Journal of Machine Learning Research x () x-xx Submitted x/xx; Published xx/xx

Manifold Fitting under Unbounded Noise

Zhigang Yao ZHIGANG.YAO@NUS.EDU.SG
Department of Statistics and Data Science ZHIGANG.YAO@CMSA.FAS.HARVARD.EDU
National University of Singapore
21 Lower Kent Ridge Road
Singapore 117546

Center of Mathematical Sciences and Applications
Harvard University
20 Garden Street
Cambridge USA 02138

Yuqing Xia STAXIAY@NUS.EDU.SG
Department of Statistics and Data Science
National University of Singapore
21 Lower Kent Ridge Road
Singapore 117546

Editor: xxxx

Abstract

In the field of non-Euclidean statistical analysis, a trend has emerged in recent times, of attempts to recover a low dimensional structure, namely a manifold, underlying the high dimensional data. Recovering the manifold requires the noise to be of a certain concentration and prevailing methods address this requirement by constructing an approximated manifold that is based on the tangent space estimation at each sample point. Although theoretical convergence for these methods is guaranteed, the samples are either noiseless or the noise is bounded. However, if the noise is unbounded, as is commonplace, the tangent space estimation at the noisy samples will be blurred – an undesirable outcome since fitting a manifold from the blurred tangent space might be more greatly compromised in terms of its accuracy. In this paper, we introduce a new manifold-fitting method, whereby the output manifold is constructed by directly estimating the tangent spaces at the projected points on the latent manifold, rather than at the sample points, thus reducing the error caused by the noise. Assuming the noise is unbounded, our new method has a high probability of achieving theoretical convergence, in terms of the upper bound of the distance between the estimated and latent manifold. The smoothness of the estimated manifold is also evaluated by bounding the supremum of twice difference above. Numerical simulations are conducted as part of this new method to help validate our theoretical findings and demonstrate the advantages of our method over other relevant manifold fitting methods. Finally, our method is applied to real data examples.

Keywords: Manifold learning, Riemannian embedding, Convergence, Smoothness

1. Introduction

Linearity has been viewed as a cornerstone in the development of statistical methodology. For decades, prominent progress in statistics has been made with regard to linearizing the data and the way we analyze them. More recently, the phenomenon of high-throughput data, which share a high dimensional characteristic in their varying forms, has become more commonplace. Although each data point usually represents itself as a long vector or a large matrix, in principle they all can be viewed as points on or near an intrinsic manifold. Moreover, modern data sets no longer comprise samples of real vectors in a real vector space but samples of much more complex structures, assuming values in spaces that are naturally not (Euclidean) vector spaces. We are verily witnessing an explosion in the volumn of "complex data" with a geometric structure and, therefore, a growing need for statistical analysis, utilizing the nature of the data space.

The manifold hypothesis has been carefully studied in Fefferman et al. (2016). Here, we only present several relevant examples to make sense of that hypothesis intuitively: the high dimensional data samples tend to lie near a lower dimensional manifold embedded in the ambient space. The classical Coil20 dataset (Nene et al., 1996), which contains images of 20 objects, may be used as an example. For each object, images are taken every 5 degrees as the object is rotated on a turntable, and each image is of size 32×32. In this case, the dimension of ambient space is the number of pixels, which is 1024, while the latent intrinsic structure can be compactly described with the angle of rotation. In addition to Coil20, such a structure can be found in many other data collections. In seismology, two-dimensional coordinates of earthquake epicenters are located along a one-dimensional fault line. In face recognition, high-dimensional facial images are dependent on lighting conditions (Georghiades et al., 2001) or head orientations (Happy et al., 2012).

Given this form of data collection, a natural problem arises: how can we fit a manifold to this data collection? The aim of manifold fitting is to represent the latent manifold as an embedded sub-manifold of the ambient space. Once the latent manifold is learned, various types of analyses can be carried out based on it, such as denoising the observed samples by projecting them to the learned manifold (Gong et al., 2010), generating new data samples from the manifold (Radford et al., 2015), classifying samples according to the manifold (Yao and Zhang, 2020), and detecting fault lines for seismological purposes (Yao et al., 2023). These manifold-based techniques represent powerful tools for understanding and working with complex data structures.

In addition to manifold fitting, dimension reduction constitutes another crucial branch of manifold learning. Over the past two decades, a litany of dimension reduction methods have emerged, each aiming to uncover the intrinsic structure of data by identifying its lower-dimensional embedding, as discussed in the review by Ma and Fu (2011). Unlike manifold fitting, however, these methods primarily focus on mapping data from the ambient space to a lower-dimensional one. Consequently, the outputs of most dimension reduction methods consist of low-dimensional embeddings rather than points in the ambient space, although for applications such as denoising and data generation, relying solely on low-dimensional embeddings may not suffice.

The limitations of dimension reduction and the potential applications of manifold fitting underline the value of formulating the manifold fitting problem, as follows. Suppose the observed data samples $X = \{x_i \in \mathbb{R}^D\}_{i=1}^N$ are in the form

$$x_i = y_i + \xi_i,$$

where $y_1, \cdots, y_N$ are unobserved variables drawn from the uniform distribution supported on the latent manifold $\mathcal{M}$ with dimension $d < D$. Generally, $\mathcal{M}$ is assumed to be a compact and smooth sub-manifold embedded in the ambient space $\mathbb{R}^D$. The precise conditions on $\mathcal{M}$ will be detailed in Section 2.1. The uniform distribution assumption of y_i sampled from $\mathcal{M}$ is the same as those used in the related works (Genovese et al., 2012c, 2014; Mohammed and Narayanan, 2017). Here, $\xi_1, \cdots \xi_N$ are drawn from a distribution G. The assumptions about the noisy distribution G differ among the related work. The simplest assumption is that the observed samples are noiseless (Fefferman et al., 2016; Mohammed and Narayanan, 2017). However, some literature assumes that the noise is distributed in a bounded region centered at the origin, which means that the observed samples are located in a tube centered at $\mathcal{M}$ (Genovese et al., 2012a). Other literature, such as Genovese et al. (2012c, 2014); Fefferman et al. (2018), assumes G to be a Gaussian distribution supported on $\mathbb{R}^D$, whose density at ξ is

$$(\frac{1}{2\pi\sigma^2})^{\frac{D}{2}} \exp(-\frac{\|\xi\|_2^2}{2\sigma^2}). \tag{1.1}$$

The tail of the Gaussian distribution might make the theoretical analysis more challenging than in the previous two cases. Strictly speaking, previous manifold fitting methods have not directly addressed this problem, nor have they proved the convergence of the fitted manifold under this assumption. In this paper, we are concerned with the Gaussian assumption of noisy distribution, denoting it as G_σ to stress the deviation parameter σ hereafter. Under the above settings, the goal is to produce a smooth manifold $\mathcal{M}_{\text{out}}$ convergent to $\mathcal{M}$. Specifically, if σ is sufficiently small, one could derive $\mathcal{M}_{\text{out}}$ such that $d(x, \mathcal{M}) \leq O(\sigma)$ holds for any arbitrary $x \in \mathcal{M}_{\text{out}}$. In particular, $\mathcal{M}_{\text{out}}$ converges to $\mathcal{M}$ when $\sigma \to 0$. The convergence with respect to σ is the domain that Fefferman et al. (2018) is built on, although its final theoretical result is expressed through sample complexity.

1.1 Related work

Methodological studies for manifold fitting can be traced back to works from several decades ago on the principal curve (Hastie and Stuetzle, 1989), with every point on the principal curve/surface defined as the conditional mean value of the points in the orthogonal subspace of the principal curve. Based on Hastie and Stuetzle (1989), many other principal-curve algorithms have been proposed, such as those of Banfield and Raftery (1992); Stanford and Raftery (2000); Verbeek et al. (2002), each attempting to achieve lower estimation bias and improved robustness. More recently, Ozertem and Erdogmus (2011) describes the principal curve in a seemingly different way albeit in a probabilistic sense. In the work by Ozertem and Erdogmus (2011), every point on the principal curve/surface is the local maximum, not the expected value, of the probability density in the local orthogonal subspace. This definition of the principal curve/surface is formulated as a ridge of the probability density. Although it has been demonstrated that these proposed methods produce acceptably accurate estimates in many simulated cases, they do not, however, provide a theoretical analysis for estimating accuracy nor the curvature of the output manifold in general cases, with the exception of special cases such as elliptical distributions.

Recently, some works have focused on the theoretical analysis for manifold fitting. In particular, Genovese et al. (2012a) and Genovese et al. (2012c) establish the upper bounds on the Hausdorff distance between the output and latent manifold under various noise settings, although they do not offer any practical estimators. Genovese et al. (2012b) proposes an estimator which is computationally simple, and whose convergence is guaranteed but its conclusions hold only when the noise is supported on a compact set. Genovese et al. (2014) focuses on the ridge of the probability density introduced by Ozertem and Erdogmus (2011), and proposes a convergent algorithm. It is worth noting that the data in Genovese et al. (2014) was assumed to be blurred by homogeneous Gaussian noise, an assumption that is more general than that made in Genovese et al. (2012b). Boissonnat and Ghosh (2014) proposes an algorithm based upon Delauney complexes, whose convergence was analyzed by Aamari and Levrard (2018). Aamari and Levrard (2019) presented an algorithm to estimate a point on the manifold, its tangent and second form. Based on these, they approximated the latent manifold by a mere union of polynomial patches and gave convergence rate for noise-free and tubular noise models. Aizenbud and Sober (2021) presents an algorithm that showed convergence to the manifold and its tangent bundle, even with tubular noise. However, none of the methods outlined above are guaranteed to output an actual d-dimensional manifold with certain smoothness.

To overcome this issue, some studies on manifold fitting have sought to determine how curved the output manifold is. In the spirit of Ozertem and Erdogmus (2011) and Genovese et al. (2014), Fefferman et al. (2016) and Mohammed and Narayanan (2017) also took the ridge set into consideration, the former focusing on theoretical analysis, the latter on practical algorithms. Specifically, rather than focusing on the probability density function, they both chose to work with the approximate square-distance functions (asdf), approximating the latent manifold by the ridge of the asdf. The theoretical bounds for the manifold fitting have also been considered in Fefferman et al. (2016) and Mohammed and Narayanan (2017), but for only noiseless data; that is, as long as the asdf meets certain regularity conditions, the researchers show that the output of the algorithm is a manifold with bounded reach, and the output manifold is arbitrarily close in Hausdorff to the latent one.

To deal with manifold fitting with noise, Fefferman et al. (2018) proposes a new approach to fit a putative manifold under Gaussian noise. Unlike other methods, which use the entire sample set, the method of Fefferman et al. (2018) involves subsampling first such that the number of used samples can be bounded above by e^D. Under this constraint, the noise is supported on a bounded set with high probability. Given this, the application of Fefferman et al. (2018) is feasible with the bounded noise although the constraint on the sample size is problematic in that the upper bound does not go to zero even with a sufficient number of available samples and the variance of Gaussian noise diminishes. Therefore, the problem is not essentially addressed when the support of noise is unbounded and so creates room for the manifold-fitting problem to arise, especially from the theoretical side.

1.2 Motivation

In this paper, we attempt to evaluate the convergence and smoothness of $\mathcal{M}_{\text{out}}$. Of the aforementioned works, it is those of Mohammed and Narayanan (2017) and Fefferman et al. (2018) are most relevant to our study. This section explains these two methods geometrically and analyzes their limitations, which impels us to establish a more accurate manifold-fitting method.

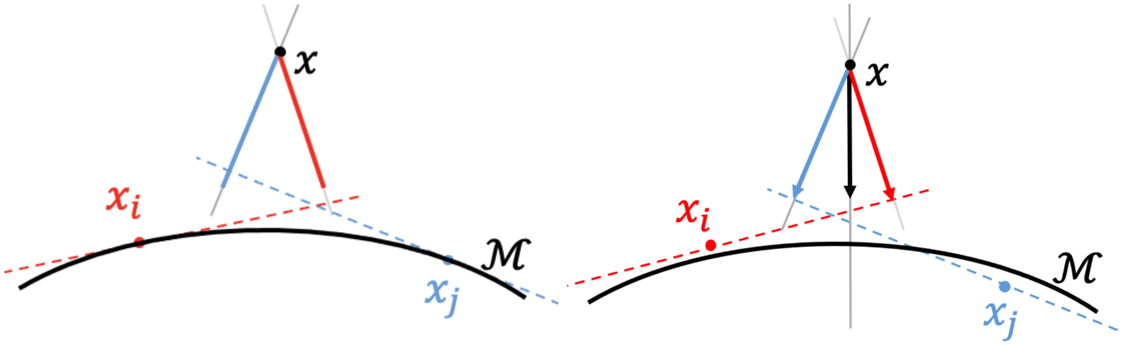

Figure 1: A toy example to illustrate the methods by Mohammed and Narayanan (2017) (left panel) and Fefferman et al. (2018) (right panel), where the black curve is a local part of $\mathcal{M}$, x is a point off $\mathcal{M}$, and the dots x_i and x_j represent two samples in the neighborhood of x. Unlike those in the right panel, the samples in the left panel are on $\mathcal{M}$, as Mohammed and Narayanan (2017) focus on the noiseless case.

The left panel of Figure 1 illustrates the method of Mohammed and Narayanan (2017), whose essence is to define an approximate squared-distance function (asdf) to $\mathcal{M}$ and estimates $\mathcal{M}$ by the ridge set of the asdf. Specifically, for a given point x, the asdf at x is defined as the weighted average of squared distances from x to the discs (the dashed lines) centered at the sample points (the blue and red dots). The right panel of Figure 1 illustrates the method of Fefferman et al. (2018). Its key idea is to approximate the bias from x to $\mathcal{M}$ for any arbitrary x and define the output manifold as points with zero bias. To obtain the approximation of bias from x to $\mathcal{M}$, Fefferman et al. (2018) calculates the weighted average bias from x to the discs (the dashed lines) centered at the sample points, and projects the average bias by the estimated orthogonal projection onto the normal space of $\mathcal{M}$ at x^* (the gray solid line).

Due to the usage of discs centered at the sample points, the effectiveness of each method depends on just how accurately these discs capture the local structure of the manifold. However, if the sample points are significantly perturbed by unbounded noise and deviate significantly from the latent manifold $\mathcal{M}$, these discs also deviate far from $\mathcal{M}$. Hence, in this scenario with unbounded noise, the methods proposed by Mohammed and Narayanan (2017) and Fefferman et al. (2018) may encounter difficulties in fitting a manifold.

Even if the sample points lie on the manifold, say $x_i \in \mathcal{M}$, the disc centered at x_i captures the local manifold at x_i rather than the local manifold at x^*. The deviation between x_i and x^* also introduces an approximation error in the distance/bias from x to $\mathcal{M}$. As shown in Figure 1, both the solid red line and the solid red arrow are shorter than the distance/bias from x to $\mathcal{M}$, and the average between the red and blue one cannot address this issue.

The limitations of the two aforementioned methods suggest that estimating the manifold based on discs centered at sample points may reduce the fitting accuracy, especially in the scenarios with unbounded noise. It is this finding that compels us to invent a new manifold-fitting method.

1.3 Main contribution

From a statistical viewpoint, there is a pressing need for the development of a practical estimator with theoretical bounds satisfying the following requirements simultaneously, and which improves on the requirements of Fefferman et al. (2018):

- The support of noise is unbounded.
- The estimator shares a similar geometric property to $\mathcal{M}$.
- For any arbitrary $x \in \mathcal{M}_{\text{out}}$, the distance between x and $\mathcal{M}$ is bounded above provided N is sufficiently large and σ is sufficiently small. In particular, the distance goes to zero as noise disappears.
- The smoothness of $\mathcal{M}_{\text{out}}$ is mathematically guaranteed.

In this paper, we propose a novel approximation $f(x)$ to the bias from any point x to $\mathcal{M}$ and fit the latent manifold $\mathcal{M}$ in the ambient space as the points with $f(x) = \mathbf{0}$, where $\mathbf{0}$ presents a zero vector. Practically, such an output manifold can be achieved by solving the minimization $\|f(x)\|_2^2$ via gradient descent. This paper provides two main contributions in this aspect, the first being the theoretical analysis satisfying the four requirements above as follows:

- The noise is assumed to be drawn from the Gaussian distribution G_σ defined in (1.1).
- Any arbitrary neighborhood of $\mathcal{M}_{\text{out}}$ is a d-dimensional manifold.
- For any $x \in \mathcal{M}_{\text{out}}$, $d(x, \mathcal{M}) \leq O(\sigma)$ given a large-enough dataset. Thus, $\mathcal{M}_{\text{out}}$ converges to $\mathcal{M}$ for an increasingly large sample size and diminishing noise.
- The twice difference of $\mathcal{M}_{\text{out}}$ is bounded above by $O(\frac{1}{\sqrt{\sigma}})$.

The second important contribution of this paper is the performance of our estimator in practice. As illustrated in Figures 1 and 2, the bias from a point x to $\mathcal{M}$ is approximated better than by the other two relevant methods. Numeric results in Section 5 demonstrate the improved performance, which further suggests that our method outputs the approximated manifold to the latent one.

1.4 Organization

The rest of the paper is organized as follows. Section 2 includes the formulation of our approximation $\mathcal{M}_{\text{out}}$ to the latent manifold $\mathcal{M}$. After that, the convergence and smoothness of $\mathcal{M}_{\text{out}}$ are analyzed in Theorem 5 and Theorem 7, respectively. Section 3 studies the function f defined in (2.6) and determines the properties of its kernel space, the first and second derivatives. Based on these properties of f, the proofs of Theorem 5 and Theorem 7 are derived in Section 4 with numeric examples listed in Section 5 .

2. Proposed method

2.1 Content and notations

Throughout this paper, the latent manifold is denoted as $\mathcal{M}$ and our approximation to $\mathcal{M}$ is denoted as $\mathcal{M}_{\text{out}}$. For a set $A \subset \mathbb{R}^D$ and a point $x \in \mathbb{R}^D$, $\Pi_A x$ denotes the projection of x onto A, namely the nearest point in A to x. Hence $\Pi_{\mathcal{M}} x$ is the projection of x onto the latent manifold. If there is no ambiguity, we might use x^* instead of $\Pi_{\mathcal{M}} x$ for simplicity. The distance between x and A, denoted by $d(x, A)$, is the Euclidean distance between $\Pi_A x$ and x. For any $x^* \in \mathcal{M}$, $T_{x^*}\mathcal{M}$ denotes the tangent space of $\mathcal{M}$ at x^* and Π_{x^*} denotes the orthogonal projection onto the normal space of $\mathcal{M}$ at x^*. We will make frequent use of the lower-cases c, c_0, c_1, etc. and upper-cases C, C_0, C_1 etc., in the rest of this paper with the lower-cases denoting generic constants less than 1, while the upper-cases denote generic constants greater than 1. Values of the generic constants may change from line to line. By constants, we mean they are independent of the radius r, the standard deviation σ or x, while the constants may depend on some other constants used to characterize the manifold, such as the reach of $\mathcal{M}$.

We denote $B_D(x, r)$ as the Euclidean ball in $\mathbb{R}^D$ centered at x of radius r, which defines a neighborhood of x. The index set $I_{x,r}$ is defined as the indices of the sample points in $B_D(x, r)$, and $|I_{x,r}|$ denotes the cardinality of $I_{x,r}$. As given in (1.1), σ represents the standard deviation of noise. Throughout this paper, we assume

$$r = O(\sqrt{\sigma}), \quad \sigma < 1 \tag{2.1}$$

without loss of generality, otherwise the data could be rescaled so that $\sigma < 1$ holds. Here $r = O(\sqrt{\sigma})$ means that there exist constants c and C such that $c\sqrt{\sigma} \le r \le C\sqrt{\sigma}$. Noticing $\sigma < 1$, we obtain

$$r \le C\sqrt{\sigma} < C. \tag{2.2}$$

This means r can be bounded above by certain constant. In subsequent proofs, under the premise that it does not affect the final precision of conclusive upper bound, we will relax some r to C in order to simplify the proof.

The latent manifold $\mathcal{M}$ is supposed to be boundaryless, compact, d-dimensional, and twice differentiable, with a reach bounded by $\tau > 0$. The concept reach is a measure of the regularity of the manifold, first introduced by Federer (Federer, 1959) as follows:

Definition 1 (Reach)**.** *Let $\mathcal{M}$ be a closed subset of $\mathbb{R}^D$. The reach of $\mathcal{M}$, denoted by* reach($\mathcal{M}$)*, is the largest number τ to have the property that any point at a distance $r < \tau$ from $\mathcal{M}$ has a unique nearest point in $\mathcal{M}$.*

An important understanding of reach is that it is a twice differentiable quantity if the manifold is treated as a function. Specifically, if γ is an arc-length parametrized geodesic of $\mathcal{M}$, then for all t, $\|\gamma''(t)\| \leq 1/\tau$ according to Niyogi et al. (2008). As a twice differentiable quantity, it is easy to understand that the reach describes how flat the manifold is locally. For example, the reach of a sharp cusp is zero, and the reach of a linear subspace is infinite. Thus, it is natural that the reach measures how close a manifold is to the tangent space locally. The following proposition by Federer (1959) explains this phenomenon:

Proposition 2.

$$\text{reach}(\mathcal{M})^{-1} = \sup\left\{\frac{2d(y, T_x\mathcal{M})}{\|x-y\|_2^2} | x, y \in \mathcal{M}, x \neq y\right\} \tag{2.3}$$

We emphasize that if reach$(\mathcal{M}) > 0$, the error between $\mathcal{M}$ and $T_x\mathcal{M}$ at y is of a higher order than $\|x-y\|_2$. Thus, in a small-enough neighbor of x, we can estimate $\mathcal{M}$ by $T_x\mathcal{M}$ with negligible error, which is the foundation of our approximation.

The approximation $\mathcal{M}_{\text{out}}$ is defined using the noisy sample points $\{x_i\}_{i=1}^N$. The number of sample points should be sufficiently large such that $B_D(x, r)$ contains enough sample points. Proposition 3 claims the relationship between $|I_{x,r}|$ and N.

Proposition 3. *Suppose x satisfies $d(x, \mathcal{M}) \leq cr$ with some $c < 1$. There exist constants c' and C such that $|I_{x,r}| \geq c' r^d N$ with probability at least $1 - C/\sqrt{N}$.*

Proof of proposition 3 is given in Appendix A.1. Based on this proposition, the requirement on $|I_{x,r}|$ can be transformed to the requirement on N for further analysis in later sections. Specifically, N is required to be a sufficiently large quantity in the order of $O(r^{-(d+2)})$.

2.2 Definition of the approximated manifold

This section introduces a novel method for estimating the bias $f(x)$ from a point x to the latent manifold $\mathcal{M}$, and defines the approximated manifold $\mathcal{M}_{\text{out}}$ as the points satisfying $f(x) = \mathbf{0}$. Unlike the aforementioned methods, which rely on discs centered at sample points to estimate the bias $f(x)$, here we build upon the fact that a Riemannian manifold can be locally treated as an affine space and calculate the bias from x to $T_{x^*}\mathcal{M}$ as an equivalent measure of the bias $f(x)$ from x to $\mathcal{M}$. Thus, the key to addressing such a bias is to find an affine space $\{x' : \Psi_x^\alpha(x' - \mathbf{b})\}$ approximating $T_{x^*}\mathcal{M}$, where Ψ_x^α estimates the orthogonal projection onto the normal space at x^* and $\mathbf{b}$ estimates one points in $T_{x^*}\mathcal{M}$.

In order to approximate the orthogonal projection onto the normal space of $\mathcal{M}$ at x^*, Ψ_x^α is defined as the weighted average of $\{P_{x_i}\}_{i\in I_{x,r}}$, where P_{x_i} is the orthogonal projection perpendicular to the first d principal components in $B_D(x_i, r')$. Mathematically, $P_{x_i} = V_\perp V_\perp^T$, where $V_\perp$ is the orthogonal component of V and V is the $D \times d$ matrix whose columns are the eigenvectors corresponding to the largest d eigenvalues of $\sum_{j\in I_{x_i,r'}}(x_j - x_i)(x_j - x_i)^T$. The radius r' should be sufficiently large, so that the intersection of $B_D(x_i, r')$ and $\mathcal{M}$ is nonempty. Further analysis in Section 3.1 explains that we need $r' \geq 2r$.

As the weighted average of $\{P_{x_i}\}_{i\in I_{x,r}}$,

$$\Psi_x^\alpha = \Pi_{\text{hi}}(A_x), \quad A_x = \sum_{i\in I_{x,r}} \alpha_i(x) P_{x_i}. \tag{2.4}$$

Here, $\Pi_{\mathrm{hi}}(A)$ denotes the projection onto the span of the eigenvectors corresponding to the largest $D-d$ eigenvalues of A. Specifically, $\Pi_{\mathrm{hi}}\big(A\big) = VV^T$, V is a $D \times (D-d)$ matrix whose columns are the eigenvectors corresponding to the largest $D-d$ eigenvalues of A. Further, the weights $\alpha_i : \mathbb{R}^D \to \mathbb{R}$ in (2.4) are defined as follows:

$$\tilde{\alpha}_i(x) = \begin{cases} \left(1 - \frac{\|x-x_i\|_2^2}{r^2}\right)^\beta, & x \in B_D(x_i, r) \\ 0, & otherwise \end{cases}, \quad \tilde{\alpha}(x) = \sum_i \tilde{\alpha}_i(x), \quad \alpha_i(x) = \frac{\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}, \tag{2.5}$$

with $\beta \geq 2$ a fixed integer guaranteeing $f(x)$ in (2.6) to be twice differentiable.

