Tangent Estimation from Point Samples*

Siu-Wing Cheng† Man-Kwun Chiu‡§

Abstract

Let $\mathcal{M}$ be an m-dimensional smooth compact manifold embedded in $\mathbb{R}^d$, where m is a constant known to us. Suppose that a dense set of points are sampled from $\mathcal{M}$ according to a Poisson process with an unknown parameter. Let p be any sample point, let ϱ be the local feature size at p, and let $\varrho\varepsilon$ be the distance from p to the $(n+1)$th nearest sample point for some n between $\binom{m+1}{2}+1$ and $\binom{d+1}{2}$. Using the n sample points nearest to p, we can estimate the tangent space at p and it holds with probability $1-O(n^{-1/3})$ that the angular error is $O(\varepsilon^2)$. The running time is bounded by the time to compute the thin SVD of an $n\times\binom{d+1}{2}$ matrix and the full SVD of an $n\times d$ matrix, which is usually $O(d^2n^2)$ in practice. We implemented the algorithm and experimentally verified its effectiveness on both noiseless and noisy data.

*Research supported by the Research Grant Council, Hong Kong, China (project no. 612109). Part of the work was done while Chiu was at HKUST.

†Dept. of Computer Science and Engineering, HKUST, Hong Kong.

‡National Institute of Informatics (NII), Tokyo, Japan.

§JST, ERATO, Kawarabayashi Large Graph Project.

1 Introduction

Data points corresponding to experimental observations commonly reside in $\mathbb{R}^d$ for some large d, but it is often postulated that the data points live on an unknown manifold $\mathcal{M}$ of much lower dimension. Indeed, manifold learning has been applied in various problems such as network anomaly detection, image segmentation, and object tracking in video [6, 29]. The goal is to learn the manifold structure from sample points, including the intrinsic dimension, tangent spaces, and ultimately a faithful reconstruction. Theoretical algorithms have been developed to obtain faithful reconstructions [3, 4, 9], but their practical performance is unclear. We reexamine the key tasks in the problem to put our result in context.

The first task is to detect the manifold dimension. Many effective methods have already been developed in machine learning (e.g. [21, 22, 25, 27, 33]) and computational geometry [8, 10, 11, 15].

The second task is to estimate the tangent space at the sample points. Approximate tangent spaces at the sample points are needed in [9] to compute the cocone complex, which becomes a faithful reconstruction after removing slivers. Approximate tangent spaces at the sample points are also needed in [3] in order to form the tangential Delaunay complex, from which a faithful reconstruction is extracted after sliver removal. Tangent space estimation also finds application in clustering data points from multiple manifolds that may intersect each other [14, 18]. The tangent space estimation at a sample point p has been explicitly or implicitly performed in many previous work by fitting an affine subspace to the sample points in a neighborhood of p [1, 2, 10, 15, 26, 30, 31, 33]. An alternative method is based on analyzing the Voronoi cell of p [11]. The error measure for tangent space estimation is the angular error, which is the maximum angle between a vector in the true tangent space at p and the projection of that vector in the estimated tangent space. Bounds on the angular error have been proved (explicitly or implicitly) for the methods in [2, 10, 11, 15, 26], and the emphasis is on how these bounds depend on the sampling density. The radii of the neighborhoods used by the methods in [2, 10, 11, 15] for tangent estimation decrease as the sampling density increases, and their angular error bounds are *linear* in the ratio of the neighborhood radius to the local feature size at p.

Let m be the dimension of the unknown manifold $\mathcal{M}$. We present a method to estimate the tangent space at a sample point p using the sample points in a local neighborhood. Let ϱ be the local feature size of $\mathcal{M}$ at p. Let r be the neighborhood radius so that all sample points at distance less than r from p are used in the tangent space estimation. Our method gives an angular error bound of $O((r/\varrho)^2)$ radians with high probability, provided that $\mathcal{M}$ is smooth and compact, the sample points are drawn from $\mathcal{M}$ according to a Poisson process with an unknown parameter, and the manifold dimension m is a constant known to us.[1] Within a neighborhood of radius r from p, the tangent space can rotate by at most $O(r/\varrho)$ radians [9], where ϱ is the local feature size at p, and the turning angle may sometimes be $\Omega(r/\varrho)$; for example, when $\mathcal{M}$ is the unit sphere $\mathbb{S}^m$. An angular error bound linear in r/ϱ is thus asymptotically as good as taking the tangent space at an arbitrary sample point in the neighborhood as the approximation. We do better as our angular error bound is $O((r/\varrho)^2)$. We elaborate on our result and compare it with previous works that provide angular error bounds in Section 1.2.

