}(\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\|))^2}{\eta^2 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(\varepsilon_i \varepsilon_j)}\right).$$

where ε_i is the distance from $\boldsymbol{x}_i$ to its ℓ-nearest neighbor. ℓ is a parameter.

Although we do not analyze their respective ways of estimating the tangent subspaces, our analysis provides essential insights into their methods, and for that matter, any other method built on spectral clustering based on tangent subspace comparisons.

3. Mathematical Analysis

While the analysis of Algorithm 4 seems within reach, there are some complications due to the fact that points near the intersection may form a cluster of their own—we were not able to discard this possibility. Instead, we study the simpler variants described in Algorithm 2 and Algorithm 3. Even then, the arguments are rather complex and interestingly involved. The theoretical guarantees that we obtain for these variants are stated in Theorem 1 and proved in Section 6. We comment on the analysis of Algorithm 4 right after that. We note that there are very few theoretical results on resolving intersecting manifolds—in fact, we are only aware of (Arias-Castro et al., 2011) (under severe restrictions on the dimension of the intersection). Some such results have been established for a number of methods for subspace clustering (affine surfaces), for example, in (Chen and Lerman, 2009b; Soltanolkotabi and Candès, 2012; Soltanolkotabi et al., 2014; Wang and Xu, 2013; Heckel and Bölcskei, 2013; Tsakiris and Vidal, 2015; Ma et al., 2008).

The generative model we assume is a natural mathematical framework for multi-manifold learning where points are sampled in the vicinity of smooth surfaces embedded in Euclidean space. For concreteness and ease of exposition, we focus on the situation where two surfaces (i.e., $K = 2$) of same dimension $1 \leq d \leq D$ intersect. This special situation already contains all the geometric intricacies of separating intersecting clusters. On the one hand, clusters of different intrinsic dimension may be separated with an accurate estimation of the local intrinsic dimension without further geometry involved (Haro et al., 2007). On the other hand, more complex intersections (3-way and higher) complicate the situation without offering truly new challenges. For simplicity of exposition, we assume that the surfaces are submanifolds without boundary, though it will be clear from the analysis (and the experiments) that the method can handle surfaces with (smooth) boundaries that may self-intersect. We discuss other possible extensions in Section 5.

Within that framework, we show that Algorithm 2 and Algorithm 3 are able to identify the clusters accurately except for points near the intersection. Specifically, with high probability with respect to the sampling distribution, Algorithm 2 divides the data points into two groups such that, except for points within distance $C\varepsilon$ of the intersection, all points from the first cluster are in one group and all points from the second cluster are in the other group. The constant C depends on the surfaces, including their curvatures, separation between them and intersection angle. The situation for Algorithm 3 is more complex, as it may return more than two clusters, but most of the two clusters (again, away from the intersection) are in separate connected components.

3.1 Generative Model

Each surface we consider is a connected, C^2 and compact submanifold without boundary and of dimension d embedded in $\mathbb{R}^D$. Any such surface has a positive reach, which is what we use to quantify smoothness. The notion of reach was introduced by Federer (1959).

Intuitively, a surface has reach exceeding r if and only if one can roll a ball of radius r on the surface without obstruction (Walther, 1997; Cuevas et al., 2012). Formally, for $\boldsymbol{x} \in \mathbb{R}^D$ and $S \subset \mathbb{R}^D$, let

$$\operatorname{dist}(\boldsymbol{x}, S) = \inf_{\boldsymbol{s} \in S} \|\boldsymbol{x} - \boldsymbol{s}\|,$$

and

$$B(S, r) = \{\boldsymbol{x} : \operatorname{dist}(\boldsymbol{x}, S) < r\},$$

which is often called the r-tubular neighborhood (or r-neighborhood) of S. The reach of S is the supremum over $r > 0$ such that, for each $\boldsymbol{x} \in B(S, r)$, there is a unique point in S nearest $\boldsymbol{x}$. It is well-known that, for C^2 submanifolds, the reach bounds the radius of curvature from below (Federer, 1959, Lem. 4.17). For submanifolds without boundaries, the reach coincides with the condition number introduced in (Niyogi et al., 2008).

When two surfaces S_1 and S_2 intersect, meaning $S_1 \cap S_2 \neq \emptyset$, we define their incidence angle as

$$\theta(S_1, S_2) := \inf \big\{\theta_{\min}(T_{S_1}(\boldsymbol{s}), T_{S_2}(\boldsymbol{s})) : \boldsymbol{s} \in S_1 \cap S_2\big\}, \tag{6}$$

where $T_S(\boldsymbol{s})$ denote the tangent subspace of submanifold S at point $\boldsymbol{s} \in S$, and $\theta_{\min}(T_1, T_2)$ is the smallest *nonzero* principal (a.k.a., canonical) angle between subspaces T_1 and T_2 (Stewart and Sun, 1990).

The clusters are generated as follows. Each data point $\boldsymbol{x}_i$ is drawn according to

$$\boldsymbol{x}_i = \boldsymbol{s}_i + \boldsymbol{z}_i, \tag{7}$$

where $\boldsymbol{s}_i$ is drawn from the uniform distribution over $S_1 \cup S_2$ and $\boldsymbol{z}_i$ is an additive noise term satisfying $\|\boldsymbol{z}_i\| \leq \tau$—thus τ represents the noise or jitter level. When $\tau = 0$ the points are sampled exactly on the surfaces. We assume the points are sampled independently of each other. We let

$$I_k = \{i : \boldsymbol{s}_i \in S_k\},$$

and the goal is to recover the groups I_1 and I_2, up to some errors.

3.2 Performance Guarantees

We state some performance guarantees for Algorithm 2 and Algorithm 3.

Theorem 1. *Consider two connected, compact, twice continuously differentiable submanifolds without boundary, of same dimension d, intersecting at a strictly positive angle, with the intersection set having strictly positive reach. Assume the parameters are set so that*

$$\tau \leq r\eta/C, \quad r \leq \varepsilon/C, \quad \varepsilon \leq \eta/C, \quad \eta \leq 1/C, \tag{8}$$

for a large-enough constant $C \geq 1$ that depends on the configuration. Then with probability at least $1 - Cn \exp\big[-nr^d\eta^2/C\big]$:

- *Algorithm 2 returns exactly two groups such that two points from different clusters are not grouped together unless one of them is within distance Cr from the intersection.*
- *Algorithm 3 returns at least two groups, and such that two points from different clusters are not grouped together unless one of them is within distance Cr from the intersection.*

Thus, as long as, (8) is satisfied, the algorithms have the above properties with high probability when $r^d\eta^2 \geq C' \log n/n$, where $C' > C$ is fixed. In particular, we may choose $\eta \asymp 1$ and $\varepsilon \asymp r \asymp \tau \vee (\log(n)/n)^{1/d}$, which lightens the computational burden.

We note that, while the constant $C > 0$ does not depend on the sample size n, it depends in somewhat complicated ways on characteristics of the surfaces and their position relative to each other, such as their reach and intersection angle, but also aspects that are harder to quantify, like their separation away from their intersection. We note, however, that it behaves as expected: C is indeed decreasing in the reach and the intersection angle, and increasing in the intrinsic dimensions of the surfaces, for example.

The algorithms may make clustering mistakes within distance Cr of the intersection, where $Cr \asymp \tau \vee (\log(n)/n)^{1/d}$ with the choice of parameters just described. Whether this is optimal in the nonparametric setting that we consider—for example, in a minimax sense—we do not know.

We now comment on the challenge of proving a similar result for Algorithm 4. This algorithm relies on knowledge of the intrinsic dimension of the surfaces d and the number of clusters (here $K = 2$), but these may be estimated as in (Arias-Castro et al., 2011), at least in theory, so we assume these parameters are known. The subsampling done in Step 0 does not pose any problem whatsoever, since the centers are well-spread when the points themselves are. The difficulty resides in the application of the spectral graph partitioning, Algorithm 1. If we were to include the intersection-removal step (Step 2 of Algorithm 2) before applying spectral graph partitioning, then a simple adaptation of arguments in (Arias-Castro, 2011) would suffice. The real difficulty, and potential pitfall of the method in this framework (without the intersection-removal step), is that the points near the intersection may form their own cluster. For example, in the simplest case of two affine surfaces intersecting at a positive angle and no sampling noise, the projection matrix at a point near the intersection—meaning a point whose r-ball contains a substantial piece of both surfaces—would be the projection matrix onto $S_1 + S_2$ seen as a linear subspace. We were not able to discard this possibility, although we do not observe this happening in practice. A possible remedy is to constrain the K-means part to only return large-enough clusters. However, a proper analysis of this would require a substantial amount of additional work and we did not engage seriously in this pursuit.

4. Numerical Experiments

4.1 Some Illustrative Examples

We started by applying our method[2] on a few artificial examples to illustrate the theory. As we argued earlier, the methods of Wang et al. (2011) and Gong et al. (2012) are quite similar to ours, and we encourage the reader to also look at the numerical experiments they performed. Our numerical experiments should be regarded as a proof of concept, only here to show that our method can be implemented and works on some stylized examples.

In all experiments, the number of clusters K and the dimension of the manifolds d are assumed known. We choose the spatial scale ε and the projection scale η automatically as

2. The code is available online at `https://math.cos.ucf.edu/tengz`.

follows: we let

$$\varepsilon = \max_{1 \le i \le n_0} \min_{j \neq i} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|, \tag{9}$$

and

$$\eta = \operatorname*{median}_{(i,j):\|\boldsymbol{y}_i - \boldsymbol{y}_j\| < \varepsilon} \|\boldsymbol{Q}_i - \boldsymbol{Q}_j\|.$$

Here, we implicitly assume that the union of all the underlying surfaces forms a connected set. In that case, the idea behind choosing ε as in (9) is that we want the ε-graph on the centers $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ to be connected. Then η is chosen so that a center $\boldsymbol{y}_i$ remains connected in the (ε, η)-graph to most of its neighbors in the ε-graph.

The neighborhood radius r is chosen by hand for each situation. Although we do not know how to choose r automatically, there are some general ad hoc guidelines. When r is too large, the local linear approximation to the underlying surfaces may not hold in neighborhoods of radius r, resulting in local PCA becoming inappropriate. When r is too small, there might not be enough points in a neighborhood of radius r to accurately estimate the local tangent subspace to a given surface at that location, resulting in local PCA becoming inaccurate. From a computational point of view, the smaller r, the larger the number of neighborhoods and the heavier the computations, particularly at the level of spectral graph partitioning. In our numerical experiments, we find that our algorithm is more sensitive to the choice of r when the clustering problem is more difficult. We note that automatic choice of tuning parameters remains a challenge in clustering, and machine learning at large, especially when no labels are available whatsoever. See (Zelnik-Manor and Perona, 2005; Zhang et al., 2012; Little et al., 2009; Kaslovsky and Meyer, 2011).

Since the algorithm is randomized (see Step 0 in Algorithm 4) we repeat each simulation 100 times and report the median misclustering rate and number of times where the misclustering rate is smaller than 5%, 10%, and 15%.

We first run Algorithm 4 on several artificial data sets, which are demonstrated in the LHS of Figures 3 and 4. Table 1 reports the local radius r used for each data set (R is the global radius of each data set), and the statistics for misclustering rates. Typical clustering results are demonstrated in the RHS of Figures 3 and 4. It is evident that Algorithm 4 performs well in these simulations.

data set	r	median misclustering rate	5%	10%	15%
Three curves	0.02 ($0.034R$)	4.16%	76	89	89
Self-intersecting curves	0.1 ($0.017R$)	1.16%	85	85	86
Two spheres	0.2 ($0.059R$)	3.98%	100	100	100
Möbius strips	0.1 ($0.028R$)	2.22%	85	86	88
Monkey saddle	0.1 ($0.069R$)	9.73%	0	67	97
Paraboloids	0.07 ($0.048R$)	10.42%	0	12	91

Table 1: Choices for r and misclustering statistics for the artificial data sets demonstrated in Figures 3 and 4. The statistics are based on 100 repeats and include the median misclustering rate and number of repeats where the misclustering rate is smaller than 5%, 10% and 15%.

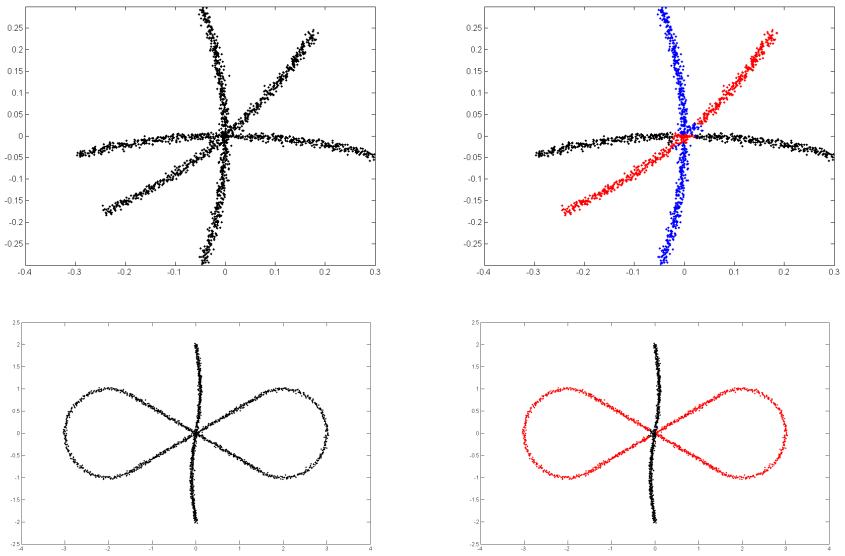

Figure 3: Performance of Algorithm 4 on data sets "Three curves" and "Self-intersecting curves". Left column is the input data sets, and right column demonstrates the typical clustering.

In another simulation, we show the dependence of the success of our algorithm on the intersecting angle between curves in Table 2 and Figure 5. Here, we fix two curves intersecting at a point, and gradually decrease the intersection angle by rotating one of them while holding the other one fixed. The angles are $\pi/2$, $\pi/4$, $\pi/6$ and $\pi/8$. From the table we can see that our algorithm performs well when the angle is $\pi/4$, but the performance deteriorates as the angle becomes smaller, and the algorithm almost always fails when the angle is $\pi/8$.

Intersecting angle	r	median misclustering rate	5%	10%	15%
$\pi/2$	0.02 ($0.034R$)	2.08%	98	98	98
$\pi/4$	0.02 ($0.034R$)	3.33%	92	94	94
$\pi/6$	0.02 ($0.034R$)	5.53%	32	59	59
$\pi/8$	0.02 ($0.033R$)	27.87%	0	2	2

Table 2: Choices for r and misclustering statistics for the instances of two intersecting curves demonstrated in Figure 5. The statistics are based on 100 repeats and include the median misclustering rate and number of repeats where the misclustering rate is smaller than 5%, 10% and 15%.

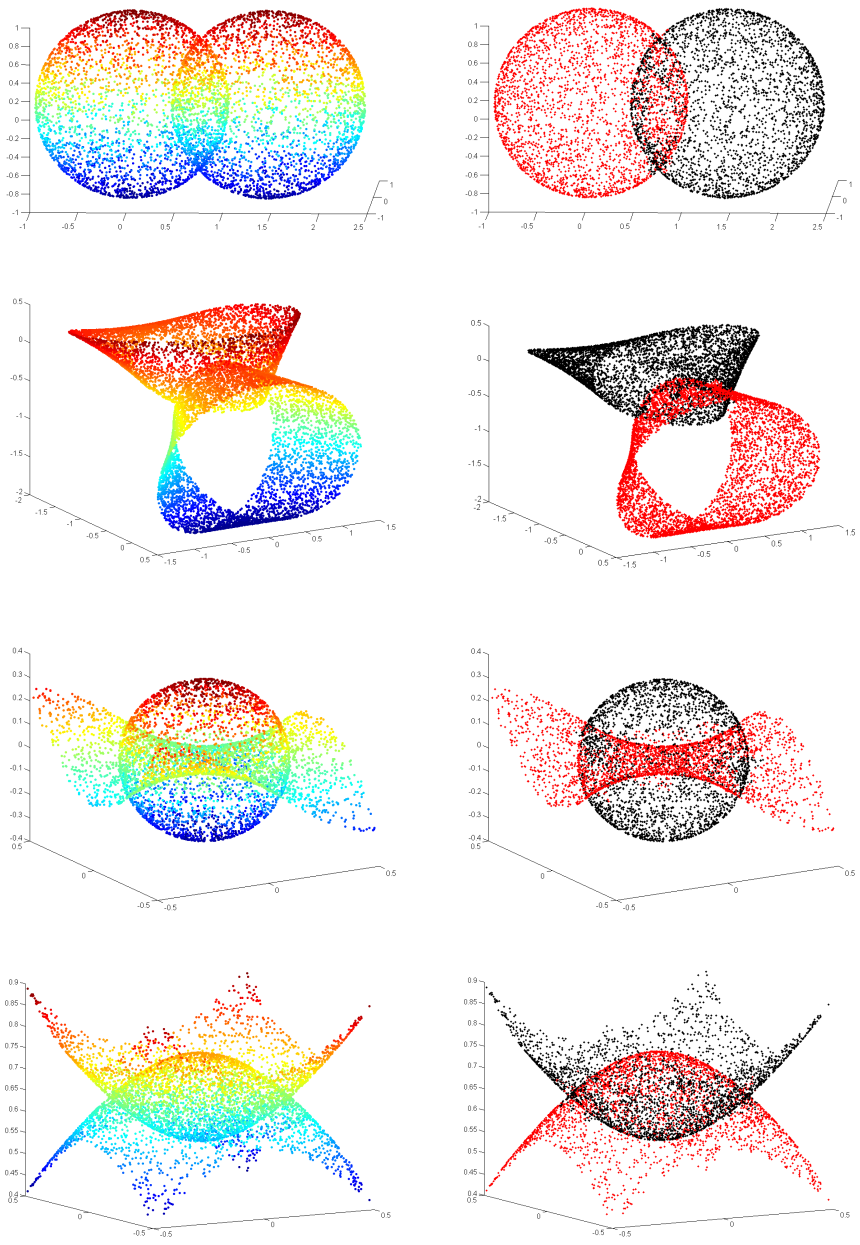

Figure 4: Performance of Algorithm 4 on data sets "Two spheres", "Möbius strips", "Monkey saddle" and "Paraboloids". Left column is the input data sets, and right column demonstrates the typical clustering.

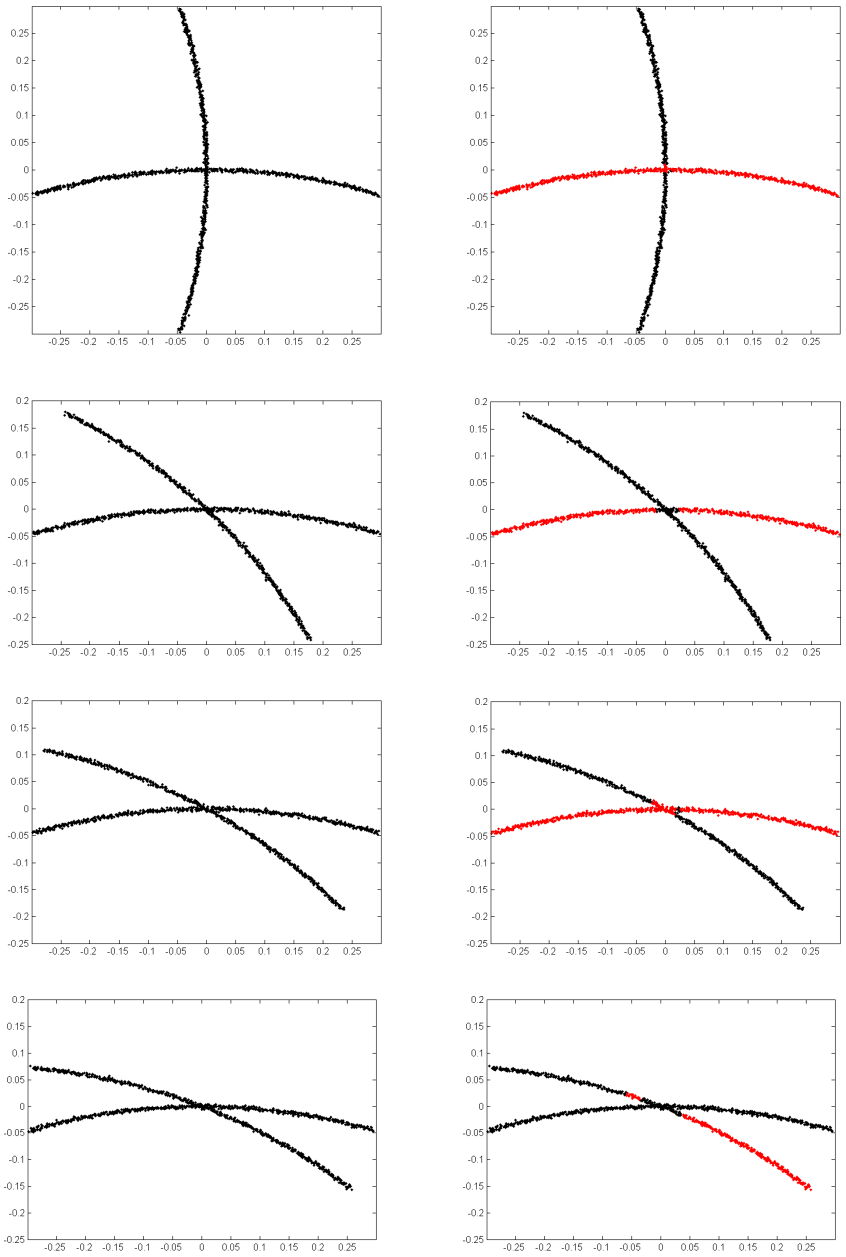

Figure 5: Performance of Algorithm 4 on two curves intersecting at angles $\frac{\pi}{2}$, $\frac{\pi}{4}$, $\frac{\pi}{6}$, $\frac{\pi}{8}$.

4.2 Comparison with Other Multi-manifold Clustering Algorithms

In this section, we compare our algorithm with several existing algorithms on multi-manifold clustering. While many algorithms have been proposed, we focus on the methods based on spectral clustering, including Sparse Manifold Clustering and Embedding (SMCE) (Elhamifar and Vidal, 2011) and Local Linear Embedding of (Polito and Perona, 2001; Goh and Vidal, 2007). Compared to these methods, a major difference of Algorithm 4 is the size of the affinity matrix $\boldsymbol{W}$: SMCE and LLE each creates an $n \times n$ affinity matrix while our method creates a smaller affinity matrix of size $n_0 \times n_0$, based on the centers chosen in step 0 of Algorithm 4. This difference enables our algorithm to handle large data sets such as $n > 10^4$, while these other two methods are computationally expensive due to eigenvalue decomposition of the $n \times n$ affinity matrix. In order to make a fair comparison, we will run simulations and experiments on small data sets. We modified our algorithm to make it more competitive in such a setting: we modify Steps 0 and 1 in Algorithm 4 slightly and use all data points $\{\boldsymbol{x}_i\}_{i=1}^n$ as centers, that is, $\boldsymbol{y}_i = \boldsymbol{x}_i$ for all $1 \le i \le n$. The motivation is that, when there is no computational constraint on the eigenvalue decomposition (due to small n), we may improve Algorithm 4 by constructing a larger affinity matrix by including all points as centers. The resulting algorithm is summarized in Algorithm 5.

Algorithm 5 Spectral Clustering Based on Local PCA (for small data sets)

Input:

Data points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$; neighborhood size $N > 0$; spatial scale $\varepsilon > 0$, projection scale $\eta > 0$; intrinsic dimension d; number of clusters K.

Steps:

1: For each $i = 1, \dots, n$, compute the sample covariance matrix $\boldsymbol{C}_i$ of from the N nearest neighbors of $\boldsymbol{x}_i$. Let $\boldsymbol{Q}_i$ denote the orthogonal projection onto the space spanned by the top d eigenvectors of $\boldsymbol{C}_i$.

2: Compute the following affinities between n data points:

$$W_{ij} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\varepsilon^2}\right) \cdot \exp\left(-\frac{\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\|^2}{\eta^2}\right).$$

3: Apply spectral graph partitioning (Algorithm 1) to $\boldsymbol{W}$.

First we test the algorithms on a simulated data set of two curves, which is subsampled from the first data set in Figure 5 with 300 data points. We plot the clustering result from Algorithm 5, SMCE, LLE in Figure 6. For Algorithm 5, K is set to be 10. For SMCE[3], $\lambda = 10$ and $L = 60$, and we remark that similar results are obtained for a wide range of parameters. For LLE, we follow the implementation in (Polito and Perona, 2001), use 10-nearest neighbors to embed the data set into $\mathbb{R}^2$ and run K-means on the embedded data set. It is clear from the figure that Algorithm 5 resolves the problem of intersection well

3. In (Elhamifar and Vidal, 2011), λ is the ℓ_1-penalty parameter and at each point the optimization is restricted to the L nearest neighbors.