Under the assumption that a manifold can be approximated well by an affine space locally, samples in the neighborhood of x lie close to $T_{x^*}\mathcal{M}$, with the exception of noise. Therefore, a convex combination of these samples also lies close to $T_{x^*}\mathcal{M}$. Thus, we can estimate $\mathbf{b}$ using the average of sample points in the neighborhood of x. Recalling the weights in (2.5), we formulate $\mathbf{b} = \sum_{i \in I_{x,r}} \alpha_i(x) x_i$ as the weighted average of sample points in the neighborhood of x. Then the bias from x to the space $\{x' : \Psi_x^\alpha(x' - \mathbf{b})\}$ is

$$f(x) : \mathbb{R}^D \to \mathbb{R}^D, \quad f(x) = \Psi_x^\alpha\bigg(x - \sum_{i \in I_{x,r}} \alpha_i(x) x_i\bigg). \tag{2.6}$$

Finally, the approximation is defined as

$$\mathcal{M}_{\mathrm{out}} = \{x : d(x, \mathcal{M}) \leq cr,\ f(x) = \mathbf{0}, c < 1\}, \tag{2.7}$$

that is, the points with zero bias. By Definition 11 of Fefferman et al. (2016), $\tilde{\mathcal{M}} = \{x : d(x, \mathcal{M}) \leq cr\}$ is a manifold. Restrict f to $\tilde{\mathcal{M}}$. When $\mathbf{0}$ is regular, the preimage $f^{-1}(\mathbf{0}) = \mathcal{M}_{\mathrm{out}} \subset \tilde{\mathcal{M}}$ is a smooth submanifold. So we call $\mathcal{M}_{\mathrm{out}}$ as the approximated manifold in the paper. Further characterization of the approximated manifold will be discussed in Theorem 4.

The definition of $\mathcal{M}_{\mathrm{out}}$ is practical. Theorem 5 in the next section claims that $\mathcal{M}_{\mathrm{out}}$ approximates $\mathcal{M}$ in the order of $O(r^2)$. This means if we have an initial estimator of $\mathcal{M}$ with error cr, then we could achieve a better estimator of $\mathcal{M}$ using the definition of $\mathcal{M}_{\mathrm{out}}$. In practice, we solve the minimization $\|f(x)\|_2^2$ via the gradient descent method given the initial estimator, and the output of the gradient descent method approximates $\mathcal{M}$ in the order of $O(r^2)$ better than the initial guess.

The advantages of our method are twofold. First, we introduce Ψ_x^α directly, thus capturing the local structure of manifold at x^*, while the aforementioned methods capture the local structure of manifold near the sample points. Second, we approximate the manifold using a space passing $\mathbf{b}$ instead of any sample point. Benefitting from the mutual offset of noise, $\mathbf{b}$ hardly deviates far away from $T_{x^*}\mathcal{M}$ even if the noise is unbounded. As a result, we can expect $\{x' : \Psi_x^\alpha(x' - \mathbf{b})\}$ to be a better approximation to the local structure of manifold at x^*, which guarantees the bias from x to $\{x' : \Psi_x^\alpha(x' - \mathbf{b})\}$ is a better approximation to the bias from x to $\mathcal{M}$. The toy example in Figure 2 illustrates the superiority of our method. The black arrow in Figure 2 is almost the bias from x to $\mathcal{M}$, while both the average length of the solid lines in the left panel of Figure 1 and the black arrow in the right panel of Figure 1 are shorter than the ideal one.

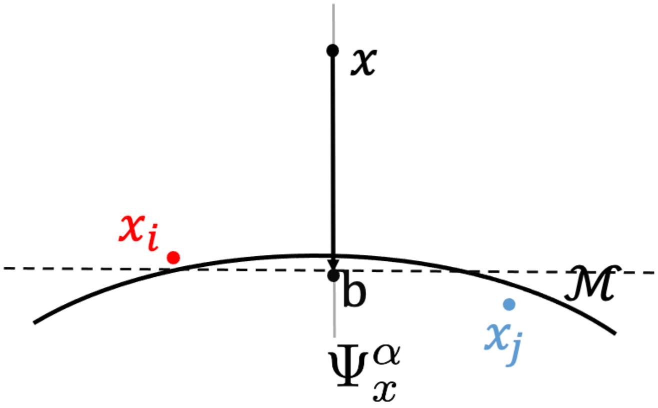

Figure 2: A toy example to illustrate the methods in our method. Ψ_x^α is used to estimate the orthogonal projection onto the normal space of $\mathcal{M}$ at x^*, the black dot $\mathbf{b}$ is used to estimate a point in $T_{x^*}\mathcal{M}$. Then the space $\{x' : \Psi_x^\alpha(x' - \mathbf{b})\}$, illustrated as the black dashed line, approximates $T_{x^*}\mathcal{M}$, and the bias from x to the black dashed line is the estimated bias from x to $\mathcal{M}$, geometrically illustrated as the black arrow.

2.3 Convergence and smoothness of the approximated manifold

In Theorem 4, we prove any arbitrary neighborhood of $\mathcal{M}_{\text{out}}$ is a d-dimensional manifold in high probability. In Theorem 5, we characterize the convergence of $\mathcal{M}_{\text{out}}$ in the probability $\delta_0(1-\delta)^2$, where we denote

$$\delta_0 = 1 - d\exp\{\frac{-cr^{d+2}N}{2\ln 2}\} \tag{2.8}$$

for convenience. When N is sufficiently large as we set, δ_0 is a high probability. Theorem 5 tells us that if $r = O(\sqrt{\sigma})$ is sufficiently small, $\mathcal{M}_{\text{out}}$ is a good estimator to $\mathcal{M}$. Moreover, Corollary 6 tells us that the approximated manifold $\mathcal{M}_{\text{out}}$ converges to the latent manifold $\mathcal{M}$ as $\sigma \to 0$.

Theorem 4. *Given $\delta > 0$ and any arbitrary $x \in \mathcal{M}_{\text{out}}$, there exists ϵ such that $\mathcal{M}_{\text{out}} \cap B_D(x,\epsilon)$ is a d-dimensional manifold with probability $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$.*

Theorem 5. *Given $\delta > 0$, there exists a constant C such that $d(x,\mathcal{M}) \leq Cr^2$ for any arbitrary $x \in \mathcal{M}_{\text{out}}$ with probability at least $\delta_0(1-\delta)^2$.*

We point out that Theorem 5 holds assuming $\sigma < 1$ and $r = O(\sqrt{\sigma})$ as (2.1) claims. If we further assume $\sigma \to 0$, we achieve the following corollary:

Corollary 6. *For any arbitrary $x \in \mathcal{M}_{\text{out}}$, $d(x,\mathcal{M}) \to 0$ as $\sigma \to 0$ with probability at least $\delta_0(1-\delta)^2$.*

Proof Given $r = O(\sqrt{\sigma})$, there exists C_0 such that $r = C_0\sqrt{\sigma}$. For any $\varepsilon > 0$, let $\sigma = \frac{\varepsilon}{CC_0^2}$, and then $d(x,\mathcal{M}) \leq Cr^2 = CC_0^2\sigma = \varepsilon$. ■

Given $x, y \in \mathcal{M}$, the fraction $d(y, T_x\mathcal{M}_{\text{out}})/\|y-x\|_2^2$ characterizes the twice differentiable quantity that controls the local flatness of $\mathcal{M}_{\text{out}}$. Therefore, the lower bound of $d(y, T_x\mathcal{M}_{\text{out}})/\|y-x\|_2^2$ guarantees the smoothness of $\mathcal{M}_{\text{out}}$. Recalling Proposition 2, such a quantity is related to the reach of a manifold, which characterizes the smoothness of a manifold.

Theorem 7. *Given $\delta > 0$, there exists constant $c_0 < 1$ and $c < 1$ such that*

$$\frac{\|z-x\|_2^2}{d(z, T_x\mathcal{M}_{\text{out}})} \geq c_0 r$$

for any arbitrary x and z in $\mathcal{M}_{\text{out}}$ with probability at least $\delta_0^2(1-\delta)^4\big(1-(1-cr^d)^N\big)$.

The proofs of Theorem 4, Theorem 5 and Theorem 7 are organized in the following way. First we explore the properties of P_{x_i} for given x_i through Theorem 11 in subsection 3.1, reveal the properties of weights $\{\alpha_i\}$ through Proposition 12 and discuss the concentration phenomenon of Gaussian noise through Lemma 13 in subsection 3.2. Based on the conclusions above, we prove in Theorem 15 an upper bound on the approximation error of Ψ_x^α, as a weighted sum of $\{P_{x_i}\}_{i\in I_{x,r}}$, to Π_{x^*}. Subsequently, in subsection 3.3 and subsection 3.4, we obtain the upper bounds on $\|f(x)\|_2$ and the first and second derivative of $f(x)$ through Theorem 16, Theorem 17 and Theorem 19. Finally, the main conclusions, namely Theorem 4, Theorem 5 and Theorem 7, are proved in Section 4, using the upper bounds regarding $f(\cdot)$ defined by (2.6). The dependency of the above theorems, lemmas, and propositions is demonstrated in Figure 3.

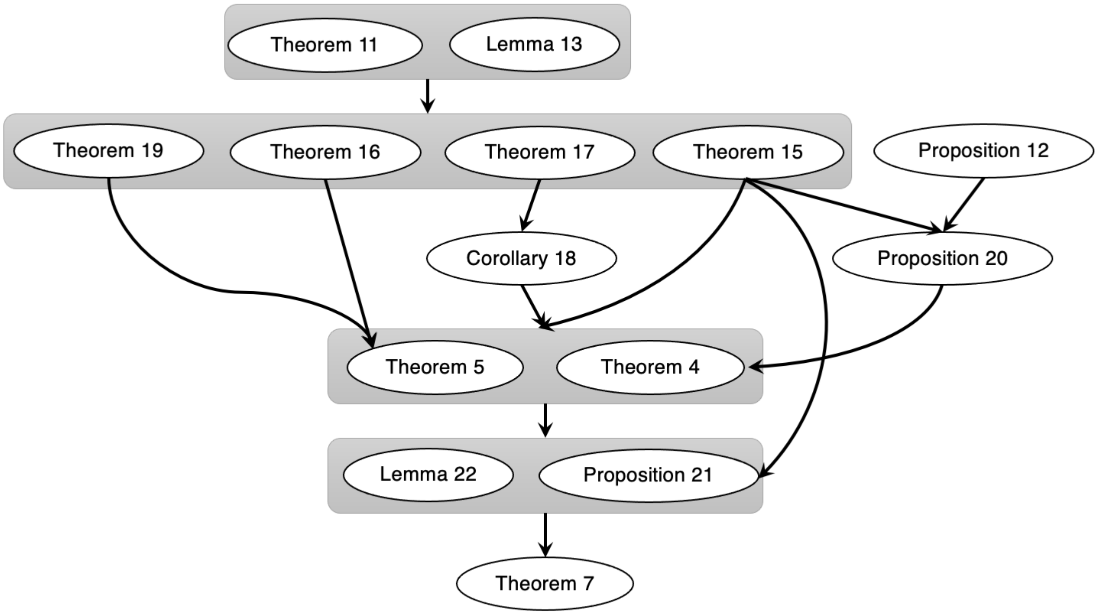

Figure 3: The dependency of the core theorems, lemmas, and propositions.

3. Bounds regarding the function f

To analyze Ψ_x^α, we first explore the properties of P_{x_i}, where x_i is any arbitrary sample point in $B_D(x, r)$. Next, properties of Ψ_x^α can be analyzed as the weighted average of $\{P_{x_i}\}_{i \in I_{x,r}}$. Finally, we successively bound $\|f(x)\|_2$, the first derivative of $f(x)$ and the second derivative of $f(x)$ above using bounds regarding Ψ_x^α.

3.1 Properties of P_{x_i}

To make the notations clearer, we replace x_i with z in this section. Recalling the notations in Section 2.1, z^* is the closest point on $\mathcal{M}$ to z and Π_{z^*} is the orthogonal projection onto the normal space of $\mathcal{M}$ at z^*. The aim of this section is to bound the error $\|P_z - \Pi_{z^*}\|_F$.

Figure 4 illustrates the variables used for the discussion of P_z and the related proof. The z (black dot) is an observed noisy point of the manifold $\mathcal{M}$ and the blue ball is $B_D(z, r')$, centered at z with radius r'. The subsequent proof requires $B_D(z, r') \cap \mathcal{M} \neq \emptyset$, which is equal to $d(z, \mathcal{M}) \leq r'$. Given $d(x, \mathcal{M}) \leq cr$ and $z \in B_D(x, r)$, we have

$$d(z, \mathcal{M}) \leq \|z - x^*\|_2 \leq \|z - x\|_2 + \|x - x^*\|_2 = \|z - x\|_2 + d(x, \mathcal{M}) \leq (c + 1)r.$$

Therefore, for any $r' \geq 2r$, $d(z, \mathcal{M}) < r'$. In this paper, we set $r' = 2r$ for convenience. The z_i (red dot) is a noisy sample located in $B_D(z, r')$, satisfying $z_i = y_i + \xi_i$, $z_i^* = \Pi_{\mathcal{M}} z_i$, and p_i is the projection of z_i onto $T_{z^*}\mathcal{M}$. The space T is the translation of $T_{z^*}\mathcal{M}$ passing z, and p_i' is the projection of z_i onto T.

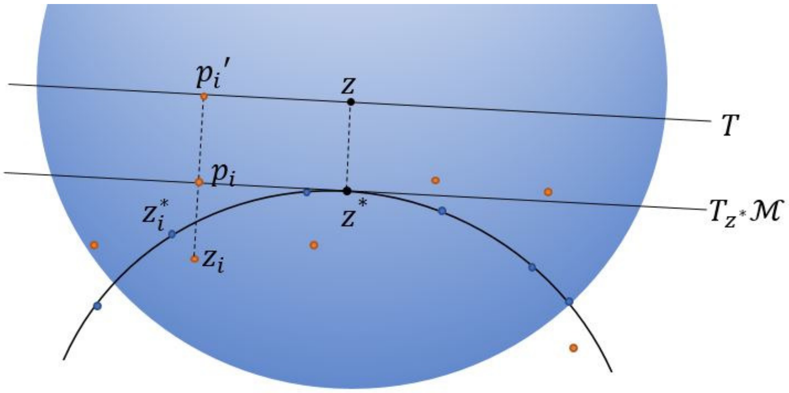

Figure 4: Diagram of variables used for the discussion of P_z.

Consider the symmetric matrix $\Lambda = \frac{1}{|I_{z,r'}|} \sum_{i \in I_{z,r'}} (p_i - z^*)(p_i - z^*)^T$. Since both p_i and z^* are located in $T_{z^*}\mathcal{M}$, the spanning space of Λ is contained in $T_{z^*}\mathcal{M}$. Thus, $\text{rank}(\Lambda) \leq \dim(T_{z^*}\mathcal{M}) = d$ and thereby the $(d+1)$-th largest eigenvalue of Λ is 0. Setting columns of U be the eigenvectors of Λ corresponding to the d largest eigenvalues, U is also a basis of $T_{z^*}\mathcal{M}$. Let $U_\perp$ be the orthogonal complement of U, and we have $\Pi_{z^*} = U_\perp U_\perp^T$. Recalling $P_z = V_\perp V_\perp^T$, where $V_\perp$ is the orthogonal component of V and columns of V are the eigenvectors corresponding to the d largest eigenvalues of

$$\hat{\Lambda} = \frac{1}{|I_{z,r'}|} \sum_{i \in I_{z,r'}} (z_i - z)(z_i - z)^T,$$

we obtain

$$\|P_z - \Pi_{z^*}\|_F = \|V_\perp V_\perp^T - U_\perp U_\perp^T\|_F = \|VV^T - UU^T\|_F \leq \frac{2\sqrt{2}\|\Lambda - \hat{\Lambda}\|_F}{\lambda_d} \tag{3.1}$$

by the following Lemma:

Lemma 8. *Let Λ, $\hat{\Lambda} \in \mathbb{R}^{n\times n}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ respectively. Let $1 \leq d \leq n$ and assume $\lambda_d > 0$, $\lambda_{d+1} = 0$. Let U, $\hat{U} \in \mathbb{R}^{n\times d}$ be eigenvectors corresponding to the first d eigenvalues of Λ and $\hat{\Lambda}$, respectively. Then*

$$\|UU^T - \hat{U}\hat{U}^T\|_F = \sqrt{2}\|\sin\theta(\hat{U}, U)\|_F \leq \frac{2\sqrt{2}\|\hat{\Lambda} - \Lambda\|_F}{\lambda_d}$$

by the Davis-Kahan $\sin\theta$ theorem, where $\theta(\hat{U}, U)$ is the $n \times n$ diagonal matrix, whose diagonal comprises the principal angles between the column spaces of $\hat{U}$ and U, and $\sin\theta(\hat{U}, U)$ is defined entrywise.

We require the upper bound on $\|\hat{\Lambda} - \Lambda\|_F$ and the lower bound on the d-th eigenvalue of Λ, deriving them both in Lemma 9 and Lemma 10 as follows:

Lemma 9. *Suppose $r' = O(\sqrt{\sigma})$ and $d(z, \mathcal{M}) \leq cr'$ with some $c < 1$. There exists C such that $\|\frac{1}{|I_{z,r'}|}\big(\sum_{i\in I_{z,r'}}(z_i - z)(z_i - z)^T - \sum_{i\in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T\big)\|_F$ is bounded above by*

$$\frac{C}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}\left(\|\xi_i\|_2^4 + \|\xi_i\|_2^3 + \|\xi_i\|_2^2 + r'\|\xi_i\|_2\right) + C\left(r'^3 + r'\|z - z^*\|_2 + \|z - z^*\|_2^2\right).$$

Lemma 10. *The d-th eigenvalue of $\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T$ is bounded below by $\lambda_d \geq cr'^2$, with probability δ_0.*

Proofs of Lemma 9 and 10 appear in Appendix A.2. Plugging the upper bound of Lemma 9 and the lower bound of Lemma 10 into (3.1), we can obtain the following theorem:

Theorem 11. *Suppose $r' = O(\sqrt{\sigma})$ and $d(z, \mathcal{M}) \leq r'$. For any given δ ,there exists C such that the difference between P_z and Π_{z^*} is bounded by*

$$\begin{aligned}\|P_z - \Pi_{z^*}\|_F \leq \frac{C}{r'^2}\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}&\left(\|\xi_i\|_2^4 + \|\xi_i\|_2^3 + \|\xi_i\|_2^2 + r'\|\xi_i\|_2\right)\\ &+ C\left(r' + \frac{\|z - z^*\|_2}{r'} + \frac{\|z - z^*\|_2^2}{r'^2}\right), \quad \text{in probability } \delta_0.\end{aligned}$$

The term $\|z - z^*\|_2^2$ in Theorem 11 tells us that P_z cannot approximate Π_{z^*} well if z is distant from $\mathcal{M}$. When the sample size N is sufficiently large and the sample points are blurred by Gaussian noise, there will always be several sample points that deviate far away from the latent manifold. If x_i represents such a sample point, given Theorem 11, P_{x_i} cannot effectively capture the local structure of the latent manifold. However, the next section will explain how the error caused by P_{x_i} can be eliminated when we calculate a weighted average over $\{P_{x_i}\}_{i\in I_{x,r}}$, denoted as Ψ_x^α.

3.2 Properties of Ψ_x^α

This section evaluates how Ψ_x^α approximates Π_{x^*} using the upper bound of $\|P_{x_i} - \Pi_{x_i^*}\|_F$ as derived in Theorem 11. As the weighted average of $\{P_{x_i}\}_{i\in I_{x,r}}$, Ψ_x^α benefits from the mutual offset of the Gaussian noise. To mathematically clarify this phenomenon, Lemma 13 below bounds the weighted averages regarding $\|\xi_i\|_2$ above by the Berry-Esseen Theorem and the properties of the weights $\{\alpha_i(x)\}$ stated in Proposition 12.

Proposition 12. *For a point x satisfying $d(x,\mathcal{M}) \le cr$, there exist constants c_0 and c_0' such that*

(i) $\tilde{\alpha}(x)$ is bounded below by $c_0|I_{x,r}|$, with probability $1 - C_0/\sqrt{|I_{x,r}|}$

(ii) $\tilde{\alpha}(x)$ is bounded below by a constant c_0' with probability $1-(1-cr^d)^N = O(Nr^d)$.

Lemma 13. *Suppose $d(x,\mathcal{M}) \le cr$ with some constant $c < 1$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C, c_0 and n_0 such that if $N \ge n_0 r^{-d}$, then $\tilde{\alpha}(x) \ge c_0|I_{x,r}|$ with probability at least $(1-\delta)$ and*

$$\sum_{i\in I_{x,r}} \alpha_i(x)\|\xi_i\|_2^k \le C\sigma^k \quad \text{and} \quad \frac{1}{|I_{x,r}|^2}\sum_{i,j\in I_{x,r}} \|\xi_i\|_2^s\|\xi_i\|_2^t \le C\sigma^{s+t} \tag{3.2}$$

hold for $k,s,t \le 4$ with probability at least $(1-\delta)^2$.

Lemma 14. *Suppose x and y are two points on $\mathcal{M}$, then*

$$\|\Pi_x^* - \Pi_y^*\|_2 \le \|\Pi_x^* - \Pi_y^*\|_F \le C\frac{\|x-y\|_2}{\tau}. \tag{3.3}$$

The proof of Lemma 13 is shown in Appendix A.3. To evaluate how Ψ_x^α approximates Π_{x^*}, we first evaluate how the tangent space changes when the point of tangency changes as Lemma 14, which is also proved in Appendix A.3. Based on the theorem above and lemmas, we obtain the following theorem to evaluate Ψ_x^α:

Theorem 15. *Suppose $d(x,\mathcal{M}) \le cr$ with some constant $c < 1$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N \ge n_0 r^{-d}$, then*

$$\|\Psi_x^\alpha - \Pi_{x^*}\|_2 \le \|\Psi_x^\alpha - \Pi_{x^*}\|_F \le Cr \tag{3.4}$$

holds with probability $\delta_0(1-\delta)^2$.

Proof By definition of A_x,

$$\begin{aligned}\|A_x - \Pi_{x^*}\|_F &= \left\| \sum_{i\in I_{x,r}} \alpha_i(x)(P_{x_i} - \Pi_{x_i^*}) + \sum_i \alpha_i(x)(\Pi_{x_i^*} - \Pi_{x^*}) \right\|_F \\ &\le \sum_{i\in I_{x,r}} \alpha_i(x)\|P_{x_i} - \Pi_{x_i^*}\|_F + \sum_{i\in I_{x,r}} \alpha_i(x)\|\Pi_{x_i^*} - \Pi_{x^*}\|_F. \end{aligned} \tag{3.5}$$

Setting z in Theorem 11 to be x_i and replacing r' by $r' = 2r$, we obtain the upper bound of $\|P_{x_i} - \Pi_{x_i^*}\|_F$ with probability $(1-\delta)^2$. Plugging the upper bound into the first term on the right-hand side of (3.5), we obtain

$$\begin{aligned}\sum_{i\in I_{x,r}} \alpha_i(x)\|P_{x_i} - \Pi_{x_i^*}\|_F \leq{}& \frac{C}{r^2}\sum_{i\in I_{x,r}}\sum_{j\in I_{x_i,2r}} \frac{\alpha_i(x)}{|I_{x_i,2r}|}\Big(\|\xi_j\|_2^4 + \|\xi_j\|_2^3 + \|\xi_j\|_2^2 + r\|\xi_j\|_2\Big)\\ &+ C\Big(r + \frac{\sum_{i\in I_{x,r}}\alpha_i(x)\|x_i - x_i^*\|_2}{r} + \frac{\sum_{i\in I_{x,r}}\alpha_i(x)\|x_i - x_i^*\|_2^2}{r^2}\Big)\\ \leq{}& \frac{C}{r^2}\sum_{i\in I_{x,r}}\sum_{j\in I_{x_i,2r}} \frac{\alpha_i(x)}{|I_{x_i,2r}|}\Big(\|\xi_j\|_2^4 + \|\xi_j\|_2^3 + \|\xi_j\|_2^2 + r\|\xi_j\|_2\Big)\\ &+ C\Big(r + \frac{\sum_{i\in I_{x,r}}\alpha_i(x)\|\xi_i\|_2}{r} + \frac{\sum_{i\in I_{x,r}}\alpha_i(x)\|\xi_i\|_2^2}{r^2}\Big).\end{aligned}$$

Plugging the upper bound of $\sum_{i\in I_{x,r}}\alpha_i(x)\|\xi_i\|_2^k$ into the last formula leads to

$$\sum_{i\in I_{x,r}} \alpha_i(x)\|P_{x_i} - \Pi_{x_i^*}\|_F \leq C\Big(\frac{\sigma}{r} + \frac{\sigma^2}{r^2} + r\Big) \leq Cr,$$

with probability $\delta_0(1-\delta)^2$, where the last inequality holds given $r = O(\sqrt{\sigma})$. As for the second term on the right-hand side of (3.5),

$$\begin{aligned}\sum_{i\in I_{x,r}} \alpha_i(x)\|\Pi_{x_i^*} - \Pi_{x^*}\|_F &\leq \frac{C}{\tau}\sum_{i\in I_{x,r}}\alpha_i(x)\|x_i^* - x^*\|_2\\ &\leq \frac{C}{\tau}\sum_{i\in I_{x,r}}\alpha_i(x)\Big(\|x_i^* - x_i\|_2 + \|x_i - x\|_2 + \|x - x^*\|_2\Big)\\ &\leq C\frac{r}{\tau} + \frac{C}{\tau}\sum_{i\in I_{x,r}}\alpha_i(x)\|x_i^* - x_i\|_2\\ &\leq C\frac{r}{\tau} + C\frac{\sigma}{\tau} \leq C\frac{r}{\tau}.\end{aligned}$$

where the first inequality is by Lemma 14, the second-to-last inequality holds given $r = O(\sqrt{\sigma})$, and the last inequality holds by (2.2). Since Ψ_x^α is the closest $(D-d)$-rank projection matrix to A_x, we have

$$\|\Psi_x^\alpha - A_x\|_F \leq \|A_x - \Pi_{x^*}\|_F \leq Cr, \quad \text{with probability } \delta_0(1-\delta)^2. \tag{3.6}$$

Hence, $\|\Psi_x^\alpha - \Pi_{x^*}\|_F \leq \|\Psi_x^\alpha - A_x\|_F + \|A_x - \Pi_{x^*}\|_F \leq Cr$ with probability $\delta_0(1-\delta)^2$. ■

3.3 A bound on $f(x)$

This section examines how $f(x)$ approximates the bias from x to $\mathcal{M}$, which is achieved by calculating $\|f(x)\|_2$ for $x \in \mathcal{M}$. If f approximates the bias well, such $\|f(x)\|_2$ should be bounded above by a small value with $x \in \mathcal{M}$.