[1] Since our method reduces to solving an eigenvalue problem, an appropriate thresholding of the eigenvalues should determine m. We do not pursue automatic dimension detection in this article in order to focus on the tangent estimation. We comment on the determination of m further in the conclusion.

1.1 Notation

An uppercase letter in `mathsf` font denotes a matrix and the corresponding italic lowercase letter with subscripts denotes the matrix elements. For example, Z denotes a matrix; z_{ij} denotes the (i, j) entry of Z; z_{i*} and z_{*j} denote the ith row vector and the jth column vector of Z, respectively. Similarly, v denotes a vector and v_i denotes the ith coordinate of v. Z^t and v^t denote the transposes. We use I_j to denote a $j \times j$ identity matrix, $\mathsf{0}_{i,j}$ an $i \times j$ zero matrix, and $\text{diag}_j(\sigma_1, \sigma_2, \ldots, \sigma_j)$ a $j \times j$ diagonal matrix with entries $\sigma_1, \ldots, \sigma_j$ in this order. We reserve $\mathsf{0}$ to denote the origin of $\mathbb{R}^d$.

The 2-norms of v and Z are $\|\mathsf{v}\| = \left(\sum_i v_i^2\right)^{1/2}$ and $\|\mathsf{Z}\| = \max\{\, \|\mathsf{Z}\mathsf{v}\| : \|\mathsf{v}\| = 1 \,\}$. If Z is symmetric, $\|\mathsf{Z}\|$ also equals $\max\left\{\, |\mathsf{v}^t\mathsf{Z}\mathsf{v}| : \|\mathsf{v}\| = 1 \,\right\}$. The Frobenius norm of Z is $\|\mathsf{Z}\|_F = \left(\sum_i \sum_j \; z_{ij}^2\right)^{1/2}$. It is known that $\|\mathsf{Z}\| \leq \|\mathsf{Z}\|_F \leq \sqrt{k}\,\|\mathsf{Z}\|$, where k is the number of rows or columns in Z, whichever is smaller [17].

Given a square matrix Z , a vector v is an eigenvector of Z if and only if $\mathsf{Z}\mathsf{v} = \lambda\mathsf{v}$ for some $\lambda \in \mathbb{R}$, and λ is known as an eigenvalue of Z. If Z has dimension k, then Z has at most k real eigenvalues. If Z is symmetric as well, it has k real eigenvalues.

The *thin* singular value decomposition (thin SVD) of a $k \times l$ matrix Z, $k \leq l$, is a product $\mathsf{L}\mathsf{D}\mathsf{R}^t$, where L is a $k \times k$ matrix consisting of unit eigenvectors of $\mathsf{Z}\mathsf{Z}^t$, D is a $k \times k$ diagonal matrix consisting of the singular values of Z (i.e., square roots of the eigenvalues of $\mathsf{Z}\mathsf{Z}^t$), and R is an $l \times k$ matrix formed by k of the unit eigenvectors of $\mathsf{Z}^t\mathsf{Z}$ corresponding to the k largest eigenvalues. We assume that the singular values of Z are in descending order on the diagonal of D. The *full* SVD of the same matrix Z is $\mathsf{L}\,(\mathsf{D} \;\; \mathsf{0}_{k,l-k})\,\bar{\mathsf{R}}^t$, where $\bar{\mathsf{R}}$ is an $l \times l$ matrix formed by the l unit eigenvectors of $\mathsf{Z}^t\mathsf{Z}$ and R is the leftmost $l \times k$ submatrix of $\bar{\mathsf{R}}$.