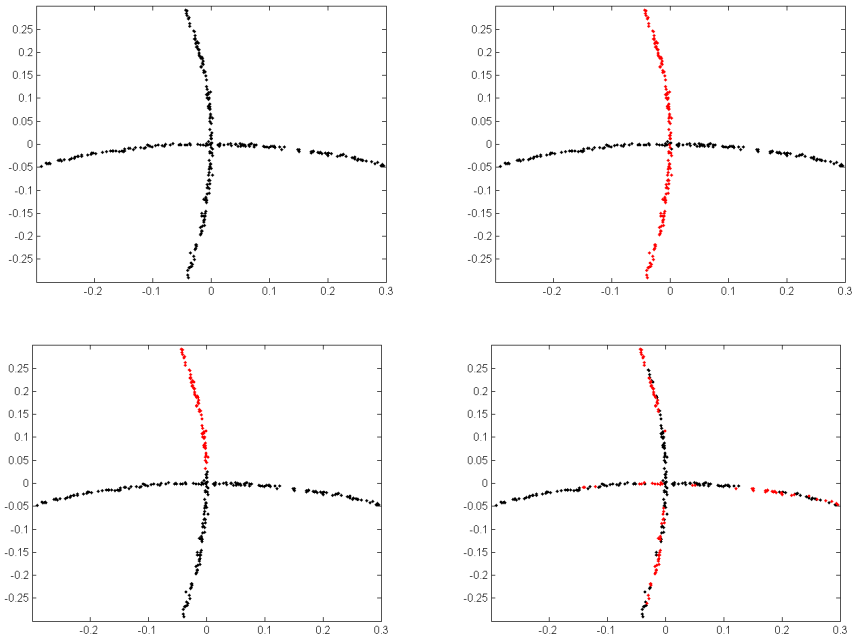

Figure 6: Performance of Algorithm 5 on two intersecting curves. Left top: the input data set. Right top: the clustering result by Algorithm 5. Left bottom: the clustering result by SMCE. Right bottom: the clustering result by LLE.

by using the affinity from estimated local subspaces, while SMCE and LLE tend to give a larger affinity between nearby data points and have difficulties in handling intersection.

Next, we run experiments on the Extended Yale Face Database B (Lee et al., 2005), with the goal of clustering face images of two different subjects. This data set contains face images from 39 subjects, and each subject has 64 images of 192 pixels under varying lightening conditions. In our experiments, we found that the images of a person in this database lie roughly in a 4-dimensional subspace. We preprocess the data set by applying PCA and reducing the dimension to 8. We also "normalize" the covariance of the data set when performing dimension reduction, such that the projected data set has a unit covariance. We record the misclustering rates of Algorithm 5, SMCE and LLE in Table 3. For SMCE, we follow (Elhamifar and Vidal, 2011) by setting $\lambda = 10$ and we let $L = 30$. For Algorithm 5, we let the neighborhood size be 40. From the table, we can see that the two methods perform similarly.

subjects	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]	[1,7]	[1,8]	[1,9]
Local PCA	8.59%	11.72%	10.94%	4.69%	8.59%	7.81%	5.47%	7.03%
SMCE	8.59%	11.72%	8.59%	0.00%	4.69%	8.59%	9.38%	4.69%

Table 3: Some misclustering rates for the Extended Yale Face Database B.

5. Discussion

We distilled the ideas of Goldberg et al. (2009) and of Kushnir et al. (2006) to cluster points sampled near smooth surfaces. The key ingredient is the use of local PCA to learn about the local spread and orientation of the data, so as to use that information in an affinity when building a neighborhood graph.

In a typical stylized setting for multi-manifold clustering, we established performance bounds for the simple variants described in Algorithm 2 and Algorithm 3, which essentially consist of connecting points that are close in space and orientation, and then extracting the connected components of the resulting graph. Both are shown to resolve general intersections as long as the incidence angle is strictly positive and the parameters are carefully chosen. As is commonly the case in such analyses, our setting can be generalized to other sampling schemes, to multiple intersections, to some features of the surfaces changing with the sample size, and so on, in the spirit of (Arias-Castro et al., 2011; Arias-Castro, 2011; Chen and Lerman, 2009b). We chose to simplify the setup as much as possible while retaining the essential features that make resolving intersecting clusters challenging. The resulting arguments are nevertheless rich enough to satisfy the mathematically thirsty reader. Whether the conditions required in Theorem 1 are optimal in some sense is an interesting and challenging open question for future research. Note that very few optimality results exist for manifold clustering; see (Arias-Castro, 2011) for an example.

We implemented a spectral version of Algorithm 3, described in Algorithm 4, that assumes the intrinsic dimensionality and the number of clusters are known. The resulting approach is very similar to what is offered by Wang et al. (2011) and Gong et al. (2012), although it was developed independently of these works. Algorithm 4 is shown to perform well in some simulated experiments, although it is somewhat sensitive to the choice of parameters. This is the case of all other methods for multi-manifold clustering we know of and choosing the parameters automatically remains an open challenge in the field.

6. Proofs

We start with some additional notation. The ambient space is $\mathbb{R}^D$ unless noted otherwise. For a vector $\boldsymbol{v} \in \mathbb{R}^D$, $\|\boldsymbol{v}\|$ denotes its Euclidean norm and for a real matrix $\boldsymbol{M} \in \mathbb{R}^{D\times D}$, $\|\boldsymbol{M}\|$ denotes the corresponding operator norm. For a point $\boldsymbol{x} \in \mathbb{R}^D$ and $r > 0$, $B(\boldsymbol{x}, r)$ denotes the open ball of center $\boldsymbol{x}$ and radius r, i.e., $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \mathbb{R}^D : \|\boldsymbol{y} - \boldsymbol{x}\| < r\}$. For a set S and a point $\boldsymbol{x}$, define $\operatorname{dist}(\boldsymbol{x}, S) = \inf\{\|\boldsymbol{x} - \boldsymbol{y}\| : \boldsymbol{y} \in S\}$. For two points $\boldsymbol{a}, \boldsymbol{b}$ in the same Euclidean space, $\boldsymbol{b} - \boldsymbol{a}$ denotes the vector moving $\boldsymbol{a}$ to $\boldsymbol{b}$. For a point $\boldsymbol{a}$ and a vector $\boldsymbol{v}$ in the same Euclidean space, $\boldsymbol{a} + \boldsymbol{v}$ denotes the translate of $\boldsymbol{a}$ by $\boldsymbol{v}$. We identify an affine subspace T with its corresponding linear subspace, for example, when saying that a vector belongs to T.

For two subspaces T and T', of possibly different dimensions, let $0 \le \theta_{\max}(T, T') \le \pi/2$ denote the largest and by $\theta_{\min}(T, T')$ the smallest nonzero principal angle between T and T' (Stewart and Sun, 1990). When $\boldsymbol{v}$ is a vector and T is a subspace, $\angle(\boldsymbol{v}, T) := \theta_{\max}(\mathbb{R}\boldsymbol{v}, T)$ this is the usual definition of the angle between $\boldsymbol{v}$ and T.

For a subset $A \subset \mathbb{R}^D$ and positive integer d, $\mathrm{vol}_d(A)$ denotes the d-dimensional Hausdorff measure of A, and $\mathrm{vol}(A)$ is defined as $\mathrm{vol}_{\dim(A)}(A)$, where $\dim(A)$ is the Hausdorff dimension of A. For a Borel set A, let λ_A denote the uniform distribution on A.

For a set $S \subset \mathbb{R}^D$ with reach at least $1/\kappa$, and $\boldsymbol{x}$ with $\mathrm{dist}(\boldsymbol{x}, S) < 1/\kappa$, let $P_S(\boldsymbol{x})$ denote the metric projection of $\boldsymbol{x}$ onto S, that is, the point on S closest to $\boldsymbol{x}$. Note that, if T is an affine subspace, then P_T is the usual orthogonal projection onto T, and we let $\boldsymbol{P}_T$ denote the orthogonal projection onto the linear subspace of same dimension and parallel to T. Let $\mathcal{S}_d(\kappa)$ denote the class of connected, C^2 and compact d-dimensional submanifolds without boundary embedded in $\mathbb{R}^D$, with reach at least $1/\kappa$. For a submanifold $S \in \mathbb{R}^D$, let $T_S(\boldsymbol{x})$ denote the tangent space of S at $\boldsymbol{x} \in S$.

We will often identify a linear map with its matrix in the canonical basis. For a symmetric (real) matrix $\boldsymbol{M}$, let $\beta_1(\boldsymbol{M}) \geq \beta_2(\boldsymbol{M}) \geq \cdots$ denote its eigenvalues in decreasing order.

We say that $f : \Omega \subset \mathbb{R}^D \to \mathbb{R}^D$ is C-Lipschitz if $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq C\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega$.

For two reals a and b, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Additional notation will be introduced as needed.

6.1 Preliminaries

This section gathers a number of general results from geometry and probability. We took time to package them into standalone lemmas that could be of potential independent interest, particularly to researchers working in machine learning and computational geometry. When needed, we use C to denote a constant that may change with each appearance.

6.1.1 Smooth Surfaces and Their Tangent Subspaces

The following comes directly from (Federer, 1959, Th. 4.18(12)). It gives us a simple criterion for identifying the metric projection of a point on a surface with given reach.

Lemma 1. *Consider $S \in \mathcal{S}_d(\kappa)$ and $\boldsymbol{x} \in \mathbb{R}^D$ such that $\mathrm{dist}(\boldsymbol{x}, S) < 1/\kappa$. Then $\boldsymbol{s} = P_S(\boldsymbol{x})$ if and only if $\|\boldsymbol{x} - \boldsymbol{s}\| < 1/\kappa$ and $\boldsymbol{x} - \boldsymbol{s} \perp T_S(\boldsymbol{s})$.*

The following result is on approximating a smooth surface near a point by the tangent subspace at that point. It is based on (Federer, 1959, Th. 4.18(2)).

Lemma 2. *For $S \in \mathcal{S}_d(\kappa)$, and any two points $\boldsymbol{s}, \boldsymbol{s}' \in S$,*

$$\mathrm{dist}(\boldsymbol{s}', T_S(\boldsymbol{s})) \leq \frac{\kappa}{2}\|\boldsymbol{s}' - \boldsymbol{s}\|^2, \tag{10}$$

and when $\mathrm{dist}(\boldsymbol{s}', T_S(\boldsymbol{s})) \leq 1/\kappa$,

$$\mathrm{dist}(\boldsymbol{s}', T_S(\boldsymbol{s})) \leq \kappa\|P_{T_S(\boldsymbol{s})}(\boldsymbol{s}') - \boldsymbol{s}\|^2. \tag{11}$$

Moreover, for $\boldsymbol{t} \in T_S(\boldsymbol{s})$ such that $\|\boldsymbol{s} - \boldsymbol{t}\| \leq \frac{1}{3\kappa}$,

$$\mathrm{dist}(\boldsymbol{t}, S) \leq \kappa\|\boldsymbol{t} - \boldsymbol{s}\|^2. \tag{12}$$

Proof. Let T be short for $T_S(\boldsymbol{s})$. (Federer, 1959, Th. 4.18(2)) says that

$$\operatorname{dist}(\boldsymbol{s}' - \boldsymbol{s}, T) \leq \frac{\kappa}{2}\|\boldsymbol{s}' - \boldsymbol{s}\|^2. \tag{13}$$

Immediately, we have

$$\operatorname{dist}(\boldsymbol{s}' - \boldsymbol{s}, T) = \|\boldsymbol{s}' - P_T(\boldsymbol{s}')\| = \operatorname{dist}(\boldsymbol{s}', T),$$

and (10) comes from that. Based on that and Pythagoras theorem, we have

$$\operatorname{dist}(\boldsymbol{s}', T) = \|P_T(\boldsymbol{s}') - \boldsymbol{s}'\| \leq \frac{\kappa}{2}\|\boldsymbol{s}' - \boldsymbol{s}\|^2 = \frac{\kappa}{2}\big(\|P_T(\boldsymbol{s}') - \boldsymbol{s}'\|^2 + \|P_T(\boldsymbol{s}') - \boldsymbol{s}\|^2\big),$$

so that

$$\operatorname{dist}(\boldsymbol{s}', T)\big(1 - \frac{\kappa}{2}\operatorname{dist}(\boldsymbol{s}', T)\big) \leq \frac{\kappa}{2}\|P_T(\boldsymbol{s}') - \boldsymbol{s}\|^2,$$

and (11) follows easily from that. For (12), let $r = 1/(3\kappa)$ and $\boldsymbol{s}' = P_T^{-1}(\boldsymbol{t})$, the latter being well-defined by Lemma 5 below and belongs to $B(\boldsymbol{s}, r(1 + \kappa r)) \subset B(\boldsymbol{s}, 4/(9\kappa))$. By (10), $\|\boldsymbol{s}' - \boldsymbol{s}\| \leq 8/(81\kappa) < 1/\kappa$, and by (11),

$$\operatorname{dist}(\boldsymbol{t}, S) \leq \|\boldsymbol{t} - \boldsymbol{s}'\| = \operatorname{dist}(\boldsymbol{s}', T) \leq \kappa\|\boldsymbol{t} - \boldsymbol{s}\|^2.$$

This concludes the proof of (12). □

We will need a bound on the angle between tangent subspaces on a smooth surface as a function of the distance between the corresponding points of contact.

Lemma 3 (Boissonnat et al. (2013))**.** *For $S \in \mathcal{S}_d(\kappa)$, and any $\boldsymbol{s}, \boldsymbol{s}' \in S$,*

$$\sin\theta_{\max}(T_S(\boldsymbol{s}), T_S(\boldsymbol{s}')) \leq 6\kappa\|\boldsymbol{s}' - \boldsymbol{s}\|. \tag{14}$$

The following bounds the difference between the metric projection onto a surface and the orthogonal projection onto one of its tangents.

Lemma 4. *Consider $S \in \mathcal{S}_d(\kappa)$ and $\boldsymbol{s} \in S$. Then for any $\boldsymbol{x} \in B(S, r)$, we have*

$$\|P_S(\boldsymbol{x}) - P_{T_S(\boldsymbol{s})}(\boldsymbol{x})\| \leq C_4\kappa\|\boldsymbol{x} - \boldsymbol{s}\|^2,$$

for a numeric constant $C_4 > 0$.

Proof. Let $T = T_S(\boldsymbol{s})$ and define $\boldsymbol{t} = P_T(\boldsymbol{x})$, $\boldsymbol{s}' = P_S(\boldsymbol{x})$ and $\tilde{\boldsymbol{t}} = P_T(\boldsymbol{s}')$, and also $T' = T_S(\boldsymbol{s}')$ and $\theta = \theta_{\max}(T, T')$. We have

$$\begin{aligned} P_S(\boldsymbol{x}) &= P_{T'}(\boldsymbol{x}) = \boldsymbol{P}_{T'}(\boldsymbol{x} - \boldsymbol{s}') + \boldsymbol{s}' \\ P_T(\boldsymbol{x}) &= \boldsymbol{P}_T(\boldsymbol{x} - \tilde{\boldsymbol{t}}) + \tilde{\boldsymbol{t}} = \boldsymbol{P}_T(\boldsymbol{x} - \boldsymbol{s}') + \boldsymbol{P}_T(\boldsymbol{s}' - \tilde{\boldsymbol{t}}) + \tilde{\boldsymbol{t}}. \end{aligned}$$

Hence,

$$\|P_S(\boldsymbol{x}) - P_T(\boldsymbol{x})\| \leq \|\boldsymbol{P}_{T'}(\boldsymbol{x} - \boldsymbol{s}') - \boldsymbol{P}_T(\boldsymbol{x} - \boldsymbol{s}')\| + \|\boldsymbol{P}_T(\boldsymbol{s}' - \tilde{\boldsymbol{t}})\| + \|\boldsymbol{s}' - \tilde{\boldsymbol{t}}\|.$$

On the one hand,

$$\|\boldsymbol{P}_T(\boldsymbol{s}' - \tilde{\boldsymbol{t}})\| \leq \|\boldsymbol{s}' - \tilde{\boldsymbol{t}}\| = \operatorname{dist}(\boldsymbol{s}', T) \leq \frac{\kappa}{2}\|\boldsymbol{s}' - \boldsymbol{s}\|^2 \leq 2\kappa\|\boldsymbol{x} - \boldsymbol{s}\|^2,$$

by (10) and the fact that

$$\|\boldsymbol{s}' - \boldsymbol{s}\| \leq \|\boldsymbol{s}' - \boldsymbol{x}\| + \|\boldsymbol{x} - \boldsymbol{s}\| = \operatorname{dist}(\boldsymbol{x}, S) + \|\boldsymbol{x} - \boldsymbol{s}\| \leq 2\|\boldsymbol{x} - \boldsymbol{s}\|.$$

On the other hand, applying Lemma 18 (see further down)

$$\|\boldsymbol{P}_{T'}(\boldsymbol{x} - \boldsymbol{s}') - \boldsymbol{P}_T(\boldsymbol{x} - \boldsymbol{s}')\| \leq \|\boldsymbol{P}_{T'} - \boldsymbol{P}_T\|\|\boldsymbol{x} - \boldsymbol{s}'\| = (\sin\theta)\|\boldsymbol{x} - \boldsymbol{s}'\|,$$

with $\|\boldsymbol{x} - \boldsymbol{s}'\| \leq \|\boldsymbol{x} - \boldsymbol{s}\|$ and, applying Lemma 3,

$$\sin\theta \leq 6\kappa\|\boldsymbol{s}' - \boldsymbol{s}\| \leq 12\kappa\|\boldsymbol{x} - \boldsymbol{s}\|.$$

All together, we conclude. □

Below we state some properties of a projection onto a tangent subspace. A result similar to the first part was proved in (Arias-Castro et al., 2011, Lem. 2) based on results in (Niyogi et al., 2008), but the arguments are simpler here and the constants are sharper.

Lemma 5. *There is a numeric constant $C_5 \geq 1$ such that the following holds. Take $S \in \mathcal{S}_d(\kappa)$, $\boldsymbol{s} \in S$ and $r \leq 1/C_5\kappa$, and let T be short for $T_S(\boldsymbol{s})$. P_T is injective on $B(\boldsymbol{s}, r) \cap S$ and its image contains $B(\boldsymbol{s}, r') \cap T$, where $r' := (1 - C_5(\kappa r)^2)r$. Moreover, P_T^{-1} has Lipschitz constant bounded by $1 + C_5(\kappa r)^2 \leq 1 + \kappa r$ over $B(\boldsymbol{s}, r) \cap T$.*

Proof. Take $\boldsymbol{s}', \boldsymbol{s}'' \in S$ distinct such that $P_T(\boldsymbol{s}') = P_T(\boldsymbol{s}'')$. Equivalently, $\boldsymbol{s}'' - \boldsymbol{s}'$ is perpendicular to T. Let T' be short for $T_S(\boldsymbol{s}')$. By (13) and the fact that $\operatorname{dist}(\boldsymbol{v}, T) = \|\boldsymbol{v}\| \sin\angle(\boldsymbol{v}, T)$ for any vector $\boldsymbol{v}$ and any linear subspace T, we have

$$\sin\angle(\boldsymbol{s}'' - \boldsymbol{s}', T') \leq \frac{\kappa}{2}\|\boldsymbol{s}'' - \boldsymbol{s}'\|,$$

and by (14),

$$\sin\theta_{\max}(T, T') \leq 6\kappa\|\boldsymbol{s} - \boldsymbol{s}'\|.$$

Now, by the triangle inequality,

$$\frac{\pi}{2} = \angle(\boldsymbol{s}'' - \boldsymbol{s}', T) \leq \angle(\boldsymbol{s}'' - \boldsymbol{s}', T') + \theta_{\max}(T, T'),$$

so that

$$\frac{\kappa}{2}\|\boldsymbol{s}'' - \boldsymbol{s}'\| \wedge 1 \geq \frac{\pi}{2} - \sin^{-1}\left(6\kappa\|\boldsymbol{s}' - \boldsymbol{s}\| \wedge 1\right).$$

When $\|\boldsymbol{s}' - \boldsymbol{s}\| \leq 1/12\kappa$, the RHS is bounded from below by $\pi/2 - \sin^{-1}(1/2)$, which then implies that $\frac{\kappa}{2}\|\boldsymbol{s}'' - \boldsymbol{s}'\| \geq \sin(\pi/2 - \sin^{-1}(1/2)) = \sqrt{3}/2$, that is, $\|\boldsymbol{s}'' - \boldsymbol{s}'\| \geq \sqrt{3}/\kappa$. This precludes the situation where $\boldsymbol{s}', \boldsymbol{s}'' \in B(\boldsymbol{s}, 1/12\kappa)$, so that P_T is injective on $B(\boldsymbol{s}, r)$ when $r \leq 1/12\kappa$.

The same arguments imply that P_T is an open map on $R := B(\boldsymbol{s}, r) \cap S$. In particular, $P_T(R)$ contains an open ball in T centered at $\boldsymbol{s}$ and $P_T(\partial R) = \partial P_T(R)$, with $\partial R = S \cap$

$\partial B(\boldsymbol{s}, r)$ since $\partial S = \emptyset$. Now take any ray out of $\boldsymbol{s}$ within T, which is necessarily of the form $\boldsymbol{s} + \mathbb{R}_+ \boldsymbol{v}$, where $\boldsymbol{v}$ is a unit vector in T. Let $\boldsymbol{t}_a = \boldsymbol{s} + a\boldsymbol{v} \in T$ for $a \in [0, \infty)$. Let a_* be the infimum over all $a > 0$ such that $\boldsymbol{t}_a \in P_T(R)$. Note that $a_* > 0$ and $\boldsymbol{t}_{a_*} \in P_T(\partial R)$, so that there is $\boldsymbol{s}_* \in \partial R$ such that $P_T(\boldsymbol{s}_*) = \boldsymbol{t}_{a_*}$. Let $\boldsymbol{s}_a = P_T^{-1}(\boldsymbol{t}_a)$, which is well-defined on $[0, a_*]$ by definition of a_* and the fact that P_T is injective on R. Let $J_{\boldsymbol{t}}$ denote the differential of P_T^{-1} at $\boldsymbol{t}$. We have that $\dot{\boldsymbol{s}}_a = J_{\boldsymbol{t}_a} \boldsymbol{v}$ is the unique vector in $T_a := T_S(\boldsymbol{s}_a)$ such that $P_T(\dot{\boldsymbol{s}}_a) = \boldsymbol{v}$. Elementary geometry shows that

$$\|P_T(\dot{\boldsymbol{s}}_a)\| = \|\dot{\boldsymbol{s}}_a\| \cos \angle(\dot{\boldsymbol{s}}_a, T) \geq \|\dot{\boldsymbol{s}}_a\| \cos \theta_{\max}(T_a, T),$$

with

$$\cos \theta_{\max}(T_a, T) \geq \cos \left[\sin^{-1}\left(6\kappa \|\boldsymbol{s}_a - \boldsymbol{s}\| \wedge 1\right)\right] \geq \zeta := 1 - (6\kappa r)^2,$$

by (14) and fact that $\|\boldsymbol{s}_a - \boldsymbol{s}\| \leq r$ (and assuming $6\kappa r \leq 1$). Since $\|P_T(\dot{\boldsymbol{s}}_a)\| = \|\boldsymbol{v}\| = 1$, we have $\|\dot{\boldsymbol{s}}_a\| \leq 1/\zeta$, and this holds for all $a < a_*$. So we can extend $\boldsymbol{s}_a$ to $[0, a_*]$ into a Lipschitz function with constant $1/\zeta$. Together with the fact that $\boldsymbol{s}_* \in \partial B(\boldsymbol{s}, r)$, this implies that

$$r = \|\boldsymbol{s}_* - \boldsymbol{s}\| = \|\boldsymbol{s}_{a_*} - \boldsymbol{s}_0\| \leq a_*/\zeta.$$

Hence, $a_* \geq \zeta r$ and therefore $P_T(R)$ contains $B(\boldsymbol{s}, \zeta r) \cap T$ as stated.