Theorem 16. *Suppose $x \in \mathcal{M}$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N \geq n_0 r^{-d}$, then $\|f(x)\|_2 \leq Cr^2$ with probability $\delta_0(1-\delta)^2$.*

Proof It is clear that $x = x^*$ when $x \in \mathcal{M}$. Accordingly, we use x instead of x^* in the following discussion for convenience. First, we bound the distance between $\sum_{i \in I_{x,r}} \alpha_i(x) x_i$ and $T_x\mathcal{M}$. By definition,

$$\begin{aligned}
d\Big(\sum_{i \in I_{x,r}} \alpha_i(x) x_i, T_x\mathcal{M}\Big) &= \Big\|\Pi_x^*\Big(\sum_{i \in I_{x,r}} \alpha_i(x) x_i - x\Big)\Big\|_2 \\
&\leq \sum_{i \in I_{x,r}} \alpha_i(x) \|\Pi_x^*(x_i - x)\|_2 \\
&\leq \sum_{i \in I_{x,r}} \alpha_i(x) \|x_i - x_i^*\|_2 + \sum_{i \in I_{x,r}} \alpha_i(x) \|\Pi_x^*(x_i^* - x)\|_2 \\
&\leq \sum_{i \in I_{x,r}} \alpha_i(x) \|\xi_i\|_2 + \sum_{i \in I_{x,r}} \alpha_i(x) \frac{\|x_i^* - x\|_2^2}{\tau} \\
&\leq C_1\sigma + C_2 \sum_{i \in I_{x,r}} \alpha_i(x) \frac{(\|x_i^* - x_i\|_2 + \|x_i - x\|_2)^2}{\tau} \\
&\leq C_1\sigma + C_2 \frac{(\sigma + r)^2}{\tau},
\end{aligned}$$

where the second-to-last inequality holds by Lemma 13 with probability $(1-\delta)^2$. The parameter r is selected in the order of $\sqrt{\sigma}$, namelyC_3 exists such that $r = C_3\sqrt{\sigma} > C_3\sigma$ since $\sigma < 1$. So $(\sigma + r)^2 < (\frac{1}{C_3} + 1)^2 r^2$ and

$$C_1\sigma + \frac{C_2}{\tau}(\sigma + r)^2 \leq \frac{C_1 r^2}{C_3^2} + \frac{C_2}{\tau}(\frac{1}{C_3} + 1)^2 r^2 = Cr^2.$$

Hence, we obtain $d(\sum_{i \in I_{x,r}} \alpha_i(x) x_i, T_x\mathcal{M}) \leq Cr^2$.

We let $a = \sum_{i \in I_{x,r}} \alpha_i(x) x_i$ and b be the projection of a onto $T_x\mathcal{M}$. Then, we have

$$\|a - b\|_2 = \|\Pi_x^*(a - b)\|_2 = d\Big(\sum_{i \in I_{x,r}} \alpha_i(x) x_i, T_x\mathcal{M}\Big) \leq Cr^2.$$

According to the definition of $f(x)$,

$$f(x) = \Psi_x^\alpha(x - a) = \Pi_x^*(x - b) + (\Psi_x^\alpha - \Pi_x^*)(x - b) + \Psi_x^\alpha(b - a),$$

where $\Pi_x^*(x - b) = \mathbf{0}$, since $x = x^* \in T_x\mathcal{M}$ and $b \in T_x\mathcal{M}$. Hence, we obtain

$$\begin{aligned}
\|f(x)\|_2 &\leq \|\Psi_x^\alpha - \Pi_x^*\|_F\big(\|x - a\|_2 + \|a - b\|_2\big) + \|\Psi_x^\alpha(a - b)\|_2 \\
&\leq \|\Psi_x^\alpha - \Pi_x^*\|_F\big(\|x - a\|_2 + \|a - b\|_2\big) + \|a - b\|_2 \\
&\leq C_1 r \times (r + r^2) + C_2 r^2 \leq Cr^2,
\end{aligned}$$

where the second-to-last inequality holds by Theorem 11 with probability δ_0 and the last inequality holds by (2.2). In summary, $\|f(x)\|_2 \leq Cr^2$ with probability $\delta_0(1-\delta)^2$. ■

3.4 A bound on the first and second derivative of $f(x)$

We now proceed to obtain an upper bound on $\|\partial_v f(x)\|_2$ with $\|v\|_2 = 1$, where

$$\partial_v f(x) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t},$$

for any $v \in \mathbb{R}^D$.

Theorem 17. *Suppose $d(x, \mathcal{M}) \leq cr$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N \geq n_0 r^{-d}$,*

$$\|\partial_v f(x) - \Psi_x^\alpha v\|_2 \leq Cr, \tag{3.7}$$

with probability $\delta_0(1-\delta)^2$.

The proof of Theorem 17 refers to Appendix A.4. This theorem claims the first derivative of $f(x)$ approximates $\Psi_x^\alpha v$ in the order of $O(r)$. Taking v in Theorem 17 as $e_1, \cdots, e_D$, we achieve the following Corollary 18.

Corollary 18. *Suppose $d(x, \mathcal{M}) \leq cr$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N \geq n_0 r^{-d}$,*

$$\|J_f(x) - \Psi_x^\alpha\|_F \leq Cr \tag{3.8}$$

with probability $\delta_0(1-\delta)^2$.

Proof Let e_i represent a D-dimensional vector where the i-th component is 1, and the other components are 0. The Jacobian matrix of function f can be represented as

$$J_f(x) = \big(\partial_{e_1} f(x), \cdots, \partial_{e_D} f(x)\big)$$

and $\Psi_x^\alpha = \big(\Psi_x^\alpha e_1, \cdots, \Psi_x^\alpha e_D\big)$. Hence,

$$\begin{aligned}\|J_f(x) - \Psi_x^\alpha\|_F &= \|\big(\partial_{e_1} f(x), \cdots, \partial_{e_D} f(x)\big) - \big(\Psi_x^\alpha e_1, \cdots, \Psi_x^\alpha e_D\big)\|_F \\ &= \sqrt{\sum_{i=1}^D \|\partial_{e_i} f(x) - \Psi_x^\alpha e_i\|_2^2} \leq \sqrt{\sum_{i=1}^D (C_1^2 r^2)} = C_1\sqrt{D} r = Cr.\end{aligned}$$

The last inequality holds by Theorem 17, which concludes $\|J_f(x) - \Psi_x^\alpha\|_F \leq Cr$ with probability $\delta_0(1-\delta)^2$. ■

We now proceed to obtain an upper bound on $\|\partial_v\big(\partial_u f(x)\big)\|_2$ with $\|v\|_2 = \|u\|_2 = 1$ in Theorem 19. This theorem proves that the second derivative of $f(x)$ is bounded above by a certain constant, which indicates the smoothness of $f(x)$. The proof of Theorem 19 refers to Appendix A.5.

Theorem 19. *Suppose $d(x, \mathcal{M}) \leq cr$ with some constant $c < 1$ and $r = O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N \geq n_0 r^{-d}$, then $\|\partial_v \partial_u f(x)\|_2 \leq C$ with probability $\delta_0(1-\delta)^2$.*

4. Proofs of Theorem 4, Theorem 5 and Theorem 7

Theorem 4 claims that the intersection of $\mathcal{M}_{\text{out}}$ and the neighborhood of $x \in \mathcal{M}_{\text{out}}$ is a d-dimensional manifold. To prove this conclusion, we first need to discuss the properties of the neighborhood of x, as stated in Proposition 20.

Proposition 20. *Let* $\epsilon = \min\{\sqrt{\frac{\alpha(x)}{|I_{x,2r}|^2}\frac{r^3}{\beta}}, r\}$ *for given* x, *then*

$$\|\Psi_x^\alpha - \Psi_z^\alpha\|_2 \leq Cr, \quad \forall z \in B_D(x,\epsilon)$$

with probability $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$.

The proof of Proposition 20 can be found in Appendix A.6. Based on Proposition 20, we construct an auxiliary function h to further characterize the neighborhood of x and obtain the proof of Theorem 4, as shown below:

Proof of Theorem 4 Let $h(z) : B_D(x,\epsilon) \subset \mathbb{R}^D \to \mathbb{R}^{D-d}$, per

$$h(z) = V_x^T f(z), \tag{4.1}$$

where V_x is the factor of Ψ_x^α such that $\Psi_x^\alpha = V_x V_x^T$. Then $h(z) = \mathbf{0}$ if $f(z) = \mathbf{0}$. Assuming there exists z such that $h(z) = \mathbf{0}$ but $f(z) \neq \mathbf{0}$, we obtain

$$\begin{aligned}\|\Psi_x^\alpha - \Psi_z^\alpha\|_2 &= \max_{v \neq 0} \frac{\big\|(\Psi_x^\alpha - \Psi_z^\alpha)v\big\|_2}{\|v\|_2} \geq \frac{\big\|(\Psi_x^\alpha - \Psi_z^\alpha)f(z)\big\|_2}{\|f(z)\|_2} \\ &= \frac{\big\|\Psi_x^\alpha f(z) - \Psi_z^\alpha f(z)\big\|_2}{\|f(z)\|_2} = \frac{\big\|V_x h(z) - f(z)\big\|_2}{\|f(z)\|_2} = \frac{\big\|0 - f(z)\big\|_2}{\|f(z)\|_2} = 1.\end{aligned}$$

However, $\|\Psi_x^\alpha - \Psi_z^\alpha\|_2 \leq Cr$ with probability $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$ via Proposition 20, which is contradictory to $\|\Psi_x^\alpha - \Psi_z^\alpha\|_2 \geq 1$. Hence, $f(z) = \mathbf{0}$ if and only if $h(z) = \mathbf{0}$, equivalently $h^{-1}(\mathbf{0}) = \mathbf{f^{-1}}(\mathbf{0})$ in $B_D(x,\epsilon)$, with probability $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$.

For $z \in B_D(x,\epsilon)$,

$$\begin{aligned}J_h(z) &= V_x^T J_f(z) \\ &= V_x^T(J_f(z) - J_f(x)) + V_x^T(J_f(x) - \Psi_x^\alpha) + V_x^T \Psi_x^\alpha.\end{aligned}$$

On the right hand side of the above equality, we have $\|J_f(x) - \Psi_x^\alpha\|_F \leq Cr$ by Corollary 18, $V_x^T \Psi_x^\alpha = V_x^T$ and $\|J_f(z) - J_f(x)\|_F \leq C \max_{i=1}^D \|J_{e_i} f(z) - J_{e_i} f(x)\|_2 \leq C\|x - z\|_2 \leq C\epsilon \leq Cr$, where the second inequality holds by Theorem 19, which implies

$$\|J_h(z) - V_x^T\|_2 \leq \|J_h(z) - V_x^T\|_F = \|V_x^T(J_f(z) - J_f(x)) + V_x^T(J_f(x) - \Psi_x^\alpha)\|_F \leq Cr.$$

Hence, the maximal difference between the singular values of $J_h(z)$ and V_x^T is bounded by Cr. Let $\sigma_1 \geq \cdots \geq \sigma_{D-d}$ be the singular values of $J_h(z)$. We obtain $|\sigma_{D-d} - 1| \leq Cr$ since the singular values of V_x^T are 1, which implies $\sigma_{D-d} \geq 1 - Cr$ and $\text{rank}\big(J_h(z)\big) = D - d$ for any $z \in B_D(x,\epsilon)$. This means the rank of h at z equals $D - d$ for any $z \in B_D(x,\epsilon)$, and thus $h^{-1}(\mathbf{0})$ is a d-dimensional submanifold of $B_D(x,\epsilon) \subset \mathbb{R}^D$. The equivalence between $h^{-1}(\mathbf{0})$ and $f^{-1}(\mathbf{0})$ in $B_D(x,\epsilon)$ guarantees that $f^{-1}(\mathbf{0})$ is also a d-dimensional submanifold of $B_D(x,\epsilon) \subset \mathbb{R}^D$.

The above proof is based on Proposition 20, Corollary 18 and Theorem 19, where Proposition 20 is proved based on Theorem 15 and Proposition 12(ii), and Corollary 18 is proved based on Theorem 17. Noting that Theorem 15, Theorem 17 and Theorem 19 are valid when Lemma 13 and Theorem 11 hold, we obtain

$$\begin{aligned}
&\mathbb{P}\big(\mathcal{M}_{\text{out}} \cap B_D(x,\epsilon) \text{ is a } d\text{−dimensional manifold}\big)\\
&\geq \mathbb{P}\big(\text{Proposition 12(ii), Lemma 13 and Theorem 11 hold for } x\big)\\
&\geq \delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big).
\end{aligned}$$

■

Proof of Theorem 5 For any fixed $x \in \mathcal{M}_{\text{out}}$, we let $V_x \in \mathbb{R}^{D\times(D-d)}$ denote the orthonormal matrix such that $\Psi_x^\alpha = V_xV_x^T$, and let U_x denote the orthogonal complement of V_x. Then, we define

$$F(z) = f(z) + U_xU_x^Tz.$$

Let x^* be the projection of x onto $\mathcal{M}$, as done previously, $\Pi_{x^*} = V_*V_*^T$, and U_* be the orthogonal complement of V_*. The difference $\|F(x^*) - F(x)\|_2$ can be evaluated as

$$\begin{aligned}
&\|F(x^*) - F(x)\|_2\\
=&\|f(x^*) + U_xU_x^Tx^* - f(x) - U_xU_x^Tx\|_2\\
=&\|f(x^*) + U_xU_x^Tx^* - U_xU_x^Tx\|_2\\
\leq&\|f(x^*)\|_2 + \|(U_xU_x^T - U_*U_*^T)(x-x^*)\|_2 + \|U_*U_*^T(x-x^*)\|_2\\
=&\|f(x^*)\|_2 + \|(\Psi_x^\alpha - \Pi_{x^*})(x-x^*)\|_2 + \|U_*U_*^T(x-x^*)\|_2\\
\leq&\|f(x^*)\|_2 + \|\Psi_x^\alpha - \Pi_{x^*}\|_F\|x-x^*\|_2 + \|U_*U_*^T(x-x^*)\|_2\\
\leq&Cr^2
\end{aligned}$$

The second equality holds because $f(x) = \mathbf{0}$ for $x \in \mathcal{M}_{\text{out}}$ while the last inequality holds because $\|f(x^*)\|_2 \leq Cr^2$ via Theorem 16, $\|\Psi_x^\alpha - \Pi_{x^*}\|_F \leq Cr$ via Theorem 15, $\|x - x^*\| = d(x,\mathcal{M}) \leq cr$ via the definition of $\mathcal{M}_{\text{out}}$, and $U_*U_*^Tx = U_*U_*^Tx^*$, since x^* is the projection of x onto $T_{x^*}\mathcal{M}$.

The Jacobian matrix of F at $z = x$, denoted by $J_F(x)$ for simplicity, is

$$J_F(x) = J_f(x) + U_xU_x^T = I_D + \big(J_f(x) - \Psi_x^\alpha\big).$$

By Corollary 18, each entry of the matrix $J_f(x) - \Psi_x^\alpha$ is bounded above by Cr. Hence, we obtain

$$J_F(x) = I_D + O(r),$$

which means that $J_F(x)$ approximates I_D with precision $O(r)$ and $J_F(x)$ is invertible. Moreover, $\|J_F(x)\|_F \leq C(1+r)$ and its inversion is $\|J_F^{-1}(x)\|_F \leq C(1+r)$.

The changing rate of J_F can also be bounded as follows: supposing x' and x'' are two arbitrary points, we have

$$\|J_F(x') - J_F(x'')\|_F = \|J_f(x') - J_f(x'')\|_F \leq C\|x' - x''\|_2 \tag{4.2}$$

by the upper bound on the second derivative of $f(x)$ in Theorem 19.

Based on the conclusions that $\|F(x)-F(x^*)\|_2 \leq Cr^2$, $J_F(x) = I_D+O(r)$, and $\|J_F(x')-J_F(x'')\|_F \leq C\|x'-x''\|_2$, we could bound $\|x-x^*\|_2$ via Theorem 2.9.4 (the inverse function theorem) in Hubbard and Hubbard (2001). Specifically,

$$\|x - x^*\|_2 \leq Cr^2.$$

The above proof is based on Theorem 15, Theorem 16, Corollary 18 and Theorem 19, which are valid when Lemma 13 and Theorem 11 hold. Hence, the conclusion $\|x - x^*\|_2 \leq Cr^2$ is drawn with probability $\delta_0(1-\delta)^2$.

■

To prove the smoothness of the estimated manifold $\mathcal{M}_{\text{out}}$, as stated in Theorem 7, we construct the following two auxiliary functions and clarify their properties.

Proposition 21. *For any fixed point $x \in \mathcal{M}_{\text{out}}$, set W_x to be the basis of the spanning space of $J_f(x)^T$ and*

$$g(z) = W_x^T f(z). \tag{4.3}$$

The following two statements

(i) g is a function from $\mathbb{R}^D$ to $\mathbb{R}^{D\times(D-d)}$

(ii) Given $z \in B_D(x, r\tau)$, $g(z) = \mathbf{0}$ if and only if $f(z) = \mathbf{0}$

hold simultaneously with probability at least $\delta_0^2(1-\delta)^4\big(1-(1-cr^d)^N\big)$.

Proposition 21 claims that $f^{-1}(\mathbf{0})$ and $g^{-1}(\mathbf{0})$ describe the same set in the neighborhood of x whose proof can be found in Appendix A.6. By W_x, we reset the coordinate system for g. Specifically, the rows of $J_f(x)$ are orthogonal to the contour surface at x, and W_x is also the basis of the normal space of $\mathcal{M}_{\text{out}}$ at x. Accordingly, we set the first d coordinates as the basis of $T_x\mathcal{M}_{\text{out}}$ and the last $D-d$ coordinates as the columns of W_x. In this coordinate system, we define an implicit function $\phi : \mathbb{R}^d \to \mathbb{R}^{D-d}$ based on $g(\cdot)$ using the implicit function theorem, such that $\big(\zeta; \phi(\zeta)\big)$ maps $\zeta \in T_x\mathcal{M}_{\text{out}}$ to a point on the manifold $\mathcal{M}_{\text{out}}$. Here, we let $(\eta; \zeta)$ denote the concatenation of column vectors η and ζ. The upper bound on the first and second derivatives of ϕ is given in Lemma 22, with its proof appearing in Appendix A.6.

Lemma 22. *Suppose function g is defined as (4.3). The implicit function $\phi : \mathbb{R}^d \to \mathbb{R}^{D-d}$ satisfying $g\big(\cdot, \phi(\cdot)\big) = \mathbf{0}$ exists, and its first and second derivatives are bounded above by*

$$\partial_s\phi(\zeta) \leq C\|\big(\zeta; \phi(\zeta)\big) - x\|_2, \quad \partial_t\partial_s\phi(\zeta) \leq C,$$

with probability at least $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$, for any $\|s\|_2 = \|t\|_2 = 1$.

Proof of Theorem 7 Let x and z be two points on $\mathcal{M}_{\text{out}}$, and $T_x\mathcal{M}_{\text{out}}$ be the tangent space to $\mathcal{M}_{\text{out}}$ at x. The proof is conducted with $\|z-x\|_2 > r\tau$ and $\|z-x\|_2 \leq r\tau$, respectively. First, when $\|z-x\|_2 > r\tau$,

$$\frac{\|z-x\|_2^2}{d(z, T_x\mathcal{M}_{\text{out}})} \geq r\tau \tag{4.4}$$

holds because $\|z-x\|_2 \geq d(z, T_x\mathcal{M}_{\text{out}})$. Second, when $\|z-x\|_2 \leq r\tau$, we have $g(z) = f(z) = g(x) = f(x) = \mathbf{0}$ by Proposition 21 with probability $\delta_0^2(1-\delta)^4\big(1-(1-cr^d)^N\big)$, since x and z are on $\mathcal{M}_{\text{out}}$. Let ζ_x and ζ_z denote the first d coordinates of x and z, respectively. We have $z = \big(\zeta_z; \phi(\zeta_z)\big)$, $x = \big(\zeta_x; \phi(\zeta_x)\big)$, $\partial_s\phi(\zeta_z) \leq C\|z-x\|_2$ and $\partial_t\partial_s\phi(\zeta_z) \leq C$ with probability at least $\delta_0(1-\delta)^2$ by Lemma 22. So,

$$\begin{aligned} d(z, T_x\mathcal{M}_{\text{out}}) &= \|\phi(\zeta_z) - \phi(\zeta_x)\|_2 \\ &\leq C\|z-x\|_2\|\zeta_z - \zeta_x\|_2 + C\|\zeta_z - \zeta_x\|_2^2 \\ &\leq C\|z-x\|_2^2 + C\|z-x\|_2^2 \leq C\|z-x\|_2^2. \end{aligned}$$

As a result,

$$\frac{\|z-x\|_2^2}{d(z, T_x\mathcal{M}_{\text{out}})} \geq \frac{\|z-x\|_2^2}{C\|z-x\|_2^2} = \frac{1}{C} := c_0.$$

Combined with (4.4), we complete this proof.

The above proof requires Proposition 21 and Lemma 22, which are valid when Theorem 4 and Theorem 5 hold simultaneously for x and when Theorem 15 holds for z. Given the dependency between the theorems as shown in Figure 3, we establish that the above theorems hold when Lemma 13 and Theorem 11 simultaneously hold for x, z, and when Proposition 12(ii) holds for x. Hence, we have

$$\begin{aligned} &\mathbb{P}\Big(\frac{\|z-x\|_2^2}{d(z, T_x\mathcal{M}_{\text{out}})} \geq c_0 r\Big) \\ &\geq \mathbb{P}\Big(\big(\text{Proposition 12(ii) holds for } x\big) \cap \big(\text{Lemma 13 and Theorem 11 hold for } x, z\big)\Big) \\ &\geq \delta_0^2(1-\delta)^4\big(1-(1-cr^d)^N\big). \end{aligned}$$

■

5. Experimental Results

This section comprises two parts. The first part provides numerical comparisons of the methods of Mohammed and Narayanan (2017), Fefferman et al. (2018), and Aizenbud and Sober (2021). Further, we apply relevant methods on several known manifolds, illustrate the output manifolds, and calculate the Hausdorff distances between the output and latent manifolds. In the second part, we focus on real applications, and use our method to denoise facial images sampled from a lengthy video recording. The results of our method are then compared to the findings of each of the other aforementioned methods.

Algorithm 1: Project x onto $\mathcal{M}_{\text{out}}$

```
Input: a point $x$, noisy data $X = [x_1, \cdots, x_N]$, bandwidth parameters $r$ and $r'$, a
 step length parameter $a$, a tolerance $\epsilon$, and the maximal number of iteration $T$.
Output: projection $\tilde{x}$ of $x$ onto $\mathcal{M}_{\text{out}}$.
  1. Calculate $P_{x_i} = I - VV^T$ for each $x_i \in X$, where $V$ is the $D \times d$ matrix whose
     columns are the eigenvectors corresponding to the largest $d$ eigenvectors of
     $\sum_{j \in I_{x_i,r'}} (x_j - x_i)(x_j - x_i)^T$.
  2. Set $t = 1$.
      (1). Calculate $\tilde{\alpha}_i(x)$ and $\alpha_i(x)$ for $i \in I_{x,r}$ by (2.5).
      (2). Plug $\{\tilde{\alpha}_i(x), \alpha_i(x), P_{x_i}\}_{i \in I_{x,r}}$ into (B.1) to obtain the gradient grad$(x)$ of
           $\|f(x)\|_2^2$.
      (3). Update $t = t + 1$ and $x = x - a \cdot \text{grad}(x)$.
      (4). Repeat (1) to (3) until the tolerance condition $\|f(x)\|_2^2 \le \epsilon$ or the maximal
           iteration $T$ is met.
  3. Output $\tilde{x} = x$.
```

Implementation: the MATLAB codes, together with all numerical examples used in this paper, are available at `https://zhigang-yao.github.io/research.html` which contains a GitHub link under the code tap. We have also implemented the related methods from Mohammed and Narayanan (2017) and Fefferman et al. (2018), since the authors of both papers have not provided implementation due to the nature of their work having been purely abstract.

5.1 Simulation

As explained in Subsection 1.2, by removing the unreliable discs which centered at the sample points as in Mohammed and Narayanan (2017) and Fefferman et al. (2018), one would expect an improved performance compared to these two methods. Assuming the data points are sampled from a tubular neighborhood, Aizenbud and Sober (2021) denoises the sample points iteratively using a local polynomial regression. As the degree increases, polynomial regression fits a manifold better when the noise is limited but on the other hand a polynomial regression exhibits sensitivity once noise increases. As a method designed for Gaussian noise, our method is expected to be more robust as noise increases. To support this claim, we test methods in Mohammed and Narayanan (2017) (marked by km17), Fefferman et al. (2018) (marked by cf18), and Aizenbud and Sober (2021) with polynomial degree 1 and 2 (marked by ya21(deg=1) and ya21(deg=2)) on manifolds with both constant and inconstant curvature, namely: a circle embedded in $\mathbb{R}^2$, a sphere embedded in $\mathbb{R}^3$, and a torus embedded in $\mathbb{R}^3$. To ensure a traceable comparison, all the tests are conducted in the following way, similar to that of Mohammed and Narayanan (2017):

- Sample N points from the latent manifold, blur the points with Gaussian noise defined in (1.1) with given σ, and use the noisy data $X = [x_1, \cdots, x_N]$ to implicitly construct output manifolds.