Given a diagonal square matrix D, its *pseudoinverse* $\mathsf{D}^\dagger$ is obtained by replacing each non-zero entry by its reciprocal and leaving the zero entries in place. The pseudoinverse of a general matrix Z with thin SVD $\mathsf{L}\mathsf{D}\mathsf{R}^t$ and full SVD $\mathsf{L}\,(\mathsf{D} \;\; \mathsf{0}_{k,l-k})\,\bar{\mathsf{R}}^t$ is $\mathsf{Z}^\dagger = \mathsf{R}\mathsf{D}^\dagger\mathsf{L}^t = \bar{\mathsf{R}}\,(\mathsf{D}^\dagger \;\; \mathsf{0}_{k,l-k})^t\,\mathsf{L}^t$. When Z is square and invertible, $\mathsf{Z}^\dagger$ is just Z^{-1}.

The largest singular value of Z is equal to $\|\mathsf{Z}\|$. The positive singular values of $\mathsf{Z}^\dagger$ are the reciprocals of the positive singular values of Z. Therefore, $\|\mathsf{Z}^\dagger\|$ is the reciprocal of the smallest positive singular value of Z.

Let $x_1, x_2, \ldots, x_d$ be a fixed set of orthogonal axes throughout this paper, forming the default coordinate system of $\mathbb{R}^d$. The coordinates of the input sample points are expressed with respect to this coordinate system.

We are given a set of sample points drawn from $\mathcal{M}$ according to a Poisson process with an unknown parameter λ: (i) for any compact subset B of $\mathcal{M}$, the probability that there are k points in B is $\frac{\lambda^k \,\text{vol}(B)^k}{k!} e^{-\lambda \,\text{vol}(B)}$, and (ii) for any disjoint compact subsets $B_1, \ldots, B_j$ of $\mathcal{M}$, the probability that there are k_i points in B_i for $i \in [1, j]$ is $\prod_{i=1}^{j} \frac{\lambda^{k_i} \,\text{vol}(B_i)^{k_i}}{(k_i)!} e^{-\lambda \,\text{vol}(B_i)}$. Given such a Poisson process and on the condition that there are k sample points in a compact subset $B \subset \mathcal{M}$, these k sample points are uniformly distributed in B [5].

By translation, we assume without loss of generality that the origin is a sample point. Let $\mathcal{T}$ denote the tangent space of $\mathcal{M}$ at the origin, which is an m-dimensional vector space in $\mathbb{R}^d$. Every vector in $\mathcal{T}$ has d coordinates although $\mathcal{T}$ has dimension m. The medial axis of $\mathcal{M}$ is the closure of the set of points in $\mathbb{R}^d$ that have two or more closest points in $\mathcal{M}$. The local feature size of a point in $\mathcal{M}$ is the distance from that point to the medial axis. Let ϱ denote the local feature size of $\mathcal{M}$ at the origin. Let $\{\mathsf{a}_p : p \in [1, n]\}$ denote the n sample points nearest to the origin. Let $\varrho\varepsilon$ denote the distance from the origin to the $(n+1)$-th nearest sample point, where $\varepsilon \in (0, 1)$ and ε decreases as the sampling density increases.

The manifold dimension m is treated as a constant. So we often absorb a function of m into the hidden constants in the big-Oh, big-Theta and big-Omega notation. We keep these

hidden constants scale independent; for example, the dependence on ϱ is explicitly stated. The ambient space dimension d is not a constant because one can embed $\mathcal{M}$ in an Euclidean space of arbitrarily high dimension.