For the last part, assume $r \leq 1/C\kappa$, with C large enough that there is a unique $h \leq 1/12\kappa$ such that $\zeta h = r$, where ζ is redefined as $\zeta := 1 - (6\kappa h)^2$. Take $\boldsymbol{t}' \in B(\boldsymbol{s}, r) \cap T$ and let $\boldsymbol{s}' = P_T^{-1}(\boldsymbol{t}')$ and $T' = T_S(\boldsymbol{s}')$. We saw that P_T^{-1} is Lipschitz with constant $1/\zeta$ on any ray emanating from $\boldsymbol{s}$ of length $\zeta h = r$, so that $\|\boldsymbol{s}' - \boldsymbol{s}\| \leq (1/\zeta)\|\boldsymbol{t}' - \boldsymbol{s}\| \leq r/\zeta = h$. The differential of P_T at $\boldsymbol{s}'$ is P_T itself, seen as a linear map between T' and T. Then for any vector $\boldsymbol{u} \in T'$, we have

$$\|P_T(\boldsymbol{u})\| = \|\boldsymbol{u}\| \cos \angle(\boldsymbol{u}, T) \geq \|\boldsymbol{u}\| \cos \theta_{\max}(T', T),$$

with

$$\cos \theta_{\max}(T', T) \geq \cos \left[\sin^{-1}\left(6\kappa \|\boldsymbol{s}' - \boldsymbol{s}\|\right)\right] \geq 1 - (6\kappa h)^2 = \zeta,$$

as before. Hence, $\|J_{\boldsymbol{t}'}\| \leq 1/\zeta$, and we proved this for all $\boldsymbol{t}' \in B(\boldsymbol{s}, r) \cap T$. This last set being convex, we can apply Taylor's theorem and get that P_T^{-1} is Lipschitz on that set with constant $1/\zeta$. We then note that $\zeta = 1 + O(\kappa r)^2$. □

6.1.2 Volumes and Uniform Distributions

Below is a result that quantifies how much the volume of a set changes when applying a Lipschitz map. This is well-known in measure theory and we only provide a proof for completeness.

Lemma 6. *Suppose Ω is a measurable subset of $\mathbb{R}^D$ and $f : \Omega \subset \mathbb{R}^D \to \mathbb{R}^D$ is C-Lipschitz. Then for any measurable set $A \subset \Omega$ and real $d > 0$, $\mathrm{vol}_d(f(A)) \leq C^d \, \mathrm{vol}_d(A)$.*

Proof. By definition,

$$\mathrm{vol}_d(A) = \lim_{t \to 0} \; V_d^t(A), \qquad V_d^t(A) := \inf_{(R_i) \in \mathcal{R}^t(A)} \; \sum_{i \in \mathbb{N}} \mathrm{diam}(R_i)^d,$$

where $\mathcal{R}^t(A)$ is the class of countable sequences $(R_i : i \in \mathbb{N})$ of subsets of $\mathbb{R}^D$ such that $A \subset \bigcup_i R_i$ and $\operatorname{diam}(R_i) < t$ for all i. Since f is C-Lipschitz, $\operatorname{diam}(f(R)) \leq C \operatorname{diam}(R)$ for any $R \subset \Omega$. Hence, for any $(R_i) \in \mathcal{R}^t(A)$, $(f(R_i)) \in \mathcal{R}^{Ct}(f(A))$. This implies that

$$V_d^{Ct}(f(A)) \leq \sum_{i \in \mathbb{N}} \operatorname{diam}(f(R_i))^d \leq C^d \sum_{i \in \mathbb{N}} \operatorname{diam}(R_i)^d.$$

Taking the infimum over $(R_i) \in \mathcal{R}^t(A)$, we get $V_d^{Ct}(f(A)) \leq C^d V_d^t(A)$, and we conclude by taking the limit as $t \to 0$, noticing that $V_d^{Ct}(f(A)) \to \operatorname{vol}_d(f(A))$ while $V_d^t(A) \to \operatorname{vol}_d(A)$. □

We compare below two uniform distributions. For two Borel probability measures P and Q on $\mathbb{R}^D$, $\operatorname{TV}(P, Q)$ denotes their total variation distance, meaning,

$$\operatorname{TV}(P, Q) = \sup\{|P(A) - Q(A)| : A \text{ Borel}\}.$$

Remember that for a Borel set A, λ_A denotes the uniform distribution on A.

Lemma 7. *Suppose A and B are two Borel subsets of $\mathbb{R}^D$. Then*

$$\operatorname{TV}(\lambda_A, \lambda_B) \leq 4 \, \frac{\operatorname{vol}(A \triangle B)}{\operatorname{vol}(A \cup B)}.$$

Proof. If A and B are not of same dimension, say $\dim(A) > \dim(B)$, then $\operatorname{TV}(\lambda_A, \lambda_B) = 1$ since $\lambda_A(B) = 0$ while $\lambda_B(B) = 1$. And we also have

$$\operatorname{vol}(A \triangle B) = \operatorname{vol}_{\dim(A)}(A \triangle B) = \operatorname{vol}_{\dim(A)}(A) = \operatorname{vol}(A),$$

and

$$\operatorname{vol}(A \cup B) = \operatorname{vol}_{\dim(A)}(A \cup B) = \operatorname{vol}_{\dim(A)}(A) = \operatorname{vol}(A),$$

in both cases because $\operatorname{vol}_{\dim(A)}(B) = 0$. So the result works in that case.

Therefore assume that A and B are of same dimension. Assume WLOG that $\operatorname{vol}(A) \geq \operatorname{vol}(B)$. For any Borel set U,

$$\lambda_A(U) - \lambda_B(U) = \frac{\operatorname{vol}(A \cap U)}{\operatorname{vol}(A)} - \frac{\operatorname{vol}(B \cap U)}{\operatorname{vol}(B)},$$

so that

$$\begin{aligned} |\lambda_A(U) - \lambda_B(U)| &= \left| \frac{\operatorname{vol}(A \cap U)}{\operatorname{vol}(A)} - \frac{\operatorname{vol}(B \cap U)}{\operatorname{vol}(A)} + \frac{\operatorname{vol}(B \cap U)}{\operatorname{vol}(A)} - \frac{\operatorname{vol}(B \cap U)}{\operatorname{vol}(B)} \right| \\ &\leq \frac{|\operatorname{vol}(A \cap U) - \operatorname{vol}(B \cap U)|}{\operatorname{vol}(A)} + \frac{\operatorname{vol}(B \cap U)}{\operatorname{vol}(B)} \frac{|\operatorname{vol}(A) - \operatorname{vol}(B)|}{\operatorname{vol}(A)} \\ &\leq \frac{2 \operatorname{vol}(A \triangle B)}{\operatorname{vol}(A)}, \end{aligned}$$

and we conclude with the fact that $\operatorname{vol}(A \cup B) \leq \operatorname{vol}(A) + \operatorname{vol}(B) \leq 2 \operatorname{vol}(A)$. □

We now look at the projection of the uniform distribution on a neighborhood of a surface onto a tangent subspace. For a Borel probability measure P and measurable function $f : \mathbb{R}^D \to \mathbb{R}^D$, P^f denotes the push-forward (Borel) measure defined by $P^f(A) = P(f^{-1}(A))$.

Lemma 8. *Suppose $A \subset \mathbb{R}^D$ is Borel and $f : A \to \mathbb{R}^D$ is invertible on $f(A)$, and that both f and f^{-1} are C-Lipschitz. Then*

$$\mathrm{TV}(\lambda_A^f, \lambda_{f(A)}) \le 2(C^{\dim(A)} - 1).$$

Proof. First, note that A and $f(A)$ are both of same dimension, and that $C \ge 1$ necessarily. Let d be short for $\dim(A)$. Take $U \subset f(A)$ Borel and let $V = f^{-1}(U)$. Then

$$\lambda_A^f(U) = \frac{\mathrm{vol}(A \cap V)}{\mathrm{vol}(A)}, \qquad \lambda_{f(A)}(U) = \frac{\mathrm{vol}(f(A) \cap U)}{\mathrm{vol}(f(A))},$$

and as in (16),

$$|\lambda_A^f(U) - \lambda_{f(A)}(U)| \le \frac{|\,\mathrm{vol}(A \cap V) - \mathrm{vol}(f(A) \cap U)|}{\mathrm{vol}(A)} + \frac{|\,\mathrm{vol}(A) - \mathrm{vol}(f(A))|}{\mathrm{vol}(A)}.$$

f being invertible, we have $f(A \cap V) = f(A) \cap U$ and $f^{-1}(f(A) \cap U) = A \cap V$. Therefore, applying Lemma 6, we get

$$C^{-d} \le \frac{\mathrm{vol}(f(A) \cap U)}{\mathrm{vol}(A \cap V)} \le C^d,$$

so that

$$|\,\mathrm{vol}(A \cap V) - \mathrm{vol}(f(A) \cap U)| \le (C^d - 1)\,\mathrm{vol}(A \cap V) \le (C^d - 1)\,\mathrm{vol}(A).$$

And taking $V = A$, we also get

$$|\,\mathrm{vol}(A) - \mathrm{vol}(f(A))| \le (C^d - 1)\,\mathrm{vol}(A).$$

From this we conclude. □

Now comes a technical result on the intersection of a smooth surface and a ball.

Lemma 9. *There is a constant $C_9 \ge 3$ depending only on d such that the following is true. Take $S \in \mathcal{S}_d(\kappa)$, $r < \frac{1}{C_9\kappa}$ and $\boldsymbol{x} \in \mathbb{R}^D$ such that $\mathrm{dist}(\boldsymbol{x}, S) < \kappa$. Let $\boldsymbol{s} = P_S(\boldsymbol{x})$ and $T = T_S(\boldsymbol{s})$. Then*

$$\mathrm{vol}\left(P_T(S \cap B(\boldsymbol{x}, r)) \triangle (T \cap B(\boldsymbol{x}, r))\right) \le C_9\kappa(\|\boldsymbol{x} - \boldsymbol{s}\| + \kappa r^2)\ \mathrm{vol}(T \cap B(\boldsymbol{x}, r)).$$

Proof. If $\mathrm{dist}(\boldsymbol{x}, S) > r$, then $S \cap B(\boldsymbol{x}, r) = T \cap B(\boldsymbol{x}, r) = \emptyset$, and if $\mathrm{dist}(\boldsymbol{x}, S) = r$, then $S \cap B(\boldsymbol{x}, r) = T \cap B(\boldsymbol{x}, r) = \{\boldsymbol{s}\}$, and in both cases the inequality holds trivially. So it suffices to consider the case where $\mathrm{dist}(\boldsymbol{x}, S) < r$.

Let $A_r = B(\boldsymbol{s}, r)$, $B_r = B(\boldsymbol{x}, r)$ and $g = P_T$ for short. Note that $T \cap B_r = T \cap A_{r_0}$ where $r_0 := (r^2 - \delta^2)^{1/2}$ and $\delta := \|\boldsymbol{x} - \boldsymbol{s}\|$. Take $\boldsymbol{s}_1 \in S \cap B_r$ such that $g(\boldsymbol{s}_1)$ is farthest from $\boldsymbol{s}$, so that $g(S \cap B_r) \subset A_{r_1}$ where $r_1 := \|\boldsymbol{s} - g(\boldsymbol{s}_1)\|$—note that $r_1 \le r$. Let $\ell_1 = \|\boldsymbol{s}_1 - g(\boldsymbol{s}_1)\|$ and $\boldsymbol{y}_1$ be the orthogonal projection of $\boldsymbol{s}_1$ onto the line $(\boldsymbol{x}, \boldsymbol{s})$. By Pythagoras theorem,

we have $\|\boldsymbol{x}-\boldsymbol{s}_1\|^2 = \|\boldsymbol{x}-\boldsymbol{y}_1\|^2 + \|\boldsymbol{y}_1-\boldsymbol{s}_1\|^2$. We have $\|\boldsymbol{x}-\boldsymbol{s}_1\| \le r$ and $\|\boldsymbol{y}_1-\boldsymbol{s}_1\| = \|\boldsymbol{s}-g(\boldsymbol{s}_1)\| = r_1$. And because $\ell_1 \le \kappa r_1^2 < r$ by (11), either $\boldsymbol{y}_1$ is between $\boldsymbol{x}$ and $\boldsymbol{s}$, in which case $\|\boldsymbol{x}-\boldsymbol{y}_1\| = \delta-\ell_1$, or $\boldsymbol{s}$ is between $\boldsymbol{x}$ and $\boldsymbol{y}_1$, in which case $\|\boldsymbol{x}-\boldsymbol{y}_1\| = \delta+\ell_1$. In any case, $r^2 \ge r_1^2 + (\delta-\ell_1)^2$, which together with $\ell_1 \le \kappa r_1^2$ implies $r_1^2 \le r^2 - \delta^2 + 2\delta\ell_1 \le r_0^2 + 2\kappa r_1^2\delta$, leading to $r_1 \le (1-2\kappa\delta)^{-1/2} r_0 \le (1+4\kappa\delta) r_0$ after noticing that $\delta \le r < 1/(3\kappa)$. From $g(S\cap B_r) \subset T \cap A_{r_1}$, we get

$$\begin{aligned}\operatorname{vol}\big(g(S\cap B_r) \setminus (T\cap B_r)\big) &\le \operatorname{vol}(T\cap A_{r_1}) - \operatorname{vol}(T\cap A_{r_0}) \\ &= ((r_1/r_0)^d - 1)\operatorname{vol}(T\cap A_{r_0}).\end{aligned}$$

We follow similar arguments to get a sort of reverse relationship. Take $\boldsymbol{s}_2 \in S\cap B_r$ such that $g(S\cap B_r) \supset T\cap A_{r_2}$, where $r_2 := \|\boldsymbol{s}-g(\boldsymbol{s}_2)\|$ is largest. Assuming r is small enough, by Lemma 5, g^{-1} is well-defined on $T\cap A_r$, so that necessarily $\boldsymbol{s}_2 \in \partial B_r$. Let $\ell_2 = \|\boldsymbol{s}_2 - g(\boldsymbol{s}_2)\|$ and $\boldsymbol{y}_2$ be the orthogonal projection of $\boldsymbol{s}_2$ onto the line $(\boldsymbol{x},\boldsymbol{s})$. By Pythagoras theorem, we have $\|\boldsymbol{x}-\boldsymbol{s}_2\|^2 = \|\boldsymbol{x}-\boldsymbol{y}_2\|^2 + \|\boldsymbol{y}_2-\boldsymbol{s}_2\|^2$. We have $\|\boldsymbol{x}-\boldsymbol{s}_2\| = r$ and $\|\boldsymbol{y}_2-\boldsymbol{s}_2\| = \|\boldsymbol{s}-g(\boldsymbol{s}_2)\| = r_2$. And by the triangle inequality, $\|\boldsymbol{x}-\boldsymbol{y}_2\| \le \|\boldsymbol{x}-\boldsymbol{s}\| + \|\boldsymbol{y}_2-\boldsymbol{s}\| = \delta+\ell_2$. Hence, $r^2 \le r_2^2 + (\delta+\ell_2)^2$, which together with $\ell_2 \le \kappa r_2^2$ by (11), implies $r_2^2 \ge r^2 - \delta^2 - 2\delta\ell_2 - \ell_2^2 \ge r_0^2 - (2\delta+\kappa r^2)\kappa r_2^2$, leading to $r_2 \ge (1+2\kappa\delta+\kappa^2 r^2)^{-1/2} r_0 \ge (1-2\kappa\delta-\kappa^2 r^2) r_0$. From $g(S\cap B_r) \supset T\cap A_{r_2}$, we get

$$\begin{aligned}\operatorname{vol}\big((T\cap B_r) \setminus g(S\cap B_r)\big) &\le \operatorname{vol}(T\cap A_{r_0}) - \operatorname{vol}(T\cap A_{r_2}) \\ &= (1-(r_2/r_0)^d)\operatorname{vol}(T\cap A_{r_0}).\end{aligned}$$

All together, we have

$$\begin{aligned}\operatorname{vol}\big(g(S\cap B_r) \triangle (T\cap B_r)\big) &\le \big((r_1/r_0)^d - (r_2/r_0)^d\big)\ \operatorname{vol}(T\cap A_{r_0}) \\ &\le \big((1+4\kappa\delta)^d - (1-2\kappa\delta-\kappa^2 r^2)^d\big)\operatorname{vol}((T\cap B_r)) \\ &\le C d\kappa(\delta+\kappa r^2)\operatorname{vol}((T\cap B_r)),\end{aligned}$$

when $\delta \le r \le 1/(C\kappa)$ and C is a large-enough numerical constant. □

We bound below the d-volume of a the intersection of a ball with a smooth surface. Although it could be obtained as a special case of Lemma 9, we provide a direct proof because this result is at the cornerstone of many results in the literature on sampling points uniformly on a smooth surface.

Lemma 10. *Suppose $S \in \mathcal{S}_d(\kappa)$. Then for any $\boldsymbol{s} \in S$ and $r < \frac{1}{(d\vee C_5)\kappa}$, we have*

$$1 - 2d\kappa r \le \frac{\operatorname{vol}(S\cap B(\boldsymbol{s},r))}{\operatorname{vol}(T\cap B(\boldsymbol{s},r))} \le 1 + 2d\kappa r,$$

where $T := T_S(\boldsymbol{s})$ is the tangent subspace of S at $\boldsymbol{s}$.

Proof. Let $T = T_S(\boldsymbol{s})$, $B_r = B(\boldsymbol{s},r)$ and $g = P_T$ for short. By Lemma 5, g is 1-Lipschitz and g^{-1} is $(1+\kappa r)$-Lipschitz on $T\cap B_r$, so by Lemma 6 we have

$$(1+\kappa r)^{-d} \le \frac{\operatorname{vol}(g(S\cap B_r))}{\operatorname{vol}(S\cap B_r)} \le 1.$$

That g^{-1} is Lipschitz with constant $1+\kappa r$ on $g(S\cap B_r)$ also implies that $g(S\cap B_r)$ contains $T\cap B_{r'}$ where $r' := r/(1+\kappa r)$. From this, and the fact that $g(S\cap B_r)\subset T\cap B_r$, we get

$$1 \leq \frac{\mathrm{vol}(T\cap B_r)}{\mathrm{vol}(g(S\cap B_r))} \leq \frac{\mathrm{vol}(T\cap B_r)}{\mathrm{vol}(T\cap B_{r'})} = \frac{r^d}{r'^d} = (1+\kappa r)^d.$$

We therefore have

$$\mathrm{vol}(S\cap B_r) \geq \mathrm{vol}(g(S\cap B_r)) \geq (1+\kappa r)^{-d}\,\mathrm{vol}(T\cap B_r),$$

and

$$\mathrm{vol}(S\cap B_r) \leq (1+\kappa r)^{d}\,\mathrm{vol}(g(S\cap B_r)) \leq (1+\kappa r)^{d}\,\mathrm{vol}(T\cap B_r).$$

And we conclude with the inequality $(1+x)^d \leq 1+2dx$ valid for any $x\in[0,1/d]$ and any $d\geq 1$. □

We now look at the density of a sample from the uniform on a smooth, compact surface.

Lemma 11. *Consider $S\in\mathcal{S}_d(\kappa)$. There is a constant $C_{11}>0$ depending on S such that the following is true. Sample n points $\boldsymbol{s}_1,\dots,\boldsymbol{s}_n$ independently and uniformly at random from S. Take $0<r<1/(C_{11}\kappa)$. Then with probability at least $1-C_{11}r^{-d}\exp(-nr^d/C_{11})$, any ball of radius r with center on S has between nr^d/C_{11} and $C_{11}nr^d$ sample points.*

Proof. For a set R, let $N(R)$ denote the number of sample points in R. For any R measurable, $N(R)\sim \mathrm{Bin}(n,p_R)$, where $p_R := \mathrm{vol}(R\cap S)/\mathrm{vol}(S)$. Let $\boldsymbol{x}_1,\dots,\boldsymbol{x}_m$ be a maximal $(r/2)$-packing of S, and let $B_i = B(\boldsymbol{x}_i, r/4)\cap S$. For any $\boldsymbol{s}\in S$, there is i such that $\|\boldsymbol{s}-\boldsymbol{x}_i\|\leq r/2$, which implies $B_i\subset B(\boldsymbol{s},r)$ by the triangle inequality. Hence, $\min_{\boldsymbol{s}\in S} N(B(\boldsymbol{s},r)) \geq \min_i N(B_i)$.

By the fact that $B_i\cap B_j=\emptyset$ for $i\neq j$,

$$\mathrm{vol}(S) \geq \sum_{i=1}^m \mathrm{vol}(B_i) \geq m\min_i \mathrm{vol}(B_i),$$

and assuming that κr is small enough, we have

$$\min_i \mathrm{vol}(B_i) \geq \frac{\omega_d}{2}(r/4)^d,$$

by Lemma 10, where ω_d is the volume of the d-dimensional unit ball. This leads to $m\leq Cr^{-d}$ and $p := \min_i p_{B_i} \geq r^d/C$, when $\kappa r\leq 1/C$, where $C>0$ depends only on S.

Now, applying Bernstein's inequality to the binomial distribution, we get

$$\mathbb{P}\left(N(B_i)\leq np/2\right) \leq \mathbb{P}\left(N(B_i)\leq np_{B_i}/2\right) \leq e^{-(3/32)np_{B_i}} \leq e^{-(3/32)np}.$$

We follow this with the union bound, to get

$$\mathbb{P}\left(\min_{\boldsymbol{s}\in S} N(B(\boldsymbol{s},r)) \leq nr^d/(2C)\right) \leq me^{-(3/32)np} \leq\ Cr^{-d}e^{-\frac{3}{32C}nr^d}.$$

From this the lower bound follows. The proof of the upper bound is similar. □

Next, we bound the volume of the symmetric difference between two balls.

Lemma 12. *Take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $0 < \delta \leq 1$. Then*

$$\frac{\operatorname{vol}(B(\boldsymbol{x}, \delta) \triangle B(\boldsymbol{y}, 1))}{2 \operatorname{vol}(B(0,1))} \leq 1 - (1 - \|\boldsymbol{x} - \boldsymbol{y}\|)_+^d \wedge \delta^d.$$

Proof. It suffices to prove the result when $\|\boldsymbol{x} - \boldsymbol{y}\| < 1$. In that case, with $\gamma := (1 - \|\boldsymbol{x} - \boldsymbol{y}\|) \wedge \delta$, we have $B(\boldsymbol{x}, \gamma) \subset B(\boldsymbol{x}, \delta) \cap B(\boldsymbol{y}, 1)$, so that

$$\begin{aligned} \operatorname{vol}(B(\boldsymbol{x}, \delta) \triangle B(\boldsymbol{y}, 1)) &= \operatorname{vol}(B(\boldsymbol{x}, \delta)) + \operatorname{vol}(B(\boldsymbol{y}, 1)) - 2 \operatorname{vol}(B(\boldsymbol{x}, \delta) \cap B(\boldsymbol{y}, 1)) \\ &\leq 2 \operatorname{vol}(B(\boldsymbol{y}, 1)) - 2 \operatorname{vol}(B(\boldsymbol{x}, \gamma)) \\ &= 2 \operatorname{vol}(B(0,1))(1 - \gamma^d). \end{aligned}$$

This concludes the proof. □

6.1.3 Covariances

The result below describes explicitly the covariance matrix of the uniform distribution over the unit ball of a subspace.

Lemma 13. *Let T be a subspace of dimension d. Then the covariance matrix of the uniform distribution on $T \cap B(0, a)$ (seen as a linear map) is equal to $a^2 c \boldsymbol{P}_T$, where $c := \frac{1}{d+2}$.*

Proof. Assume WLOG that $T = \mathbb{R}^d \times \{0\}$. Let X be distributed according to the uniform distribution on $T \cap B(0,1)$ and let $R = \|X\|$. Note that

$$\mathbb{P}(R \leq r) = \frac{\operatorname{vol}(T \cap B(0, r))}{\operatorname{vol}(T \cap B(0,1))} = r^d, \quad \forall r \in [0,1].$$

By symmetry, $\mathbb{E}(X_i X_j) = 0$ if $i \neq j$, while

$$\mathbb{E}(X_1^2) = \frac{1}{d} \mathbb{E}(X_1^2 + \cdots + X_d^2) = \frac{1}{d} \mathbb{E}(R^2) = \frac{1}{d} \int_0^t r^2 (d r^{d-1}) \mathrm{d}r = \frac{1}{d+2}.$$

Now the covariance matrix of the uniform distribution on $T \cap B(0, a)$ equals $\mathbb{E}[(aX)(aX)^\top] = a^2 \mathbb{E}[XX^\top]$, which is exactly the representation of $a^2 c \boldsymbol{P}_T$ in the canonical basis of $\mathbb{R}^D$. □

We now show that a bound on the total variation distance between two compactly supported distributions implies a bound on the difference between their covariance matrices. For a measure P on $\mathbb{R}^D$ and an integrable function f, let $P(f)$ denote the integral of f with respect to P, that is,

$$P(f) = \int f(x) P(dx),$$

and let $\mathbb{E}(P) = P(\boldsymbol{x})$ and $\operatorname{Cov}(P) = P(\boldsymbol{x}\boldsymbol{x}^\top) - P(\boldsymbol{x}) P(\boldsymbol{x})^\top$ denote the mean and covariance matrix of P, respectively.