- Initialize a collection of points $P = [p_1, \cdots, p_{N_0}]$ around the latent manifold.

- Project each p_i to the constructed output manifolds via km17, cf18, ya21(deg=1), ya21(deg=2)) and our method, respectively. We will then obtain $\tilde{P}$ as the projection of P for each method.

- Calculate the Hausdorff distance between each $\tilde{P}$ and $\mathcal{M}$ to estimate the Hausdorff distance between the corresponding $\mathcal{M}_{\text{out}}$ and $\mathcal{M}$.

As projections, points in $\tilde{P}$ lie on the corresponding $\mathcal{M}_{\text{out}}$, and the Hausdorff distance $H(\tilde{P}, \mathcal{M})$ could estimate $H(\mathcal{M}_{\text{out}}, \mathcal{M})$ when $\tilde{P}$ are dense enough. This motivates us to evaluate the approximation error of $\mathcal{M}_{\text{out}}$ to $\mathcal{M}$ by $H(\tilde{P}, \mathcal{M})$. To project a point p onto a manifold defined by (2.7), we design algorithm 1. Taking $x = p$ and f in algorithm 1 as (2.6), we could project p onto our output manifold. It should be noted that the difficulty of calculating such a gradient lies in calculating a gradient of orthogonal projection, which can be addressed, according to Shapiro and Fan (1995). Detailed formula refers to Appendix B. Mohammed and Narayanan (2017) suggested a subspace-constrained gradient descent algorithm to project a point onto $\mathcal{M}_{\text{out}}$ constructed by km17. Thus, we adopt this algorithm to implement km17 in this simulation. Although Fefferman et al. (2018) did not consider the issue, we nevertheless implement their method too via algorithm 1, treating $f(x)$ as the approximated bias at x defined by Fefferman et al. (2018).

The details of this simulation are as follows: we uniformly sample N points denoted by $y_1, \cdots, y_N$ from each target manifold and i.i.d. sample $\xi_1, \cdots, \xi_N$ from a Gaussian distribution (1.1) with a given standard derivation σ. Then, the noisy data $X = \{x_i\}_{i=1}^N$ is constructed by $x_i = y_i + \xi_i$. The initial points P are sampled from the tube centered at $\mathcal{M}$ with radius $\frac{1}{2}\sqrt{\frac{\sigma}{D}}$, so that $d(p_i, \mathcal{M}) \le \sqrt{\sigma}$ for each p_i. According to Theorem 5, $d(\tilde{p}_i, \mathcal{M}) \le O(\sigma)$, which means the output points should be much closer to the latent manifold than the initial points. Again, we take $N_0 = N$ initial points for each test in the simulation.

To implicitly construct the output manifolds, the methods–km17, cf18, and our method,each require a bandwidth parameter r. According to the theoretical analysis, $r = O(\sqrt{\sigma})$. So we take $r = \lambda\sqrt{\sigma}$ in this simulation, where λ is tuned in a large range for each method and each σ. All the results reported in this section are the ones using the best λ. The method ya21 also requires a bandwidth parameter h, which is again selected as the best one tuned from a large range. In constructing $\tilde{\alpha}_i(x)$, our method requires $\beta \ge 2$. We take $\beta = d + 2$ in the simulation, as Fefferman et al. (2018) did.

5.1.1 Manifold with constant curvature

This part tests the manifold fitting methods for the circle in $\mathbb{R}^2$ and the sphere embedded in $\mathbb{R}^3$. For the circle, we set $N = N_0 = 300$, while for the sphere, we set $N = N_0 = 1000$. The different sample-size settings guarantee comparable density in each case, as Figure 5 illustrates that the $\tilde{P}$ (black dots) and their projection onto $\mathcal{M}$ (red dots) obtained by our

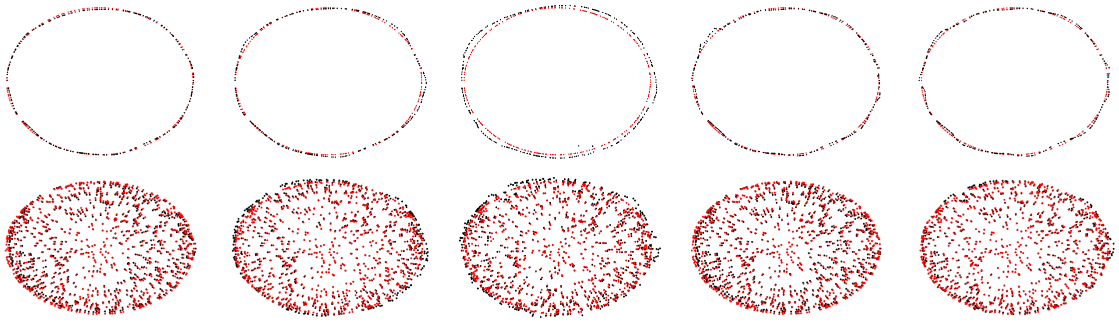

Figure 5: The performance of our method, km17, cf18, ya21(deg=1) and ya21(deg=2) when fitting a circle (top row) and a sphere (bottom row), where black points represent points in $\tilde{P}$(black dots) and red points represents their projections onto $\mathcal{M}$.

method, cf18, cf18, ya21(deg=1) and ya21(deg=2), from left to right. The black dots and red dots can be treated as the discretized versions of $\mathcal{M}_{\text{out}}$ and $\mathcal{M}$, respectively. Thus, a larger overlap of the two sets of dots means the manifold is better fitted. For the circle embedded in $\mathbb{R}^2$, we show the entire space in the left column, while for the sphere embedded in $\mathbb{R}^3$, we show the view from the positive z axis. Figure 5 shows that km17 clearly performs worse than the other methods in terms of fitting error. From the two estimated circles by ya21(deg=1) and ya21(deg=2), we observe that there are sharp corners – both at the top left and at the bottom right – an observation that confirms that the estimator by ya21 is not smooth. From the right edge of the circle and the sphere, we can also observe that our method preforms slightly better than cf18 in this experiment.

To confirm the superiority of our method, we repeat each test for 20 trials, and list the results of $H(\tilde{P}, \mathcal{M})$ using the different methods in Figure 6. Generally speaking, our method outperforms cf18, km17 and ya21(deg=1) in the compared cases and although ya21(deg=2) performs slightly better than our method in instances of very low noise, it is much more sensitive than our method. As the σ increases, ya21(deg=2) fails to outperform other methods. From Figure 6, $H(\mathcal{M}, \mathcal{M}_{\text{out}}) = O(\sigma)$ for our method, which supports Theorem 5.

5.1.2 Manifold with inconstant curvature

We also implement the compared methods in the torus case, which is a type of manifold with inconstant curvature. Figure 7 illustrates the case with $N = N_0 = 800$ and $\sigma = 0.04$, and the torus embedded in $\mathbb{R}^3$ is shown from the positive z axis. Here, the sample points in $\tilde{P}$ are marked by black dots and their projection onto $\mathcal{M}$ are marked by red dots. The five subfigures are obtained from our method, cf18, km17, ya21(deg=1) and ya21(deg=2), from left to right. From the top and right edges of the torus, we can observe that our method performs better than both cf18 and km17. From the fourth subfigure, we can identify a clear gap between the red and black dots around the edge of the torus, which means ya21(deg=1) failed to fit these points but using a second degree polynomial, ya21(deg=2) achieves a better fitting as the right subfigure shows.

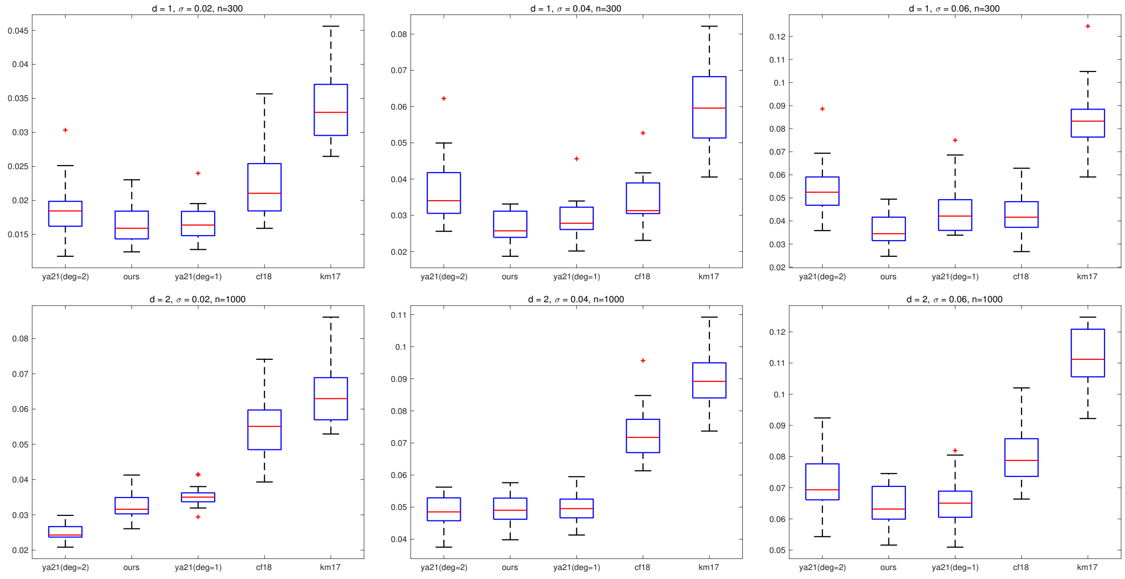

Figure 6: The Hausdorff distance of fitting a circle (top row) and a sphere (bottom row) with $\sigma = 0.02$ (left column), $\sigma = 0.04$ (middle column) and $\sigma = 0.06$ (right column) using ya21(deg=2), our method, ya21(deg=1), cf18 and km17 respectively.

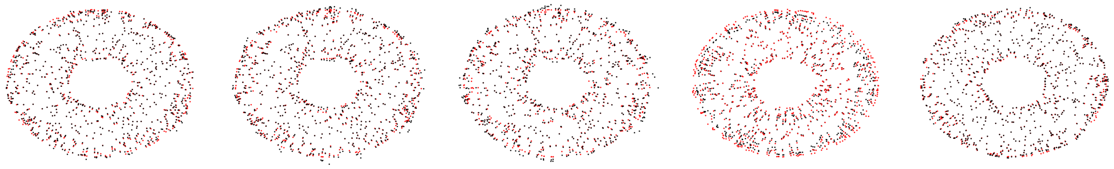

Figure 7: The performance of our method, km17, cf18, ya21(deg=1) and ya21(deg=2) when fitting a torus with $N = N_0 = 800$ and $\sigma = 0.04$, where black points represent points in $\tilde{P}$(black dots) and red points represent their projections onto $\mathcal{M}$.

We also repeat each test for 20 trials and list the results of $H(\tilde{P}, \mathcal{M})$ using the different methods shown in Figure 8. When $\sigma = 0.02$ and $\sigma = 0.04$, our method performs better than cf18, km17 and ya21(deg=1) but as σ increases to 0.06, the fitting problem becomes more difficult and the performance of km17, cf18 and our method are similar, which further demonstrates the sensitivity of ya21(deg=2). When σ is small and the sample size is adequate, ya21(deg=2) outperforms the other methods but when the sample size decreases and σ increases, the performance of ya21(deg=2) deteriorates rapidly.

5.2 Facial image denoising

This subsection considers a concrete case - denoising facial images selected from the video database in Happy et al. (2012). We select 1,000 images of an individual turning his head

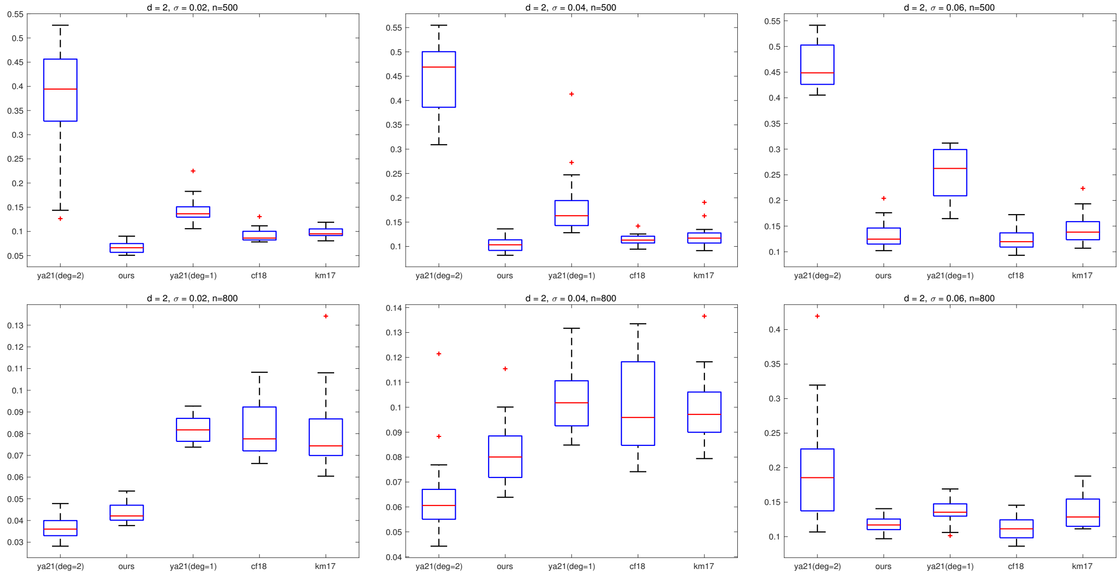

Figure 8: The Hausdorff distance of fitting a torus given 500 (top row) and 1000 (bottom row) samples with $\sigma = 0.02$ (left column), $\sigma = 0.04$ (middle column) and $\sigma = 0.06$ (right column) using ya21(deg=2), our method, ya21(deg=1), cf18 and km17 respectively.

around, then blurring those images via a Gaussian distribution with a different standard derivation σ. In this experiment, σ is set to be the average of all pixels in 1,000 images multiplied by $\rho = 0.2, 0.3$, or 0.4. The size of each facial image is 80×80, which means $D = 6400$. The dimension d of the latent manifold is tuned from $\{1, 5, 10, 15, 20, 50, 75, 100\}$ for each method and we choose $d = 10$ because of its outperformance.

From the 1,000 facial images, we select 5 with different head orientations. The top row of Figure 9 displays these five original images, while the second row of Figure 9 shows these five images blurred, with $\rho = 0.3$. The goal of this experiment is to denoise these five blurred images by projecting them to the manifold learnt by the remaining 995 blurred images, which are treated as the noisy samples. To achieve the denoising, we use km17, cf18, ya21(deg=1), ya21(deg=2) and our method to construct the output manifold with the 995 noisy samples, and project the five tested images to each output manifold. When the output manifold correctly fits the latent one, projecting blurred images to the output manifold denoises these facial images. In this experiment, we take $\beta = 2$ for our method to construct $\tilde{\alpha}_i(x)$. If cf18 uses $\tilde{\alpha}_i(x)$ as Fefferman et al. (2018) has suggested, it would not work quite satisfactorily, because of the over-large power $d + 2$ rather than β. Therefore, we take the same $\tilde{\alpha}_i(x)$ for cf18 and our method to make the results comparable.

The last three rows of Figure 9 show the denoised images obtained by km17, cf18, ya21(deg=1), ya21(deg=2) and our method, respectively. The first and third facial images were not recovered by km17. Although the faces in the other three images obtained by km17 can be distinguished, they are still very noisy. Cf18 could not recover the third image

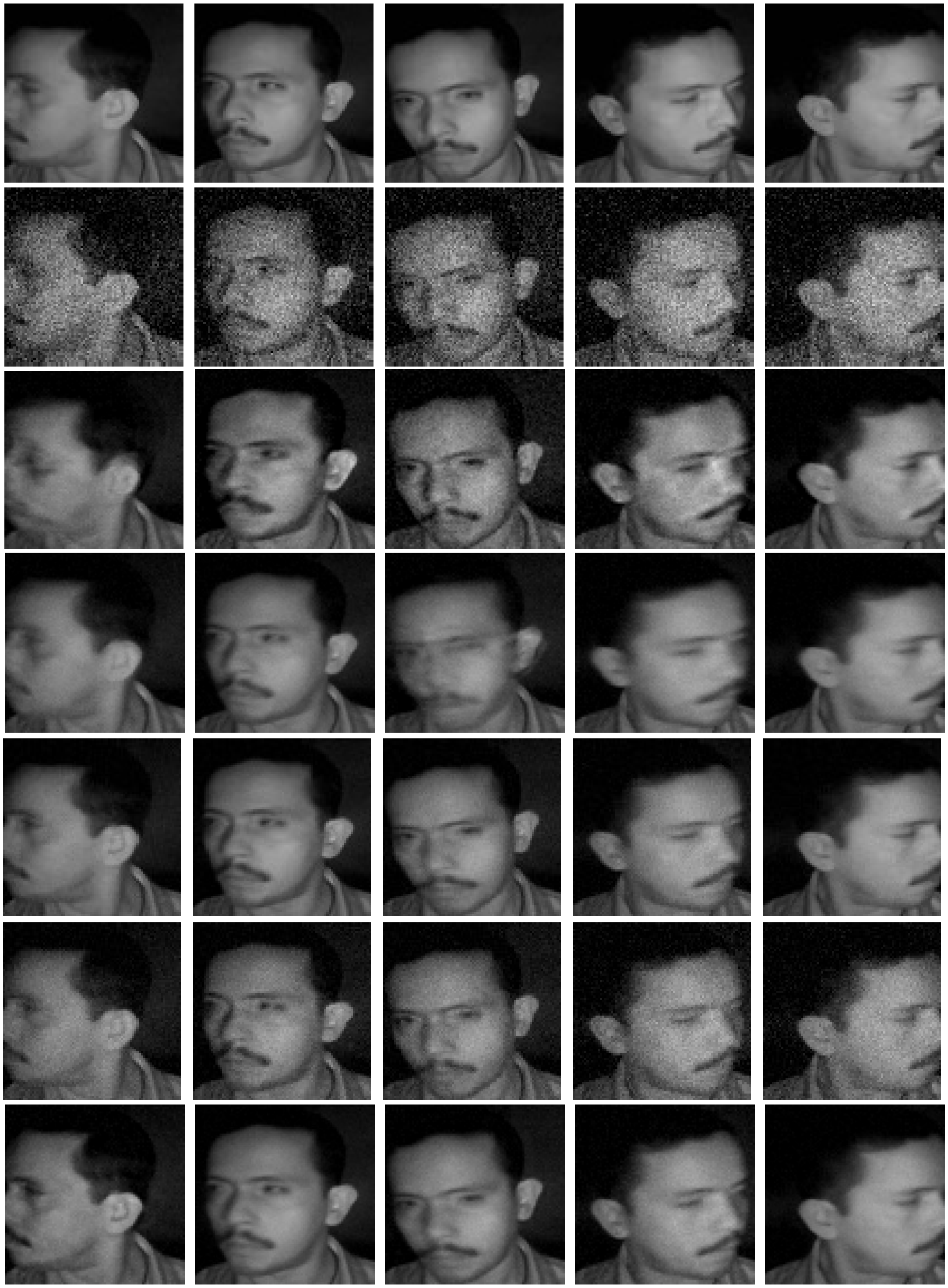

Figure 9: Performance of facial image denoising with $\rho = 0.3$. The first row consists of original images while the second row consists of blurred images. The third to seventh rows contain deblurred images using km17, cf18, ya21(deg=1), ya21(deg=2) and our method, respectively.

either, although the other four images obtained by cf18 are better than the ones obtained by km17. Both ya21(deg=1) and ya21(deg=2) can recover these five faces. However, the faces obtained by ya21(deg=2) are still somewhat fuzzy, compared with the ones obtained by ya21(deg=1) and our method. Our method recovered all the five faces, with the third face of much better quality than the faces from km17 and cf18.

The results with the settings $\rho = 0.2$ or $\rho = 0.4$ are listed in Figure 10 and Figure 11 (Appendix C). When we take $\rho = 0.2$, the results of all three methods provide fairly good results. However, the results from km17 are somewhat noisy, with the obtained faces darker than the original ones. When $\rho = 0.4$, km17 is barely able to recover the faces, cf18 fails at the first and third ones, but our method can still provide acceptable faces.

6. Discussion

We have proposed a new output manifold $\mathcal{M}_{\text{out}}$ to fit data collection with Gaussian noise. The theoretical analysis of $\mathcal{M}_{\text{out}}$ has two main components: (1) the upper bound on $d(x, \mathcal{M})$ for arbitrary x, which guarantees $\mathcal{M}_{\text{out}}$ approximates $\mathcal{M}$ well, and (2) the upper bound on the second-order difference of $\mathcal{M}_{\text{out}}$, which guarantees the smoothness of $\mathcal{M}_{\text{out}}$.

To demonstrate the contribution of this paper, we compared our theoretical results to relevant works presented in Mohammed and Narayanan (2017) and Fefferman et al. (2018). All three of these works aim to fit data collection by a smooth manifold, while the difference among these works lies in the assumptions about noise. Mohammed and Narayanan (2017) requires the data to be noiseless, which is the most strict assumption of the three. As mentioned in the Introduction, Fefferman et al. (2018) essentially requires the noise of data to be bounded, that is, the data collection X satisfying $H(X, \mathcal{M}) \leq O(r^2)$, where $H(\cdot, \cdot)$ denotes the Hausdorff distance. If the noise of data obeys a Gaussian distribution, the researchers would select a subset from the entire dataset, assume the noise of the subset is bounded, and implement their proof on this subset of data. However, their sample selection step imposes a lower bound on r, meaning that the upper bound of $H(\mathcal{M}, \mathcal{M}_{\text{out}})$ cannot tend to 0. This paper, therefore, proposes a method to address the problem of Gaussian noise, which is commonly assumed but unsolved in relevant works. Unlike the bounded noise, X with Gaussian noise are not required to satisfy $H(X, \mathcal{M}) \leq O(r^2)$, which increases the difficulty of manifold fitting.

According to the discussion in Subsection 1.2 and the experiment results, our method could achieve a smaller approximating error than the methods presented in Mohammed and Narayanan (2017) and Fefferman et al. (2018). One possible reason is that we use the weighted average $\sum_{i \in I_{x,r}} \alpha_i(x) P_{x_i}$ to estimate Π_{x^*} rather than using each P_{x_i} separately. To explain this claim, we consider the following expression:

$$\sum_{i \in I_{x,r}} \alpha_i(x) P_{x_i} - \Pi_{x^*} = \sum_{i \in I_{x,r}} \alpha_i(x)(P_{x_i} - \Pi_{x_i^*}) + \Big(\sum_{i \in I_{x,r}} \alpha_i(x) \Pi_{x_i^*} - \Pi_{x^*} \Big). \tag{6.1}$$

For certain “symmetric” manifolds, the second term in the right hand side of (6.1) might be much closer to zero matrix than $(\Pi_{x_i^*} - \Pi_{x^*})$.

A circle may be considered as an example. Suppose x, x_1, and x_2 are points on the circle satisfying $x_1 - x = x - x_2$; then, the average of orthogonal projections onto the normal spaces at x_1 and x_2 equals the orthogonal projection onto the normal space at x, while the projection onto the normal space at x_1 (or x_2) differs from that at x with an error in the order of $\|x - x_1\|_2$ (or $\|x - x_2\|_2$) by Lemma 14.

This phenomenon illustrates that the average of $\{P_{x_i}\}_{i \in I_{x,r}}$ approximates Ψ_x^α better than each P_{x_i} for certain manifolds. We benefit from this fact by using $\sum_{i \in I_{x,r}} \alpha_i(x) P_{x_i}$ to construct our output manifold, while Mohammed and Narayanan (2017) and Fefferman et al. (2018) use each P_{x_i} separately instead. Characterizing the "symmetric" property mentioned above and using this property in the methodology of manifold fitting is an attractive and promising topic, and our work on it will continue.

Acknowledgments

ZY and YX were supported by the MOE Tier 1 A-0004809-00-00 and Tier 2 R-155-000-184-112 at the National University of Singapore. ZY is also supported by Tier 1 A-8002931-00-00 and A-8001562-00-00. ZY thanks Professor Charles Fefferman and Professor Hariharan Narayanan for their helpful discussions on some details of Mohammed and Narayanan (2017) and Fefferman et al. (2018) which we found very useful. ZY thanks Professor Shing-Tung Yau for his sagacious comments and the support from the Center of Mathematical Sciences and Applications at Harvard University.