1.2 Main result and comparison with previous work

The intuition behind our strategy is to compute a smooth approximation of $\mathcal{M}$ locally around the origin. Let $\gamma_1, \ldots, \gamma_d$ be any d orthogonal coordinate axes of $\mathbb{R}^d$ such that $\gamma_1, \ldots, \gamma_m$ span $\mathcal{T}$. Let ψ be an (unknown) orthonormal transformation such that for every point $\mathsf{y} \in \mathcal{M}$, $\psi(\mathsf{y})$ are the coordinates of y with respect to the coordinate system $(\gamma_1, \ldots, \gamma_d)$. By the implicit function theorem, for every point $\mathsf{y} \in \mathcal{M}$ close enough to the origin and every $\ell \in [m+1, d]$, the ℓ-th coordinate of $\psi(\mathsf{y})$ can be expressed as a function $f_\ell : \mathbb{R}^m \to \mathbb{R}$ in the first m coordinates of $\psi(\mathsf{y})$. We call $\{f_\ell : \ell \in [m+1, d]\}$ the *coordinate functions* of $\mathcal{M}$ at the origin with respect to $(\gamma_1, \ldots, \gamma_d)$.

We will approximate f_ℓ by an "almost quadratic" function $\widehat{F}_\ell : \mathbb{R}^d \to \mathbb{R}$ via solving an eigenvalue problem. There is not enough data to define the $\widehat{F}_\ell$'s unambiguously because there are only $n \leq \binom{d+1}{2}$ sample points. A popular approach is to add a penalty function, but a penalty function usually involves some parameter(s) and it is unclear how to tune them to obtain guarantees on the angular error. This parameter tuning phase may also be time-consuming. We also use a penalty function that involves a positive parameter. Our innovation is pushing this parameter to zero in the limit and obtain a modified eigenvalue problem. Hence, no parameter needs to be tuned and no training is required in the end. Solving this modified eigenvalue problem is equivalent to minimizing a measure of "curviness" of the fitting solution, which implies a theoretical guarantee on the angular error.

Our main result is stated in the following theorem. Let $T_{\text{tsvd}}(i, j)$ and $T_{\text{fsvd}}(i, j)$ denote the time to construct the thin and full singular value decompositions of an $i \times j$ matrix, respectively.

Theorem 1.1 *Suppose that $\mathcal{M}$ is a smooth compact m-dimensional manifold in $\mathbb{R}^d$, where m is a constant known to us, and that points are sampled from $\mathcal{M}$ according to a Poisson process with an unknown parameter. Assume that the origin is a sample point and its nearest n sample points are given, where $\binom{m+1}{2} + 1 \leq n \leq \binom{d+1}{2}$. Let $\varrho\varepsilon$ be the distance from the origin to the $(n+1)$-th nearest sample point, where ϱ is the local feature size of $\mathcal{M}$ at the origin and ε is a value in $(0, 1)$. We can compute in $O\big(T_{\text{tsvd}}\big(n, \binom{d+1}{2}\big) + T_{\text{fsvd}}(n, d) + d^2 n\big)$ time m orthogonal coordinate axes that span the approximate tangent space at the origin. If ε is sufficiently small, then with probability $1 - O\left(n^{-1/3}\right)$, the angular error is $O(\varepsilon^2)$.*

The running time $O(T_{\text{tsvd}}\big(n, \binom{d+1}{2}\big) + T_{\text{fsvd}}(n, d) + d^2 n)$ of our tangent estimation algorithm is $O(d^2 n^2)$ in practice [7, 16, 17]. The worst-case running time is asymptotically bounded by the worst-case running time of multiplying an $n \times n^r$ matrix with an $n^r \times n$ matrix for some r. The exact bound has a sophisticated expression depending on r [24]. As two examples in our case, if $d = O(n)$, the time bound is $O(n^{3.256689})$, and if $d = O(n^2)$, the time bound is $O(n^{5.180715})$. In general, the worst-case running time is slightly better than $O(d^2 n^2)$. Although the local feature size ϱ at the origin is used to obtain ε for expressing the angular error bound, our algorithm does not need to know ϱ.

In addition to developing an algorithm for the tangent estimation problem, we also develop some useful results along the way that may be of independent interests. Taubin gave a method for converting the curve reconstruction problem to an eigenvalue problem when there are enough sample points. We generalize this method for manifold reconstruction in high dimensions when there are insufficient sample points (Section 4). We also derive some concentration bounds on sums of powers of the coordinates of the sample points, which may be useful for other statistical analysis (Section 5).