Lemma 14. *Suppose λ and ν are two Borel probability measures on $\mathbb{R}^d$ supported on $B(0,1)$. Then*

$$\|\mathbb{E}(\lambda) - \mathbb{E}(\nu)\| \leq \sqrt{d}\, \mathrm{TV}(\lambda, \nu), \qquad \|\operatorname{Cov}(\lambda) - \operatorname{Cov}(\nu)\| \leq 3d\, \mathrm{TV}(\lambda, \nu).$$

Proof. Let $f_k(\boldsymbol{t}) = t_k$ when $\boldsymbol{t} = (t_1, \ldots, t_d)$, and note that $|f_k(\boldsymbol{t})| \leq 1$ for all k and all $\boldsymbol{t} \in B(0,1)$. By the fact that

$$\mathrm{TV}(\lambda, \nu) = \sup\{\lambda(f) - \nu(f) : f : \mathbb{R}^d \to \mathbb{R} \text{ measurable with } |f| \leq 1\},$$

we have

$$|\lambda(f_k) - \nu(f_k)| \leq \mathrm{TV}(\lambda, \nu), \quad \forall k = 1, \ldots, d.$$

Therefore,

$$\|\,\mathbb{E}(\lambda) - \mathbb{E}(\nu)\|^2 = \sum_{k=1}^{d} (\lambda(f_k) - \nu(f_k))^2 \leq d\,\mathrm{TV}(\lambda, \nu)^2,$$

which proves the first part.

Similarly, let $f_{k\ell}(\boldsymbol{t}) = t_k t_\ell$. Since $|f_{k\ell}(\boldsymbol{t})| \leq 1$ for all k, ℓ and all $\boldsymbol{t} \in B(0,1)$, we have

$$|\lambda(f_{k\ell}) - \nu(f_{k\ell})| \leq \mathrm{TV}(\lambda, \nu), \quad \forall k, \ell = 1, \ldots, d.$$

Since for any probability measure μ on $\mathbb{R}^d$,

$$\mathrm{Cov}(\mu) = \big(\mu(f_{k\ell}) - \mu(f_k)\mu(f_\ell) : k, \ell = 1, \ldots, d\big),$$

we have

$$\begin{aligned}
\|\,\mathrm{Cov}(\lambda) - \mathrm{Cov}(\nu)\| &\leq d \max_{k,\ell} \big(|\lambda(f_{k\ell}) - \nu(f_{k\ell})| + |\lambda(f_k)\lambda(f_\ell) - \nu(f_k)\nu(f_\ell)|\big) \\
&\leq d \max_{k,\ell} \big(|\lambda(f_{k\ell}) - \nu(f_{k\ell})| + |\lambda(f_k)||\lambda(f_\ell) - \nu(f_\ell)| \\
&\qquad\qquad + |\nu(f_\ell)||\lambda(f_k) - \nu(f_k)|\big) \\
&\leq 3d\,\mathrm{TV}(\lambda, \nu),
\end{aligned}$$

using the fact that $|\lambda(f_k)| \leq 1$ and $|\nu(f_k)| \leq 1$ for all k. □

The following compares the mean and covariance matrix of a distribution before and after transformation by a function.

Lemma 15. *Let λ be a Borel distribution with compact support $A \subset \mathbb{R}^D$ and consider a measurable function $g : A \mapsto \mathbb{R}^D$ such that $\|g(x) - x\| \leq \eta$. Then*

$$\|\,\mathbb{E}(\lambda^g) - \mathbb{E}(\lambda)\| \leq \eta, \quad \|\,\mathrm{Cov}(\lambda^g) - \mathrm{Cov}(\lambda)\| \leq \eta(\mathrm{diam}(A) + \eta).$$

Proof. Take $X \sim \lambda$ and let $Y = g(X) \sim \lambda^g$. For the means, by Jensen's inequality,

$$\|\,\mathbb{E}\,X - \mathbb{E}\,Y\| \leq \mathbb{E}\,\|X - Y\| = \mathbb{E}\,\|X - g(X)\| \leq \eta.$$

For the covariances, we first note that

$$\mathrm{Cov}(X) - \mathrm{Cov}(Y) = \frac{1}{2}\big(\,\mathrm{Cov}(X - Y, X + Y) + \mathrm{Cov}(X + Y, X - Y)\big),$$

where $\operatorname{Cov}(U,V) := \mathbb{E}((U-\mathbb{E}\,U)(V-\mathbb{E}\,V)^T)$ is the cross-covariance of random vectors U and V. By Jensen's inequality, the fact $\|\boldsymbol{u}\boldsymbol{v}^T\| = \|\boldsymbol{u}\|\|\boldsymbol{v}\|$ for any pair of vectors $\boldsymbol{u},\boldsymbol{v}$, and then the Cauchy-Schwarz inequality,

$$\|\operatorname{Cov}(U,V)\| \le \mathbb{E}(\|U-\mathbb{E}\,U\|\cdot\|V-\mathbb{E}\,V\|) \le \mathbb{E}(\|U-\mathbb{E}\,U\|^2)^{1/2}\cdot\mathbb{E}(\|V-\mathbb{E}\,V\|^2)^{1/2}.$$

Hence,

$$\begin{aligned}\|\operatorname{Cov}(X)-\operatorname{Cov}(Y)\| &\le \|\operatorname{Cov}(X-Y,X+Y)\| \\ &\le \mathbb{E}\left[\|X-Y-\mathbb{E}\,X+\mathbb{E}\,Y\|^2\right]^{1/2}\mathbb{E}\left[\|X+Y-\mathbb{E}\,X-\mathbb{E}\,Y\|^2\right]^{1/2}.\end{aligned}$$

By the fact that the mean minimizes the mean-squared error,

$$\mathbb{E}\left[\|X-Y-\mathbb{E}\,X+\mathbb{E}\,Y\|^2\right]^{1/2} \le \mathbb{E}\left[\|X-Y\|^2\right]^{1/2} = \mathbb{E}\left[\|X-g(X)\|^2\right]^{1/2} \le \eta.$$

In the same vein, letting $\boldsymbol{z}\in\mathbb{R}^D$ be such that $\|\boldsymbol{x}-\boldsymbol{z}\| \le \frac{1}{2}\operatorname{diam}(A)$ for all $\boldsymbol{x}\in A$, we get

$$\begin{aligned}\mathbb{E}\left[\|X+Y-\mathbb{E}\,X-\mathbb{E}\,Y\|^2\right]^{1/2} &\le \mathbb{E}\left[\|X-\boldsymbol{z}+Y-\boldsymbol{z}\|^2\right]^{1/2} \\ &\le \mathbb{E}\left[\|X-\boldsymbol{z}\|^2\right]^{1/2} + \mathbb{E}\left[\|Y-\boldsymbol{z}\|^2\right]^{1/2} \\ &\le 2\,\mathbb{E}\left[\|X-\boldsymbol{z}\|^2\right]^{1/2} + \mathbb{E}\left[\|X-g(X)\|^2\right]^{1/2} \\ &\le \operatorname{diam}(A)+\eta.\end{aligned}$$

Using the triangle inequality twice. From this, we conclude. □

Next we compare the covariance matrix of the uniform distribution on a small piece of smooth surface with that of the uniform distribution on the projection of that piece onto a nearby tangent subspace.

Lemma 16. *There is a constant $C_{16}>0$ depending only on d such that the following is true. Take $S\in\mathcal{S}_d(\kappa)$, $r<\frac{1}{C_{16}\kappa}$ and $\boldsymbol{x}\in\mathbb{R}^D$ such that $\operatorname{dist}(\boldsymbol{x},S)\le r$. Let $\boldsymbol{s}=P_S(\boldsymbol{x})$ and $T=T_S(\boldsymbol{s})$. If $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ are the means, and $\boldsymbol{M}$ and $\boldsymbol{N}$ are the covariance matrices, of the uniform distributions on $S\cap B(\boldsymbol{x},r)$ and $T\cap B(\boldsymbol{x},r)$, respectively, then*

$$\|\boldsymbol{\zeta}-\boldsymbol{\xi}\| \le C_{16}\kappa r^2, \qquad \|\boldsymbol{M}-\boldsymbol{N}\| \le C_{16}\kappa r^3.$$

Proof. We focus on proving the bound on the covariances, and leave the bound on the means—whose proof is both similar and simpler—as an exercise to the reader. Let $T=T_S(\boldsymbol{s})$, $B_r=B(\boldsymbol{x},r)$ and $g=P_T$ for short. Let $A=S\cap B_r$ and $A'=T\cap B_r$.

First, applying Lemma 15, assuming κr is sufficiently small, we get that

$$\|\operatorname{Cov}(\lambda_A)-\operatorname{Cov}(\lambda_A^g)\| \le \frac{\kappa}{2}r^2\Big(2r+\frac{\kappa}{2}r^2\Big) \le 2\kappa r^3, \tag{38}$$

because $\operatorname{diam}(A)\le 2r$ and $\|g(\boldsymbol{x})-\boldsymbol{x}\| \le \frac{\kappa}{2}\|\boldsymbol{x}-\boldsymbol{s}\|^2 \le \frac{\kappa}{2}r^2$ for all $\boldsymbol{x}\in A$ by (10).

Let $\lambda_{g(A)}$ denote the uniform distribution on $g(A)$. λ_A^g and $\lambda_{g(A)}$ are both supported on B_r, so that applying Lemma 14 with proper scaling, we get

$$\|\operatorname{Cov}(\lambda_A^g)-\operatorname{Cov}(\lambda_{g(A)})\| \le 3dr^2\operatorname{TV}(\lambda_A^g,\lambda_{g(A)}).$$

When κr is small enough, we know that g is 1-Lipschitz, and by Lemma 5, g^{-1} is well-defined and is $(1+\kappa r)$-Lipschitz on A'. Hence, by Lemma 8 and the fact that $\dim(A) = d$, we have

$$\mathrm{TV}(\lambda_A^g, \lambda_{g(A)}) \leq 2((1+\kappa r)^d - 1) \leq 4d\kappa r,$$

using the inequality $(1+x)^d \leq 1 + 2dx$, valid for any $x \in [0, 1/d]$ and any $d \geq 1$.

Noting that $\lambda_{A'}$ is also supported on B_r, applying Lemma 14 with proper scaling, we get

$$\|\operatorname{Cov}(\lambda_{g(A)}) - \operatorname{Cov}(\lambda_{A'})\| \leq 3dr^2 \,\mathrm{TV}(\lambda_{g(A)}, \lambda_{A'}),$$

with

$$\mathrm{TV}(\lambda_{g(A)}, \lambda_{A'}) \leq 4\frac{\mathrm{vol}(g(A) \triangle A')}{\mathrm{vol}(A')} \leq 4C_9\kappa(r + \kappa r^2) \leq 8C_9\kappa r,$$

by Lemma 7 and Lemma 9, and assuming that κr is small enough.

By the triangle inequality,

$$\begin{aligned} \|\boldsymbol{M} - \boldsymbol{N}\| &= \|\operatorname{Cov}(\lambda_A) - \operatorname{Cov}(\lambda_{A'})\| \\ &\leq \|\operatorname{Cov}(\lambda_A) - \operatorname{Cov}(\lambda_A^g)\| + \|\operatorname{Cov}(\lambda_A^g) - \operatorname{Cov}(\lambda_{g(A)})\| \\ &\qquad + \|\operatorname{Cov}(\lambda_{g(A)}) - \operatorname{Cov}(\lambda_{A'})\| \\ &\leq 2\kappa r^3 + 12d^2\kappa r^3 + 24dC_9\kappa r^3. \end{aligned}$$

From this, we conclude. □

Next is a lemma on the estimation of a covariance matrix. The result is a simple consequence of the matrix Hoeffding inequality of Tropp (2012). Note that simply bounding the operator norm by the Frobenius norm, and then applying the classical Hoeffding inequality (Hoeffding, 1963) would yield a bound sufficient for our purposes, but this is a good opportunity to use a more recent and sophisticated result.

Lemma 17. *Let $\boldsymbol{C}_m$ denote the empirical covariance matrix based on an i.i.d. sample of size m from a distribution with support in $B(0,r) \subset \mathbb{R}^d$ with covariance $\boldsymbol{\Sigma}$. Then*

$$\mathbb{P}\left(\|\boldsymbol{C}_m - \boldsymbol{\Sigma}\| > r^2 t\right) \leq 4d \exp\left(-\frac{mt}{16}\min\left(\frac{t}{32}, \frac{m}{d}\right)\right), \quad \forall t \geq 0. \tag{43}$$

Proof. Without loss of generality, we assume that the distribution has zero mean and is now supported on $B(0, 2r)$. In fact, by a simple rescaling, we may also assume that $r = 1$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ denote the sample, with $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$. We have

$$\boldsymbol{C}_m = \boldsymbol{C}_m^\star - \frac{1}{m}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T,$$

where

$$\boldsymbol{C}_m^\star := \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T, \quad \bar{\boldsymbol{x}} := \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_i.$$

Note that

$$\|\boldsymbol{C}_m - \boldsymbol{\Sigma}\| \leq \|\boldsymbol{C}_m^\star - \boldsymbol{\Sigma}\| + \frac{1}{m}\|\bar{\boldsymbol{x}}\|^2.$$

Applying the union bound and then Hoeffding's inequality to each coordinate—which is in $[-2,2]$—we get

$$\mathbb{P}(\|\bar{\boldsymbol{x}}\| > t) \leq \sum_{j=1}^{d} \mathbb{P}(|\bar{x}_j| > t/\sqrt{d}) \leq 2d \exp\left(-\frac{mt^2}{8d}\right).$$

Noting that $\frac{1}{m}(\boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{\Sigma}), i = 1, \ldots, m$, are independent, zero-mean, self-adjoint matrices with spectral norm bounded by $4/m$, we may apply the matrix Hoeffding inequality (Tropp, 2012, Th. 1.3), to get

$$\mathbb{P}\left(\|\boldsymbol{C}_m^\star - \boldsymbol{\Sigma}\| > t\right) \leq 2d \exp\left(-\frac{t^2}{8\sigma^2}\right), \quad \sigma^2 := m(4/m)^2 = 16/m.$$

Applying the union bound and using the previous inequalities, we arrive at

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{C}_m - \boldsymbol{\Sigma}\| > t\right) &\leq \mathbb{P}\left(\|\boldsymbol{C}_m^\star - \boldsymbol{\Sigma}\| > t/2\right) + \mathbb{P}\left(\|\bar{\boldsymbol{x}}\| > \sqrt{mt/2}\right) \\
&\leq 2d \exp\left(-\frac{mt^2}{512}\right) + 2d \exp\left(-\frac{m^2 t}{16d}\right) \\
&\leq 4d \exp\left(-\frac{mt}{16} \min\left(\frac{t}{32}, \frac{m}{d}\right)\right),
\end{aligned}$$

which concludes the proof. □

6.1.4 Projections

We relate below the difference of two orthogonal projections with the largest principal angle between the corresponding subspaces.

Lemma 18. *For two affine non-null subspaces T, T',*

$$\|\boldsymbol{P}_T - \boldsymbol{P}_{T'}\| = \begin{cases} \sin\theta_{\max}(T, T'), & \text{if } \dim(T) = \dim(T'), \\ 1, & \text{otherwise.} \end{cases}$$

Proof. For two affine subspaces $T, T' \subset \mathbb{R}^D$ of same dimension, let $\frac{\pi}{2} \geq \theta_1 \geq \cdots \geq \theta_D \geq 0$, denote the principal angles between them. By (Stewart and Sun, 1990, Th. I.5.5), the singular values of $\boldsymbol{P}_T - \boldsymbol{P}_{T'}$ are $\{\sin\theta_j : j = 1, \ldots, q\}$, so that $\|\boldsymbol{P}_T - \boldsymbol{P}_{T'}\| = \max_j \sin\theta_j = \sin\theta_1 = \sin\theta_{\max}(T, T')$. Suppose now that T and T' are of different dimension, say $\dim(T) > \dim(T')$. We have $\|\boldsymbol{P}_T - \boldsymbol{P}_{T'}\| \leq \|\boldsymbol{P}_T\| \vee \|\boldsymbol{P}_{T'}\| = 1$, since $\boldsymbol{P}_T$ and $\boldsymbol{P}_{T'}$ are orthogonal projections and therefore positive semidefinite with operator norm equal to 1. Let $L = \boldsymbol{P}_T(T')$. Since $\dim(L) \leq \dim(T') < \dim(T)$, there is $\boldsymbol{u} \in T \cap L^\perp$ with $\boldsymbol{u} \neq 0$. Then $\boldsymbol{v}^\top \boldsymbol{u} = \boldsymbol{P}_T(\boldsymbol{v})^\top \boldsymbol{u} = 0$ for all $\boldsymbol{v} \in T'$, implying that $\boldsymbol{P}_{T'}(\boldsymbol{u}) = 0$ and consequently $(\boldsymbol{P}_T - \boldsymbol{P}_{T'})\boldsymbol{u} = \boldsymbol{u}$, so that $\|\boldsymbol{P}_T - \boldsymbol{P}_{T'}\| \geq 1$. □

The lemma below is a perturbation result for eigenspaces and widely known as the $\sin\Theta$ Theorem of Davis and Kahan (1970). See also (Luxburg, 2007, Th. 7) or (Stewart and Sun, 1990, Th. V.3.6).

Lemma 19. *Let $\boldsymbol{M}$ be positive semi-definite with eigenvalues $\beta_1 \geq \beta_2 \geq \cdots$. Suppose that $\beta_d - \beta_{d+1} > 0$. Then for any other positive semi-definite matrix $\boldsymbol{N}$,*

$$\|\boldsymbol{P}_{\boldsymbol{N}}^{(d)} - \boldsymbol{P}_{\boldsymbol{M}}^{(d)}\| \leq \frac{\sqrt{2}\|\boldsymbol{N} - \boldsymbol{M}\|}{\beta_d - \beta_{d+1}},$$

where $\boldsymbol{P}_{\boldsymbol{M}}^{(d)}$ and $\boldsymbol{P}_{\boldsymbol{M}}^{(d)}$ denote the orthogonal projections onto the top d eigenvectors of $\boldsymbol{M}$ and $\boldsymbol{N}$, respectively.

6.1.5 Intersections

We start with an elementary result on points near the intersection of two affine subspaces.

Lemma 20. *Take any two linear subspaces $T_1, T_2 \subset \mathbb{R}^D$. For any point $\boldsymbol{t}_1 \in T_1 \setminus T_2$, we have*

$$\operatorname{dist}(\boldsymbol{t}_1, T_2) \geq \operatorname{dist}(\boldsymbol{t}_1, T_1 \cap T_2) \sin\theta_{\min}(T_1, T_2).$$

Proof. We may reduce the problem to the case where $T_1 \cap T_2 = \{0\}$. Indeed, let $\tilde{T}_1 = T_1 \cap T_2^\perp$, $\tilde{T}_2 = T_1^\perp \cap T_2$ and $\tilde{\boldsymbol{t}}_1 = \boldsymbol{t}_1 - P_{T_1 \cap T_2}(\boldsymbol{t}_1)$. Then

$$\|\boldsymbol{t}_1 - P_{T_2}(\boldsymbol{t}_1)\| = \|\tilde{\boldsymbol{t}}_1 - P_{\tilde{T}_2}(\tilde{\boldsymbol{t}}_1)\|, \quad \|\boldsymbol{t}_1 - P_{T_1 \cap T_2}(\boldsymbol{t}_1)\| = \|\tilde{\boldsymbol{t}}_1\|,$$

and

$$\sin\theta_{\min}(T_1, T_2) = \sin\theta_{\min}(\tilde{T}_1, \tilde{T}_2).$$

So assume that $T_1 \cap T_2 = \{0\}$. By (Afriat, 1957, Th. 10.1), the angle formed by $\boldsymbol{t}_1$ and $P_{T_2}(\boldsymbol{t}_1)$ is at least as large as the smallest principal angle between T_1 and T_2, which is $\theta_{\min}(T_1, T_2)$ since $T_1 \cap T_2 = \{0\}$. From this the result follows immediately. □

The following result says that a point cannot be close to two compact and smooth surfaces intersecting at a positive angle without being close to their intersection. Note that the constant there cannot be solely characterized by κ, as it also depends on the separation between the surfaces away from their intersection.

Lemma 21. *Suppose $S_1, S_2 \in \mathcal{S}_d(\kappa)$ intersect at a strictly positive angle and that* $\operatorname{reach}(S_1 \cap S_2) \geq 1/\kappa$. *Then there is a constant C_{21} such that*

$$\operatorname{dist}(\boldsymbol{x}, S_1 \cap S_2) \leq C_{21} \max\big\{ \operatorname{dist}(\boldsymbol{x}, S_1), \operatorname{dist}(\boldsymbol{x}, S_2) \big\}, \quad \forall \boldsymbol{x} \in \mathbb{R}^D.$$

Proof. Assume the result is not true, so there is a sequence $(\boldsymbol{x}_n) \subset \mathbb{R}^D$ such that $\operatorname{dist}(\boldsymbol{x}_n, S_1 \cap S_2) > n \max_k \operatorname{dist}(\boldsymbol{x}_n, S_k)$. Because the surfaces are bounded, we may assume WLOG that the sequence is bounded. Then $\operatorname{dist}(\boldsymbol{x}_n, S_1 \cap S_2)$ is bounded, which implies that $\max_k \operatorname{dist}(\boldsymbol{x}_n, S_k) = O(1/n)$. This also forces $\operatorname{dist}(\boldsymbol{x}_n, S_1 \cap S_2) \to 0$. Indeed, otherwise there is a constant $C > 0$ and a subsequence $(\boldsymbol{x}_{n'}) \subset (\boldsymbol{x}_n)$ such that $\operatorname{dist}(\boldsymbol{x}_{n'}, S_1 \cap S_2) \geq C$. Since $(\boldsymbol{x}_{n'})$ is bounded, there is a subsequence $(\boldsymbol{x}_{n''}) \subset (\boldsymbol{x}_{n'})$ that converges, and by the fact that $\max_k \operatorname{dist}(\boldsymbol{x}_{n''}, S_k) = o(1)$, and by compactness of S_k, the limit is necessarily in $S_1 \cap S_2$, which is incompatible with the fact that $\operatorname{dist}(\boldsymbol{x}_{n''}, S_1 \cap S_2) \geq C$. Since $\operatorname{dist}(\boldsymbol{x}_n, S_1 \cap S_2) \to 1$, we have

$$n \max_k \operatorname{dist}(\boldsymbol{x}_n, S_k) < \operatorname{dist}(\boldsymbol{x}_n, S_1 \cap S_2) = o(1). \tag{47}$$

Assume n is large enough that $\mathrm{dist}(\boldsymbol{x}_n, S_1 \cap S_2) < 1/\kappa$ and let $\boldsymbol{s}_n^k$ be the projection of $\boldsymbol{x}_n$ onto S_k, and $\boldsymbol{s}_n^\ddagger$ the projection of $\boldsymbol{x}_n$ onto $S_1 \cap S_2$. Let $T_k = T_{S_k}(\boldsymbol{s}_n^\ddagger)$ and note that $\theta_{\min}(T_1, T_2) \geq \theta := \theta(S_1, S_2)$—see definition (6). Let $\boldsymbol{t}_n^k$ be the projection of $\boldsymbol{s}_n^k$ onto T_k. Assume WLOG that $\|\boldsymbol{t}_n^1 - \boldsymbol{s}_n^1\| \geq \|\boldsymbol{t}_n^2 - \boldsymbol{s}_n^2\|$. Let $\boldsymbol{t}_n$ denote the projection of $\boldsymbol{t}_n^1$ onto $T_1 \cap T_2$, and then let $\boldsymbol{s}_n = P_{S_1 \cap S_2}(\boldsymbol{t}_n)$.