Appendix A. Proofs

A.1 Proof of Proposition 3

Lemma 23. *If $d(x,\mathcal{M}) \leq cr$ with some $c < 1$ and c_1 satisfies $c < c_1 \leq 1$, then there exists a constant c' such that $\mathbb{P}(i \in I_{x,c_1 r}) \geq c' r^d$.*

Proof Setting c_2 be a constant satisfying $c < c_2 < c_1$, then

$$\begin{aligned}\mathbb{P}(i \in I_{x,c_1 r}) &\geq \mathbb{P}\big(y_i \in \mathcal{M} \cap B_D(x, c_2 r), \|\xi_i\|_2 \leq (c_1 - c_2)r\big)\\ &= \mathbb{P}\big(y_i \in \mathcal{M} \cap B_D(x, c_2 r)\big)\mathbb{P}\big(\|\xi_i\|_2 \leq (c_1 - c_2)r\big).\end{aligned}$$

In order to bound $\mathbb{P}(i \in I_{x,c_1 r})$ below, we bound the two probability expressions $\mathbb{P}(y_i \in \mathcal{M} \cap B_D(x, c_2 r))$ and $\mathbb{P}(\|\xi_i\|_2 \leq (c_1 - c_2)r)$, respectively. Since $d(x,\mathcal{M}) \leq cr < c_2 r$, there exists c_3 such that

$$\mathbb{P}\big(y_i \in \mathcal{M} \cap B_D(x, c_2 r)\big) = \frac{\mathrm{Vol}\big(\mathcal{M} \cap B_D(x, c_2 r)\big)}{\mathrm{Vol}(\mathcal{M})} = c_3 r^d.$$

By the assumptions in (2.1), there exists $0 < c_4 < 1$ satisfying $r \geq c_4\sqrt{\sigma} \geq c_4 \sigma$, which leads to $r/\sigma \geq c_4$ and

$$\mathbb{P}\big(\|\xi_i\|_2 \leq (c_1 - c_2)r\big) = \mathbb{P}\left(\frac{\|\xi_i\|_2^2}{\sigma^2} \leq \frac{(c_1 - c_2)^2 r^2}{\sigma^2}\right) \geq \mathbb{P}\left(\frac{\|\xi_i\|_2^2}{\sigma^2} \leq (c_1 - c_2)^2 c_4^2\right) := c_5.$$

Since $\|\xi_i\|_2^2/\sigma^2$ obeys Chi-square distribution and the constant $(c_1 - c_2)^2 c_4^2 > 0$, c_5 is positive. Calculating the product of $\mathbb{P}\big(y_i \in \mathcal{M} \cap B_D(x, c_2 r)\big)$ and $\mathbb{P}\big(\|\xi_i\|_2 \leq (c_1 - c_2)r\big)$, we have

$$\mathbb{P}(i \in I_{x,c_1 r}) \geq \mathbb{P}\big(y_i \in \mathcal{M} \cap B_D(x, c_2 r)\big)\mathbb{P}\big(\|\xi_i\|_2 \leq (c_1 - c_2)r\big) \geq c_3 c_5 r^d = c' r^d$$

which, completes this proof. ■

Proof of Proposition 3 Setting $c_1 = 1$ in Lemma 23, we obtain $\mathbb{P}(i \in I_{x,r}) \geq c' r^d$. Hence, whether $i \in I_{x,r}$ or not, can be treated as a Bernoulli distribution with the expectation of $c' r^d$. Applying the Berry-Esseen theorem to the N Bernoulli trials, there exists $c' < 1$ such that $|I_{x,r}| \geq c' r^d N$ with probability $1 - C/\sqrt{N}$. ■

A.2 Proof of Lemma 9 and Lemma 10

The following proof is derived from the notations illustrated in Figure 4 and the settings $\sigma < 1$, $r' = 2r$ and $r = O(\sqrt{\sigma})$, which imply that there exist constants C and C' independent on σ such that $r < C$ and $r' < C'$ by (2.2).

Proof of Lemma 9 Let $p_i - z^* = q_i$; then, $p_i' - z = p_i - z^* = q_i$. Considering $z_i - z = z_i - p_i' + p_i' - z = z_i - p_i' + q_i := \delta_i + q_i$, we can rewrite $\sum_{i \in I_{z,r'}}(z_i - z)(z_i - z)^T - \sum_{i \in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T$ as

$$\sum_{i \in I_{z,r'}} q_i \delta_i^T + \delta_i q_i^T + \delta_i \delta_i^T. \tag{A.1}$$

To begin with, we bound $\|\delta_i\|_2$. Recalling that the projection onto the normal space at z^* is Π_{z^*},

$$\begin{aligned}\|\delta_i\|_2 = \|\Pi_{z^*}(z_i - z)\|_2 &\leq \|\Pi_{z^*}\big((z_i - z_i^*) + (z_i^* - z^*) + (z^* - z)\big)\|_2 \\ &\leq \|\Pi_{z^*}(z_i^* - z^*)\|_2 + \|z_i - z_i^*\|_2 + \|z^* - z\|_2 \\ &\leq \frac{\|z_i^* - z^*\|_2^2}{\tau} + \|z_i - z_i^*\|_2 + \|z - z^*\|_2 \\ &\leq \frac{(\|z_i^* - z_i\|_2 + \|z_i - z\|_2 + \|z - z^*\|_2)^2}{\tau} + \|z_i - z_i^*\|_2 + \|z - z^*\|_2.\end{aligned}$$

The last but one inequality holds in accordance with Proposition 2. As established previously, each z_i is generated as $y_i + \xi_i$ with $y_i \in \mathcal{M}$ and $\xi_i \sim N(0, \sigma^2 I_D)$. Then, $\|\xi_i\|_2 = \|z_i - y_i\|_2 \geq \|z_i - z_i^*\|_2$ since z_i^* is the projection of z_i onto $\mathcal{M}$. Thus, $\|\delta_i\|_2$ can be bounded by

$$\|\delta_i\|_2 \leq \frac{(\|\xi_i\|_2 + r' + \|z - z^*\|_2)^2}{\tau} + \|\xi_i\|_2 + \|z - z^*\|_2 \leq C_1\big(\|\xi_i\|_2^2 + \|\xi_i\|_2 + r'^2 + \|z - z^*\|_2\big).$$

The last inequality is achieved by replacing certain $\|z - z^*\|_2$ by its upper bound r' and replacing certain r' by a constant independent on σ , since $r' < C'$ by $r' = 2r$ and (2.2). Considering the average over $I_{z,r'}$, we obtain

$$\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}} \|\delta_i\|_2 \leq C_1\big(\psi_2 + \psi_1 + r^2 + \|z - z^*\|_2\big),$$

and

$$\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}} \|\delta_i\|_2^2 \leq C_2\big(\psi_4 + \psi_3 + \psi_2 + r\psi_1 + r^4 + r^2\|z - z^*\|_2 + \|z - z^*\|_2^2\big).$$

where $\psi_k := \frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}} \|\xi_i\|_2^k$, the above bounds are then plugged into the bound of (A.1) as follows:

$$\begin{aligned}&\|\frac{1}{|I_{z,r'}|}\Big(\sum_{i\in I_{z,r'}}(z_i - z)(z_i - z)^T - \sum_{i\in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T\Big)\|_F \\ \leq&\|\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}\big(q_i\delta_i^T + \delta_i q_i^T + \delta_i\delta_i^T\big)\|_F \\ \leq&\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}\big(2\|q_i\|_2\|\delta_i\|_2 + \|\delta_i\|_2^2\big) \\ \leq&\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}\big(2r'\|\delta_i\|_2 + \|\delta_i\|_2^2\big) \\ \leq&C\big(\psi_4 + \psi_3 + \psi_2 + r'\psi_1 + r'^3 + r'\|z - z^*\|_2 + \|z - z^*\|_2^2\big)\end{aligned}$$

The last but one inequality holds since $\|q_i\|_2 \leq \|z_i - z\|_2 \leq r'$. Replacing ψ_k by corresponding summation completes the proof. ■

Lemma 24 (Theorem 21 in Mohammed and Narayanan (2017))**.** *Let $\Lambda_1, \cdots, \Lambda_k$ be i.i.d. random positive semidefinite $D \times D$ matrices with expected value $\mathbb{E}[\Lambda_i] = M \succeq \mu I$ and $\Lambda_i \preceq I$. Then for all $\epsilon \in [0, 1/2]$,*

$$\mathbb{P}\Big[\frac{1}{k}\sum_{i=1}^{k}\Lambda_i \notin [(1-\epsilon)M, (1+\epsilon)M]\Big] \leq 2D\exp\big\{\frac{-\epsilon^2\mu k}{2\ln 2}\big\}.$$

Here, the matrix interval $A \in [B, C]$ means $a_{ij} \in [b_{ij}, c_{ij}]$ holds for any i, j and the matrix ordering $A \succeq B$ means $A - B$ is a positive semidefinite.

Proof of Lemma 10 Before the proof of Lemma 10, we provide the useful notations and contents. For convenience, z^* is set to be the origin of the local coordinate system, and the coordinates in $T_{z^*}\mathcal{M}$ are set to be the first d coordinates of the D coordinates. We let $\mathcal{P}_d : \mathbb{R}^D \to \mathbb{R}^D$ be an operator, setting the last $(D-d)$ entries of a vector to be zeros, that is, $\mathcal{P}_d(v) = [v_1, \cdots, v_d, 0, \cdots, 0]^T$. We also let $\bar{\mathcal{P}}_d$ be the operator, setting the first d entries of a vector to be zeros, that is, $\bar{\mathcal{P}}_d = \mathcal{I} - \mathcal{P}_d$, with $\mathcal{I}$ being the identity operator. Notations $\bar{v} := \mathcal{P}_d(v)$ and $\hat{v} = \bar{\mathcal{P}}_d(v)$ are also used without confusion.

Based on these notations, we calculate the useful bound on $\|\hat{\eta}\|_2$ for $\eta \in \mathcal{M} \cap B_D(z, r')$. Using the definition of $\bar{\eta}$, we obtain $\langle z^* - \bar{\eta}, z - z^*\rangle = 0$, $\langle z^* - \bar{\eta}, \hat{\eta}\rangle = 0$, and, therefore

$$\begin{aligned} r'^2 &\geq \|z - \eta\|_2^2 = \|(z - z^*) + (z^* - \bar{\eta}) - \hat{\eta}\|_2^2 \\ &\geq \|z - z^*\|_2^2 - 2\|z - z^*\|_2\|\hat{\eta}\|_2 + \|z^* - \bar{\eta}\|_2^2 + \|\hat{\eta}\|_2^2 \\ &= \|z - z^*\|_2^2 - 2\|z - z^*\|_2\|\hat{\eta}\|_2 + \|z^* - \eta\|_2^2. \end{aligned}$$

Moreover, in accordance with Proposition 2, $\|z^* - \eta\|_2^2 \geq 2\tau\|\hat{\eta}\|_2$. Combining these two inequalities, we obtain

$$r'^2 - \|z - z^*\|_2^2 + 2\|z - z^*\|_2\|\hat{\eta}\|_2 \geq \|z^* - \eta\|_2^2 \geq 2\tau\|\hat{\eta}\|_2$$

and, hence,

$$\|\hat{\eta}\| \leq \frac{r'^2 - \|z - z^*\|^2}{2(\tau - \|z - z^*\|)}. \tag{A.2}$$

We are now ready to prove Lemma 10. Let $\lambda_1 \geq \cdots \geq \lambda_D$ be the eigenvalues of matrix $\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T$ and $\mu_1 \cdots \geq \mu_D$ be the eigenvalues of the population covariance matrix M, that is,

$$M := \mathbb{E}[\frac{1}{|I_{z,r'}|}\sum_{i\in I_{z,r'}}(p_i - z^*)(p_i - z^*)^T].$$

We see that $\lambda_{d+1} = \cdots = \lambda_D = \mu_{d+1} = \cdots = \mu_D = 0$. Therefore, we need only a lower bound for λ_d, which can be obtained by relating its value to μ_d through a concentration inequality given in Lemma 24. Assuming the first d coordinates are aligned with the eigenvectors corresponding to the d largest eigenvalues of M, μ_d is the variance in the d-th direction. Clearly, the first d coordinates are located in $T_{z^*}\mathcal{M}$. Let $\mathbb{P}$ be the probability measure on $T_{z^*}\mathcal{M} \cap B_D(z, r')$. For any $q \in T_{z^*}\mathcal{M} \cap B_D(z, r')$, we first bound $\mathbb{P}(q)$ above.

We set $S(q) = \{\zeta' : \bar{\zeta}' = q\} \cap B_D(z, r')$, and $\hat{S}(q) = \cup_{\zeta' \in S(q)} \{\eta' : |\eta'(i) - \zeta'(i)| \leq 3\sigma, \forall i = 1, \cdots, D\}$, where $\eta(i)$ and $\zeta(i)$ represent the i-th element of η and ζ, respectively. Then, we have $\cup_{q \in T_{z^*}\mathcal{M} \cap B_D(z,r')} S(q) \subset B_D(z, r')$ and

$$\begin{aligned}\cup_{q \in T_{z^*}\mathcal{M} \cap B_D(z,r')} \hat{S}(q) &\subset \cup_{\zeta' \in B_D(z,r')} \{\eta' : |\eta'(i) - \zeta'(i)| \leq 3\sigma, \forall i = 1, \cdots, D\} \\ &\subset B_D(z, r' + 3\sigma\sqrt{D}).\end{aligned}$$

The probability at q is

$$\begin{aligned}\mathbb{P}(q) &= \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{S(q)} d\zeta' \int_{\mathcal{M}} e^{-\|\eta' - \zeta'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &= \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{S(q)} d\zeta' \int_{\mathcal{M} \cap \hat{S}(q)} e^{-\|\eta' - \zeta'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \qquad \text{(A.3)} \\ &+ \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{S(q)} d\zeta' \int_{\mathcal{M} \setminus \hat{S}(q)} e^{-\|\eta' - \zeta'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta'). \qquad \text{(A.4)}\end{aligned}$$

We bound $\mathbb{P}(q)$ above by bounding (A.3) and (A.4).

$$\begin{aligned}(A.3) &= \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{S(q)} d\zeta' \int_{\mathcal{M} \cap \hat{S}(q)} e^{-\|\bar{\eta}' - q\|_2^2/2\sigma^2} e^{-\|\hat{\eta}' - \hat{\zeta}'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &\leq \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{\mathcal{M} \cap \hat{S}(q)} e^{-\|\bar{\eta}' - q\|_2^2/2\sigma^2} \Big(\int_{0_d \times \mathbb{R}^{D-d}} e^{-\|\hat{\eta}' - \hat{\zeta}'\|_2^2/2\sigma^2} d\hat{\zeta}' \Big) d\mu_{\mathcal{M}}(\eta') \\ &= \frac{(2\pi\sigma)^{-d/2}}{\text{Vol}(\mathcal{M})} \int_{\mathcal{M} \cap \hat{S}(q)} e^{-\|\bar{\eta}' - q\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &= \frac{(2\pi\sigma)^{-d/2}}{\text{Vol}(\mathcal{M})} \int_{\mathcal{P}_d(\mathcal{M} \cap \hat{S}(q))} e^{-\|\bar{\eta}' - q\|_2^2/2\sigma^2} \sqrt{\det\big(I + J(\bar{\eta}')^T J(\bar{\eta}')\big)} d\bar{\eta}' \\ &\leq \frac{(2\pi\sigma)^{-d/2}}{\text{Vol}(\mathcal{M})} \Big(1 + \frac{C^2(r' + 3\sigma\sqrt{D})^2}{\tau^2}\Big)^{d/2} \int_{\mathbb{R}^d \times 0_{D-d}} e^{-\|\bar{\eta}' - q\|_2^2/2\sigma^2} d\bar{\eta}' \\ &= \frac{1}{\text{Vol}(\mathcal{M})} \Big(1 + \frac{C^2(r' + 3\sigma\sqrt{D})^2}{\tau^2}\Big)^{d/2}.\end{aligned}$$

The last inequality holds since $\|J(\bar{\eta}')\|_F \leq (C(r' + 3\sigma\sqrt{D}))/\tau$ with $\eta' \in B_D(z, r' + 3\sigma\sqrt{D})$. According to the definition of $S(q)$ and $\hat{S}(q)$, we have for any $\eta' \in \mathcal{M} \setminus \hat{S}(q)$ and $\zeta' \in S(q)$ the formula $|\eta(i)' - \zeta(i)'| \leq 3\sigma$, which implies

$$(2\pi\sigma^2)^{-D/2} \int_{S(q)} e^{-\|\eta' - \zeta'\|/2\sigma^2} d\zeta' \leq (0.01)^D \quad \forall \eta' \in \mathcal{M} \setminus \hat{S}(q).$$

Hence,

$$\begin{aligned}(A.4) &= \frac{(2\pi\sigma)^{-D/2}}{\text{Vol}(\mathcal{M})} \int_{\mathcal{M} \setminus \hat{S}(q)} \Big(\int_{S(q)} e^{-\|\eta' - \zeta'\|_2^2/2\sigma^2} d\zeta' \Big) d\mu_{\mathcal{M}}(\eta') \\ &\leq \frac{\text{Vol}(\mathcal{M} \setminus \hat{S}(q))}{\text{Vol}(\mathcal{M})} (0.01)^D \leq (0.01)^D.\end{aligned}$$

In summary, we have

$$\mathbb{P}(q) \leq \frac{1}{\mathrm{Vol}(\mathcal{M})}\Big(1+\frac{C^2(r'+3\sigma\sqrt{D})^2}{\tau^2}\Big)^{d/2}+(0.01)^D \tag{A.5}$$

for any $q \in T_{z^*}\mathcal{M} \cap B_D(z, r')$.

We consider only the lower bound for q in a subset of $T_{z^*}\mathcal{M} \cap B_D(z, r')$, namely $T_{z^*}\mathcal{M} \cap B_D(z^*, r_0)$, where r_0 is set as

$$\begin{aligned} r_0 = \min\{&\sqrt{r'^2 - \big(\frac{r'^2 - \|z - z^*\|_2^2}{2(\tau - \|z - z^*\|_2)} + \|z - z^*\|_2 + 3\sigma\sqrt{D-d}\big)^2}, \\ &\sqrt{r'^2 - \big(\frac{r'^2 - \|z - z^*\|_2^2}{2(\tau - \|z - z^*\|_2)} + \|z - z^*\|_2\big)^2} - 3\sigma\sqrt{d}\} \end{aligned} \tag{A.6}$$

For any $q \in T_{z^*}\mathcal{M} \cap B_D(z^*, r_0)$ and $\eta \in \mathcal{M} \cap B_D(z, r')$, we can verify the following conclusions via (A.2):

(i) The d-dimensional cube

$$\begin{aligned} &\{q' : q'(i) = 0 \ \forall i \geq d+1, \ |q'(j) - q(j)| \leq 3\sigma \ \forall j \leq d\} \\ &\qquad \subset B_D(q, 3\sigma\sqrt{d}) \cap T_{z^*}\mathcal{M} \subset \{\bar{\eta}' : \eta' \in \mathcal{M} \cap B_D(z, r')\}, \end{aligned}$$

(ii) The $(D-d)$-dimensional cube

$$\{\eta' : \eta'(i) = 0 \ \forall i \leq d, \ |\eta'(j) - \eta(j)| \leq 3\sigma \ \forall j \geq d+1\} \subset \{\hat{\zeta}' : \zeta' \in S(q)\}.$$

Now, we are ready to bound $\mathbb{P}(q)$ below for any $q \in B_D(z^*, r_0) \cap T_{z^*}\mathcal{M}$.

$$\begin{aligned} \mathbb{P}(q) &= \frac{(2\pi\sigma)^{-D/2}}{\mathrm{Vol}(\mathcal{M})}\int_{S(q)} d\zeta' \int_{\mathcal{M}} e^{-\|\eta'-\zeta'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &\geq \frac{(2\pi\sigma)^{-D/2}}{\mathrm{Vol}(\mathcal{M})}\int_{S(q)} d\zeta' \int_{\mathcal{M}\cap B_D(z,r')} e^{-\|\bar{\eta}'-q\|_2^2/2\sigma^2} e^{-\|\hat{\eta}'-\hat{\zeta}'\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &\geq \frac{(0.99)^{D-d}(2\pi\sigma^2)^{-d/2}}{\mathrm{Vol}(\mathcal{M})}\int_{\mathcal{M}\cap B_D(z,r')} e^{-\|\bar{\eta}'-q\|_2^2/2\sigma^2} d\mu_{\mathcal{M}}(\eta') \\ &= \frac{(0.99)^{D-d}(2\pi\sigma^2)^{-d/2}}{\mathrm{Vol}(\mathcal{M})}\int_{\mathcal{P}_d(\mathcal{M}\cap B_D(z,r'))} e^{-\|\bar{\eta}'-q\|_2^2/2\sigma^2} \sqrt{\det\big(I + J(\bar{\eta}')^T J(\bar{\eta}')\big)} d\bar{\eta}' \\ &= \frac{(0.99)^{D-d}(2\pi\sigma^2)^{-d/2}}{\mathrm{Vol}(\mathcal{M})}\int_{\mathcal{P}_d(\mathcal{M}\cap B_D(z,r'))} e^{-\|\bar{\eta}'-q\|_2^2/2\sigma^2} d\bar{\eta}' \\ &\geq \frac{(0.99)^D}{\mathrm{Vol}(\mathcal{M})} \end{aligned}$$

The last but one inequality holds since $\sqrt{\det\big(I + J(\bar{\eta}')^T J(\bar{\eta}')\big)} \geq 1$.

Since μ_d is the variance in the d-th direction, we have

$$\begin{aligned}
\mu_d &= \frac{1}{\int_{T_{z^*}\mathcal{M}\cap B_D(z,r')} \mathbb{P}(q')d\mathcal{L}_d(q')} \int_{T_{z^*}\mathcal{M}\cap B_D(z,r')} q_d^2 \mathbb{P}(q) d\mathcal{L}_d(q) \\
&\geq \frac{\alpha}{\mathrm{Vol}(B_d(r'))} \int_{T_{z^*}\mathcal{M}\cap B_D(z^*,r_0)} q_d^2 d\mathcal{L}_d(q) \\
&\geq \frac{\alpha}{\mathrm{Vol}(B_d(r'))} \int_0^{r_0} \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{2\pi} \left(\ell \Pi_{j=1}^{d-1} \phi_j\right)^2 dV \\
&= \frac{\Gamma(d/2+1)\alpha}{\pi^{d/2}(r')^d} \int_0^{r_0} \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{2\pi} \ell^{d+1} \prod_{j=1}^{d-1} \sin^{d-j+1} \phi_j d\ell \prod_{j=1}^{d-1} d\phi_j.
\end{aligned}$$

where α is the ratio between the lower bound and upper bound of $\mathbb{P}(q)$, namely,

$$\alpha = \frac{(0.99)^D}{(1+\frac{C^2(r'+3\sigma\sqrt{D})^2}{\tau^2})^{d/2} + (0.01)^D \mathrm{Vol}(\mathcal{M})},$$

and the third line follows with a change of coordinates. Substitute

$$\left\{ q_1 \to \ell \cos \phi_1, q_{2\leq i\leq d-1} \to \ell \cos \phi_i \Pi_{j=1}^{T} i - 1 \sin \phi_j, q_d \to \ell \sin \Pi_{j=1}^{d-1} \phi_j \right\}$$

with $\phi_{d-1} \in [0, 2\pi], \phi_{i\leq d-2} \in [0,\pi], \ell \in [0, r_0]$, and let

$$dV := \ell^{d-1} \Pi_{j=1}^{d-2} \sin^{d-j-i} \phi_j d\ell d\phi_1 \cdots d\phi_{d-1}.$$

The integral in the fourth line can be evaluated by noting that $\int_0^{r_0} \ell^{d+1} d\ell = {r_0}^{d+2}/(d+2)$, $\int_0^{2\pi} \sin^2 \phi_{d-1} d\phi_{d-1} = \pi$ and $\int_0^{\pi} \sin^{d-j+1} \phi_j d\phi_j = \frac{\sqrt{\pi}\Gamma((d-j+2)/2)}{\Gamma(1+(d-j+1)/2}$ for $1 \leq j \leq d-2$. Simplifying as Mohammed and Narayanan (2017) did, we get

$$\mu_d \geq \frac{\alpha}{d+2} \frac{{r_0}^{d+2}}{(r')^d} = c_0 \frac{(0.99)^D}{d+2} \frac{(r_0)^{d+2}}{(r')^d}.$$

According to Lemma 24, for any $\epsilon \in [0, 1/2]$, $\lambda_d \geq (1-\epsilon)\mu_d$ with probability $1 - d\exp\{\frac{-\epsilon^2 \mu_d |I_{z,r'}|}{2\ln 2}\}$. Taking $\epsilon = 1/2$, we have

$$\lambda_d \geq c_0 \frac{(0.99)^D}{d+2} \frac{(r_0)^{d+2}}{(r')^d}.$$

with probability $1 - d\exp\{\frac{-\epsilon^2 \mu_d |I_{z,r'}|}{2\ln 2}\}$. Using $r = O(\sqrt{\sigma})$ and $\|z - z^*\| \leq (1+c)r$, we can simplify r_0 and find c_0 satisfying $r_0 \geq c_0 r$. Hence, there exists a constant c independent on r such that $\lambda_d \geq cr^2$, which completes this proof. ■

A.3 Proof of Proposition 12, Lemma 13 and Lemma 14

Proof of Proposition 12 To show that $\tilde{\alpha}(x)$ is bounded below by $c_0|I_{x,r}|$ is equivalent to showing that there exists constant $c_1 > c$ and c_2 such that among the $|I_{x,r}|$ samples there are $c_2|I_{x,r}|$ ones lying in $B_D(x, c_1r)$, where c_0 in the lower bound is $c_0 = c_2(1-c_1^2)^{\beta}$. To quantify the number of samples $\{x_i\}_{i\in I_{x,r}}$ lying in $B_D(x, c_1r)$, we bound the conditional probability $\mathbb{P}(\|x_i - x\|_2 \le c_1r|i \in I_{x,r})$ below by calculating the lower bound of $\mathbb{P}(\|x_i - x\|_2 \le c_1r)$ and the upper bound of $\mathbb{P}(i \in I_{x,r})$, respectively.