Consider the condition $\binom{m+1}{2}+1 \leq n \leq \binom{d+1}{2}$ in Theorem 1.1. The formulation of our approach in Section 2 requires that $n \geq \binom{m+1}{2}+1$. Notice that $\binom{m+1}{2}+1 < \binom{d+1}{2}$ for $m \leq d-1$. Our techniques are not designed for the case of $n > \binom{d+1}{2}$. In a manifold learning context, it is predominantly the case that d is large. Therefore, the requirement of $n \leq \binom{d+1}{2}$ is not an issue because it is very likely that there are fewer than $\binom{d+1}{2}$ sample points nearby, and even if there are so many sample points nearby, it is computationally less efficient to use them all. If d is not large and $n > \binom{d+1}{2}$, our result can still be applied by increasing d and padding zeros to the coordinates of the sample points. Increasing d keeps a zero fitting error which allows our approach to minimize the "curviness".

Our theoretical result should hold when each sample point is perturbed in a random direction in $\mathbb{R}^d$ by a distance $O(\varrho\varepsilon^3)$, but we have not pursued the analysis as $O(\varrho\varepsilon^3)$ is rather small. We experimented with a fair amount of noise and the estimates are satisfactory. Refer to Section 3 for details.

How does our result compare with those in the literature? In [10, 11, 15], the sample points are required to satisfy two conditions: (i) for every point $\mathsf{y} \in \mathcal{M}$, the distance between y and the nearest sample point is at most the local feature size at y times μ for some sufficiently small $\mu \in (0,1)$, and (ii) for every pair of sample points p and q, the distance between p and q is at least the local feature size at p times δ for some $\delta \in (0,\mu)$. To estimate the tangent space at a sample point p, the methods in [10, 15] use the sample points no farther from p than the local feature size at p times $c\mu$ for some $c \geq 2$. It follows that at least a^m sample points are needed for some constant $a > 1$ depending on $\mathcal{M}$. The method in [11] uses the Voronoi cell of p for tangent space estimation. Using local information only, it is impossible to obtain the Voronoi cell of p, and it is unclear to obtain an appropriate approximate Voronoi cell. In the worst case, p can have at least a^m Voronoi neighbors for some constant $a > 1$ depending on $\mathcal{M}$. The angular error bounds given in [10, 11, 15] are $O(\mu)$. The running times are $O(d2^{O(m^7 \log m)})$ in [15], $O(d2^{O(m)})$ in [10], and $O(N^{\lceil (d+1)/2 \rceil})$ in [11], where N is the total number of sample points. In [2], the sampling is required to satisfy the condition that for every point $\mathsf{y} \in \mathcal{M}$, the distance between y and the nearest sample point is at most μ for some sufficiently small $\mu \in (0,1)$. The tangent space at a sample point p is estimated using the sample points within a distance r from p, where r can be any value in $[10\mu, 1/2)$. Thus, at least a^m sample points are needed for some constant $a > 1$ depending on $\mathcal{M}$. The angular error is $O(r/\varrho)$. The running time is $O(dn^{O(m^6 \log m)})$, where n is the number of sample points in the neighborhood. The work by Little et al. [26] is a multiscale analysis of the local covariance matrix. Noise is allowed and only roughly $O(m \log m)$ points in a local neighborhood are required for computation.

In our case, although n can be as small as $\binom{m+1}{2}+1$ for the algorithm to be applied, the probability bound $1 - O(n^{-1/3})$ is only meaningful for larger values of n because the hidden constant in the probability bound is a polynomial in m. Nevertheless, a polynomial in m is asymptotically smaller than a^m for any constant $a > 1$. This makes our neighborhood radius smaller than those in [2, 10, 11, 15] for large m, but we require more sample points than the approach in [26]. The angular error bounds in [2, 10, 11, 15] are $O(r/\varrho)$. (Note that $r/\varrho < 1$.) Roughly speaking, the angular error bound in [26] is linear in r, but the bound has a sophisticated expression and the reader is referred to [26] for details. Our angular error bound is $O(\varepsilon^2) = O((r/\varrho)^2)$. The hidden constant in our angular error bound depends on $\mathcal{M}$ and a polynomial in m.