By (47), we have

$$n \max_k \|\boldsymbol{x}_n - \boldsymbol{s}_n^k\| \leq \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\| = o(1). \tag{48}$$

We start with the RHS:

$$\|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\| = \min_{\boldsymbol{s} \in S_1 \cap S_2} \|\boldsymbol{x}_n - \boldsymbol{s}\| \leq \|\boldsymbol{x}_n - \boldsymbol{s}_n\|, \tag{49}$$

and first show that $\|\boldsymbol{x}_n - \boldsymbol{s}_n\| = o(1)$ too. We use the triangle inequality multiple times in what follows. We have

$$\|\boldsymbol{x}_n - \boldsymbol{s}_n\| \leq \|\boldsymbol{x}_n - \boldsymbol{s}_n^1\| + \|\boldsymbol{s}_n^1 - \boldsymbol{t}_n^1\| + \|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| + \|\boldsymbol{t}_n - \boldsymbol{s}_n\|. \tag{50}$$

From (48), $\|\boldsymbol{x}_n - \boldsymbol{s}_n^1\| \leq \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\| = o(1)$, so that $\|\boldsymbol{s}_n^1 - \boldsymbol{s}_n^\ddagger\| = o(1)$, and by (10),

$$\|\boldsymbol{s}_n^1 - \boldsymbol{t}_n^1\| \leq \frac{\kappa}{2}\|\boldsymbol{s}_n^1 - \boldsymbol{s}_n^\ddagger\|^2 \leq \kappa(\|\boldsymbol{s}_n^1 - \boldsymbol{x}_n\|^2 + \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\|^2) = o(1). \tag{51}$$

We also have

$$\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| = \min_{\boldsymbol{t} \in T_1 \cap T_2} \|\boldsymbol{t}_n^1 - \boldsymbol{t}\| \leq \|\boldsymbol{t}_n^1 - \boldsymbol{s}_n^\ddagger\| \leq \|\boldsymbol{t}_n^1 - \boldsymbol{s}_n^1\| + \|\boldsymbol{s}_n^1 - \boldsymbol{x}_n\| + \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\| = o(1). \tag{52}$$

Finally,

$$\|\boldsymbol{t}_n - \boldsymbol{s}_n\| = \min_{\boldsymbol{s} \in S_1 \cap S_2} \|\boldsymbol{t}_n - \boldsymbol{s}\| \leq \|\boldsymbol{t}_n - \boldsymbol{s}_n^\ddagger\| \leq \|\boldsymbol{t}_n - \boldsymbol{t}_n^1\| + \|\boldsymbol{t}_n^1 - \boldsymbol{s}_n^\ddagger\| = o(1).$$

We now proceed. The last upper bound is rather crude. Indeed, using (12) for $S = S_1 \cap S_2$ and $\boldsymbol{s} = \boldsymbol{s}_n^\ddagger$, and noting that $T_{S_1 \cap S_2}(\boldsymbol{s}_n^\ddagger) = T_1 \cap T_2$, and using the fact that $\|\boldsymbol{t}_n - \boldsymbol{s}_n^\ddagger\| = o(1)$, we get

$$\|\boldsymbol{t}_n - \boldsymbol{s}_n\| \leq \kappa\|\boldsymbol{t}_n - \boldsymbol{s}_n^\ddagger\|^2 \leq \kappa(\|\boldsymbol{t}_n - \boldsymbol{s}_n\| + \|\boldsymbol{s}_n - \boldsymbol{x}_n\| + \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\|)^2.$$

Using (49),

$$\|\boldsymbol{t}_n - \boldsymbol{s}_n\| \leq \kappa(\|\boldsymbol{t}_n - \boldsymbol{s}_n\| + 2\|\boldsymbol{s}_n - \boldsymbol{x}_n\|)^2 \leq 5\kappa\|\boldsymbol{x}_n - \boldsymbol{s}_n\|^2, \tag{53}$$

eventually, since $\|\boldsymbol{t}_n - \boldsymbol{s}_n\| = o(1)$.

Combining (49), (50), (51) and (53), we get

$$\|\boldsymbol{x}_n - \boldsymbol{s}_n\| \leq \|\boldsymbol{x}_n - \boldsymbol{s}_n^1\| + O(\|\boldsymbol{x}_n - \boldsymbol{s}_n^1\|^2 + \|\boldsymbol{x}_n - \boldsymbol{s}_n\|^2) + \|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| + O(\|\boldsymbol{x}_n - \boldsymbol{s}_n\|^2),$$

which leads to

$$\|\boldsymbol{x}_n - \boldsymbol{s}_n\| \leq 2\|\boldsymbol{x}_n - \boldsymbol{s}_n^1\| + 2\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\|, \tag{54}$$

when n is large enough. Using this bound in (48) combined with (49), we get

$$\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| \geq \frac{n-2}{2} \max_k \|\boldsymbol{x}_n - \boldsymbol{s}_n^k\|.$$

We then have

$$\begin{aligned}\max_k \|\boldsymbol{x}_n - \boldsymbol{s}_n^k\| &\geq \frac{1}{2}\|\boldsymbol{s}_n^1 - \boldsymbol{s}_n^2\| \\ &\geq \frac{1}{2}(\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n^2\| - \|\boldsymbol{s}_n^1 - \boldsymbol{t}_n^1\| - \|\boldsymbol{s}_n^2 - \boldsymbol{t}_n^2\|) \\ &\geq \frac{1}{2}\operatorname{dist}(\boldsymbol{t}_n^1, T_2) - \|\boldsymbol{s}_n^1 - \boldsymbol{t}_n^1\|,\end{aligned}$$

with

$$\|\boldsymbol{s}_n^1 - \boldsymbol{t}_n^1\| = O(\|\boldsymbol{x}_n - \boldsymbol{s}_n^1\|^2 + \|\boldsymbol{x}_n - \boldsymbol{s}_n^\ddagger\|^2) = O(\|\boldsymbol{x}_n - \boldsymbol{s}_n\|^2) = O(\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\|^2),$$

due (in the same order) to (51), (48)-(49), and (54). Recalling that $\|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| = \operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2)$, we conclude that

$$\operatorname{dist}(\boldsymbol{t}_n^1, T_2) = O(1/n)\operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2) + O(1)\operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2)^2.$$

However, by Lemma 20, we have $\operatorname{dist}(\boldsymbol{t}_n^1, T_2) \geq (\sin\theta)\operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2)$, so that dividing by $\operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2)$ above leads to

$$1 = O(1/n) + O(1)\operatorname{dist}(\boldsymbol{t}_n^1, T_1 \cap T_2),$$

which is in contradiction with the fact that $\operatorname{dist}(\boldsymbol{t}_n^1, T_2) \leq \|\boldsymbol{t}_n^1 - \boldsymbol{t}_n\| = o(1)$, established in (52). □

6.1.6 Covariances near an Intersection

The following compares a covariance matrix of a distribution supported on the union of two smooth surfaces with that of the projection of that distribution on tangent subspaces.

Lemma 22. *Consider $S_1, S_2 \in \mathcal{S}_d(\kappa)$ intersecting at a positive angle, with $\operatorname{reach}(S_1 \cap S_2) \geq 1/\kappa$. Fix $\boldsymbol{s} \in S_1$ and let $A = B(\boldsymbol{s}, r) \cap (S_1 \cup S_2)$, where $r < 1/\kappa$. Let $\boldsymbol{C}$ denote the empirical covariance of an i.i.d. sample from a distribution ν supported on A of size m and define $\boldsymbol{\Sigma} = \operatorname{Cov}(\nu)$. Then*

$$\mathbb{P}\left(\|\boldsymbol{C}_m - \boldsymbol{\Sigma}\| > r^2 t + 3\kappa r^3\right) \leq C_{22}\exp\left(-\frac{mt}{C_{22}}\min(t, m)\right), \quad \forall t \geq 0,$$

where C_{22} depends only on d.

Proof. Let $T_1 = T_{S_1}(\boldsymbol{s})$, and if $\operatorname{dist}(\boldsymbol{s}, S_2) \leq r$, let $\boldsymbol{s}_2 = P_{S_2}(\boldsymbol{s})$ and $T_2 = T_{S_2}(\boldsymbol{s}_2)$. Define $g : A \mapsto T_1 \cup T_2$ where $g(\boldsymbol{x}) = P_{T_1}(\boldsymbol{x})$ if $\boldsymbol{x} \in S_1$, and $g(\boldsymbol{x}) = P_{T_2}(\boldsymbol{x})$ otherwise. By (10), $\|g(\boldsymbol{x}) - \boldsymbol{x}\| \leq \frac{\kappa}{2}r^2$ for all $\boldsymbol{x} \in A$. Let $\boldsymbol{\Sigma}^g = \operatorname{Cov}(\nu^g)$. Also, let ν_m denote the sample distribution and note that $\boldsymbol{C}_m = \operatorname{Cov}(\nu_m)$. Let $\boldsymbol{C}_m^g = \operatorname{Cov}(\nu_m^g)$. Note that ν_m^g is the empirical distribution of an i.i.d. sample of size m from ν^g, so the notation is congruent. If $\operatorname{dist}(\boldsymbol{s}, S_2) > r$, ν^g is supported on $B(\boldsymbol{s}, r) \cap T_1$, which is a d-dimensional ball of radius r, so that (43) applies to $\boldsymbol{C}_m^g$ and $\boldsymbol{\Sigma}^g$. If $\operatorname{dist}(\boldsymbol{s}, S_2) \leq r$, ν^g is supported on

$$(B(\boldsymbol{s}, r) \cap T_1) \cup (B(\boldsymbol{s}_2, r) \cap T_2) \subset B(\boldsymbol{s}, 2r) \cap (T_1 \cup T_2) \subset B(\boldsymbol{s}, 2r) \cap (T_1 + T_2),$$

where the latter is a ball in dimension $d' := \dim(T_1 + T_2) \le 2d$ of radius $2r$, so that (43)—with $2r$ in place of r and $2d$ in place of d—applies to $\boldsymbol{C}_m^g$ and $\boldsymbol{\Sigma}^g$. Then, by the triangle inequality and Lemma 15,

$$\begin{aligned}\|\boldsymbol{C}_m - \boldsymbol{\Sigma}\| &\le \|\boldsymbol{C}_m - \boldsymbol{C}_m^g\| + \|\boldsymbol{C}_m^g - \boldsymbol{\Sigma}^g\| + \|\boldsymbol{\Sigma}^g - \boldsymbol{\Sigma}\| \\ &\le 2\kappa r^3 + \kappa^2 r^4/2 + \|\boldsymbol{C}_m^g - \boldsymbol{\Sigma}^g\|,\end{aligned}$$

as in (38). □

Here is a continuity result.

Lemma 23. *Let T_1 and T_2 be two linear subspaces of same dimension d. For $\boldsymbol{x} \in T_1$, denote by $\boldsymbol{\Sigma}_T(\boldsymbol{x})$ the covariance matrix of the uniform distribution over $B(\boldsymbol{x}, r) \cap (T_1 \cup T_2)$. Then, for all $\boldsymbol{x}, \boldsymbol{y} \in T_1$,*

$$\|\boldsymbol{\Sigma}_T(\boldsymbol{x}) - \boldsymbol{\Sigma}_T(\boldsymbol{y})\| \le \begin{cases} 5dr\,\|\boldsymbol{x} - \boldsymbol{y}\|, & \text{if } d \ge 2, \\ r\,\|\boldsymbol{x} - \boldsymbol{y}\| + 2\sqrt{r^3\,\|\boldsymbol{x} - \boldsymbol{y}\|}, & \text{if } d = 1. \end{cases}$$

We note that, when $d = 1$, we also have the bound

$$\|\boldsymbol{\Sigma}_T(\boldsymbol{x}) - \boldsymbol{\Sigma}_T(\boldsymbol{y})\| \le 4\gamma r\,\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \text{if } \min(\operatorname{dist}(\boldsymbol{x}, T_2), \operatorname{dist}(\boldsymbol{y}, T_2)) \le r\sqrt{1 - 1/\gamma^2}.$$

Proof. Assume without loss of generality that $r = 1$. Let $A_{\boldsymbol{x}}^j = B(\boldsymbol{x}, 1) \cap T_j$ for any $\boldsymbol{x}$ and $j = 1, 2$. By Lemma 14 and then Lemma 7, we have

$$\begin{aligned}\|\boldsymbol{\Sigma}_T(\boldsymbol{x}) - \boldsymbol{\Sigma}_T(\boldsymbol{y})\| &= \|\operatorname{Cov}(\lambda_{A_{\boldsymbol{x}}^1 \cup A_{\boldsymbol{x}}^2}) - \operatorname{Cov}(\lambda_{A_{\boldsymbol{y}}^1 \cup A_{\boldsymbol{y}}^2})\| \\ &\le 3d\,\mathrm{TV}(\lambda_{A_{\boldsymbol{x}}^1 \cup A_{\boldsymbol{x}}^2}, \lambda_{A_{\boldsymbol{y}}^1 \cup A_{\boldsymbol{y}}^2}) \\ &\le 12d\,\frac{\operatorname{vol}\left((A_{\boldsymbol{x}}^1 \cup A_{\boldsymbol{x}}^2) \triangle (A_{\boldsymbol{y}}^1 \cup A_{\boldsymbol{y}}^2)\right)}{\operatorname{vol}\left((A_{\boldsymbol{x}}^1 \cup A_{\boldsymbol{x}}^2) \cup (A_{\boldsymbol{y}}^1 \cup A_{\boldsymbol{y}}^2)\right)} \\ &\le 12d\,\frac{\operatorname{vol}(A_{\boldsymbol{x}}^1 \triangle A_{\boldsymbol{y}}^1) + \operatorname{vol}(A_{\boldsymbol{x}}^2 \triangle A_{\boldsymbol{y}}^2)}{\operatorname{vol}(A_{\boldsymbol{x}}^1)}.\end{aligned}$$

Note that $A_{\boldsymbol{x}}^2 = B(\boldsymbol{x}_2, \eta) \cap T_2$ where $\boldsymbol{x}_2 := P_{T_2}(\boldsymbol{x})$ and $\eta := \sqrt{1 - \|\boldsymbol{x} - \boldsymbol{x}_2\|^2 \wedge 1}$. Similarly, $A_{\boldsymbol{y}}^2 = B(\boldsymbol{y}_2, \delta) \cap T_2$ where $\boldsymbol{y}_2 := P_{T_2}(\boldsymbol{y})$ and $\delta := \sqrt{1 - \|\boldsymbol{y} - \boldsymbol{y}_2\|^2 \wedge 1}$. Therefore, applying Lemma 12, we get

$$\frac{\operatorname{vol}(A_{\boldsymbol{x}}^1 \triangle A_{\boldsymbol{y}}^1)}{2\operatorname{vol}(A_{\boldsymbol{x}}^1)} \le 1 - (1 - t)_+^d, \tag{64}$$

and assuming WLOG that $\delta \le \eta$ (i.e., $\|\boldsymbol{y} - \boldsymbol{y}_2\| \ge \|\boldsymbol{x} - \boldsymbol{x}_2\|$) and after proper scaling, we get

$$\frac{\operatorname{vol}(A_{\boldsymbol{x}}^2 \triangle A_{\boldsymbol{y}}^2)}{2\operatorname{vol}(A_{\boldsymbol{x}}^1)} \le \zeta := \eta^d - (\eta - t_2)_+^d \wedge \delta^d, \tag{65}$$

where $t := \|\boldsymbol{x} - \boldsymbol{y}\|$ and $t_2 := \|\boldsymbol{x}_2 - \boldsymbol{y}_2\|$. Note that $t_2 \le t$ by the fact that P_{T_2} is 1-Lipschitz.

For (64), we have $1 - (1 - t)_+^d \le dt$. This is obvious when $t \ge 1$, while when $t \le 1$ it is obtained using the fact that, for any $0 \le a < b \le 1$,

$$b^d - a^d = (b - a)(b^{d-1} + ab^{d-2} + \cdots + a^{d-2}b + a^{d-1}) \le db^{d-1}(b - a) \le d(b - a). \tag{66}$$

For (65), we consider several cases.

- When $\eta \le t_2$, then $\zeta = \eta^d \le \eta \le t_2 \le t$.
- When $t_2 < \eta \le t_2 + \delta$, then $\zeta = \eta^d - (\eta - t_2)^d \le dt_2 \le dt$, by (66).
- When $\eta \ge t_2 + \delta$ and $d \ge 2$, we have

$$\begin{aligned}\zeta &= \eta^d - \delta^d \le d\eta^{d-1}(\eta - \delta) \le d\eta(\eta - \delta) \le d(\eta^2 - \delta^2) \\ &= d(\|\boldsymbol{y} - \boldsymbol{y}_2\|^2 \wedge 1 - \|\boldsymbol{x} - \boldsymbol{x}_2\|^2 \wedge 1) \\ &= d(\|\boldsymbol{y} - \boldsymbol{y}_2\| \wedge 1 + \|\boldsymbol{x} - \boldsymbol{x}_2\| \wedge 1)(\|\boldsymbol{y} - \boldsymbol{y}_2\| \wedge 1 - \|\boldsymbol{x} - \boldsymbol{x}_2\| \wedge 1) \\ &\le 2d(t + t_2) \le 4dt,\end{aligned}$$

  where (66) was applied in the first line and the triangle inequality was applied in the last line, in the form of

$$\|\boldsymbol{y} - \boldsymbol{y}_2\| \le \|\boldsymbol{y} - \boldsymbol{x}\| + \|\boldsymbol{x} - \boldsymbol{x}_2\| + \|\boldsymbol{x}_2 - \boldsymbol{y}_2\| = \|\boldsymbol{x} - \boldsymbol{x}_2\| + t + t_2.$$

- When $\eta \ge t_2 + \delta$ and $d = 1$, we have

$$\zeta = \eta - \delta \le \sqrt{\|\boldsymbol{y} - \boldsymbol{y}_2\|^2 \wedge 1 - \|\boldsymbol{x} - \boldsymbol{x}_2\|^2 \wedge 1} \le 2\sqrt{t},$$

  using (70), preceded by the fact that, for any $0 \le a < b \le 1$,

$$0 \le \sqrt{1-a} - \sqrt{1-b} = \frac{b-a}{\sqrt{1-a} + \sqrt{1-b}} \le \frac{b-a}{\sqrt{1-b+b-a}} \le \frac{b-a}{\sqrt{b-a}} = \sqrt{b-a}.$$

In summary, when $d \ge 2$, we can therefore bound $\|\boldsymbol{\Sigma}_T(\boldsymbol{x}) - \boldsymbol{\Sigma}_T(\boldsymbol{y})\|$ by $dt + 4dt = 5dt$; and when $d = 1$, we bound that by $t + 2\sqrt{t}$. □

The following is in some sense a converse to Lemma 23, in that we lower-bound the distance between covariance matrices near an intersection of linear subspaces. Note that the covariance matrix does not change when moving parallel to the intersection; however, it does when moving perpendicular to the intersection.

Lemma 24. *Let T_1 and T_2 be two linear subspaces of same dimension d with $\theta_{\min}(T_1, T_2) \ge \theta_0 > 0$. Fix a unit norm vector $\boldsymbol{v} \in T_1 \cap (T_1 \cap T_2)^\perp$. With $\boldsymbol{\Sigma}_T(h\boldsymbol{v})$ denoting the covariance of the uniform distribution over $B(h\boldsymbol{v}, r) \cap (T_1 \cup T_2)$, we have*

$$\inf_h \sup_\ell \|\boldsymbol{\Sigma}_T(h\boldsymbol{v}) - \boldsymbol{\Sigma}_T(\ell\boldsymbol{v})\| \ge r^2/C_{24},$$

where the infimum is over $0 \le h \le 1/\sin\theta_0$ and the supremum over $\max(0, h - 1/2) \le \ell \le \min(1/\sin\theta_0, h + 1/2)$, and $C_{24} \ge 1$ depends only on d and θ_0.

Proof. Assume without loss of generality that $r = 1$. If the statement of the lemma is not true, there are subspaces T_1 and T_2 of same dimension d, a unit length vector $\boldsymbol{v} \in T_1 \cap (T_1 \cap T_2)^\perp$ and $0 \le h \le 1/\sin\theta_0$, such that

$$\boldsymbol{\Sigma}_T(\ell\boldsymbol{v}) = \boldsymbol{\Sigma}_T(h\boldsymbol{v}) \text{ for all } \max(0, h - 1/2) \le \ell \le \min(1/\sin\theta_0, h + 1/2). \tag{71}$$

This is because $\ell \mapsto \mathbf{\Sigma}_T(\ell \boldsymbol{v})$ is continuous once T_1, T_2 and $\boldsymbol{v}$ are chosen. By projecting onto $(T_1 \cap T_2)^\perp$, we may assume that $T_1 \cap T_2 = \{0\}$ without loss of generality. Let $\theta = \angle(\boldsymbol{v}, T_2)$ and note that $\theta \geq \theta_0$ since $T_1 \cap T_2 = \{0\}$. Define $\boldsymbol{u} = (\boldsymbol{v} - P_{T_2}\boldsymbol{v})/\sin\theta$ and also $\boldsymbol{w} = P_{T_2}\boldsymbol{v}/\cos\theta$ when $\theta < \pi/2$, and $\boldsymbol{w} \in T_2$ is any vector perpendicular to $\boldsymbol{v}$ when $\theta = \pi/2$. $B(h\boldsymbol{v}, 1) \cap T_1$ is the d-dimensional ball of T_1 of radius 1 and center $h\boldsymbol{v}$, while—using Pythagoras theorem—$B(h\boldsymbol{v}, 1) \cap T_2$ is the d-dimensional ball of T_2 of radius $t := (1 - (h\sin\theta)^2)^{1/2}$ and center $(h\cos\theta)\boldsymbol{w}$. Let λ denote the uniform distribution over $B(h\boldsymbol{v}, 1) \cap (T_1 \cup T_2)$ and λ_k the uniform distribution over $B(h\boldsymbol{v}, 1) \cap T_k$. Note that $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2$, where

$$\alpha := \frac{\mathrm{vol}(B(h\boldsymbol{v}, 1) \cap T_1)}{\mathrm{vol}(B(h\boldsymbol{v}, 1) \cap (T_1 \cup T_2))} = \frac{\mathrm{vol}(B(h\boldsymbol{v}, 1) \cap T_1)}{\mathrm{vol}(B(h\boldsymbol{v}, 1) \cap T_1) + \mathrm{vol}(B((h\cos\theta)\boldsymbol{w}, t) \cap T_2)} = \frac{1}{1 + t^d}.$$

By the law of total covariance,

$$\mathrm{Cov}(\lambda) = \alpha\,\mathrm{Cov}(\lambda_1) + (1-\alpha)\,\mathrm{Cov}(\lambda_2) + \alpha(1-\alpha)(\mathbb{E}(\lambda_1) - \mathbb{E}(\lambda_2))(\mathbb{E}(\lambda_1) - \mathbb{E}(\lambda_2))^\top. \quad (72)$$

with $\mathbb{E}(\lambda_1) = h\boldsymbol{v}$ and $\mathbb{E}(\lambda_2) = h\cos(\theta)\boldsymbol{w}$, and $\mathrm{Cov}(\lambda_1) = c\boldsymbol{P}_{T_1}$ and $\mathrm{Cov}(\lambda_2) = t^2 c\boldsymbol{P}_{T_2}$, by Lemma 13. Hence,

$$\mathbf{\Sigma}_T(h\boldsymbol{v}) = \alpha c\boldsymbol{P}_{T_1} + (1-\alpha)t^2 c\boldsymbol{P}_{T_2} + \alpha(1-\alpha)(1-t^2)\boldsymbol{u}\boldsymbol{u}^\top,$$

using the fact that $\boldsymbol{v} - (\cos\theta)\boldsymbol{w} = (\sin\theta)\boldsymbol{u}$ and the definition of t. Let $\theta_1 = \theta_{\max}(T_1, T_2)$ and let $\boldsymbol{v}_1 \in T_1$ be of unit length and such that $\angle(\boldsymbol{v}_1, T_2) = \theta_1$. Then for any $0 \leq h, \ell \leq 1/\sin\theta_0$, we have

$$\|\mathbf{\Sigma}_T(h\boldsymbol{v}) - \mathbf{\Sigma}_T(\ell\boldsymbol{v})\| \geq |\boldsymbol{v}_1^\top \mathbf{\Sigma}_T(h\boldsymbol{v})\boldsymbol{v}_1 - \boldsymbol{v}_1^\top \mathbf{\Sigma}_T(\ell\boldsymbol{v})\boldsymbol{v}_1| = |f(t_h) - f(t_\ell)|, \quad (73)$$

where $t_h := (1 - (h\sin\theta)^2)^{1/2}$ and

$$f(t) \;=\; \frac{c}{1+t^d} + \frac{ct^{d+2}(\cos\theta_1)^2}{1+t^d} + \frac{t^d(1-t^2)(\boldsymbol{u}^\top\boldsymbol{v}_1)^2}{(1+t^d)^2}.$$

It is easy to see that the interval

$$I_h = \{t_\ell : (h - 1/2)_+ \leq \ell \leq (1/\sin\theta_0) \wedge (h + 1/2)\}$$

is non empty. Because of (71) and (73), $f(t)$ is constant over $t \in I_h$, but this is not possible since f is a rational function not equal to a constant and therefore cannot be constant over an interval of positive length. □

We now look at the eigenvalues of the covariance matrix.

Lemma 25. *Let T_1 and T_2 be two linear subspaces of same dimension d. For $\boldsymbol{x} \in T_1$, denote by $\mathbf{\Sigma}_T(\boldsymbol{x})$ the covariance matrix of the uniform distribution over $B(\boldsymbol{x}, 1) \cap (T_1 \cup T_2)$. Then, for all $\boldsymbol{x} \in T_1$,*

$$c\big(1 - (1 - \delta^2(\boldsymbol{x}))_+^{d/2}\big) \leq \beta_d(\mathbf{\Sigma}_T(\boldsymbol{x})), \qquad \beta_1(\mathbf{\Sigma}_T(\boldsymbol{x})) \leq c + \delta^2(\boldsymbol{x})(1 - \delta^2(\boldsymbol{x}))_+^{d/2}, \quad (75)$$

$$\frac{c}{8}(1 - \cos\theta_{\max}(T_1, T_2))^2(1 - \delta^2(\boldsymbol{x}))_+^{d/2+1} \leq \beta_{d+1}(\mathbf{\Sigma}_T(\boldsymbol{x})) \leq (c + \delta^2(\boldsymbol{x}))(1 - \delta^2(\boldsymbol{x}))_+^{d/2}, \quad (76)$$

where $c := 1/(d+2)$ and $\delta(\boldsymbol{x}) := \mathrm{dist}(\boldsymbol{x}, T_2)$.