By Lemma 23, we have $\mathbb{P}(i \in I_{x,c_1r}) \ge c_3r^d$. For the probability $\mathbb{P}(i \in I_{x,r})$, we have

$$\begin{aligned}\mathbb{P}(i \in I_{x,r}) = \mathbb{P}(&\|x_i - x\|_2 \le r,\ y_i \in \mathcal{M} \setminus B_D(x, Cr)) \\ &+ \mathbb{P}(\|x_i - x\|_2 \le r,\ y_i \in \mathcal{M} \cap B_D(x, Cr)),\end{aligned} \tag{A.7}$$

where

$$\begin{aligned}\mathbb{P}(\|x_i - x\|_2 \le r,\ y_i \in \mathcal{M} \cap B_D(x, Cr)) &\le \mathbb{P}(y_i \in \mathcal{M} \cap B_D(x, Cr)) \\ &= \frac{\mathrm{Vol}(\mathcal{M} \cap B_D(x, C_0r))}{\mathrm{Vol}(\mathcal{M})} = Cr^d,\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\|x_i - x\|_2 \le r,\ y_i \in \mathcal{M} \setminus B_D(x, Cr)) &\le \mathbb{P}(\|\xi_i\|_2 \ge (C-1)r) \\ &\le \frac{C_1}{r}e^{-\frac{C_2}{r^2}} \le Cr^d,\end{aligned}$$

where the second-last inequality holds by Chernoff bound, and the last inequality holds since $r = O(\sqrt{\sigma})$ is sufficiently small. Plugging the above bounds into (A.7), we obtain

$$\mathbb{P}(i \in I_{x,r}) \le Cr^d.$$

Hence, for any $i \in I_{x,r}$, we have $\|x_i - x\|_1 \le c_1r$ with probability $\rho = (cr^d)/(Cr^d) < 1$ for being a constant independent on r.

Applying the Berry-Esseen theorem to the $|I_{x,r}|$ Bernoulli trials, we conclude that there exists $c_2|I_{x,r}|$ i' in $I_{x,r}$ such that $\|x_i - x\|_2 \le c_1r$ with probability $1 - C/\sqrt{|I_{x,r}|}$, which proves (i).

To show (ii), we recall Lemma 23 that $\mathbb{P}(i \in I_{x,c_1r}) \ge cr^d$. Thus there is a sample among N samples lying in $B_D(x, c_1r)$ with probability

$$1 - (1 - cr^d)^N = O(Nr^d).$$

Then, $\tilde{\alpha}(x) \ge (1 - c_1^2)^{\beta} := c_0'$ with the same probability. ■

Lemma 25. *Suppose $\xi \sim N(0, \sigma^2 I_D)$; then we have, for any positive integer k:*

(i) $\mathbb{E}(\|\xi\|_2^k) = C_1\sigma^k$

(ii) $\mathrm{Var}(\|\xi\|_2^k) = C_2\sigma^{2k}$

(iii) $\mathbb{E}[\big(\|\xi\|_2^k - \mathbb{E}(\|\xi\|_2^k)\big)^3] = C_3\sigma^{3k}$

(iv) $\|\xi_i\|_2^k$ *and* $\|\xi_j\|_2^k$ *are independent if* ξ_i *and* ξ_j *are independent,*

(v) $\mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t) = C_4\sigma^{s+t}$

(vi) $\mathrm{Var}(\|\xi_i\|_2^s\|\xi_j\|_2^t) = C_5\sigma^{2(s+t)}$

(vii) $\mathbb{E}[\big(\|\xi_i\|_2^s\|\xi_j\|_2^t - \mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)\big)^3] = C_6\sigma^{3(s+t)}$

where C_n*,* $n = 1, \cdots, 6$ *are constants depending on* D *and* k*.*

Proof Letting the i-th element of ξ be denoted by $\xi^{(i)}$, we have the following qualities:

$$
\begin{aligned}
\mathbb{E}(\|\xi\|_2^k) &= \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(\sum_{i=1}^{D}\xi^{(i)2})^{k/2}e^{-\frac{\sum_{i=1}^{D}\xi^{(i)2}}{2\sigma^2}}d\xi^{(1)}\cdots d\xi^{(D)}\\
&= \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}}\int_{r=0}^{+\infty} r^k e^{-r^2/(2\sigma^2)}S_D(r)dr\\
&= \frac{2\pi^{\frac{D}{2}}}{(2\pi\sigma^2)^{\frac{D}{2}}\Gamma(\frac{D}{2})}\int_{r=0}^{+\infty} r^{D+k-1}e^{-r^2/(2\sigma^2)}dr\\
&= \frac{\pi^{\frac{D}{2}}}{(2\pi\sigma^2)^{\frac{D}{2}}\Gamma(\frac{D}{2})}\int_{r=0}^{+\infty} r^{D+k-2}e^{-r^2/(2\sigma^2)}dr^2\\
&= \frac{\pi^{\frac{D}{2}}(2\sigma^2)^{\frac{D+k}{2}}}{(2\pi\sigma^2)^{\frac{D}{2}}\Gamma(\frac{D}{2})}\int_{z=0}^{+\infty}(\frac{z}{2\sigma^2})^{\frac{D+k}{2}-1}e^{-z/(2\sigma^2)}d\frac{z}{2\sigma^2}\\
&= \frac{2^{k/2}\sigma^k}{\Gamma(\frac{D}{2})}\int_{z=0}^{+\infty} z^{\frac{D+k}{2}-1}e^{-z}dz = \frac{2^{k/2}\Gamma(\frac{D+k}{2})}{\Gamma(\frac{D}{2})}\sigma^k
\end{aligned}
$$

where $\Gamma(t) = \int_0^{+\infty} s^{t-1}e^{-s}ds$ is the Gamma function. Plugging the above equality into $\mathrm{Var}(\|\xi\|_2^k) = \mathbb{E}(\|\xi\|_2^{2k}) - \mathbb{E}(\|\xi\|_2^k)^2$, and

$$
\begin{aligned}
\mathbb{E}[\big(\|\xi\|_2^k - \mathbb{E}(\|\xi\|_2^k)\big)^3] =& \mathbb{E}(\|\xi\|_2^{3k}) - 3\mathbb{E}(\|\xi\|_2^{2k})\mathbb{E}(\|\xi\|_2^k)\\
&+ 3\mathbb{E}(\|\xi\|_2^k)\mathbb{E}(\|\xi\|_2^k)^2 - \mathbb{E}(\|\xi\|_2^k)^3,
\end{aligned}
$$

will yield the variance and third moment.

To show the independence, we set F_X as the cumulative distribution function of X, $S_t(\zeta) = \{\xi_t : \|\xi_t\|_2^k \le \zeta\}$ and $\eta_t = \|\xi_t\|_2^k$ with $t = i, j$. Then

$$
\begin{aligned}
F_{\eta_i,\eta_j}(\zeta_i,\zeta_j) &= P(\eta_i \le \zeta_i, \eta_j \le \zeta_j)\\
&= P(\xi_i \in S_i(\zeta_i), \xi_j \in S_j(\zeta_j))\\
&= P(\xi_i \in S_i(\zeta_i))P(\xi_j \in S_j(\zeta_j))\\
&= P(\eta_i \le \zeta_i)P(\eta_j \le \zeta_j)\\
&= F_{\eta_i}(\zeta_i)F_{\eta_j}(\zeta_j),
\end{aligned}
$$

which completes the proof of independence by definition. Based on the independence, we obtain

$$\mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t) = \mathbb{E}(\|\xi_i\|_2^s)\mathbb{E}(\|\xi_i\|_2^s) = (C_1\sigma^s) \times C_1(\sigma^t) = C_4\sigma^{s+t}.$$

Plugging the above equality into

$$\mathrm{Var}(\|\xi_i\|_2^s\|\xi_j\|_2^t) = \mathbb{E}(\|\xi_i\|_2^{2s}\|\xi_j\|_2^{2t}) - \mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)^2$$

and

$$\begin{aligned}\mathbb{E}[\big(\|\xi_i\|_2^s\|\xi_j\|_2^t - \mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)\big)^3] =& \mathbb{E}(\|\xi_i\|_2^{3s}\|\xi_j\|_2^{3t}) - 3\mathbb{E}(\|\xi_i\|_2^{2s}\|\xi_j\|_2^{2t})\mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t) \\ &+ 3\mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)\mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)^2 - \mathbb{E}(\|\xi_i\|_2^s\|\xi_j\|_2^t)^3,\end{aligned}$$

will produce the variance and the third moment. ■

Proposition 26. *Suppose $\{\xi_i\}_{i=1}^n$ are i.i.d. drawn from $N(0, \sigma^2 I_D)$, $\sum_{i=1}^n \alpha_i = 1$ and $\max_{i\in\{1,\cdots,n\}} \alpha_i \leq C_\alpha/n$ with certain constant C_α. For any δ, there exist constants C, depending on D, k, δ, and n_1, depending on δ and C_α such that if $n \geq n_1$, then*

$$\sum_{i=1}^n \alpha_i\|\xi_i\|_2^k \leq C\sigma^k \quad \text{and} \quad \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\xi_i\|_2^s\|\xi_j\|_2^t \leq C\sigma^{s+t}$$

holds for $k, s, t \leq 4$ with probability at least $1-\delta$.

Proof By Lemma 25, $\|\xi_1\|_2^k, \cdots, \|\xi_n\|_2^k$ are i.i.d. random variables drawn from a distribution whose expectation is $\mathbb{E}(\|\xi\|_2^k)$ and variance is $\mathrm{Var}(\|\xi\|_2^k)$. Using the Berry-Esseen Theorem, the cumulative distribution function of the variable

$$(\sum_{i=1}^n \alpha_i\|\xi_i\|_2^k - \mathbb{E}(\|\xi\|_2^k)) \;/\; (C_2^{1/2}\sigma^k\sqrt{\sum_{i=1}^n \alpha_i^2})$$

denoted by F_n satisfies

$$|F_n(t) - \Phi(t)| \leq \frac{C_0\rho\sum_{i=1}^n \alpha_i^3}{\sigma^{3k}(\sum_{i=1}^n \alpha_i^2)^{3/2}} = C_0'\frac{\sum_{i=1}^n \alpha_i^3}{(\sum_{i=1}^n \alpha_i^2)^{3/2}} \leq C_0'\frac{\sum_{i=1}^n \alpha_i^3}{\left(\frac{1}{n}(\sum_{i=1}^n \alpha_i)^2\right)^{3/2}} = C_0' n^{\frac{3}{2}}\sum_{i=1}^n \alpha_i^3$$

where Φ is the cumulative distribution function of standard normal distribution, ρ is the third moment of $\|\xi\|_2^k$, which is in the order of σ^{3k} according to Lemma 25(iii), and the last inequality holds in accordance with Cauchy's inequality.

Since $\alpha_i \leq C_\alpha/n$, we obtain $\sum_{i=1}^n \alpha_i^3 \leq n(\frac{C_\alpha}{n})^3 = C_\alpha^3 n^{-2}$ and therefore $|F_n(t) - \Phi(t)| \leq C'/\sqrt{n}$. So there exists a constant C depending on D, k, and δ such that

$$\sum_{i=1}^n \alpha_i\|\xi_i\|_2^k \leq C\sigma^k,$$

with probability $1-\frac{\delta}{2}-C'/\sqrt{n}$. Taking $n_1=\frac{4C'}{\delta^2}$, $\sum_{i=1}^{n}\alpha_i\|\xi_i\|_2^k\le C\sigma^k$ with probability at least $1-\delta$ when $n\ge n_1$. Analogously, there exist C and n_0 such that

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\xi_i\|_2^s\|\xi_i\|_2^t\le C\sigma^{s+t}.$$

with probability at least $1-\delta$ when $n\ge n_1$. ■

Proof of Lemma 13 By (2.2), we have $r<C_1$. For any given δ, let

$$n_0=\max\left\{\frac{4C^2C_1^{2d}}{\delta^2},\frac{\max\{n_1,\frac{4C_0^2}{\delta^2}\}}{c'}\right\},$$

where C and c' are the two constants in Proposition 3, n_1 is the constant in Proposition 26, and C_0 is the constant in Proposition 12. Plugging $N\ge n_0r^{-d}$ into Proposition 3, we obtain $|I_{x,r}|\ge\max\{n_1,\frac{4C_0^2}{\delta^2}\}$ with probability at least $1-\frac{\delta}{2}$. Recalling Proposition 12 (i) and the definition of α_i in (2.5), $\alpha_i\le\frac{1}{\tilde{\alpha}}\le\frac{C_\alpha}{|I_{x,r}|}$ with probability at least $1-\frac{\delta}{2}$ and $C_\alpha=\frac{1}{c_0}$ since $1-\frac{C_0}{\sqrt{|I_{x,r}|}}\ge1-\frac{\delta}{2}$ by $|I_{x,r}|\ge\frac{4C_0^2}{\delta^2}$. As a result, conditions of Proposition 26 hold with probability at least $(1-\frac{\delta}{2})^2\ge1-\delta$. Using Proposition 3, we are able to complete the proof. ■

Proof of Lemma 14 Corollary 12 of Boissonnat et al. (2018) shows that

$$\left\|\sin\frac{\theta(U_x,U_y)}{2}\right\|_2\le\frac{\|x-y\|_2}{2\mathrm{reach}(\mathcal{M})},$$

where U_x and U_y are the basis of $T_x\mathcal{M}$ and $T_y\mathcal{M}$, respectively. Letting the orthogonal complements of U_x and U_y be denoted by V_x and V_y, respectively, we obtain $\Pi_x=V_xV_x^T$ and $\Pi_y=V_yV_y^T$. Then, in accordance with (ii) of Lemma 8,

$$\begin{aligned}\|\Pi_x-\Pi_y\|_F=\|V_xV_x^T-V_yV_y^T\|_F=\|U_xU_x-U_yU_y\|_F\le C\Big\|\sin\theta(U_x,U_y)\Big\|_2\\ \le 2C\Big\|\sin\frac{\theta(U_x,U_y)}{2}\Big\|_2\le C\frac{\|x-y\|_2}{\tau}.\end{aligned}$$

■

A.4 Proof of Theorem 17

To prove Theorem 17, we first introduce two lemmas, namely Lemma 27 and Lemma 28.

Lemma 27. *Suppose $d(x,\mathcal{M})\le cr$ with some constant $c<1$ and $r=O(\sqrt{\sigma})$. For any given δ, there exist constants C and n_0 such that if $N\ge n_0r^{-d}$, then the following inequalities hold:*

(i) $\|(\|P_{x_i}-\Pi_{x_i^*}\|_2)_{i\in I_{x,r}}\|_2\le Cr|I_{x,r}|^{\frac{1}{2}}$ *with probability* $\delta_0(1-\delta)^2$,

(ii) $\|\big(\|x_i - x_i^*\|_2\big)_{I_{x,r}}\|_2 \le Cr^2|I_{x,r}|^{\frac{1}{2}}$ *with probability* $1-\delta$,

(iii) $\|\big(\|x_i^* - x^*\|_2\big)_{I_{x,r}}\|_2 \le Cr|I_{x,r}|^{\frac{1}{2}}$ *with probability* $1-\delta$.

Proof We begin with (i). Plugging $z = x_i$ into Theorem 11, we obtain

$$\Big\|\big(\|P_{x_i} - \Pi_{x_i^*}\|_2\big)_{i\in I_{x,r}}\Big\|_2^2 \le \sum_{i\in I_{x,r}} (A^2 + 2*AB + B^2) \textit{ with probability } \delta_0,$$

where $A = \frac{C}{r^2}\frac{1}{|I_{x_i,2r}|}\sum_{j\in I_{x_i,2r}}\big(\|\xi_j\|_2^4 + \|\xi_j\|_2^3 + \|\xi_j\|_2^2 + r\|\xi_j\|_2\big)$ and $B = C(r + \frac{\|\xi_i\|_2}{r} + \frac{\|\xi_i\|_2^2}{r^2})$. In accordance with Lemma 13, there exist C and n_0 such that

$$\begin{aligned}\sum_{i\in I_{x,r}} AB = {} & \frac{C|I_{x,r}|}{r}\frac{1}{|I_{x,r}|\times|I_{x_i,2r}|}\sum_{i\in I_{x,r}}\sum_{j\in I_{x_i,2r}}\big(\|\xi_j\|^4 + \|\xi_j\|^3 + \|\xi_j\|^2 + r\|\xi_j\|\big)\\ & + \frac{C|I_{x,r}|}{r^3}\frac{1}{|I_{x,r}|\times|I_{x_i,2r}|}\sum_{i\in I_{x,r}}\sum_{j\in I_{x_i,2r}}\big(\|\xi_j\|^4\|\xi_i\| + \|\xi_j\|^3\|\xi_i\| + \|\xi_j\|^2\|\xi_i\| + r\|\xi_j\|\|\xi_i\|\big)\\ & + \frac{C|I_{x,r}|}{r^4}\frac{1}{|I_{x,r}|\times|I_{x_i,2r}|}\sum_{i\in I_{x,r}}\sum_{j\in I_{x_i,2r}}\big(\|\xi_j\|^4\|\xi_i\|^2 + \|\xi_j\|^3\|\xi_i\|^2 + \|\xi_j\|^2\|\xi_i\|^2 + r\|\xi_j\|\|\xi_i\|^2\big)\\ & \le C|I_{x,r}|(r^2 + r^2 + r^3) \le Cr^2|I_{x,r}|,\end{aligned}$$

$$\begin{aligned}A^2 &= \frac{C}{r^4}\frac{1}{|I_{x_i,2r}|^2}\sum_{j,k\in I_{x_i,2r}}\big(\|\xi_j\|^4\|\xi_k\|^4 + \|\xi_j\|^4\|\xi_k\|^3 + \cdots + r^2\|\xi_j\|\|\xi_k\|\big)\\ &\le \frac{C}{r^4}\big(\sum_{k=4}^{8}\sigma^k + r\sum_{k=3}^{5}\sigma^k + r^2\sigma^2\big) \le Cr^2,\end{aligned}$$

$$\sum_{i\in I_{x,r}} B^2 = \sum_{i\in I_{x,r}}\Big(r^2 + \frac{\|\xi_i\|_2^2}{r^2} + \frac{\|\xi_i\|_2^4}{r^4} + 2\|\xi_i\|_2 + 2\frac{\|\xi_i\|_2^2}{r} + 2\frac{\|\xi_i\|_2^3}{r^3}\Big) \le Cr^2|I_{x,r}|,$$

with probability $1-\delta/3$ respectively. The above bounds amount to $\Big\|\big(\|P_{x_i} - \Pi_{x_i^*}\|_2\big)_{i\in I_{x,r}}\Big\|_2^2 \le Cr^2|I_{x,r}|$, which leads to

$$\Big\|\big(\|P_{x_i} - \Pi_{x_i^*}\|_2\big)_{i\in I_{x,r}}\Big\|_2 \le Cr|I_{x,r}|^{\frac{1}{2}}, \textit{ with probability } \delta_0(1-\delta)^2.$$

As for (ii),

$$\Big\|\big(\|x_i - x_i^*\|_2\big)_{I_{x,r}}\Big\|_2^2 = \sum_{i\in I_{x,r}}\|x_i - x_i^*\|_2^2 \le \sum_{i\in I_{x,r}}\|\xi_i\|_2^2 \le |C|I_{x,r}|\sigma^2,$$

with probability $1-\delta$, which implies $\big\|\big(\|x_i - x_i^*\|_2\big)_{I_{x,r}}\big\|_2 \le C|I_{x,r}|^{\frac{1}{2}}\sigma = Cr^2|I_{x,r}|^{\frac{1}{2}}$. We derive (iii) based on

$$\|x_i^* - x^*\|_2 \le \|x_i^* - x_i\|_2 + \|x_i - x\|_2 + \|x - x^*\|_2 \le \|\xi_i\|_2 + 2r.$$

Thus we have

$$\Big\|\big(\|x_i^* - x^*\|_2\big)_{I_{x,r}}\Big\|_2^2 \le \sum_{i\in I_{x,r}} \big(\|\xi_i\|_2^2 + 4r^2 + 2r\|\xi_i\|_2\big) \le C|I_{x,r}|\big(\sigma^2 + 4r^2 + 2r\sigma\big) \le Cr^2|I_{x,r}|$$

with probability $1-\delta$, which implies $\Big\|\big(\|x_i^* - x^*\|_2\big)_{I_{x,r}}\Big\|_2 \le Cr|I_{x,r}|^{\frac{1}{2}}$. ■

Lemma 28. *Suppose $d(x,\mathcal{M}) \le cr$ with some constant $c<1$, $r = O(\sqrt{\sigma})$ and $\beta \ge 2$. For any given δ, there exist constants C and n_0 such that if $N \ge n_0 r^{-d}$, then*

$$\|\big(\partial_v \alpha_i(x)\big)_{i\in I_{x,r}}\|_2 \le \frac{C}{r}|I_{x,r}|^{-\frac{1}{2}} \text{ with probability } 1-\delta.$$

Proof By Lemma 13, $\tilde{\alpha}(x) \ge c_0|I_{x,r}|$ with probability at least $1-\delta$. Based on this, we obtain the following inequalities given $0 \le \tilde{\alpha}_i(x) \le 1$:

$$\begin{aligned}
\Big\|\big(\partial_v \alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2 &\le \Big\|\Big(\frac{\partial_v \tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\Big)_{i\in I_{x,r}}\Big\|_2 + \Big\|\Big(\frac{(\partial_v\tilde{\alpha}(x))\tilde{\alpha}_i(x)}{\tilde{\alpha}^2(x)}\Big)_{i\in I_{x,r}}\Big\|_2 \\
&\le \frac{C}{r}\Big\|\Big(\frac{\tilde{\alpha}_i(x)^{\frac{\beta-1}{\beta}}}{\tilde{\alpha}(x)}\Big)_{i\in I_{x,r}}\Big\|_2 + \Big|\frac{\partial_v\tilde{\alpha}(x)}{\tilde{\alpha}^2(x)}\Big|\|\big(\tilde{\alpha}_i(x)\big)_{i\in I_{x,r}}\|_2 \\
&\le \frac{C}{r}\Big\|\Big(\frac{1}{\tilde{\alpha}(x)}\Big)_{i\in I_{x,r}}\Big\|_2 + |\frac{\partial_v\tilde{\alpha}(x)}{\tilde{\alpha}^2(x)}|\|(1)_{i\in I_{x,r}}\|_2 \\
&\le \frac{C}{r}|I_{x,r}|^{-\frac{1}{2}} + \frac{C}{r}|I_{x,r}|^{\frac{1}{2}}\frac{\sum_{i\in I_{x,r}}\tilde{\alpha}_i(x)^{\frac{\beta-1}{\beta}}}{\tilde{\alpha}^2(x)} \\
&\le \frac{C}{r}|I_{x,r}|^{-\frac{1}{2}} + \frac{C}{r}|I_{x,r}|^{\frac{1}{2}}\frac{|I_{x,r}|}{|I_{x,r}|^2} \le \frac{C}{r}|I_{x,r}|^{-\frac{1}{2}}.
\end{aligned}$$

■

Proof of Theorem17

We rewrite (2.6) as

$$f(x) = \Psi_x^\alpha \sum_{i\in I_{x,r}} \alpha_i(x)(x - x_i), \tag{A.8}$$

and calculate the first derivative of $f(x)$ as

$$\begin{aligned}
\partial_v f(x) = &\sum_{i\in I_{x,r}} \alpha_i(x)\Psi_x^\alpha\big(\partial_v(x-x_i)\big) \\
&+ \sum_{i\in I_{x,r}} \alpha_i(x)(\partial_v\Psi_x^\alpha)(x-x_i) \\
&+ \sum_{i\in I_{x,r}} (\partial_v\alpha_i(x))\Psi_x^\alpha(x-x_i).
\end{aligned} \tag{A.9}$$

We deal with the three terms one by one. First,

$$\sum_{i\in I_{x,r}} \alpha_i(x)\Psi_x^\alpha\big((\partial_v(x-x_i)\big) = \sum_{i\in I_{x,r}} \alpha_i(x)\Psi_x^\alpha v = \Psi_x^\alpha v.$$

To bound the second term of (A.9), we proceed to bound $\|\partial_v \Psi_x^\alpha\|_2$. In accordance with (26) of Fefferman et al. (2018), we establish the relationship between $\|\partial_v \Psi_x^\alpha\|_2$ and $\|\partial_v A_x\|_2$ as follows:

$$\begin{aligned}
\|\partial_v \Psi_x^\alpha\|_2 &\le 8\|\partial_v A_x\|_2 \\
&= C\Big\| \sum_i \partial_v \alpha_i(x)\big((P_{x_i} - \Pi_{x_i^*}) + (\Pi_{x_i^*} - \Pi_{x^*})\big) + \Pi_{x^*}\big(\partial_v \sum_i \alpha_i(x)\big)\Big\|_2 \\
&\le C\sum_i |\partial_v \alpha_i(x)|\|P_{x_i} - \Pi_{x_i^*}\|_2 + \frac{C}{\tau}\sum_i |\partial_v \alpha_i(x)|\|x_i^* - x^*\|_2 + 0 \\
&\le C\Big\|\big(\partial_v \alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2 \Big\|\big(\|P_{x_i} - \Pi_{x_i^*}\|_2\big)_{i\in I_{x,r}}\Big\|_2 \\
&+ \frac{C}{\tau}\Big\|\big(\partial_v \alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2 \Big\|\big(\|x_i^* - x^*\|_2\big)_{i\in I_{x,r}}\Big\|_2 \\
&\le Cr\Big\|\big(\partial_v \alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2 |I_{x,r}|^{\frac{1}{2}},
\end{aligned}$$

where the second to the last inequality holds by Cauchy-Schwarz inequality, and the last inequality holds by Lemma 27 and Lemma 28. As a result,

$$\|\partial_v \Psi_x^\alpha\|_2 \le 8\|\partial_v A_x\|_2 \le C. \tag{A.10}$$

Therefore, the second term of (A.9) is bounded as

$$\Big\| \sum_{i\in I_{x,r}} \alpha_i(x)(\partial_v \Psi_x^\alpha)(x-x_i)\Big\|_2 \le \sum_{i\in I_{x,r}} \alpha_i(x)\|\partial_v \Psi_x^\alpha\|_2\|x-x_i\|_2 \le \sum_{i\in I_{x,r}} \alpha_i(x)Cr = Cr.$$

As for the last term in (A.9), we have

$$\begin{aligned}
\Big\| \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Psi_x^\alpha(x-x_i)\Big\|_2 &\le \Big\| \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Psi_x^\alpha(x^*-x_i^*)\Big\|_2 \\
&+ \Big\|\big(\Psi_x^\alpha(x-x^*)\big) \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Big\|_2 \\
&+ \Big\| \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Psi_x^\alpha(x_i^*-x_i)\Big\|_2 \\
&= \Big\| \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Psi_x^\alpha(x^*-x_i^*)\Big\|_2 + 0 \\
&+ \Big\| \sum_{i\in I_{x,r}} \partial_v \alpha_i(x)\Psi_x^\alpha(x_i^*-x_i)\Big\|_2,
\end{aligned}$$

where

$$\begin{aligned}\Big\|\sum_{i\in I_{x,r}} \partial_v\alpha_i(x)\Psi_x^\alpha(x^*-x_i^*)\Big\|_2 &\le \sum_{i\in I_{x,r}} |\partial_v\alpha_i(x)|\|\Psi_x^\alpha(x_i^*-x^*)\|_2\\ &\le \Big\|\big(\partial_v\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\Big\|\big(\|\Psi_x^\alpha(x_i^*-x^*)\|\big)_{i\in I_{x,r}}\Big\|_2 \le Cr\end{aligned}$$

and

$$\begin{aligned}\Big\|\sum_{i\in I_{x,r}} \partial_v\alpha_i(x)\Psi_x^\alpha(x_i^*-x_i)\Big\|_2 &\le \sum_{i\in I_{x,r}} |\partial_v\alpha_i(x)|\|\Psi_x^\alpha(x_i^*-x_i)\|_2\\ &\le \Big\|\big(\partial_v\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\Big\|\big(\|\Psi_x^\alpha(x_i^*-x_i)\|\big)_{i\in I_{x,r}}\Big\|_2 \le Cr\end{aligned}$$

based on

$$\begin{aligned}\|\Psi_x^\alpha(x_i^*-x^*)\|_2 &\le \|\Psi_x^\alpha-\Pi_{x^*}\|_2\|x_i^*-x^*\|_2+\|\Pi_{x^*}(x_i^*-x^*)\|_2\\ &\le Cr^2+C\frac{\|x_i^*-x^*\|_2^2}{\tau}\le Cr^2,\end{aligned}$$

where the second inequality holds in probability via Theorem 15 and Proposition 2. The above bounds amount to the bound on the first derivative, that is, $\|\partial_v f(x)-\Psi_x^\alpha v\|_2\le Cr$.