The probability bound $1 - O(n^{-1/3})$ appears in many places in our analysis, where it is also implicitly assumed that n is greater than or equal to some appropriate polynomial in m. In practice, we suggest setting $n \geq \binom{m+1}{2} + m$ because our approach is based on locally fitting a quadratic function, and the minimum number of variables in such a quadratic function is $\binom{m+1}{2} + m$ when $d = m+1$. In our experiments (Section 3), setting $n = \binom{m+1}{2} + m + 30$ gives

good results in both the noiseless and noisy cases.

2 Problem formulation, algorithm and overview

2.1 Modeling

We discuss in this section how to model the local neighborhood of the origin using some implicit functions F_ℓ, $\ell \in [m+1, d]$, with domain $\mathbb{R}^d$. Recall that $\{ \mathsf{a}_p : p \in [1,n] \}$ are the n sample points nearest to the origin (which is also a sample point). The functions F_ℓ are constructed so that $F_\ell(\mathsf{a}_p) = 0$ for every $p \in [1,n]$ and every $\ell \in [m+1, d]$. The goal is to obtain the compact representation of $\big(F_\ell(\mathsf{a}_1) \quad \cdots \quad F_\ell(\mathsf{a}_n)\big)$ in (2.3) below.

Let $\gamma_1, \ldots, \gamma_d$ be any d orthogonal coordinate axes of $\mathbb{R}^d$ (with the same origin) such that $\gamma_1, \ldots, \gamma_m$ span $\mathcal{T}$.

First, we apply an (unknown) orthonormal transformation ψ so that the coordinates of each point $\psi(\mathsf{a}_p)$ is expressed with respect to the coordinate system $(\gamma_1, \ldots, \gamma_d)$. Recall that the ℓ-th coordinate of $\psi(\mathsf{a}_p)$, $\ell \in [m+1, d]$, is the value of the coordinate function f_ℓ on the first m coordinates of $\psi(\mathsf{a}_p)$. Figure 1(a) shows an example of a manifold and the coordinate system before applying the transformation ψ. Figure 1(b) shows the corresponding f_2 and f_3 after applying ψ.

For every positive integer k, let $\mathrm{D}^k f_\ell|_{\mathsf{0}}$ denote the k-th derivative of f_ℓ at the origin, which is a map that sends k vectors from $\mathbb{R}^m$ to a real number. The domain of $\mathrm{D}^k f_\ell|_{\mathsf{0}}$ consists of k copies of $\mathbb{R}^m$ spanned by $(\gamma_1, \ldots, \gamma_m)$. When $k = 2$, one can view $\mathrm{D}^2 f_\ell|_{\mathsf{0}}$ as an $m \times m$ matrix, and then $\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v}, \mathsf{v})$ is equal to $\mathsf{v}^t \cdot \mathrm{D}^2 f_\ell|_{\mathsf{0}} \cdot \mathsf{v}$ for every vector $\mathsf{v} \in \mathbb{R}^m$. The matrix $\mathrm{D}^2 f_\ell|_{\mathsf{0}}$ is known as the *Hessian matrix*.

For every vector $\mathsf{v} \in \mathbb{R}^m$ with a small enough $\|\mathsf{v}\|$, the Taylor expansion of $f_\ell(\mathsf{v})$ is

$$f_\ell(\mathsf{v}) = \frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v}, \mathsf{v}) + \frac{1}{6}\mathrm{D}^3 f_\ell|_{\mathsf{0}}(\mathsf{v}, \mathsf{v}, \mathsf{v}) + \cdots .$$