Proof. Applying (72), we have

$$\boldsymbol{\Sigma}_T(x) = \alpha c \boldsymbol{P}_{T_1} + (1-\alpha) c t^2 \boldsymbol{P}_{T_2} + \alpha(1-\alpha)(\boldsymbol{x} - \boldsymbol{x}_2)(\boldsymbol{x} - \boldsymbol{x}_2)^\top, \tag{77}$$

where $\boldsymbol{x}_2 := P_{T_2}(\boldsymbol{x})$ and $\alpha := (1+t^d)^{-1}$ with $t := (1 - \delta^2(\boldsymbol{x}))_+^{1/2}$. Because all the matrices in this display are positive semidefinite, we have

$$\beta_d(\boldsymbol{\Sigma}_T(\boldsymbol{x})) \geq \alpha c \|\boldsymbol{P}_{T_1}\| = \alpha c,$$

with $\alpha \geq 1 - t^d$. And because of the triangle inequality, we have

$$\beta_1(\boldsymbol{\Sigma}_T(\boldsymbol{x})) \leq \alpha c \|\boldsymbol{P}_{T_1}\| + (1-\alpha) c t^2 \|\boldsymbol{P}_{T_2}\| + \alpha(1-\alpha)\|\boldsymbol{x} - \boldsymbol{x}_2\|^2 \leq c + \alpha(1-\alpha)\delta^2(\boldsymbol{x}),$$

with $\alpha(1-\alpha) \leq t^d$. Hence, (75) is proved.

For the upper bound in (76), by Weyl's inequality (Stewart and Sun, 1990, Cor. IV.4.9) and the fact that $\beta_{d+1}(\boldsymbol{P}_{T_1}) = 0$, and then the triangle inequality, we get

$$\begin{aligned} \beta_{d+1}(\boldsymbol{\Sigma}_T(\boldsymbol{x})) &\leq \|\boldsymbol{\Sigma}_T(\boldsymbol{x}) - \alpha c \boldsymbol{P}_{T_1}\| \\ &\leq c(1-\alpha) t^2 \|\boldsymbol{P}_{T_2}\| + \alpha(1-\alpha)\delta^2(\boldsymbol{x}) \\ &\leq (1-\alpha)(c + \delta^2(\boldsymbol{x})), \end{aligned}$$

and we then use the fact that $1 - \alpha \leq t^d$.

For the lower bound in (76), let $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_d$ denote the principal angles between T_1 and T_2. By the definition of principal angles, there are orthonormal bases for T_1 and T_2, denoted $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$, such that $\boldsymbol{v}_j^\top \boldsymbol{w}_k = \mathbb{1}_{j=k} \cdot \cos\theta_j$. Take $\boldsymbol{u} \in \operatorname{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d, \boldsymbol{w}_1)$, that is, of the form $\boldsymbol{u} = a\boldsymbol{v}_1 + \boldsymbol{v} + b\boldsymbol{w}_1$, with $\boldsymbol{v} \in \operatorname{span}(\boldsymbol{v}_2, \ldots, \boldsymbol{v}_d)$. Since $\boldsymbol{P}_{T_1} = \boldsymbol{v}_1\boldsymbol{v}_1^\top + \cdots + \boldsymbol{v}_d\boldsymbol{v}_d^\top$ and $\boldsymbol{P}_{T_2} = \boldsymbol{w}_1\boldsymbol{w}_1^\top + \cdots + \boldsymbol{w}_d\boldsymbol{w}_d^\top$, when $\|\boldsymbol{u}\| = 1$, we have

$$\frac{1}{c}\boldsymbol{u}^\top \boldsymbol{\Sigma}_T(\boldsymbol{x})\boldsymbol{u} \geq \alpha \boldsymbol{u}^\top \boldsymbol{P}_{T_1}\boldsymbol{u} + (1-\alpha)t^2 \boldsymbol{u}^\top \boldsymbol{P}_{T_2}\boldsymbol{u}$$

with

$$\boldsymbol{u}^\top \boldsymbol{P}_{T_1}\boldsymbol{u} = a^2 + \|\boldsymbol{v}\|^2 + 2ab\cos\theta_1 + b^2\cos^2\theta_1 = 1 - b^2\sin^2\theta_1 \geq 0,$$

and

$$\boldsymbol{u}^\top \boldsymbol{P}_{T_2}\boldsymbol{u} = b^2 + 2ab\cos\theta_1 + a^2\cos^2\theta_1 = (a\cos\theta_1 + b)^2.$$

If $|b| \leq 1/2$, then $\boldsymbol{u}^\top \boldsymbol{P}_{T_1}\boldsymbol{u} \geq 3/4$, implying that $\frac{1}{c}\boldsymbol{u}^\top \boldsymbol{\Sigma}_T(\boldsymbol{x})\boldsymbol{u} \geq 3\alpha/4 \geq 3/8$. Otherwise,

$$\begin{aligned} \boldsymbol{u}^\top \boldsymbol{P}_{T_1}\boldsymbol{u} + \boldsymbol{u}^\top \boldsymbol{P}_{T_2}\boldsymbol{u} &\geq (a + b\cos\theta_1)^2 + (a\cos\theta_1 + b)^2 \\ &\geq (1 - \cos\theta_1)^2(a^2 + b^2) \geq (1-\cos\theta_1)^2/4, \end{aligned}$$

and using the fact that $\alpha \geq 1 - \alpha \geq (1-\alpha)t^2$, we get

$$\frac{1}{c}\boldsymbol{u}^\top \boldsymbol{\Sigma}_T(\boldsymbol{x})\boldsymbol{u} \geq (1-\alpha)t^2(1-\cos\theta_1)^2/4.$$

This last bound is always worst. Hence, by the Courant-Fischer theorem (Stewart and Sun, 1990, Cor. IV.4.7), we have

$$\beta_{d+1}(\boldsymbol{\Sigma}_T(\boldsymbol{x})) \geq \frac{c}{4}(1-\alpha)t^2(1-\cos\theta_1)^2,$$

and we conclude with $1 - \alpha \geq t^d/2$. □

Below are two technical results on the covariance matrix of the uniform distribution on the intersection of a ball and the union of two smooth surfaces, near where the surfaces intersect. The first one is in fact a stepping stone to the second. The latter will enable us to apply Lemma 23.

Lemma 26. *There is a constant $C_{26} \geq 3$ such that the following holds. Consider $S_1, S_2 \in \mathcal{S}_d(\kappa)$. Fix $r \leq \frac{1}{C_{26}\kappa}$, and for $\boldsymbol{s} \in S_1$ with $\operatorname{dist}(\boldsymbol{s}, S_2) < 1/\kappa$, let $\boldsymbol{\Sigma}(\boldsymbol{s})$ and $\boldsymbol{\Sigma}_T(\boldsymbol{s})$ denote the covariance matrices of the uniform distributions over $B(\boldsymbol{s}, r) \cap (S_1 \cup S_2)$ and $B(\boldsymbol{s}, r) \cap (T_1 \cup T_2)$, respectively, where $T_1 := T_{S_1}(\boldsymbol{s})$ and $T_2 := T_{S_2}(P_{S_2}(\boldsymbol{s}))$. Then*

$$\|\boldsymbol{\Sigma}(\boldsymbol{s}) - \boldsymbol{\Sigma}_T(\boldsymbol{s})\| \leq C_{26}\kappa r^3. \tag{83}$$

Proof. We note that it is enough to prove the result when r is small enough. Let $\delta = \operatorname{dist}(\boldsymbol{s}, S_2)$ and let $\boldsymbol{s}_2 = P_{S_2}(\boldsymbol{s})$, so that $\|\boldsymbol{s} - \boldsymbol{s}_2\| = \delta$. We first treat the case where $\delta \leq r$. Let B_r be short for $B(\boldsymbol{s}, r)$ and define $A_k = B_r \cap S_k$, $\boldsymbol{\mu}_k = \mathbb{E}(\lambda_{A_k})$ and $\boldsymbol{D}_k = \operatorname{Cov}(\lambda_{A_k})$, for $k = 1, 2$. Applying (72), we have

$$\boldsymbol{\Sigma}(\boldsymbol{s}) = \alpha \boldsymbol{D}_1 + (1-\alpha)\boldsymbol{D}_2 + \alpha(1-\alpha)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top,$$

where

$$\alpha := \frac{\operatorname{vol}(A_1)}{\operatorname{vol}(A_1) + \operatorname{vol}(A_2)}.$$

Recall that $T_1 = T_{S_1}(\boldsymbol{s})$ and $T_2 = T_{S_2}(\boldsymbol{s}_2)$, and define $A'_k = B_r \cap T_k$, so that $B_r \cap (T_1 \cup T_2) = A'_1 \cup A'_2$. Note that $\mathbb{E}(\lambda_{A'_1}) = \boldsymbol{s}$ and $\mathbb{E}(\lambda_{A'_2}) = \boldsymbol{s}_2$, and by Lemma 13, $\boldsymbol{D}'_1 := \operatorname{Cov}(\lambda_{A'_1}) = cr^2 \boldsymbol{P}_{T_1}$ and $\boldsymbol{D}'_2 := \operatorname{Cov}(\lambda_{A'_2}) = c(r^2 - \delta^2)\boldsymbol{P}_{T_2}$, where $c := 1/(d+2)$. Applying (72), we have

$$\boldsymbol{\Sigma}_T(\boldsymbol{s}) = \alpha' \boldsymbol{D}'_1 + (1-\alpha')\boldsymbol{D}'_2 + \alpha'(1-\alpha')(\boldsymbol{s} - \boldsymbol{s}_2)(\boldsymbol{s} - \boldsymbol{s}_2)^\top,$$

where

$$\alpha' := \frac{\operatorname{vol}(A'_1)}{\operatorname{vol}(A'_1) + \operatorname{vol}(A'_2)}.$$

By the triangle inequality, we have

$$\|\alpha \boldsymbol{D}_1 - \alpha' \boldsymbol{D}'_1\| \leq |\alpha' - \alpha| \|\boldsymbol{D}'_1\| + \alpha \|\boldsymbol{D}_1 - \boldsymbol{D}'_1\| \leq cr^2 |\alpha' - \alpha| + \|\boldsymbol{D}_1 - \boldsymbol{D}'_1\|,$$

and similarly,

$$\|(1-\alpha)\boldsymbol{D}_2 - (1-\alpha')\boldsymbol{D}'_2\| \leq cr^2 |\alpha' - \alpha| + (1-\alpha)\|\boldsymbol{D}_2 - \boldsymbol{D}'_2\|.$$

Assuming that $\kappa r \leq 1/C_{16}$, by Lemma 16, we have $\|\boldsymbol{D}_1 - \boldsymbol{D}'_1\| \vee \|\boldsymbol{D}_2 - \boldsymbol{D}'_2\| \leq C_{16}\kappa r^3$. Because $|\alpha'(1-\alpha') - \alpha(1-\alpha)| \leq |\alpha' - \alpha|$ and the identity $\|\boldsymbol{a}\boldsymbol{a}^\top - \boldsymbol{b}\boldsymbol{b}^\top\| \leq (\|\boldsymbol{a}\| + \|\boldsymbol{b}\|)\|\boldsymbol{a} - \boldsymbol{b}\|$, we also have

$$\begin{aligned}
&\|\alpha(1-\alpha)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top - \alpha'(1-\alpha')(\boldsymbol{s} - \boldsymbol{s}_2)(\boldsymbol{s} - \boldsymbol{s}_2)^\top\| \\
&\leq |\alpha' - \alpha| \|\boldsymbol{s} - \boldsymbol{s}_2\|^2 + \alpha(1-\alpha) 3r\big(\|\boldsymbol{\mu}_1 - \boldsymbol{s}\| + \|\boldsymbol{\mu}_2 - \boldsymbol{s}_2\|\big).
\end{aligned}$$

By Lemma 16, we have $\|\boldsymbol{\mu}_1 - \boldsymbol{s}\| \vee \|\boldsymbol{\mu}_2 - \boldsymbol{s}_2\| \leq C_{16}\kappa r^2$. Assuming that $\kappa r \leq 1/3$, $P_{T_k}^{-1}$ is well-defined and $(1+\kappa r)$-Lipschitz on $S_k \cap B_r$. And being an orthogonal projection, P_{T_k} is 1-Lipschitz . Hence, applying Lemma 6, we have

$$1 \leq \frac{\mathrm{vol}(A_k)}{\mathrm{vol}(P_{T_k}(A_k))} \leq 1 + \kappa r, \quad k = 1, 2.$$

Then by Lemma 9,

$$\frac{\mathrm{vol}(P_{T_k}(A_k) \triangle A_k')}{\mathrm{vol}(A_k')} \leq C_9\kappa(r + \kappa r^2) \leq 2C_9\kappa r, \quad k = 1, 2,$$

when κr is small enough, implying

$$1 - 2C_9\kappa r \leq \frac{\mathrm{vol}(P_{T_k}(A_k))}{\mathrm{vol}(A_k')} \leq 1 + 2C_9\kappa r, \quad k = 1, 2.$$

Hence,

$$1 - C\kappa r \leq 1 - 2C_9\kappa r \leq \frac{\mathrm{vol}(A_k)}{\mathrm{vol}(A_k')} \leq (1 + \kappa r)(1 + 2C_9\kappa r) \leq 1 + C\kappa r, \quad k = 1, 2,$$

for some constant $C > 0$. Since for all $a, b, a', b' > 0$ we have

$$\left| \frac{a}{a+b} - \frac{a'}{a'+b'} \right| \leq \frac{|a - a'| \vee |b - b'|}{(a+b) \vee (a'+b')} \leq |1 - a/a'| \vee |1 - b/b'|,$$

we then get

$$|\alpha - \alpha'| \leq C\kappa r.$$

Applying the triangle inequality to $\|\boldsymbol{\Sigma}(\boldsymbol{s}) - \boldsymbol{\Sigma}_T(\boldsymbol{s})\|$ and using the bounds above, we conclude.

We assumed above that $\delta \leq r$. When $\delta > r$, $A_2 = A_2' = \emptyset$ and $\alpha = \alpha' = 1$. In this case, $\boldsymbol{\mu}_2, \boldsymbol{D}_2, \mathrm{D}_2'$ are not well-defined, but one can easily check that the above calculations, and the resulting bound, remain valid. In fact, in this case $\boldsymbol{\Sigma}(\boldsymbol{s}) = \boldsymbol{D}_1$ and $\boldsymbol{\Sigma}_T(\boldsymbol{s}) = \boldsymbol{D}_1'$. □

Lemma 27. *Consider $S_1, S_2 \in \mathcal{S}_d(\kappa)$ intersecting at a positive angle, with* $\mathrm{reach}(S_1 \cap S_2) \geq 1/\kappa$. *There is a constant $C_{27} \geq 1$, depending on S_1 and S_2, such that the following holds. For $\boldsymbol{s} \in S_1$ with $\mathrm{dist}(\boldsymbol{s}, S_2) \leq 2r$, assuming $r \leq \frac{1}{C_{27}}$, let $\boldsymbol{\Sigma}(\boldsymbol{s})$ and $\boldsymbol{\Sigma}_T^0(\boldsymbol{s})$ denote the covariance matrices of the uniform distributions over $B(\boldsymbol{s}, r) \cap (S_1 \cup S_2)$ and $B(P_{T_1^0}(\boldsymbol{s}), r) \cap (T_1^0 \cup T_2^0)$, where $T_k^0 := T_{S_k}(\boldsymbol{s}^0)$ and $\boldsymbol{s}^0 := P_{S_1 \cap S_2}(\boldsymbol{s})$. Then*

$$\|\boldsymbol{\Sigma}(\boldsymbol{s}) - \boldsymbol{\Sigma}_T^0(\boldsymbol{s})\| \leq C_{27}\, r^{5/2}.$$

Proof. We use the notation of Lemma 26 and its proof. In addition, we let $C > 0$ denote a constant that depends only on S_1 and S_2 whose value increases with each appearance.

Define $\boldsymbol{t}_1 = P_{T_1^0}(\boldsymbol{s})$ and $\boldsymbol{t}_2 = P_{T_2^0}(\boldsymbol{t}_1)$. Let $\delta_0 = \|\boldsymbol{t}_1 - \boldsymbol{t}_2\|$, $\delta_1 = \|\boldsymbol{s} - \boldsymbol{t}_1\|$, $\delta_2 = \|\boldsymbol{s}_2 - \boldsymbol{t}_2\|$. Define $A_1^0 = T_1^0 \cap B(\boldsymbol{t}_1, r)$ and $A_2^0 = T_2^0 \cap B(\boldsymbol{t}_1, r)$. Note that $A_2^0 = T_2^0 \cap B(\boldsymbol{t}_2, \sqrt{r^2 - \delta_0^2})$

when $\delta_0 \le r$, and $A_2^0 = \emptyset$ when $\delta_0 > r$. We have $\|\boldsymbol{s} - \boldsymbol{s}^0\| \le Cr$ by Lemma 21, which then implies $\delta_1 \le Cr^2$ by Lemma 2. Assuming that κr is small enough,

$$\begin{aligned}\delta_2 = \|\boldsymbol{s}_2 - \boldsymbol{t}_2\| &\le \|P_{S_2}(\boldsymbol{s}) - P_{T_2^0}(\boldsymbol{s})\| + \|P_{T_2^0}(\boldsymbol{s}) - P_{T_2^0}(\boldsymbol{t}_1)\| \\ &\le C_4\kappa\|\boldsymbol{s} - \boldsymbol{s}^0\|^2 + \|\boldsymbol{s} - \boldsymbol{t}_1\| \le Cr^2,\end{aligned}$$

where in the second line we used Lemma 4 and the fact that any metric projection is 1-Lipschitz. Hence,

$$|\delta_0 - \delta| \le \big|\|\boldsymbol{t}_1 - \boldsymbol{t}_2\| - \|\boldsymbol{s} - \boldsymbol{s}_2\|\big| \le \|\boldsymbol{t}_1 - \boldsymbol{s}\| + \|\boldsymbol{t}_2 - \boldsymbol{s}_2\| = \delta_1 + \delta_2 \le Cr^2.$$

Note that $\mathbb{E}(\lambda_{A_1^0}) = \boldsymbol{t}_1$ and, by Lemma 13 (with c denoting the constant defined there), $\boldsymbol{D}_1^0 := \mathrm{Cov}(\lambda_{A_1^0}) = cr^2\boldsymbol{P}_{T_1^0}$. If $\delta_0 \le r$, $\mathbb{E}(\lambda_{A_2^0}) = \boldsymbol{t}_2$ and $\boldsymbol{D}_2^0 := \mathrm{Cov}(\lambda_{A_2^0}) = c(r^2 - \delta_0^2)\boldsymbol{P}_{T_2^0}$; if $\delta_0 > r$, then $A_2^0 = \emptyset$, and we define $\boldsymbol{D}_2^0 = 0$ in that case. Applying (72), we have

$$\boldsymbol{\Sigma}_T^0(\boldsymbol{s}) = \alpha^0\boldsymbol{D}_1^0 + (1-\alpha^0)\boldsymbol{D}_2^0 + \alpha^0(1-\alpha^0)(\boldsymbol{t}_1 - \boldsymbol{t}_2)(\boldsymbol{t}_1 - \boldsymbol{t}_2)^\top,$$

where

$$\alpha^0 := \frac{\mathrm{vol}(A_1^0)}{\mathrm{vol}(A_1^0) + \mathrm{vol}(A_2^0)}.$$

Note that $\alpha^0 = 1$ when $\delta_0 > r$ (i.e., when $A_2^0 = \emptyset$). We focus on the case where $\max(\delta, \delta_0) \le r$, which is the most complex one. We have

$$\|\boldsymbol{D}_1' - \boldsymbol{D}_1^0\| = \|cr^2\boldsymbol{P}_{T_1} - cr^2\boldsymbol{P}_{T_1^0}\| = cr^2\|\boldsymbol{P}_{T_1} - \boldsymbol{P}_{T_1^0}\|,$$

and

$$\|\boldsymbol{D}_2' - \boldsymbol{D}_2^0\| = \|c(r^2 - \delta^2)\boldsymbol{P}_{T_2} - c(r^2 - \delta_0^2)_+\boldsymbol{P}_{T_2^0}\| \le cr^2\|\boldsymbol{P}_{T_2} - \boldsymbol{P}_{T_2^0}\| + c|\delta^2 - \delta_0^2|,$$

by the triangle inequality and the fact that $\|\boldsymbol{P}_T\| \le 1$ for any affine subspace T. By Lemma 3 and Lemma 18, we have

$$\|\boldsymbol{P}_{T_1} - \boldsymbol{P}_{T_1^0}\| \le 6\kappa\|\boldsymbol{s} - \boldsymbol{s}^0\| \le Cr, \qquad \|\boldsymbol{P}_{T_2} - \boldsymbol{P}_{T_2^0}\| \le 6\kappa\|\boldsymbol{s}_2 - \boldsymbol{s}^0\| \le Cr,$$

since

$$\|\boldsymbol{s}_2 - \boldsymbol{s}^0\| \le \|\boldsymbol{s}_2 - \boldsymbol{s}\| + \|\boldsymbol{s} - \boldsymbol{s}^0\| \le 2r + Cr.$$

And since $|\delta^2 - \delta_0^2| \le 2r|\delta - \delta_0| \le Cr^3$, we have $\|\boldsymbol{D}_k' - \boldsymbol{D}_k^0\| \le Cr^3$ for $k = 1, 2$. Let ω_d denote the volume of the d-dimensional unit ball. Then

$$\mathrm{vol}(A_1') = \omega_d r^d, \quad \mathrm{vol}(A_2') = \omega_d(r^2 - \delta^2)^{d/2}, \quad \mathrm{vol}(A_1^0) = \omega_d r^2, \quad \mathrm{vol}(A_2^0) = \omega_d(r^2 - {\delta_0}^2)^{d/2},$$

so that

$$\begin{aligned}|\alpha' - \alpha^0| &= \left|\frac{1}{1 + (1 - \delta/r)^{d/2}} - \frac{1}{1 + (1 - \delta_0/r)^{d/2}}\right| \\ &\le \big|(1 - \delta/r)^{d/2} - (1 - \delta_0/r)^{d/2}\big|.\end{aligned}$$

Proceeding as when we bounded ζ in the proof of Lemma 23, we get

$$|\alpha' - \alpha^0| \le d\sqrt{|\delta - \delta_0|/r} \le C\sqrt{r}.$$

Hence, we proved that

$$\|\boldsymbol{\Sigma}_T(\boldsymbol{s}) - \boldsymbol{\Sigma}_T^0(\boldsymbol{s})\| \le Cr^{5/2}.$$

We use this inequality, (83), and the triangle inequality, to get

$$\|\boldsymbol{\Sigma}(\boldsymbol{s}) - \boldsymbol{\Sigma}_T^0(\boldsymbol{s})\| \le \|\boldsymbol{\Sigma}(\boldsymbol{s}) - \boldsymbol{\Sigma}_T(\boldsymbol{s})\| + \|\boldsymbol{\Sigma}_T(\boldsymbol{s}) - \boldsymbol{\Sigma}_T^0(\boldsymbol{s})\| \le Cr^{5/2},$$

which is what we needed to prove. □

6.2 Performance Guarantees for Algorithm 2

We deal with the case where there is no noise, that is, $\tau = 0$ in (7), so that the data points are $\boldsymbol{s}_1, \dots, \boldsymbol{s}_N$ and sampled exactly on $S_1 \cup S_2$ according to the uniform distribution. We explain how things change when there is noise, meaning $\tau > 0$, in Section 6.4.

We start with some notation. Let $\Xi_i = \{j \ne i : \boldsymbol{s}_j \in N_r(\boldsymbol{s}_i)\}$, with (random) cardinality $N_i = |\Xi_i|$. For $i \in [n]$, let $K_i = 1$ if $\boldsymbol{s}_i \in S_1$ and $= 2$ otherwise, and let $T_i = T_{S_{K_i}}(\boldsymbol{s}_i)$, which is the tangent subspace associated with data point $\boldsymbol{s}_i$. Let $\boldsymbol{P}_i$ be short for $\boldsymbol{P}_{T_i}$. Let

$$I_\star = \{i : K_j = K_i, \forall j \in \Xi_i\},$$

or equivalently,

$$I_\star^c = \{i : \exists j \text{ s.t. } K_j \ne K_i \text{ and } \|\boldsymbol{s}_j - \boldsymbol{s}_i\| \le r\}.$$

Thus $I_\star$ indexes the points whose neighborhoods do not contain points from the other cluster.