The above proof is based on Lemma 27, Lemma 28 and Theorem 15, which are valid when Lemma 13 and Theorem 11 hold. Hence, the conclusion obtained from the above proof is valid when Lemma 13 and Theorem 11 simultaneously hold, whose probability is at least $\delta_0(1-\delta)^2$. ■

A.5 Proof of Theorem 19

To prove Theorem 19, we first introduce Lemma 29.

Lemma 29. *Suppose $d(x,\mathcal{M})\le cr$ with some constant $c<1$, $r=O(\sqrt{\sigma})$ and $\beta\ge 2$. For any given δ, there exist constants C and n_0 such that if $N\ge n_0r^{-d}$, then*

$$\Big\|\big(\partial_v\partial_u\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\le\frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}\ \textit{with probability}\ 1-\delta.$$

Proof

$$\begin{aligned}\Big\|\big(\partial_v\partial_u\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2 \le{}& \Big\|\Big(\frac{\partial_v\partial_u\tilde\alpha_i(x)}{\tilde\alpha(x)}\Big)_{i\in I_{x,r}}\Big\|_2+\Big\|\Big(\frac{\partial_v\partial_u\tilde\alpha(x)}{\tilde\alpha^2(x)}\tilde\alpha_i(x)\Big)_{i\in I_{x,r}}\Big\|_2\\ &+\Big\|\Big(\frac{(\partial_v\tilde\alpha_i(x))(\partial_u\tilde\alpha(x))}{\tilde\alpha^2(x)}\Big)_{i\in I_{x,r}}\Big\|_2+\Big\|\Big(\frac{(\partial_u\tilde\alpha_i(x))(\partial_v\tilde\alpha(x))}{\tilde\alpha^2(x)}\Big)_{i\in I_{x,r}}\Big\|_2\\ &+2\Big\|\Big(\big(\frac{\partial_v\tilde\alpha(x)}{\tilde\alpha(x)}\big)\big(\frac{\partial_u\tilde\alpha(x)}{\tilde\alpha(x)}\big)\big(\frac{\tilde\alpha_i(x)}{\tilde\alpha(x)}\big)\Big)_{i\in I_{x,r}}\Big\|_2.\end{aligned}$$

We bound these five terms one-by-one using $\tilde{\alpha}(x) \geq c|I_{x,r}|$ which holds with probability $1-\delta$ by Lemma 13 and $0 \leq \tilde{\alpha}_i(x) \leq 1$. For the first term,

$$\begin{aligned}\Big\|\Big(\frac{\partial_v\partial_u\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\Big)_{i\in I_{x,r}}\Big\|_2 &\leq \frac{C}{\tilde{\alpha}(x)}\Big\|\big(\tilde{\alpha}_i(x)^{\frac{\beta-2}{\beta}}\frac{\|x-x_i\|_2^2}{r^4}+\tilde{\alpha}_i(x)^{\frac{\beta-1}{\beta}}\frac{|v^Tu|}{r^2}\big)_{i\in I_{x,r}}\Big\|_2\\ &\leq \frac{C}{\tilde{\alpha}(x)}\Big\|\big(\frac{2}{r^2}\big)_{i\in I_{x,r}}\Big\|_2 \leq \frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}.\end{aligned}$$

For the second term,

$$\begin{aligned}\Big\|\Big(\frac{\partial_v\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}^2(x)}\tilde{\alpha}_i(x)\Big)_{i\in I_{x,r}}\Big\|_2 &\leq \Big|\frac{\partial_v\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}^2(x)}\Big|\big\|\big(\tilde{\alpha}_i(x)\big)_{i\in I_{x,r}}\big\|_2\\ &\leq \Big|\frac{\partial_v\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}^2(x)}\Big|\big\|(1)_{i\in I_{x,r}}\big\|_2\\ &\leq \frac{1}{\tilde{\alpha}(x)}\Big\|\Big(\frac{\partial_v\partial_u\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\Big)_{i\in I_{x,r}}\Big\|_2\|(1)_{i\in I_{x,r}}\|_2^2\\ &\leq \frac{C}{r^2}|I_{x,r}|^{-1}|I_{x,r}|^{-\frac{1}{2}}|I_{x,r}| = \frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}.\end{aligned}$$

The third and fourth terms are similar, where the third term is bounded by

$$\begin{aligned}\|\big(\frac{(\partial_v\tilde{\alpha}_i(x))(\partial_u\tilde{\alpha}(x))}{\tilde{\alpha}^2(x)}\big)_{i\in I_{x,r}}\|_2 &\leq |\frac{\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}(x)}|\|\big(\frac{\partial_v\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\big)_{i\in I_{x,r}}\|_2\\ &\leq \|\big(\frac{\partial_u\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\big)_{i\in I_{x,r}}\|_2\|(1)_{i\in I_{x,r}}\|_2\|\big(\frac{\partial_v\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\big)_{i\in I_{x,r}}\|_2\\ &\leq \frac{C}{r}|I_{x,r}|^{-\frac{1}{2}}|I_{x,r}|^{\frac{1}{2}}\frac{C}{r}|I_{x,r}|^{-\frac{1}{2}} = \frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}},\end{aligned}$$

and analogically, the fourth is bounded by

$$\Big\|\Big(\frac{(\partial_u\tilde{\alpha}_i(x))(\partial_v\tilde{\alpha}(x))}{\tilde{\alpha}^2(x)}\Big)_{i\in I_{x,r}}\Big\|_2 \leq \frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}$$

Finally, the fifth term:

$$\begin{aligned}\Big\|\Big(\big(\frac{\partial_v\tilde{\alpha}(x)}{\tilde{\alpha}(x)}\big)\big(\frac{\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}(x)}\big)\big(\frac{\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\big)\Big)_{i\in I_{x,r}}\Big\|_2 &= \big(\frac{\partial_v\tilde{\alpha}(x)}{\tilde{\alpha}(x)}\big)\big(\frac{\partial_u\tilde{\alpha}(x)}{\tilde{\alpha}(x)}\big)\Big\|\big(\frac{\tilde{\alpha}_i(x)}{\tilde{\alpha}(x)}\big)_{i\in I_{x,r}}\Big\|_2\\ &\leq \frac{C}{r}\times\frac{C}{r}\times|I_{x,r}|^{-\frac{1}{2}} = \frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}.\end{aligned}$$

Summing the above five terms up amounts to the proof. ■

Proof of Theorem19 Letting $G(x) = \sum_{i\in I_{x,r}}\alpha_i(x)(x-x_i)$, we obtain the following bound on the second derivative of $f(x)$

$$\begin{aligned}\big\|\partial_v\big(\partial_u f(x)\big)\big\|_2 \leq{}& \big\|(\partial_v\partial_u\Psi_x^{\alpha})G(x)\big\|_2 + \big\|(\partial_v\Psi_x^{\alpha})(\partial_u G(x))\big\|_2\\ &+ \big\|(\partial_u\Psi_x^{\alpha})(\partial_v G(x))\big\|_2 + \big\|\Psi_x^{\alpha}(\partial_v\partial_u G(x))\big\|_2.\end{aligned} \tag{A.11}$$

For the first term, we have

$$\begin{aligned}
\|\partial_v\partial_u\Psi_x^\alpha\|_2 &\le C\big(\|\partial_v A_x\|_2\|\partial_u A_x\|_2 + \|\partial_v\partial_u A_x\|_2\big)\\
&\le C + C\sum_i |\partial_v\partial_u\alpha_i(x)|\big(\|P_{x_i}-\Pi_{x_i^*}\|_2 + \|\Pi_{x_i^*}-\Pi_{x^*}\|_2\big)\\
&\quad + C\Big\|\Pi_{x^*}\big(\partial_v\partial_u\sum_i\alpha_i(x)\big)\Big\|_2\\
&\le C + C\Big\|\big(\partial_v\partial_u\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\Big\|\big(\|P_{x_i}-\Pi_{x_i^*}\|_2\big)_{i\in I_{x,r}}\Big\|_2\\
&\quad + C\Big\|\big(\partial_v\partial_u\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\Big\|\big(\frac{\|x_i^*-x^*\|_2}{\tau}\big)_{i\in I_{x,r}}\Big\|_2 + 0\\
&\le C + \Big(\frac{C}{r^2}|I_{x,r}|^{-\frac{1}{2}}\Big)\times\Big(Cr|I_{x,r}|^{\frac{1}{2}}\Big) \le \frac{C}{r},
\end{aligned}$$

where the second to the last inequality holds by Lemma 27 while the last inequality holds by Lemma 29, and therefore

$$\|(\partial_v\partial_u\Psi_x^\alpha)G(x)\|_2 \le \frac{C}{r}\times r = C. \tag{A.12}$$

For the second and third terms,

$$\begin{aligned}
\|\partial_v G(x)\|_2 &= \Big\|v + \sum_i\partial_v\alpha_i(x)(x_i-x_1) + \big(\sum_i\partial_v\alpha_i(x)\big)x_1\Big\|_2\\
&\le 1 + \big\|\big(\partial_v\alpha_i(x)\big)_{i\in I_{x,r}}\big\|_2\big\|(2r)_{i\in I_{x,r}}\big\|_2\\
&\le 1 + \Big(\frac{C}{r}|I_{x,r}|^{-\frac{1}{2}}\Big)\times\Big(2r|I_{x,r}|^{\frac{1}{2}}\Big) = 1 + C,
\end{aligned}$$

and by (A.10) we obtain

$$\big\|(\partial_v\Psi_x^\alpha)\big(\partial_u G(x)\big)\big\|_2 \le C, \quad \big\|\big(\partial_u G(x)\big)(\partial_v\Psi_x^\alpha)\big\|_2 \le C.$$

For the fourth term, we have

$$\begin{aligned}
&\big\|\Psi_x^\alpha\big(\partial_v\partial_u G(x)\big)\big\|_2\\
\le&\Big\|\Psi_x^\alpha\sum_i(\partial_v\partial_u\alpha_i(x))x_i\Big\|_2\\
\le&\|\Psi_x^\alpha\|_2\sum_i|\partial_v\partial_u\alpha_i(x)|\|x_i-x_i^*\|_2 + \sum_i|\partial_v\partial_u\alpha_i(x)|\|\Psi_x^\alpha(x_i^*-x^*)\|_2 + 0\\
\le&\Big\|\big(\partial_v\partial_u\alpha_i(x)\big)_{i\in I_{x,r}}\Big\|_2\Big(\big\|(\|x_i-x_i^*\|)_{i\in I_{x,r}}\big\|_2 + \big\|(\|\Psi_x^\alpha(x_i^*-x^*)\|)_{i\in I_{x,r}}\big\|_2\Big)\\
\le&C\Big(\frac{1}{r^2}|I_{x,r}|^{-\frac{1}{2}}\Big)\times\Big((\sigma+r^2)|I_{x,r}|^{\frac{1}{2}}\Big) = C.
\end{aligned}$$

The above proof is based on Lemma 27 and Lemma 29, which are valid when Lemma 13 and Theorem 11 simultaneously hold. Hence, $\|\partial_v\partial_u f(x)\|_2 \le C$ when Lemma 13 and Theorem 11 simultaneously hold, whose probability is at least $\delta_0(1-\delta)^2$. ■

A.6 Proof of Proposition 20, Proposition 21 and Lemma 22

Proof of Proposition 20 Considering the function $\phi(d) = (1 - \frac{d}{r^2})^\beta$ for $t \geq 0$, whose derivative is $\phi'(d) = \frac{\beta}{r^2}\big(1 - \frac{d}{r^2}\big)^{\beta - 1}$, we obtain $|\phi'(d)| \leq \frac{\beta}{r^2}$. This implies

$$|\tilde{\alpha}_i(x) - \tilde{\alpha}_i(z)| \leq \frac{\beta}{r^2}\|x - z\|_2^2 \leq \frac{\beta}{r^2}\epsilon^2 \leq \frac{\alpha(x)}{|I_{x,2r}|^2}r \leq \frac{r}{|I_{x,2r}|},$$

where the last inequality holds since $\alpha(x) = \sum_{i \in I_{x,r}} \tilde{\alpha}_i(x) \leq \sum_{i \in I_{x,r}} 1 = |I_{x,r}| \leq |I_{x,2r}|$. For any $z \in B_D(x, \epsilon)$, we have $\|z - x\|_2 \leq r$ and $I_{z,r} \subset I_{x,2r}$. By the definition of $\tilde{\alpha}_i(z)$, we have $\tilde{\alpha}_i(z) = 0$ for $i \notin I_{z,r}$, and therefore

$$\begin{aligned}
\alpha(z) &= \sum_{i \in I_{z,r}} \tilde{\alpha}_i(z) = \sum_{i \in I_{x,2r}} \tilde{\alpha}_i(z) \\
&= \sum_{i \in \mathcal{I}_{x,2r}} \big(\tilde{\alpha}_i(x) + \tilde{\alpha}_i(z) - \tilde{\alpha}_i(x)\big) = \alpha(x) + \sum_{i \in \mathcal{I}_{x,2r}} \big(\tilde{\alpha}_j(z) - \tilde{\alpha}_j(x)\big).
\end{aligned}$$

Plug $\alpha(z)$ into the following denominator,

$$\begin{aligned}
|\alpha_i(z) - \alpha_i(x)| &= \Big|\frac{\tilde{\alpha}_i(z)}{\alpha(z)} - \frac{\tilde{\alpha}_i(x)}{\alpha(x)}\Big| \\
&\leq \max\left\{\frac{\tilde{\alpha}_i(x) \pm |\tilde{\alpha}_i(z) - \tilde{\alpha}_i(x)|}{\alpha(x) \mp \sum_{j \in I_{x,2r}} |\tilde{\alpha}_i(z) - \tilde{\alpha}_i(x)|} - \frac{\tilde{\alpha}_i(x)}{\alpha(x)}\right\} \\
&\leq \max\left\{\frac{\tilde{\alpha}_i(x) + \frac{\alpha(x)}{|I_{x,2r}|^2}r}{\alpha(x) - |I_{x,2r}|\frac{\alpha(x)}{|I_{x,2r}|^2}r} - \frac{\tilde{\alpha}_i(x)}{\alpha(x)}, \frac{\tilde{\alpha}_i(x)}{\alpha(x)} - \frac{\tilde{\alpha}_i(x) - \frac{\alpha(x)}{|I_{x,2r}|^2}r}{\alpha(x) + |I_{x,2r}|\frac{\alpha(x)}{|I_{x,2r}|^2}r}\right\} \\
&\leq \max\left\{\frac{\tilde{\alpha}_i(x) + \frac{r}{|I_{x,2r}|}}{\alpha(x) - \alpha(x)\frac{r}{|I_{x,2r}|}} - \frac{\tilde{\alpha}_i(x)}{\alpha(x)}, \frac{\tilde{\alpha}_i(x)}{\alpha(x)} - \frac{\tilde{\alpha}_i(x) - \frac{r}{|I_{x,2r}|}}{\alpha(x) + \alpha(x)\frac{r}{|I_{x,2r}|}}\right\} \\
&\leq \max\left\{\frac{\big(\tilde{\alpha}_i(x) + \frac{r}{|I_{x,2r}|}\big)(1 + C\frac{r}{|I_{x,2r}|}) - \tilde{\alpha}_i(x)}{\alpha(x)}, \frac{\tilde{\alpha}_i(x) - \big(\tilde{\alpha}_i(x) - \frac{r}{|I_{x,2r}|}\big)(1 - C\frac{r}{|I_{x,2r}|})}{\alpha(x)}\right\} \\
&\leq \frac{\tilde{\alpha}_i(x)r + Cr + Cr^2}{\alpha(x)|I_{x,2r}|} \leq \frac{r + Cr + Cr^2}{\alpha(x)|I_{x,2r}|} \leq \frac{r + Cr + Cr^2}{c_0|I_{x,2r}|} = C'\frac{r}{|I_{x,2r}|},
\end{aligned}$$

the second-to-last inequality holds since $\tilde{\alpha}_i(x) \leq 1$ while the last inequality holds with probability $1 - (1 - cr^d)^N$ by Proposition 12(ii).

Based on the upper bound of $|\alpha_i(x) - \alpha_i(z)|$, we obtain

$$\begin{aligned}
\|A_x - A_z\|_2 = \|\sum_{i \in I_{x,r}} \alpha_i(x)P_{x_i} - \sum_{i \in I_{z,r}} \alpha_i(z)P_{x_i}\|_2 &= \|\sum_{i \in I_{x,2r}} \alpha_i(x)P_{x_i} - \sum_{i \in I_{x,2r}} \alpha_i(z)P_{x_i}\|_2 \\
\leq \sum_{i \in I_{x,2r}} |\alpha_i(x) - \alpha_i(z)|\|P_{x_i}\|_2 &\leq \sum_{i \in I_{x,2r}} C'\frac{r}{|I_{x,2r}|} \cdot 1 = C'r.
\end{aligned}$$

Noting $\|\Psi_x^\alpha - A_x\|_2 \leq Cr$ with probability $\delta_0(1 - \delta)^2$ by (3.6) in Theorem 15, we have

$$\|\Psi_z^\alpha - A_z\|_2 \leq \|\Psi_x^\alpha - A_z\|_2 \leq \|\Psi_x^\alpha - A_x\|_2 + \|A_x - A_z\|_2 \leq Cr,$$

and hence $\|\Psi_x^\alpha - \Psi_z^\alpha\| \le \|\Psi_x^\alpha - A_z\|_2 + \|A_z - \Psi_z^\alpha\|_2 \le Cr$ with probability $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$, which completes this proof. ■

Proof of Proposition 21 We first prove (i). Since the rows of $J_f(x)$ are orthogonal to the contour surface at x, as the basis of the spanning space of $J_f(x)^T$, W_x is also the basis of the normal space of $\mathcal{M}_{\text{out}}$ at x and thereby $W_x \in \mathbb{R}^{D\times(D-d)}$ by Theorem 4. This implies

$$\mathbb{P}\big(\text{statement (i) holds}\big) = \mathbb{P}\big(W_x \in \mathbb{R}^{D\times(D-d)}\big) \ge \mathbb{P}\big(\text{Theorem 4 holds}\big) \tag{A.13}$$

Now we proceed to prove (ii). It is clear that $g(z) = \mathbf{0}$ if $f(z) = \mathbf{0}$. Thus, we only need to prove that $g(z) = \mathbf{0}$ implies $f(z) = \mathbf{0}$. To do this, we first assume the reverse, $f(z) \ne 0$ and $g(z) = W_x^T f(z) = \mathbf{0}$. Since W_x^T is the basis of $\operatorname{span}\big(J_f(x)^T\big)$, $J_f(x)$ can be rewritten as $J_f(x) = YW_x^T$ and $J_f(x)f(z) = Y\big(W_x^T f(z)\big) = Yg(z) = \mathbf{0}$. By the definition of $f(z)$ in equality (2.6), $\Psi_z^\alpha f(z) = f(z)$. Hence, we obtain

$$\|J_f(x) - \Psi_z^\alpha\|_2 = \max_{v\ne 0} \frac{\big\|\big(J_f(x) - \Psi_z^\alpha\big)v\big\|_2}{\|v\|_2} \ge \frac{\big\|\big(J_f(x) - \Psi_z^\alpha\big)f(z)\big\|_2}{\|f(z)\|_2} = \frac{\|0 - f(z)\|_2}{\|f(z)\|_2} = 1.$$

However,

$$\|J_f(x) - \Psi_z^\alpha\|_2 \le \|J_f(x) - \Psi_x^\alpha\|_2 + \|\Psi_x^\alpha - \Pi_{x^*}\|_F + \|\Pi_{x^*} - \Pi_{z^*}\|_F + \|\Pi_{z^*} - \Psi_z^\alpha\|_F \le Cr$$

where the first term is bounded by (3.8) in Corollary 18, the second and fourth terms are bounded by applying Theorem 15 for x and z, respectively, and the third term is bounded by Lemma 14. We conduct contradictory bounds of $\|J_f(x) - \Psi_z^\alpha\|_2$. Hence, $f(z) = \mathbf{0}$ if $g(z) = \mathbf{0}$. The statement (ii) is proved when Corollary 18, Theorem 15 and Lemma 14 hold simultaneously. Noticing that Corollary 18 holds when Lemma 13 and Theorem 11 hold, Theorem 4 holds when Theorem 15 and Proposition 12(ii) hold, and Theorem 15 holds when Lemma 13 and Theorem 11 hold, we obtain

$$\begin{aligned}
&\mathbb{P}\big(\text{statement (i) and (ii) hold}\big)\\
&\ge \mathbb{P}\Big(\big(\text{Theorem 4 and Corollary 18 hold for } x\big) \cap \big(\text{ Theorem 15 holds for } x, z\big)\Big)\\
&\ge \mathbb{P}\big(\text{Proposition 12(ii) holds for } x\big)\mathbb{P}\big(\text{Lemma 13 and Theorem 11 hold for } x, z\big)\\
&\ge \delta_0^2(1-\delta)^4\big(1-(1-cr^d)^N\big).
\end{aligned}$$

The proof is, therefore, complete. ■

Proposition 30. *Letting $\sigma_1 \ge \cdots \ge \sigma_D$ be the singular values of $J_f(x)$, then with probability at least $\delta_0(1-\delta)^2$,*

$$1 + O(r) \ge \sigma_1 \ge \sigma_{D-d} \ge 1 - O(r).$$

Proof Let $\Psi_x^\alpha = V_x V_x^T$ and $J_f(x) = U_x \Sigma_x W_x^T$ be the thin singular value decomposition of $J_f(x)$, where $U_x, W_x \in \mathbb{R}^{D\times(D-d)}$ and $\Sigma_x \in \mathbb{R}^{(D-d)\times(D-d)}$ by Theorem 4.

To begin with, we bound σ_{D-d} below. Let $S_1 = \operatorname{span}(V_x)$ and $S_2 = \operatorname{span}\{w_1, \cdots w_{D-d-1}\}$, where $w_1, \cdots w_{D-d-1}$ are the first $(D-d-1)$ columns of W_x. Since $\dim(S_1) > \dim(S_2)$, there exists $\eta \neq 0 \in S_1 \cap S_2^{\perp}$, which implies $\Psi_x^{\alpha}\eta = \eta$ and $w_i^T\eta = 0$ for $i = 1, \cdots, D-d-1$. Hence,

$$\big(\Psi_x^{\alpha} - J_f(x)\big)\eta = \eta - U_x\Sigma_x W_x^T\eta = \eta - u_{D-d}\sigma_{D-d}w_{D-d}^T\eta,$$

where u_{D-d} is the $(D-d)$-th column of U_x. This leads to

$$\begin{aligned}\|\big(\Psi_x^{\alpha} - J_f(x)\big)\eta\|_2 &= \|\eta - u_{D-d}\sigma_{D-d}w_{D-d}^T\eta\|_2\\ &\geq \Big|\|\eta\|_2 - \|u_{D-d}\sigma_{D-d}w_{D-d}^T\eta\|_2\Big| = |1-\sigma_{D-d}|\|\eta\|_2.\end{aligned}$$

We obtain

$$Cr \geq \|\Psi_x^{\alpha} - J_f(x)\|_2 \geq \frac{\Big\|\big(\Psi_x^{\alpha} - J_f(x)\big)\eta\Big\|_2}{\|\eta\|_2} = |1-\sigma_{D-d}|,$$

where the first inequality holds by (3.8) in Corollary 18. So, $\sigma_{D-d} \geq 1 - O(r)$.