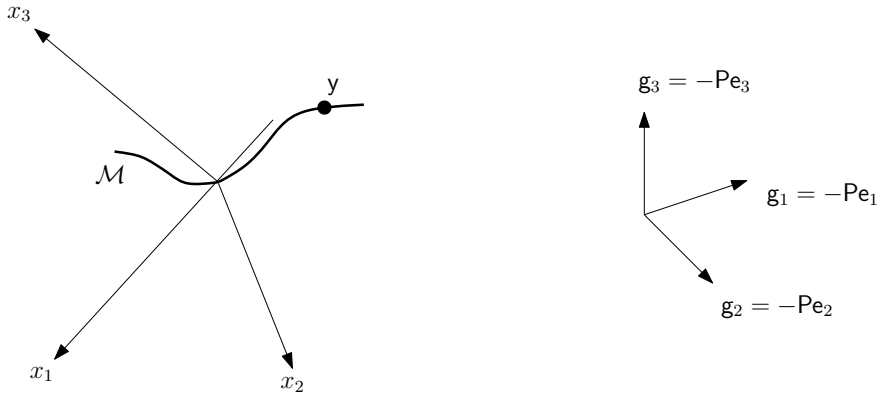
There is no constant term in the Taylor expansion above because $\mathcal{M}$ passes through the origin. There is no linear term because $(\gamma_1, \cdots, \gamma_m)$ span $\mathcal{T}$ and so $\mathrm{D} f_\ell|_{\mathsf{0}}$ vanishes. We extend the domain of f_ℓ from $\mathbb{R}^m$ to $\mathbb{R}^d$ by ignoring the last $d-m$ coordinates of the input vector. That is, the vector v can be paired with any vector $\mathsf{w} \in \mathbb{R}^{d-m}$ to yield the following extended expansion:

$$\frac{1}{2}(\mathsf{v}^t \ \mathsf{w}^t) \begin{pmatrix} \mathrm{D}^2 f_\ell|_{\mathsf{0}} & \mathsf{0}_{m,d-m} \\ \mathsf{0}_{d-m,m} & \mathsf{0}_{d-m,d-m} \end{pmatrix} \begin{pmatrix} \mathsf{v} \\ \mathsf{w} \end{pmatrix} + \cdots .$$

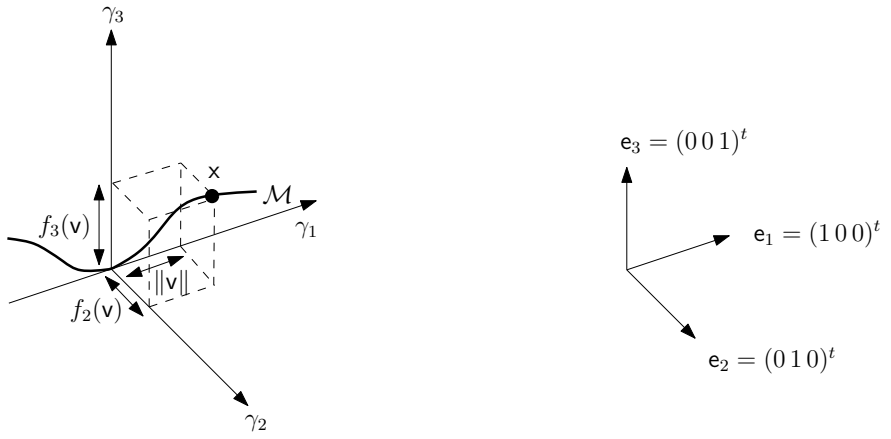
Transforming this extended expansion back to the coordinate system $(x_1, \ldots, x_d)$ gives the following function $F_\ell : \mathbb{R}^d \to \mathbb{R}$ whose zero-set contains the origin and a_p for $p \in [1,n]$.

$$\begin{aligned} \forall\, \ell \in [m+1, d], \quad F_\ell(\mathsf{y}) &= \mathsf{y}^t \mathsf{g}_\ell + \tfrac{1}{2}\mathsf{y}^t \mathsf{Q}_\ell\, \mathsf{y} + \cdots, \quad \text{such that} \\ \mathsf{g}_\ell &= -\mathsf{P}\,(\mathsf{0}_{1,\ell-1}\ 1\ \ \mathsf{0}_{1,d-\ell})^t, \\ \mathsf{Q}_\ell &= \mathsf{P} \begin{pmatrix} \mathrm{D}^2 f_\ell|_{\mathsf{0}} & \mathsf{0}_{m,d-m} \\ \mathsf{0}_{d-m,m} & \mathsf{0}_{d-m,d-m} \end{pmatrix} \mathsf{P}^t. \end{aligned} \tag{2.1}$$

The matrix P^t is the unknown $d \times d$ orthonormal matrix that realizes the transformation ψ. The vector $(\mathsf{0}_{1,\ell-1}\ 1\ \ \mathsf{0}_{1,d-\ell})^t$ is the "vertical direction" for f_ℓ at the origin with respect to the coordinate system $(\gamma_1 \cdots \gamma_d)$. It means that g_ℓ is the gradient of F_ℓ and also a normal vector to $\mathcal{M}$ at the origin. Q_ℓ is a $d \times d$ matrix.