We will soon define a constant $C_\bullet$ and will require the following

$$C_\bullet \ge C, \quad r \le \varepsilon/C, \quad \varepsilon \le \eta/C, \quad \eta \le 1/C, \tag{90}$$

for a large enough constant $C \ge 1$ that depends only on S_1 and S_2.

6.2.1 A Concentration Bound for Local Covariances

Objective. *We derive a concentration bound for the local covariances.*

Let $\Xi_i = \{j \ne i : \boldsymbol{s}_j \in N_r(\boldsymbol{s}_i)\}$, with (random) cardinality $N_i = |\Xi_i|$. When there is no noise, $\boldsymbol{C}_i$ is the sample covariance of $\{\boldsymbol{s}_j : j \in \Xi_i\}$. Also, let $\boldsymbol{\Sigma}(\boldsymbol{s})$ denote the covariance matrix of the uniform distribution over $B(\boldsymbol{s}, r) \cap (S_1 \cup S_2)$, and note that $\boldsymbol{\Sigma}_i := \boldsymbol{\Sigma}(\boldsymbol{s}_i) = \mathbb{E}(\boldsymbol{C}_i|\boldsymbol{s}_i)$. Given N_i, $\{\boldsymbol{s}_j : j \in \Xi_i\}$ are uniformly distributed on $B(\boldsymbol{s}_i, r) \cap (S_1 \cup S_2)$, and applying Lemma 22, we get that for any $t > 0$

$$\mathbb{P}\left(\|\boldsymbol{C}_i - \boldsymbol{\Sigma}_i\| > r^2 t + 3\kappa r^3 \,\middle|\, \boldsymbol{s}_i, N_i\right) \le C_{22} \exp\left(-\frac{N_i t}{C_{22}} \min\left(t, N_i\right)\right). \tag{91}$$

Assume that $r < 1/(C_{11}\kappa)$ and let $n_\ddagger := nr^d/(2C_{11}/\chi)$ where $\chi := \frac{\mathrm{vol}_d(S_1) \wedge \mathrm{vol}_d(S_2)}{\mathrm{vol}_d(S_1) + \mathrm{vol}_d(S_2)}$. Define

$$\Omega_1 = \{\forall i : N_i > n_\ddagger\}.$$

Note that, for any i,

$$N_i \geq N_i' := \#\{j \neq i : K_i = K_j \text{ and } \boldsymbol{s}_j \in B(\boldsymbol{s}_i, r)\},$$

meaning that N_i' counts the number of neighboring points from the same surface. We have that $\#\{i : K_i = k\} \sim \text{Bin}(n, \chi_k)$ where $\chi_k := \frac{\text{vol}_d(S_k)}{\text{vol}_d(S_1)+\text{vol}_d(S_2)}$, and using a standard concentration inequality for the binomial distribution, there is a numeric constant $C > 0$ such that

$$\mathbb{P}(\#\{i : K_i = k\} \geq \chi n/2) \geq 1 - \exp(-\chi n/C).$$

(We also used the fact that $\chi = \chi_1 \wedge \chi_2$.) With this, we may apply Lemma 11, to get

$$\begin{aligned} \mathbb{P}(\Omega_1^c) &\leq 2\exp(-\chi n/C) + C_{11} r^{-d} \exp(-(\chi n/2) r^d / C_{11}) \\ &\leq Cn \exp(-nr^d/C), \end{aligned}$$

for some other constant $C > 0$. (We used the fact that the left-hand side can be made greater than 1 when $nr^d \leq 1$ by choosing C large enough, so that we may focus on the case where $nr^d \geq 1$.) Define the event

$$\Omega_2 = \Big\{\forall i : \|\boldsymbol{C}_i - \boldsymbol{\Sigma}_i\| \leq r^2\eta/C_\bullet + 3\kappa r^3\Big\},$$

where $C_\bullet$ will be chosen large enough, but fixed, later on. Assume that $n_\ddagger \geq \eta/C_\bullet$. Note that $\eta \leq 1$ by assumption, and we may assume $n_\ddagger$ to be larger than any given constant, for if not, the statement in Theorem 1 is void. Based on the union bound and (91), we have

$$\begin{aligned} \mathbb{P}(\Omega_2^c) &\leq \mathbb{P}(\Omega_2^c|\Omega_1) + \mathbb{P}(\Omega_1^c) \\ &\leq nC_{22} \exp(-n_\ddagger(\eta/C_\bullet)^2/C_{22}) + Cn\exp(-nr^d/C) \\ &\leq Cn\exp(-nr^d\eta^2/C), \end{aligned}$$

for another constant C that depends on $C_\bullet$ and d.

6.2.2 Away from the Intersection

Objective. *We show that there is an appropriate choice of parameters as in Theorem 1 such that, away from the intersection—when confined to $I_\star$—two points belonging to the same surface and within distance ε are neighbors, while two points belonging to different surfaces are not neighbors.*

For $i \in [n]$, let $K_i = 1$ if $\boldsymbol{s}_i \in S_1$ and $= 2$ otherwise. Let $T_i = T_{S_{K_i}}(\boldsymbol{s}_i)$, which is the tangent subspace associated with data point $\boldsymbol{s}_i$. Let

$$I_\star = \{i : K_j = K_i, \forall j \in \Xi_i\},$$

or equivalently,

$$I_\star^c = \{i : \exists j \text{ s.t. } K_j \neq K_i \text{ and } \|\boldsymbol{s}_j - \boldsymbol{s}_i\| \leq r\}.$$

By definition, $I_\star$ indexes the points whose neighborhoods do not contain points from the other cluster.

For $i \in I_\star$, let $\mathbf{\Sigma}_{T,i}$ denote the covariance of the uniform distribution on $T_i \cap B(\boldsymbol{s}_i, r)$. Applying Lemma 16, this leads to

$$\|\mathbf{\Sigma}_i - \mathbf{\Sigma}_{T,i}\| \le C_{16}\kappa r^3, \quad \forall i \in I_\star. \tag{95}$$

Under Ω_2, by the triangle inequality and (95), we have

$$\|\boldsymbol{C}_i - \mathbf{\Sigma}_{T,i}\| \le \|\boldsymbol{C}_i - \mathbf{\Sigma}_i\| + \|\mathbf{\Sigma}_i - \mathbf{\Sigma}_{T,i}\| \le \zeta r^2, \quad \forall i \in I_\star, \tag{96}$$

where

$$\zeta := \eta/C_\bullet + (3 + C_{16})\kappa r. \tag{97}$$

The inequality (96) leads, via the triangle inequality, to the decisive bound

$$\|\boldsymbol{C}_i - \boldsymbol{C}_j\| \le \|\mathbf{\Sigma}_{T,i} - \mathbf{\Sigma}_{T,j}\| + 2\zeta r^2, \qquad \forall i, j \in I_\star. \tag{98}$$

Take $i, j \in I_\star$ such that $K_i = K_j$ and $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le \varepsilon$. Then by Lemma 13, Lemma 18 and Lemma 3, and the triangle inequality, we have

$$\frac{1}{cr^2}\|\mathbf{\Sigma}_{T,i} - \mathbf{\Sigma}_{T,j}\| = \sin\theta_{\max}(T_i, T_j) \le 6\kappa\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le 6\kappa\varepsilon, \tag{99}$$

where $c := 1/(d+2)$. This implies that

$$\frac{1}{r^2}\|\boldsymbol{C}_i - \boldsymbol{C}_j\| \le 6c\kappa\varepsilon + 2\zeta,$$

by the triangle inequality and (98). Therefore, if $\eta > 6c\kappa\varepsilon + 2\zeta$, then any pair of points indexed by $i, j \in I_\star$ from the same cluster and within distance ε are direct neighbors in the graph built by Algorithm 2.

Take $i, j \in I_\star$ such that $K_i \ne K_j$ and $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le \varepsilon$. By Lemma 21,

$$\max\big[\operatorname{dist}(\boldsymbol{s}_i, S_1 \cap S_2), \operatorname{dist}(\boldsymbol{s}_j, S_1 \cap S_2)\big] \le C_{21}\|\boldsymbol{s}_i - \boldsymbol{s}_j\|.$$

Let $\boldsymbol{z}$ be the mid-point of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. By convexity and the display above,

$$\operatorname{dist}(\boldsymbol{z}, S_1 \cap S_2) \le \frac{1}{2}\operatorname{dist}(\boldsymbol{s}_i, S_1 \cap S_2) + \frac{1}{2}\operatorname{dist}(\boldsymbol{s}_j, S_1 \cap S_2) \le C_{21}\varepsilon.$$

Assuming $C_{21}\varepsilon < 1/\kappa$, let $\boldsymbol{s} = P_{S_1 \cap S_2}(\boldsymbol{z})$. Then, by the triangle inequality again,

$$\max\big[\|\boldsymbol{s} - \boldsymbol{s}_i\|, \|\boldsymbol{s} - \boldsymbol{s}_j\|\big] \le \operatorname{dist}(\boldsymbol{z}, S_1 \cap S_2) + \frac{1}{2}\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le C_{21}\varepsilon + \frac{1}{2}\varepsilon.$$

Let T_i' denote the tangent subspace of S_{K_i} at $\boldsymbol{s}$ and let $\mathbf{\Sigma}_i'$ be the covariance of the uniform distribution over $T_i' \cap B(\boldsymbol{s}, r)$. Define T_j' and $\mathbf{\Sigma}_j'$ similarly. Then, as in (99) we have

$$\frac{1}{cr^2}\|\mathbf{\Sigma}_{T,i} - \mathbf{\Sigma}_i'\| \le 6\kappa\|\boldsymbol{s}_i - \boldsymbol{s}\| \le 6\kappa(C_{21} + 1/2)\varepsilon,$$

and similarly,

$$\frac{1}{cr^2}\|\mathbf{\Sigma}_{T,j} - \mathbf{\Sigma}_j'\| \le 6\kappa(C_{21} + 1/2)\varepsilon.$$

Moreover, by Lemma 13 and Lemma 18,

$$\frac{1}{cr^2}\|\boldsymbol{\Sigma}_i' - \boldsymbol{\Sigma}_j'\| = \sin\theta_{\max}(T_i', T_j') \geq \sin\theta_S,$$

where θ_S is short for $\theta(S_1, S_2)$, defined in (6). Hence, by the triangle inequality,

$$\frac{1}{cr^2}\|\boldsymbol{\Sigma}_{T,i} - \boldsymbol{\Sigma}_{T,j}\| \geq \sin\theta_S - 12\kappa(C_{21} + 1/2)\varepsilon,$$

and then

$$\frac{1}{r^2}\|\boldsymbol{C}_i - \boldsymbol{C}_j\| \geq c\sin\theta_S - 12c\kappa(C_{21} + 1/2)\varepsilon - 2\zeta, \tag{100}$$

by the triangle inequality and (98). Therefore, if $\eta < c\sin\theta_S - 12c\kappa(C_{21} + 1/2)\varepsilon - 2\zeta$, then any pair of points indexed by $i, j \in I_\star$ from different clusters are *not* direct neighbors in the graph built by Algorithm 2.

In summary, we would like to choose η such that

$$c\kappa\varepsilon + 2\zeta < \eta < c\sin\theta_S - 12c\kappa(C_{21} + 1/2)\varepsilon - 2\zeta.$$

Recalling the definition of ζ in (97), this holds under (90).

6.2.3 The Different Clusters are Disconnected in the Graph

Objective. *We show that Step 2 in Algorithm 2 eliminates all points $i \notin I_\star$, implying by our conclusion in Section 6.2.2 that the two clusters are not connected in the graph.*

Hence, take $i \notin I_\star$ with $K_i = 1$ (say), so that $\mathrm{dist}(\boldsymbol{s}_i, S_2) \leq r$. By Lemma 21, we have $\mathrm{dist}(\boldsymbol{s}_i, S_1 \cap S_2) \leq C_{21}r$. Assume r is small enough that $C_{21}\kappa r < 1/2$. Let $\boldsymbol{s}^0 = P_{S_1\cap S_2}(\boldsymbol{s}_i)$ and define $T_k^0 = T_{S_k}(\boldsymbol{s}^0)$ for $k \in \{1, 2\}$. For $\boldsymbol{s} \in S_1$ such that $\mathrm{dist}(\boldsymbol{s}, S_2) \leq 2r$, define $\boldsymbol{\Sigma}_T^0(\boldsymbol{s})$ as in Lemma 27.

Assuming that $\boldsymbol{s}_i \neq \boldsymbol{s}^0$ (which is true with probability one) and $\boldsymbol{s}^0 = 0$ (by translation invariance), let $h = \|\boldsymbol{s}_i - \boldsymbol{s}^0\|$ and $\boldsymbol{v} = (\boldsymbol{s}_i - \boldsymbol{s}^0)/h$. Note that $\boldsymbol{s}_i = h\boldsymbol{v}$. Because $\boldsymbol{v} \perp T_1^0 \cap T_2^0$ by Lemma 1, and that $\theta_{\min}(T_1^0, T_2^0) \geq \theta_S$, we apply Lemma 24 to find $\ell \in h \pm r/2$ such that $\|\boldsymbol{\Sigma}_T^0(\ell\boldsymbol{v}) - \boldsymbol{\Sigma}_T^0(h\boldsymbol{v})\| \geq r^2/C_{24}$, where $C_{24} \geq 1$ depends only on θ_S and d. Letting $\tilde{\boldsymbol{s}} = \ell\boldsymbol{v}$, we have $\|\tilde{\boldsymbol{s}} - \boldsymbol{s}_i\| = |h - \ell| \leq r/2$, so that

$$\mathrm{dist}(\tilde{\boldsymbol{s}}, S_2) \leq \|\tilde{\boldsymbol{s}} - \boldsymbol{s}_i\| + \mathrm{dist}(\boldsymbol{s}_i, S_2) \leq r/2 + r \leq 3r/2.$$

By Lemma 21, we have $\mathrm{dist}(\tilde{\boldsymbol{s}}, S_1 \cap S_2) \leq C_{21}2r < 1/\kappa$, and consequently, $P_{S_1\cap S_2}(\tilde{\boldsymbol{s}}) = \boldsymbol{s}^0$, by Lemma 1. Hence, by the triangle inequality and Lemma 27,

$$\begin{aligned}\|\boldsymbol{\Sigma}(\boldsymbol{s}_i) - \boldsymbol{\Sigma}(\tilde{\boldsymbol{s}})\| &\geq \|\boldsymbol{\Sigma}_T^0(\boldsymbol{s}) - \boldsymbol{\Sigma}_T^0(\tilde{\boldsymbol{s}})\| - \|\boldsymbol{\Sigma}(\boldsymbol{s}_i) - \boldsymbol{\Sigma}_T^0(\boldsymbol{s}_i)\| - \|\boldsymbol{\Sigma}(\tilde{\boldsymbol{s}}) - \boldsymbol{\Sigma}_T^0(\tilde{\boldsymbol{s}})\| \\ &\geq r^2/C_{24} - 2C_{27}r^{5/2}.\end{aligned}$$

Take any $\bar{\boldsymbol{s}} \in S_1$ such that $\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\| \leq r/2$. By Lemma 23,

$$\|\boldsymbol{\Sigma}_T^0(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}_T^0(\tilde{\boldsymbol{s}})\| \leq 5dr\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\| + 2\sqrt{r^3\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\|} \leq (5d + 2)r^{3/2}\sqrt{\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\|}.$$

Noting that

$$\mathrm{dist}(\bar{\boldsymbol{s}}, S_2) \leq \|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\| + \mathrm{dist}(\tilde{\boldsymbol{s}}, S_2) \leq r/2 + 3r/2 \leq 2r,$$

we apply the triangle inequality and Lemma 27, and we get

$$\begin{aligned}\|\boldsymbol{\Sigma}(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}(\tilde{\boldsymbol{s}})\| &\leq \|\boldsymbol{\Sigma}_T^0(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}_T^0(\tilde{\boldsymbol{s}})\| + \|\boldsymbol{\Sigma}(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}_T^0(\bar{\boldsymbol{s}})\| + \|\boldsymbol{\Sigma}(\tilde{\boldsymbol{s}}) - \boldsymbol{\Sigma}_T^0(\tilde{\boldsymbol{s}})\| \\ &\leq (5d+2)r^{3/2}\sqrt{\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\|} + 2C_{27}r^{5/2}.\end{aligned}$$

When $\|\bar{\boldsymbol{s}} - \tilde{\boldsymbol{s}}\| \leq r/C_\ddagger$, we have

$$\begin{aligned}\|\boldsymbol{\Sigma}(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}(\boldsymbol{s}_i)\| &\geq \|\boldsymbol{\Sigma}(\boldsymbol{s}_i) - \boldsymbol{\Sigma}(\tilde{\boldsymbol{s}})\| - \|\boldsymbol{\Sigma}(\bar{\boldsymbol{s}}) - \boldsymbol{\Sigma}(\tilde{\boldsymbol{s}})\| \\ &\geq r^2\big(1/C_{24} - (5d+2)/\sqrt{C_\ddagger} - 4C_{27}r^{1/2}\big) \geq r^2/(2C_{24}),\end{aligned}$$

assuming $C_\ddagger$ is large enough and r is small enough. Fixing such a constant $C_\ddagger \geq 2$, define the event

$$\Omega_3 = \bigcap_{k=1}^{2} \Big\{\forall \boldsymbol{s} \in S_k : \#\{i : K_i = k \text{ and } \boldsymbol{s}_i \in B(\boldsymbol{s}, r/C_\ddagger)\} \geq 2\Big\}.$$

By the same arguments bounding $\mathbb{P}(\Omega_1^c)$ in Section 6.2.1, we have that there is a constant C such that, when $r \leq 1/(C\kappa)$,

$$\mathbb{P}(\Omega_3^c) \leq Cn\exp(-nr^d/C).$$

Under Ω_3, there is $\boldsymbol{s}_j \in S_1 \cap B(\tilde{\boldsymbol{s}}, r/C_\ddagger)$. Taking $\bar{\boldsymbol{s}} = \boldsymbol{s}_j$ above, we have that $\|\boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}_i\| \geq r^2/(2C_{24})$. Under Ω_2, by the triangle inequality,

$$\begin{aligned}\|\boldsymbol{C}_j - \boldsymbol{C}_i\| &\geq \|\boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}_i\| - \|\boldsymbol{C}_i - \boldsymbol{\Sigma}_i\| - \|\boldsymbol{C}_j - \boldsymbol{\Sigma}_j\| \\ &\geq r^2/(2C_{24}) - 2r^2\eta/C_\bullet \geq r^2/(3C_{24}),\end{aligned}$$

choosing $C_\bullet$ sufficiently large. (Recall that $\eta \leq 1$ by assumption.) Therefore, choosing η such that $\eta < 1/(3C_{24})$, we see that $\|\boldsymbol{C}_j - \boldsymbol{C}_i\| > \eta r^2$, even though

$$\|\boldsymbol{s}_j - \boldsymbol{s}_i\| \leq \|\boldsymbol{s}_j - \tilde{\boldsymbol{s}}\| + \|\tilde{\boldsymbol{s}} - \boldsymbol{s}_i\| \leq r/C_\ddagger + r/2 \leq r.$$

6.2.4 The Points that are Removed are close to the Intersection

Objective. *We show that the points removed by Step 2 are within distance Cr of the intersection.*

Let $I_\circ = \{i \in I_\star : \Xi_i \subset I_\star\}$. By our choice of parameters in (90), we see that $i \in I_\circ$ is neighbor with any $j \in \Xi_i$, so that i survives Step 2. Hence, the nodes removed at Step 2 are in $I_\circ^c = \{i \in I_\star : \Xi_i \cap I_\star^c \neq \emptyset\} \cup I_\star^c$, with the possibility that some nodes in $I_\circ^c$ survive. Now, for any $i \in I_\star^c$, there is j with $K_j \neq K_i$ such that $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq r$, so by Lemma 21,

$$\operatorname{dist}(\boldsymbol{s}_i, S_1 \cap S_2) \leq C_{21}\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq C_{21}r.$$

And for any $i \in I_\circ \setminus I_\star^c$, there is $j \in I_\star^c$ such that $\|\boldsymbol{s}_j - \boldsymbol{s}_i\| \leq r$, by definition of Ξ_i, so that $\operatorname{dist}(\boldsymbol{s}_i, S_1 \cap S_2) \leq C_{21}r + r$, by the triangle inequality. In fact, we just proved that the last inequality holds for all $i \in I_\circ^c$. Hence, the points removed in Step 2 are within distance $(C_{21}+1)r$ of $S_1 \cap S_2$.

6.2.5 Each Cluster is (essentially) Connected in the Graph

Objective. *We show that the points that survive Step 2 and belong to the same cluster are connected in the graph, except for possible spurious points within distance Cr of the intersection.*

Take, for example, the cluster generated from sampling S_1. Before we start, we recall that $I_\star^c$ was eliminated in Step 2, so that by our choice of η in (90), to show that two points $\boldsymbol{s}_i, \boldsymbol{s}_j$ sampled from S_1 are neighbors it suffices to show that $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le \varepsilon$.

Define $R = S_1 \setminus B(S_1 \cap S_2, r')$, where $r' := (C_{21}+2)r$. By our conclusion in Section 6.2.4, we have that

$$\text{any } \boldsymbol{s}_i \in S_1 \text{ such that } \operatorname{dist}(\boldsymbol{s}_i, R) < r \text{ survives Step 2,} \tag{109}$$

so that it suffices to show that the subset of nodes $\{i : \boldsymbol{s}_i \in R\}$ is connected.

Define $W = S_1 \setminus S_2$ and let $\{W_m\}$ denote the connected components of W. (Note that there might only be one, for example, when S_1 and S_2 are circles in dimension $D = 3$, not in the same plane, that intersect at a single point.) Define $R_m = W_m \setminus B(S_1 \cap S_2, r')$. Since $W = \sqcup_m W_m = S_1 \setminus (S_1 \cap S_2)$, we have $R = \sqcup_m R_m$. In fact, when r (and therefore r') is small enough, R_m is connected, and assuming this is the case, $\{R_m\}$ are the connected components of R.

Suppose R_1 and R_2 are two adjacent connected components of R, meaning that $R_1 \cup R_2 \cup S_2$ is connected. We show that there is at least one pair j_1, j_2 of direct neighbors in the graph such that $\boldsymbol{s}_{j_m} \in R_m$. Take $\boldsymbol{s}$ on the connected component of $S_1 \cap S_2$ separating R_1 and R_2. We construct two points, $\boldsymbol{s}^1 \in R_1$ and $\boldsymbol{s}^2 \in R_2$, that are within distance of order ε from $\boldsymbol{s}$, and then select $\boldsymbol{s}_{j_m}$ as the closest data point to $\boldsymbol{s}^m$. Let $T^k = T_{S_k}(\boldsymbol{s})$ and let H be the affine subspace parallel to $(T^1 \cap T^2)^\perp$ passing through $\boldsymbol{s}$. Take $\boldsymbol{t}^m \in P_{T^1}(R_m) \cap H \cap \partial B(\boldsymbol{s}, \varepsilon_1)$, where $\varepsilon_1 := \varepsilon/4$, which exists when ε is sufficiently small. Note that $\boldsymbol{t}^1, \boldsymbol{t}^2 \in T^1$ and assuming that ε is sufficiently small that $\varepsilon_1 < 1/\kappa$, we have $P_{S_1 \cap S_2}(\boldsymbol{t}^m) = \boldsymbol{s}$ by Lemma 1. Define $\boldsymbol{s}^m = P_{T^1}^{-1}(\boldsymbol{t}^m)$ and note that $\boldsymbol{s}^1, \boldsymbol{s}^2 \in S_1$. Lemma 5 not only justifies this construction when $\kappa\varepsilon_1 < 1/C_5$, it also says that $P_{T^1}^{-1}$ has Lipschitz constant bounded by $1 + \kappa\varepsilon_1$, which implies that

$$\|\boldsymbol{s}^m - \boldsymbol{s}\| \le (1 + \kappa\varepsilon_1)\|\boldsymbol{t}^m - \boldsymbol{s}\| = (1 + \kappa\varepsilon_1)\varepsilon_1 \le \varepsilon/3.$$

We also have

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{s}^m, S_1 \cap S_2) &\ge \operatorname{dist}(\boldsymbol{t}^m, S_1 \cap S_2) - \|\boldsymbol{s}^m - \boldsymbol{t}^m\| \\
&= \|\boldsymbol{t}^m - \boldsymbol{s}\| - \|\boldsymbol{s}^m - \boldsymbol{t}^m\| \\
&\ge \varepsilon_1 - \frac{\kappa}{2}\|\boldsymbol{s}^m - \boldsymbol{s}\|^2 \\
&\ge \big(1 - \frac{\kappa}{2}(1 + \kappa\varepsilon_1)^2\varepsilon_1\big)\varepsilon_1 \\
&\ge \varepsilon/5,
\end{aligned}$$

when ε is sufficiently small. We used (10) in the second inequality. We assume r/ε is sufficiently small that $\varepsilon/5 \ge r' + r$. Then under Ω_3, there are j_1, j_2 such that $\boldsymbol{s}_{j_m} \in B(\boldsymbol{s}^m, r) \cap S_1$, and by the triangle inequality, we then have that $\operatorname{dist}(\boldsymbol{s}_{j_m}, S_1 \cap S_2) \ge \varepsilon/5 - r \ge r'$, so

that $\boldsymbol{s}_{j_m} \in R_m$. Moreover,

$$\begin{aligned} \|\boldsymbol{s}_{j_1} - \boldsymbol{s}_{j_2}\| &\leq \|\boldsymbol{s}_{j_1} - \boldsymbol{s}^1\| + \|\boldsymbol{s}^1 - \boldsymbol{s}\| + \|\boldsymbol{s} - \boldsymbol{s}^2\| + \|\boldsymbol{s}^2 - \boldsymbol{s}_{j_2}\| \\ &\leq r + \varepsilon/3 + \varepsilon/3 + r \\ &= \frac{2}{3}\varepsilon + 2r \leq \varepsilon, \end{aligned}$$

assuming r/ε is sufficiently small.