Now, we turn to the upper bound of σ_1. Let $\eta = w_1$, then $\|\eta\|_2 = 1$ and $w_i^T\eta = 0$ for any $i \geq 2$. Hence,

$$\big(\Psi_x^{\alpha} - J_f(x)\big)\eta = \Psi_x^{\alpha}\eta - U_x\Sigma_x W_x^T\eta = \Psi_x^{\alpha}\eta - \sigma_1 u_1.$$

This leads to

$$Cr \geq \|\Psi_x^{\alpha} - J_f(x)\|_2 \geq \Big\|\big(\Psi_x^{\alpha} - J_f(x)\big)\eta\Big\|_2 = \|\Psi_x^{\alpha}\eta - \sigma_1 u_1\|_2 \geq \big|\|\Psi_x^{\alpha}\eta\|_2 - \sigma_1\big|.$$

So, $\sigma_1 \leq \|\Psi_x^{\alpha}\eta\|_2 + Cr \leq 1 + Cr$. Note that the above proof relies on Theorem 4 and Corollary 18, where Theorem 4 and Corollary 18 hold when Lemma 13, Theorem 11 and Proposition 12(ii) hold. This proof is completed with probability at least $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$. ■

Proposition 31. *$\|W_x^T\Psi_x^{\alpha}W_x^T - I_{D-d}\|_2 \leq Cr$ with probability at least $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$.*

Proof By Theorem 4, we obtain $W_x \in \mathbb{R}^{D\times(D-d)}$. Let the singular value decomposition of $W_x^T V_x = \sum_{i=1}^{D-d} s_i a_i b_i^T$, where s_i is the i-th singular value of $W_x^T V_x$ and a_i and b_i are the singular vectors corresponding to s_i. Let $\eta = V_x b_{D-d}$, then

$$\begin{aligned}Cr \geq \|\Psi_x^{\alpha} - J_f(x)\|_2 \geq \Big\|\big(\Psi_x^{\alpha} - J_f(x)\big)\eta\|_2 &= \|V_x b_{D-d} - U_x\Sigma_x W_x^T V_x b_{D-d}\|_2\\ &\big|1 - \|U_x\Sigma_x(\sum_{i=1}^{D-d} s_i a_i b_i^T)b_{D-d}\|_2\big| = \big|1 - s_{D-d}\|\Sigma_x a_{D-d}\|_2\big|,\end{aligned}$$

where the first inequality holds by (3.8) in Corollary 18. This leads to

$$\frac{1-Cr}{\|\Sigma_x a_{D-d}\|_2} \leq s_{D-d} \leq \frac{1+Cr}{\|\Sigma_x a_{D-d}\|_2}.$$

Noticing $1-O(r) \le \|\sigma_x a_{D-d}\|_2 \le 1+O(r)$ by Proposition 30, we conclude $1-O(r) \le s_{D-d} \le s_1 \le 1$ since $\|W_x^T V_x\|_2 \le 1$. So,

$$\begin{aligned}\|W_x^T \Psi_x^\alpha W_x - I_{D-d}\|_2 &= \|W_x^T V_x V_x^T W_x - I_{D-d}\|_2 \\ &= \|ASS^TA^T - AA^T\|_2 = \|A(SS^T - I_{D-d})A^T\|_2 \le Cr,\end{aligned}$$

where $A = [a_1, \cdots, a_{D-d}]$ and S is a diagonal matrix with $(s_1, \cdots, s_{D-d})$ as the diagonal entries. Note that the above proof relies on Theorem 4 and Corollary 18, where Theorem 4 and Corollary 18 hold when Lemma 13, Theorem 11 and Proposition 12(ii) hold. This proof is completed with probability at least $\delta_0(1-\delta)^2\big(1-(1-cr^d)^N\big)$. ■

Proof of Lemma 22 Under the settings that the first d coordinates are the basis of $T_x\mathcal{M}_{\text{out}}$, and the last $D-d$ coordinates are the columns of W_x, W_x can be rewritten as $W_x = (\mathbf{0}, I_{D-d})^T$. Hence, we obtain

$$\begin{aligned}&J_g(z)(\mathbf{0}, I_{D-d})^T = W_x^T J_f(z) W_x \\ &=W_x^T\big(J_f(z) - J_f(x)\big)W_x + W_x^T\big(J_f(x) - \Psi_x^\alpha\big)W_x + \big(W_x^T\Psi_x^\alpha W_x - I_{D-d}\big) + I_{D-d}.\end{aligned}$$

This leads to

$$\begin{aligned}&\|J_g(z)(\mathbf{0}, I_{D-d})^T - I_{D-d}\|_2 \\ &\qquad \le \|J_f(z) - J_f(x)\|_2 + \|J_f(x) - \Psi_x^\alpha\|_2 + \|W_x^T\Psi_x^\alpha W_x - I_{D-d}\|_2 \\ &\qquad \le C_1 r + C_2 r + C_3 r \le Cr,\end{aligned}$$

where $\|J_f(z) - J_f(x)\|_2 \le \|J_f(z) - J_f(x)\|_F \le C_1 r$ by (4.2) in Theorem 5, $\|J_f(x) - \Psi_x^\alpha\|_2 \le \|J_f(x) - \Psi_x^\alpha\|_F \le C_2 r$ by Corollary 18 and $\|W_x^T\Psi_x^\alpha W_x - I_{D-d}\|_2 \le C_3 r$ by Proposition 31. Using Theorem 2.9.10 (the implicit function theorem) in Hubbard and Hubbard (2001), ϕ exits. Carrying out the first derivative on $g\big(\zeta, \phi(\zeta)\big) = \mathbf{0}$, we obtain

$$\begin{aligned}\mathbf{0} = \partial_s g(\zeta, \phi(\zeta)) &= J_g(\zeta, \phi(\zeta))\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix} \\ &= W_x^T\Big(J_f(\zeta, \phi(\zeta)) - J_f(x)\Big)\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix} + W_x^T J_f(x)\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix} \\ &= W_x^T\Big(J_f(\zeta, \phi(\zeta)) - J_f(x)\Big)\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix} + W_x^T U_x \Sigma_x(\mathbf{0}, I_{D-d})\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix}.\end{aligned}$$

This implies that

$$\partial_s\phi(\zeta) = -\Sigma_x^{-1}\big(W_x^T U_x\big)^{-1} W_x^T\Big(J_f(\zeta, \phi(\zeta)) - J_f(x)\Big)\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix}.$$

Calculating ℓ_2-norm of the two sides of the above equality, we obtain

$$\begin{aligned}\|\partial_s\phi(\zeta)\|_2 &= \Big\|\Sigma_x^{-1}\big(W_x^T U_x\big)^{-1} W_x^T\Big(J_f(\zeta, \phi(\zeta)) - J_f(x)\Big)\begin{pmatrix}\partial_s\zeta \\ \partial_s\phi(\zeta)\end{pmatrix}\Big\|_2 \\ &\le \big(1+O(r)\big)C\Big\|J_f(\zeta, \phi(\zeta)) - J_f(x)\Big\|_2 \le C\|(\zeta, \phi(\zeta)) - x\|_2\end{aligned}$$

Carrying out the second derivative on $g(\zeta, \phi(\zeta)) = \mathbf{0}$, we obtain

$$\mathbf{0} = \partial_t J_g(\zeta, \phi(\zeta)) \left(\begin{array}{c} \partial_s \zeta \\ \partial_s \phi(\zeta) \end{array} \right) + J_g(\zeta, \phi(\zeta)) \left(\begin{array}{c} \mathbf{0} \\ \partial_t \partial_s \phi(\zeta) \end{array} \right).$$

Letting e_i denote the i-th column of I_D and

$$u = \left(\begin{array}{c} \partial_t \zeta \\ \partial_t \phi(\zeta) \end{array} \right),$$

the i-th column of $\partial_t J_g(\zeta, \phi(\zeta))$ is

$$\partial_t \partial_{e_i} g(\zeta, \phi(\zeta)) = \|u\|_2 \partial_{\frac{u}{\|u\|_2}} \partial_{e_i} g(\zeta, \phi(\zeta)) = \|u\|_2 W_x^T \partial_{\frac{u}{\|u\|_2}} \partial_{e_i} f(\zeta, \phi(\zeta)).$$

In conjunction with $\|\partial_{\frac{u}{\|u\|_2}} \partial_{e_i} f(\zeta, \phi(\zeta))\|_2 \le C$, as proved in Theorem 19, $\|\partial_t \partial_{e_i} g(\zeta, \phi(\zeta))\| \le C$, and therefore

$$\|\partial_t J_g(\zeta, \phi(\zeta)\|_2 \le C.$$

Hence,

$$\partial_t \partial_s \phi(\zeta) = -\Sigma_x^{-1} \big(W_x^T U_x\big)^{-1} \left(\partial_t J_g(\zeta, \phi(\zeta)) \left(\begin{array}{c} \partial_s \zeta \\ \partial_s \phi(\zeta) \end{array} \right) + W_x^T \Big(J_f(\zeta, \phi(\zeta)) - J_f(x) \Big) \left(\begin{array}{c} \partial_s \zeta \\ \partial_s \phi(\zeta) \end{array} \right) \right),$$

which implies

$$\|\partial_t \partial_s \phi(\zeta)\|_2 \le C(1 + O(r)) \big(C + C\|z - x\|_2\big) \le C.$$

Note that the above proof relies on Theorem 5, Corollary 18, Proposition 31 and Theorem 19, which are valid when Lemma 13, Theorem 11 and Proposition 12(ii) hold. Hence, this proof is completed with probability at least $\delta_0(1-\delta)^2\big(1 - (1 - cr^d)^N\big)$. ■

Appendix B. Gradient of $\|f(x)\|_2^2$

Let $F(x) = \|f(x)\|_2^2$, $d\cdot$ denote the differential and $G(x) = x - \sum_{i \in I_{x,r}} \alpha_i(x) x_i$, then

$$\begin{aligned} dF(x) &= 2\langle f(x), df(x) \rangle = 2\big\langle \Psi_x^\alpha G(x), d\big(\Psi_x^\alpha G(x)\big)\big\rangle \\ &= 2\langle \Psi_x^\alpha G(x) G(x)^T, d\Psi_x^\alpha \rangle + 2\langle \Psi_x^\alpha G(x), dG(x) \rangle \\ &= 2\langle \Psi_x^\alpha G(x) G(x)^T, d\Psi_x^\alpha \rangle + 2\langle \Psi_x^\alpha G(x), dx - \sum_{i \in I_{x,r}} \big(d\alpha_i(x) x_i\big) \rangle, \end{aligned}$$

where

$$\begin{aligned} d\alpha_i(x) &= \frac{d\tilde{\alpha}_i(x)}{\alpha(x)} - \frac{\tilde{\alpha}_i(x) d\alpha(x)}{\alpha(x)^2} = \frac{d\tilde{\alpha}_i(x)}{\alpha(x)} - \frac{\alpha_i(x) d\alpha(x)}{\alpha(x)} \\ &= -\frac{2(d+2)}{r^2 \alpha(x)} \langle \tilde{\alpha}_i(x)^{\frac{d+1}{d+2}} (x - x_i) - \alpha_i(x) \sum_i \tilde{\alpha}_i(x)^{\frac{d+1}{d+2}} (x - x_i), dx \rangle \\ &:= \langle \frac{d\alpha_i(x)}{dx}, dx \rangle \end{aligned}$$

and $d\Psi_x^\alpha$ can be calculated as below. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A_x and $\mu_1 > \cdots > \mu_s$ are the different values of $\{\lambda_i\}$. Suppose $\lambda_{n-d} > \lambda_{n-d+1}$, and $\mu_1 > \cdots > \mu_t$ are the different values of $\lambda_1 \geq \cdots \geq \lambda_{n-d}$. $P_{i,x} = V_{i,x}V_{i,x}^T$ is an orthogonal projection and columns of $V_{i,x}$ are the eigenvectors corresponding to μ_i. Then, we have $\Psi_x^\alpha = \sum_{i=1}^t P_{i,x}$. By Shapiro and Fan (1995),

$$dP_{i,x} = \sum_{j=1}^{s} \frac{1}{\mu_j - \mu_i} P_{i,x}(dA_x)P_{j,x} + P_{j,x}(dA_x)P_{i,x},$$

and thereby

$$d\Psi_x^\alpha = \sum_{i=1}^{t} dP_{i,x} = \sum_{i=1}^{t}\sum_{j=t+1}^{s} \frac{1}{\mu_j - \mu_i} P_{i,x}(dA_x)P_{j,x} + P_{j,x}(dA_x)P_{i,x}$$

Plug $d\Psi_x^\alpha$ into the first term of $dF(x)$,

$$\big\langle \Psi_x^\alpha G(x)G(x)^T, d\Psi_x^\alpha \big\rangle = \langle T, dA_x\rangle = \sum_{i\in I_{x,r}} \langle T, P_{x_i}\rangle\langle \frac{d\alpha_i(x)}{dx}, dx\rangle,$$

where $T = \sum_{i=1}^t \sum_{j=t+1}^s \frac{1}{\mu_j-\mu_i} P_{i,x}\big(\Psi_x^\alpha G(x)G(x)^T\big)P_{j,x} + P_{j,x}\big(\Psi_x^\alpha G(x)G(x)^T\big)P_{i,x}$. Plugging $d\alpha_i(x)$ into the second term of $dF(x)$, we obtain

$$\langle \Psi_x^\alpha G(x), dx - \sum_{i\in I_{x,r}} \big(d\alpha_i(x)x_i\big)\rangle = \langle \Psi_x^\alpha G(x), dx\rangle - \sum_{i\in I_{x,r}} \langle \Psi_x^\alpha G(x), x_i\rangle\langle \frac{d\alpha_i(x)}{dx}, dx\rangle.$$

As the summation of the first and second term,

$$dF(x) = \Big\langle 2\sum_{i\in I_{x,r}} \big(\langle T, P_{x_i}\rangle + \langle \Psi_x^\alpha G(x), x_i\rangle\big)\frac{d\alpha_i(x)}{dx} + \Psi_x^\alpha G(x), dx \Big\rangle$$

So the gradient of $F(x)$ is

$$\mathrm{grad}(x) = 2\sum_{i\in I_{x,r}} \big(\langle T, P_{x_i}\rangle + \langle \Psi_x^\alpha G(x), x_i\rangle\big)\frac{d\alpha_i(x)}{dx} + \Psi_x^\alpha G(x). \tag{B.1}$$

Appendix C. Results of Facial Image Denoising

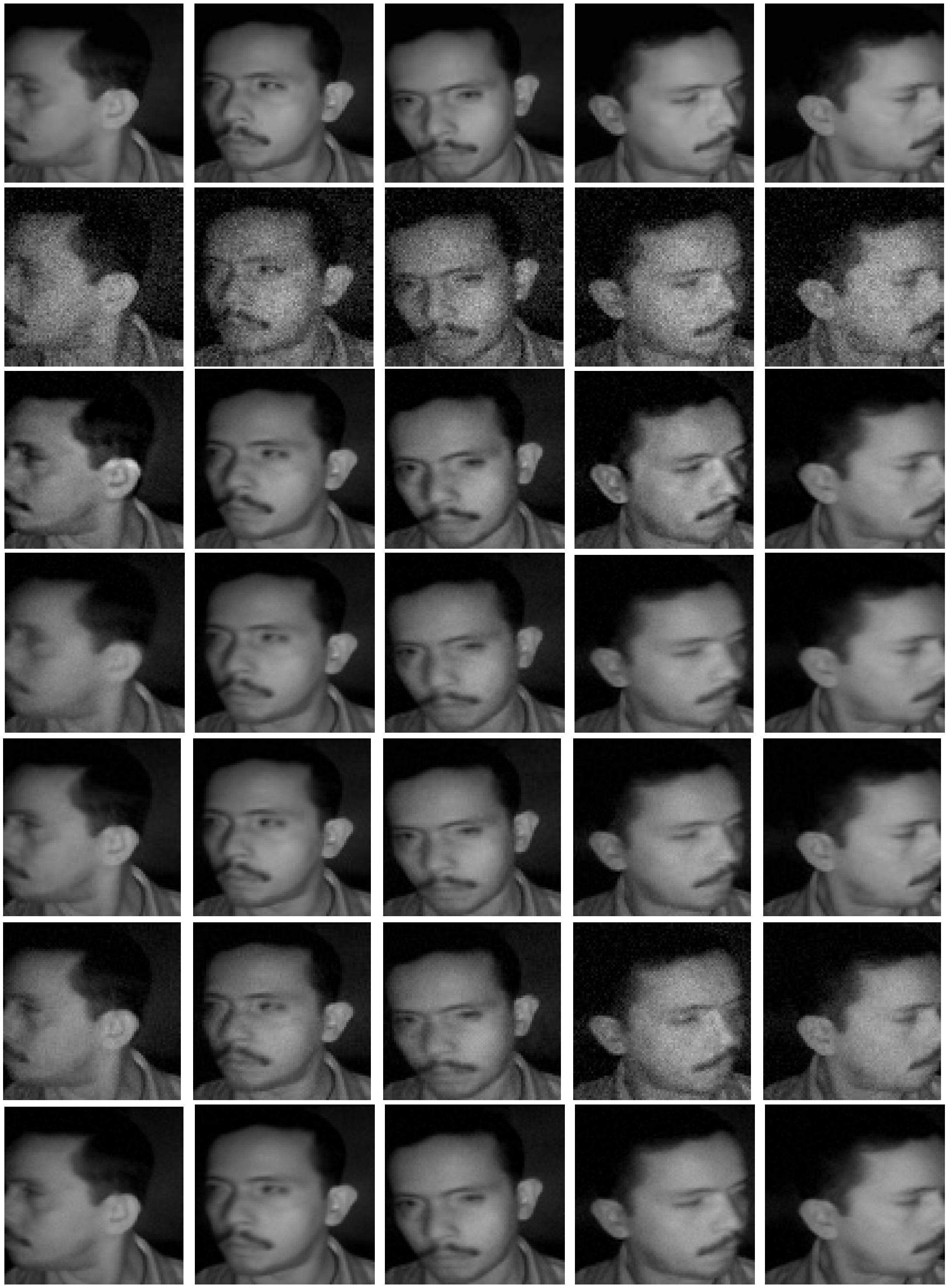

Figure 10: Performance of facial image denoising with $\rho = 0.2$. The first row consists of original images while the second row features blurred images. The third to seventh rows contain deblurred images using km17, cf18, ya21(deg=1), ya21(deg=2) and our method, respectively.

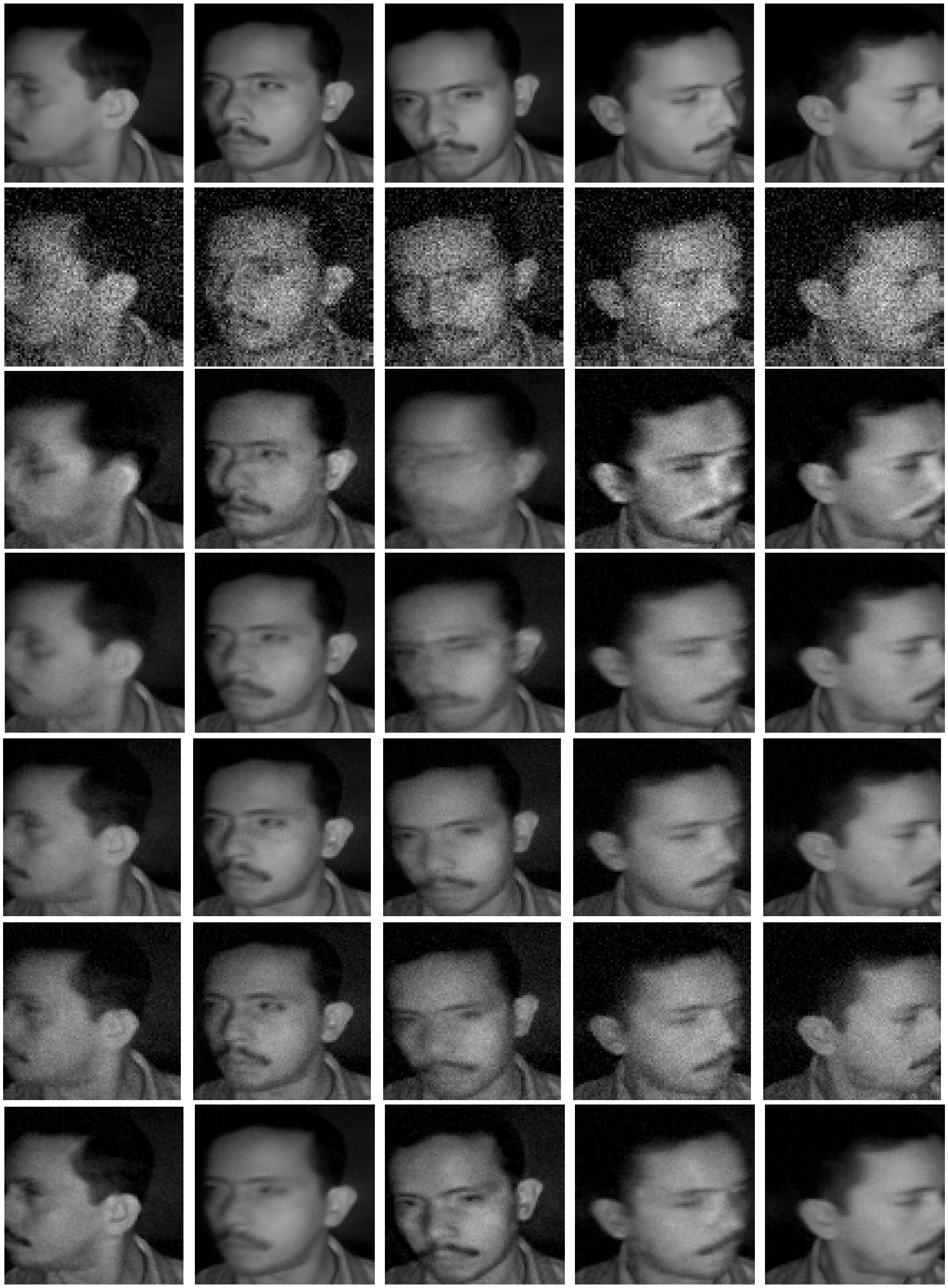

Figure 11: Performance of facial image denoising with $\rho = 0.4$. The first row consists of original images while the second row again shows blurred images. The third to seventh rows contain deblurred images using km17, cf18, ya21(deg=1), ya21(deg=2) and our method, respectively.

References

Eddie Aamari and Clément Levrard. Stability and minimax optimality of tangential delaunay complexes for manifold reconstruction. Discrete & Computational Geometry, 59:923–971, 2018. ISSN 1432-0444. doi: 10.1007/s00454-017-9962-z. URL https://doi.org/10.1007/s00454-017-9962-z.

Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. Annals of Statistics, 47:177–204, 2 2019. ISSN 00905364. doi: 10.1214/18-AOS1685.

Yariv Aizenbud and Barak Sober. Non-parametric estimation of manifolds from noisy data. 5 2021. URL http://arxiv.org/abs/2105.04754.

Jeffrey D Banfield and Adrian E Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association, 87(417):7–16, 1992.

Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential delaunay complexes. Discrete & Computational Geometry, 51:221–267, 2014. ISSN 1432-0444. doi: 10.1007/s00454-013-9557-2. URL https://doi.org/10.1007/s00454-013-9557-2.

Jean-Daniel Boissonnat, André Lieutier, and Mathijs Wintraecken. The Reach, Metric Distortion, Geodesic Convexity and the Variation of Tangent Spaces. In 34th International Symposium on Computational Geometry (SoCG 2018), volume 99, pages 10:1–10:14, 2018. doi: 10.4230/LIPIcs.SoCG.2018.10.

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

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. Journal of the American Mathematical Society, 29(4):983–1049, 2016.

Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Fitting a putative manifold to noisy data. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 688–720. PMLR, 06–09 Jul 2018.

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. Journal of machine learning research, 13(May):1263–1291, 2012a.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. The geometry of nonparametric filament estimation. Journal of the American Statistical Association, 107(498):788–799, 2012b.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Manifold estimation and singular deconvolution under hausdorff loss. The Annals of Statistics, 40(2):941–963, 2012c.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Nonparametric ridge estimation. The Annals of Statistics, 42(4):1511–1545, 2014.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intelligence, 23(6):643–660, 2001.

Dian Gong, Fei Sha, and Gérard Medioni. Locally linear denoising on image manifolds. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 265–272, 2010.

SL Happy, Anirban Dasgupta, Anjith George, and Aurobinda Routray. A video database of human faces under near infra-red illumination for human computer interaction applications. In 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), pages 1–4. IEEE, 2012.

Trevor Hastie and Werner Stuetzle. Principal curves. Journal of the American Statistical Association, 84(406):502–516, 1989.

John Hubbard and Barbara Hubbard. Vector analysis, linear algebra, and differential forms: A unified approach. Ithaca: Matrix Editions, 2001.

Yunqian Ma and Yun Fu. Manifold learning theory and applications. CRC press, 2011.

Kitty Mohammed and Hariharan Narayanan. Manifold learning using kernel density estimation and local principal components analysis. arXiv preprint arXiv:1709.03615, 2017.

Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). 1996.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. Discrete & Computational Geometry, 39(1-3):419–441, 2008.

Umut Ozertem and Deniz Erdogmus. Locally defined principal curves and surfaces. Journal of Machine learning research, 12(Apr):1249–1286, 2011.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

Alexander Shapiro and Michael Fan. On eigenvalue optimization. SIAM Journal on Optimization, 5(3):552–569, 1995. URL `https://doi.org/10.1137/0805028`.

Derek C Stanford and Adrian E Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(6):601–609, 2000.

Jakob J Verbeek, Nikos Vlassis, and B Kröse. A k-segments algorithm for finding principal curves. Pattern Recognition Letters, 23(8):1009–1017, 2002.

Zhigang Yao and Zhenyue Zhang. Principal Boundary on Riemannian Manifolds. Journal of the American Statistical Association, 115:1435–1448, 2020.

Zhigang Yao, Yuqing Xia, and Zengyan Fan. Random Fixed Boundary Flows. Journal of the American Statistical Association, 0(0):1–13, 2023. URL https://doi.org/10.1080/01621459.2023.2257892.