Now, we show that the points sampled from any connected component, say R_1, form a connected subgraph. Take $\boldsymbol{s}^1, \dots, \boldsymbol{s}^M$ an r-packing of R_1, so that

$$\bigsqcup_m (R_1 \cap B(\boldsymbol{s}^m, r/2)) \subset R_1 \subset \bigcup_m (R_1 \cap B(\boldsymbol{s}^m, r)).$$

Because R_1 is connected, $\cup_m B(\boldsymbol{s}^m, r)$ is necessarily connected. Under Ω_1, and $C_{26} \geq 2$, all the balls $B(\boldsymbol{s}^m, r), m = 1, \dots, M$, contain at least one $\boldsymbol{s}_i \in S_1$, and any such point survives Step 2, because of $\operatorname{dist}(\boldsymbol{s}_i, R_1) < r$ and (109). Assume $r/\varepsilon \leq 1/4$. Then two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ such that $\boldsymbol{s}_i, \boldsymbol{s}_j \in B(\boldsymbol{s}^m, r)$ are connected, since $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq 2r \leq \varepsilon$. And when $B(\boldsymbol{s}^{m_1}, r) \cap B(\boldsymbol{s}^{m_2}, r) \neq \emptyset$, $\boldsymbol{s}_i \in B(\boldsymbol{s}^{m_1}, r)$ and $\boldsymbol{s}_j \in B(\boldsymbol{s}^{m_2}, r)$ are such that $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq 4r \leq \varepsilon$. Hence, the points sampled from R_1 are connected in the graph under Ω_1.

6.2.6 Summary and Conclusion

We worked under the assumption that (90), with a constant C large enough and depending only on S_1 and S_2. This is compatible with the statement of Theorem 1. In Section 6.2.1, we derived a concentration inequality for local covariances. This lead to a uniform bound (defining the event Ω_2) which holds with probability at least $1 - Cn \exp(-nr^d\eta^2/C)$ for some constant $C > 0$. In Section 6.2.2, we showed that among points away from the intersection (indexed by $I_\star$), those that are in the same cluster are neighbors in the graph if they are within distance ε, while those in different clusters cannot be neighbors in the graph. In Section 6.2.3, we showed that Step 2 removes points not indexed by $I_\star$, thus implying that the two clusters are not connected in the graph. In the process, we assumed that an event (denoted Ω_3) held, which happens with probability at least $1 - Cn \exp(-nr^d/C)$ for some constant $C > 0$. In Section 6.2.4, we showed that the points removed in Step 2 are within distance $O(r)$ from the intersection of the surfaces. In Section 6.2.5, we showed that, except for points within distance $O(r)$ of the intersection, two points in the same cluster are connected in the graph.

We conclude that, after Step 4, Algorithm 2 returns two groups, and each group covers an entire cluster except for points within distance $O(r)$ of the intersection. In Step 5, each point removed in Step 2 is included in the group which contains its closest point (in Euclidean distance). Theorem 1 is silent on the accuracy of doing this step. So the proof of Theorem 1, as it relates to Algorithm 2, is complete.

6.3 Performance Guarantees for Algorithm 3

We keep the same notation as in Section 6.2 and go a little faster here as the arguments are parallel. Let d_i denote the estimated dimensionality at point $\boldsymbol{s}_i$, meaning the number

of eigenvalues of $\boldsymbol{C}_i$ exceeding $\sqrt{\eta}\,\|\boldsymbol{C}_i\|$. Recall that $\boldsymbol{Q}_i$ denotes the orthogonal projection onto the top d_i eigenvectors of $\boldsymbol{C}_i$.

6.3.1 Estimation of the Intrinsic Dimension

Objective. *We show that, for $i \in I_\star$, $d_i = d$ (the correct intrinsic dimension) with high probability.*

Take $i \in I_\star$. Under Ω_2, (96) holds, and applying Weyl's inequality (Stewart and Sun, 1990, Cor. IV.4.9), we have

$$|\beta_m(\boldsymbol{C}_i) - \beta_m(\boldsymbol{\Sigma}_{T,i})| \le \zeta r^2, \quad \forall m = 1, \dots, D.$$

By Lemma 13, $\boldsymbol{\Sigma}_{T,i} = cr^2\boldsymbol{P}_i$, so that $\beta_m(\boldsymbol{\Sigma}_{T,i}) = cr^2$ when $m \le d$ and $\beta_m(\boldsymbol{\Sigma}_{T,i}) = 0$ when $m > d$. Hence,

$$\beta_1(\boldsymbol{C}_i) \le (c+\zeta)r^2, \qquad \beta_d(\boldsymbol{C}_i) \ge (c-\zeta)r^2, \qquad \beta_{d+1}(\boldsymbol{C}_i) \le \zeta r^2.$$

This implies that

$$\frac{\beta_d(\boldsymbol{C}_i)}{\beta_1(\boldsymbol{C}_i)} \ge \frac{c-\zeta}{c+\zeta} > \sqrt{\eta}, \qquad \frac{\beta_{d+1}(\boldsymbol{C}_i)}{\beta_1(\boldsymbol{C}_i)} \le \frac{\zeta}{c+\zeta} < \sqrt{\eta},$$

when $\zeta \le \eta/2$ and η is sufficiently small, which is the case under (90). When this is so, $d_i = d$ by definition of d_i. (Note that $\|\boldsymbol{C}_i\| = \beta_1(\boldsymbol{C}_i)$.)

6.3.2 Each Cluster is (essentially) Connected in the Graph

Objective. *We show that each cluster is connected in the graph, except possibly for some points near the intersection.*

Note that the top d eigenvectors of $\boldsymbol{\Sigma}_{T,i}$ generate T_i. Hence, applying Lemma 19, and (96) again, we get that (recall that $c = 1/(d+1)$)

$$\|\boldsymbol{Q}_i - \boldsymbol{P}_i\| \le \frac{\sqrt{2}\,\zeta r^2}{cr^2} = \zeta' := \sqrt{2}(d+2)\zeta, \quad \forall i \in I_\star.$$

This is the equivalent of (96), which leads to the equivalent of (98):

$$\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| \le \frac{1}{cr^2}\|\boldsymbol{\Sigma}_{T,i} - \boldsymbol{\Sigma}_{T,j}\| + 2\zeta', \qquad \forall i, j \in I_\star,$$

using the fact that $\boldsymbol{\Sigma}_{T,i} = cr^2\boldsymbol{P}_i$. When $i, j \in I_\star$ are such that $K_i = K_j$, based on (99), we have

$$\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| \le 6\kappa\varepsilon + 2\zeta'.$$

Hence, when $\eta > 6\kappa\varepsilon + 2\zeta'$ (which is the case under (90)), two nodes $i, j \in I_\star$ such that $K_i = K_j$ and $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le \varepsilon$ are neighbors in the graph. The arguments provided in Section 6.2.5 now apply in exactly the same way to show that nodes $i \in I_\circ$ such that $K_i = 1$ are connected in the graph. The same is true of nodes $i \in I_\circ$ such that $K_i = 2$.

6.3.3 The Clusters are (essentially) Disconnected in the Graph

Objective. *We show that the two sets of nodes* $\{i \in I_\circ : K_i = 1\}$ *and* $\{i \in I_\circ : K_i = 2\}$*—shown to be connected above—are disconnected in the graph.*

When we take $i, j \in I_\star$ such that $K_i \neq K_j$, we have the equivalent of (100), meaning,

$$\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| \geq \sin\theta_S - 12\kappa(C_{21} + 1/2)\varepsilon - 2\zeta'.$$

We choose η smaller than the RHS, so that these nodes are not neighbors in the graph.

We next prove that a point indexed by $I_\star$ is not neighbor to a point near the intersection because of different estimates for the local dimension. Let C denote a positive constant which increases with each appearance. Take $\boldsymbol{s} \in S_1$ such that $\delta(\boldsymbol{s}) := \mathrm{dist}(\boldsymbol{s}, S_2) \leq r$. Reinstate the notation used in Section 6.2.3, in particular $\boldsymbol{s}^0$, T_1^0 and T_2^0, as well as $\boldsymbol{\Sigma}_T^0(\boldsymbol{s})$. Define $\boldsymbol{\Sigma}_T(\boldsymbol{s})$ as in Lemma 26. Also, let $T_1' = T_{S_1}(\boldsymbol{s})$ and $T_2' = T_{S_2}(\boldsymbol{s}^2)$ where $\boldsymbol{s}^2 := P_{S_2}(\boldsymbol{s})$. By Lemma 26 and Weyl's inequality, we have

$$\beta_{d+1}(\boldsymbol{\Sigma}(\boldsymbol{s})) \geq \beta_{d+1}(\boldsymbol{\Sigma}_T(\boldsymbol{s})) - C_{26}r^3, \qquad \beta_1(\boldsymbol{\Sigma}(\boldsymbol{s})) \leq \beta_1(\boldsymbol{\Sigma}_T(\boldsymbol{s})) + C_{26}r^3,$$

which together with Lemma 25 (and proper scaling), implies that

$$\frac{\beta_{d+1}(\boldsymbol{\Sigma}(\boldsymbol{s}))}{\beta_1(\boldsymbol{\Sigma}(\boldsymbol{s}))} \geq \frac{\frac{c}{8}(1 - \cos\theta_{\max}(T_1', T_2'))^2(1 - \delta(\boldsymbol{s})^2/r^2)^{d/2+1} - C_{26}r^3}{c + (\delta(\boldsymbol{s})/r)^2(1 - \delta(\boldsymbol{s})^2/r^2)^{d/2} + C_{26}r^3}.$$

Then, by the triangle inequality,

$$\theta_{\max}(T_1', T_2') \geq \theta_{\max}(T_1^0, T_2^0) - \theta_{\max}(T_1', T_1^0) - \theta_{\max}(T_2', T_2^0).$$

By definition, $\theta_{\max}(T_1^0, T_2^0) \geq \theta_S$, and by Lemma 3,

$$\theta_{\max}(T_1', T_1^0) \leq \sin^{-1}\left(6\kappa\|\boldsymbol{s} - \boldsymbol{s}^0\| \wedge 1\right) \leq Cr,$$

and similarly,

$$\theta_{\max}(T_2', T_2^0) \leq \sin^{-1}\left(6\kappa\|\boldsymbol{s}^2 - \boldsymbol{s}^0\| \wedge 1\right) \leq Cr,$$

because $\|\boldsymbol{s} - \boldsymbol{s}^0\| \leq Cr$ by Lemma 21 and then $\|\boldsymbol{s}^2 - \boldsymbol{s}^0\| \leq r + \|\boldsymbol{s} - \boldsymbol{s}^0\| \leq Cr$. Hence, for r small enough, $\theta_{\max}(T_1', T_2') \geq \theta_S/2$, and furthermore,

$$\frac{\beta_{d+1}(\boldsymbol{\Sigma}(\boldsymbol{s}))}{\beta_1(\boldsymbol{\Sigma}(\boldsymbol{s}))} \geq \sqrt{\eta} \quad \text{when} \quad 1 - \delta(\boldsymbol{s})^2/r^2 \geq \xi := C_*\eta^{1/(d+2)}, \tag{118}$$

where C_* is a large enough constant, assuming r/η and η are small enough. The same is true for points on $\boldsymbol{s} \in S_2$ if we redefine $\delta(\boldsymbol{s}) = \mathrm{dist}(\boldsymbol{s}, S_1)$. Hence, for $\boldsymbol{s}_i$ close enough to the intersection that $\delta(\boldsymbol{s}_i)$ satisfies (118), $d_i > d$. Then, by Lemma 18, $\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| = 1$ for any $j \in I_\star$. By our choice of $\eta < 1$, this means that i and j are not neighbors.

So the only way $\{i \in I_\star : K_i = 1\}$ and $\{i \in I_\star : K_i = 2\}$ are connected in the graph is if there are $\boldsymbol{s}_i \in S_1$ and $\boldsymbol{s}_j \in S_2$ such that $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq \varepsilon$ and both $\delta(\boldsymbol{s}_i)$ and $\delta(\boldsymbol{s}_j)$ fail to satisfy (118). We now show this is not possible. Let $\boldsymbol{\Sigma}_{T,i} = \boldsymbol{\Sigma}_T(\boldsymbol{s}_i)$ and define $\boldsymbol{\Sigma}_{T,j}$ similarly. (These are well-defined when $\varepsilon < 1/\kappa$, which we assume.) By Lemma 26, we have

$$\|\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_{T,i}\| \leq C_{26}\, r^3. \tag{119}$$

Define $\alpha_i = (1+t_i^d)^{-1}$ with $t_i := (1-\delta^2(\boldsymbol{s}_i))_+^{1/2}$. As in (77), with the triangle inequality, we get

$$\begin{aligned}\|\boldsymbol{\Sigma}_{T,i} - \alpha_i c r^2 \boldsymbol{P}_i\| &\le c(1-\alpha_i)t_i^2 r^2 + \alpha_i(1-\alpha_i)\delta^2(\boldsymbol{s}_i) \\ &\le 2(1-\alpha_i)r^2 \le 2(1-\delta(\boldsymbol{s}_i)^2/r^2)_+^{d/2} r^2 \le 2\xi^{d/2} r^2,\end{aligned}$$

where the second inequality we used the fact that $\alpha_i = 1$ if $\delta(\boldsymbol{s}_i) > r$, and in very last inequality comes from $\delta(\boldsymbol{s}_i)$ not satisfying (118). Hence, under Ω_2, with (119) and the triangle inequality, we get

$$\|\boldsymbol{C}_i - \alpha_i c r^2 \boldsymbol{P}_i\| \le r^2\eta/C_\bullet + 3\kappa r^3 + 2\xi^{d/2} r^2 + C_{26}\, r^3.$$

Since $\beta_d(\boldsymbol{P}_i) = 1$ and $\beta_{d+1}(\boldsymbol{P}_i) = 0$, by Lemma 19, we have

$$\|\boldsymbol{Q}_i - \boldsymbol{P}_i\| \le \frac{1}{\alpha_i c r^2}\big[r^2\eta/C_\bullet + 2\xi^{d/2} r^2 + (3\kappa + C_{26}) r^3\big] \le C(\eta/C_\bullet + \xi^{d/2} + r).$$

Similarly,

$$\|\boldsymbol{Q}_j - \boldsymbol{P}_j\| \le C(\eta/C_\bullet + \xi^{d/2} + r).$$

By Lemma 18, $\|\boldsymbol{P}_i - \boldsymbol{P}_j\| = \sin\theta_{\max}(T_i, T_j)$. Let $\boldsymbol{s}^0 = P_{S_1\cap S_2}(\boldsymbol{s}_i)$, and define T_1^0 and T_2^0 as before. Not that $\|\boldsymbol{s}_i - \boldsymbol{s}^0\| \le C\varepsilon$ by Lemma 21 and the fact that $\mathrm{dist}(\boldsymbol{s}_i, S_2) \le \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \le \varepsilon$. Also, $\|\boldsymbol{s}_j - \boldsymbol{s}^0\| \le \|\boldsymbol{s}_i - \boldsymbol{s}^0\| + \|\boldsymbol{s}_j - \boldsymbol{s}_i\| \le C\varepsilon$. We then have

$$\theta_{\max}(T_i, T_j) \ge \theta_{\max}(T_1^0, T_2^0) - \theta_{\max}(T_i, T_1^0) - \theta_{\max}(T_j, T_2^0) \ge \theta_S - C\varepsilon,$$

by Lemma 3. Hence, by the triangle inequality,

$$\begin{aligned}\|\boldsymbol{Q}_i - \boldsymbol{Q}_j\| &\ge \|\boldsymbol{P}_i - \boldsymbol{P}_j\| - \|\boldsymbol{Q}_i - \boldsymbol{P}_i\| - \|\boldsymbol{Q}_j - \boldsymbol{P}_j\| \\ &\ge \sin(\theta_S - C\varepsilon) - C(\eta/C_\bullet + \xi^{d/2} + r) \ge \frac{1}{2}\sin\theta_S > \eta,\end{aligned}$$

when the constant in (90) is large enough. (Recall the definition of ξ above.) Therefore i and j are not neighbors, as we needed to show.

6.4 Noisy Case

So far we only dealt with the case where $\tau = 0$ in (7). When $\tau > 0$, a sample point $\boldsymbol{x}_i$ is in general different than its corresponding point $\boldsymbol{s}_i$ sampled from one of the surfaces. However, when τ/r is small, this does not change things much. For one thing, the points are close to each other, since we have $\|\boldsymbol{x}_i - \boldsymbol{s}_i\| \le \tau$ by assumption, and τ is small compared to r. And the corresponding covariance matrices are also close to each other. To see this, redefine $\Xi_i = \{j \ne i : \boldsymbol{x}_j \in N_r(\boldsymbol{x}_i)\}$ and $\boldsymbol{C}_i$ as the sample covariance of $\{\boldsymbol{x}_j : j \in \Xi_i\}$. Let $\boldsymbol{D}_i$ denote the sample covariance of $\{\boldsymbol{s}_j : j \in \Xi_i\}$. Let X be uniform over $\{\boldsymbol{x}_j : j \in \Xi_i\}$ and define $Y = \sum_j \boldsymbol{s}_j \mathbb{1}_{\{X = \boldsymbol{x}_j\}}$. Starting from (33), we have

$$\begin{aligned}\|\boldsymbol{D}_i - \boldsymbol{C}_i\| &= \|\operatorname{Cov}(X) - \operatorname{Cov}(Y)\| \\ &\le \mathbb{E}\left[\|X - Y\|^2\right]^{1/2} \cdot \Big(\mathbb{E}\left[\|X - \boldsymbol{x}_i\|^2\right]^{1/2} + \mathbb{E}\left[\|Y - \boldsymbol{x}_i\|^2\right]^{1/2}\Big) \\ &\le \tau\cdot(r + (r+\tau)) = r^2\big(2\tau/r + (\tau/r)^2\big),\end{aligned}$$

which is small compared to r^2, and r^2 is the operating scale for covariance matrices in our setting.

Using these facts, the arguments are virtually the same, except for some additional terms due to triangle inequalities, for example, $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| - 2\tau \le \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \le \|\boldsymbol{s}_i - \boldsymbol{s}_j\| + 2\tau$. In particular, this results in ζ in (97) being now redefined as $\zeta = \frac{3\tau}{r} + \eta/C_\bullet + (3 + C_{16})\kappa r$. We omit further technical details.

Acknowledgments

This work was partially supported by grants from the National Science Foundation (DMS 0915160, 0915064, 0956072, 1418386, 1513465). We would like to thank Jan Rataj for helpful discussion around Lemma 3 and Xu Wang for his sharp proofreading. We also gratefully acknowledge the comments, suggestions, and scrutiny of an anonymous referee. We would also like to acknowledge support from the Institute for Mathematics and its Applications (IMA). For one thing, the authors first learned about the research of Goldberg et al. (2009) there, at the *Multi-Manifold Data Modeling and Applications* workshop in the Fall of 2008, and this was the main inspiration for our paper. Also, part of our work was performed while TZ was a postdoctoral fellow at the IMA, and also while EAC and GL were visiting the IMA.

References

S. N. Afriat. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society*, 53:800–816, 1957.

S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. Beyond pairwise clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 838–845, 2005.

S. Agarwal, K. Branson, and S. Belongie. Higher order learning with graphs. In *International Conference on Machine Learning (ICML)*, pages 17–24, 2006.

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.

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

J.-D. Boissonnat, R. Dyer, and A. Ghosh. Constructing intrinsic delaunay triangulations of submanifolds. *arXiv preprint arXiv:1303.6493*, 2013.

P. Bradley, K. Bennett, and A. Demiriz. Constrained k-means clustering. Technical Report MSR-TR-2000-65, Microsoft Research, 2000.

M. R. Brito, E. L. Chávez, A. J. Quiroz, and J. E. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters*, 35(1):33–42, 1997.

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

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, 2009b.

G. Chen, S. Atev, and G. Lerman. Kernel spectral curvature clustering (KSCC). In *IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 765–772, 2009.

A. Cuevas, R. Fraiman, and B. Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Advances in Applied Probability*, 44(2):311–329, 2012.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7:1–46, 1970.

E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 55–63, 2011.

R. Epstein, P. Hallinan, and A. Yuille. 5 ± 2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *IEEE Workshop on Physics-based Modeling in Computer Vision*, pages 108–116, 1995.

Z. Fan, J. Zhou, and Y. Wu. Multibody grouping by inference of multiple subspaces from high-dimensional data using oriented-frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):91–105, 2006.

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

Z. Fu, W. Hu, and T. Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing*, volume 2, pages 602–605, 2005.

A. Gionis, A. Hinneburg, S. Papadimitriou, and P. Tsaparas. Dimension induced clustering. In *International Conference on Knowledge Discovery in Data Mining (KDD)*, pages 51–60, 2005.

A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2007.

A.B. Goldberg, X. Zhu, A. Singh, Z. Xu, and R. Nowak. Multi-manifold semi-supervised learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 169–176, 2009.

D. Gong, X. Zhao, and G. Medioni. Robust multiple manifolds structure learning. In *International Conference on Machine Learning (ICML)*, pages 321–328, 2012.

Q. Guo, H. Li, W. Chen, I-F. Shen, and J. Parkkinen. Manifold clustering via energy minimization. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 375–380, 2007.

G. Haro, G. Randall, and G. Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. *Advances in Neural Information Processing Systems (NIPS)*, 19:553, 2007.

R. Heckel and H. Bölcskei. Noisy subspace clustering via thresholding. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1382–1386, 2013.

J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 1–11, 2003.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

D. N. Kaslovsky and F. G. Meyer. Optimal tangent plane recovery from noisy manifold samples. arXiv preprint arXiv:1111.4601v2, 2011.

D. Kushnir, M. Galun, and A. Brandt. Fast multiscale clustering and manifold identification. *Pattern Recognition*, 39(10):1876–1891, 2006.

K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (5):684–698, 2005.

A. V. Little, Y. M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *Manifold Learning and its Applications*, pages 26–33, 2009.

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

Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review*, 50(3):413–458, 2008.

M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410(19):1749–1764, 2009.

V. J. Martínez and E. Saar. *Statistics of the Galaxy Distribution*. CRC press, Boca Raton, 2002.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 2:849–856, 2002.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

M. Polito and P. Perona. Grouping and dimensionality reduction by locally linear embedding. *Advances in Neural Information Processing Systems (NIPS)*, 14:1255–1262, 2001.

L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

A. Shashua, R. Zass, and T. Hazan. Multi-way clustering using super-symmetric non-negative tensor factorization. In *European Conference on Computer Vision (ECCV)*, pages 595–608, 2006.

M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *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.

R. Souvenir and R. Pless. Manifold clustering. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 648–653, 2005.

G. W. Stewart and J. G. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990.

M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

M. C. Tsakiris and R. Vidal. Filtrated spectral algebraic subspace clustering. In *IEEE International Conference on Computer Vision Workshops*, pages 28–36, 2015.

R. Valdarnini. Detection of non-random patterns in cosmological gravitational clustering. *Astronomy & Astrophysics*, 366:376–386, 2001.

R. Vidal and Y. Ma. A unified algebraic approach to 2-D and 3-D motion segmentation and estimation. *Journal of Mathematical Imaging and Vision*, 25(3):403–421, 2006.

G. Walther. Granulometric smoothing. *Annals of Statistics*, 25(6):2273–2299, 1997.

Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks*, 22(7):1149–1161, 2011.

Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In *International Conference on Machine Learning (ICML)*, pages 89–97, 2013.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2005.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, pages 1–24, 2012.