(a)



(b)

Figure 1: The manifold $\mathcal{M}$ is a curve shown in bold in (a) and (b). The coordinate system in (a) is (x_1, x_2, x_3), whereas $(\gamma_1, \gamma_2, \gamma_3)$ is the coordinate system in (b). The right figure in (b) shows an orthonormal basis $(\mathsf{e}_1, \mathsf{e}_2, \mathsf{e}_3)$ of the coordinate system $(\gamma_1, \gamma_2, \gamma_3)$. The left figure in (a) is mapped by the orthonormal transformation ψ to the left figure in (b). The transformation ψ is realized by the orthonormal matrix P^t. Therefore, $\mathsf{x} = \mathsf{P}^t\mathsf{y}$. Each basis vector e_i with respect to the coordinate system $(\gamma_1, \gamma_2, \gamma_3)$ is mapped to the vector $\mathsf{g}_i = -\mathsf{P}\mathsf{e}_i$ with respect to the coordinate system (x_1, x_2, x_3). The tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin is spanned by the vector g_1 with respect to the coordinate system (x_1, x_2, x_3) in (a), or equivalently, the vector e_1 with respect to the coordinate system $(\gamma_1, \gamma_2, \gamma_3)$ in (b).

Let $q_{\ell,ij}$ denote the (i,j) entry of Q_ℓ. Since Q_ℓ is symmetric, $q_{\ell,ij}$ and $q_{\ell,ji}$ are equal. Expanding the terms $\mathsf{a}_p^t\mathsf{g}_\ell$ and $\frac{1}{2}\mathsf{a}_p^t\mathsf{Q}_\ell\mathsf{a}_p$ in $F_\ell(\mathsf{a}_p)$ gives:

$$\begin{aligned} F_\ell(\mathsf{a}_p) \;=\; & \begin{pmatrix} a_{p1} & a_{p2} & \cdots & a_{pd} \end{pmatrix} \cdot \mathsf{g}_\ell + \\ & \begin{pmatrix} \frac{1}{\sqrt{2}}a_{p1}^2 & a_{p1}a_{p2} & \cdots & a_{p1}a_{pd} & \frac{1}{\sqrt{2}}a_{p2}^2 & a_{p2}a_{p3} & \cdots & a_{p2}a_{pd} & \cdots & \frac{1}{\sqrt{2}}a_{pd}^2 \end{pmatrix} \cdot \\ & \begin{pmatrix} \frac{1}{\sqrt{2}}q_{\ell,11} & q_{\ell,12} & \cdots & q_{\ell,1d} & \frac{1}{\sqrt{2}}q_{\ell,22} & q_{\ell,23} & \cdots & q_{\ell,2d} & \cdots & \frac{1}{\sqrt{2}}q_{\ell,dd} \end{pmatrix}^t + \;\cdots \end{aligned}$$

This motivates us to define:

$$\begin{aligned} \mathsf{c}_\ell &\stackrel{\text{def}}{=} \begin{pmatrix} \frac{1}{\sqrt{2}}q_{\ell,11} & q_{\ell,12} & \cdots & q_{\ell,1d} & \frac{1}{\sqrt{2}}q_{\ell,22} & q_{\ell,23} & \cdots & q_{\ell,2d} & \cdots & \frac{1}{\sqrt{2}}q_{\ell,dd} \end{pmatrix}^t \\ \mathsf{A} &\stackrel{\text{def}}{=} \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nd} \end{pmatrix} \\ \mathsf{B} &\stackrel{\text{def}}{=} \begin{pmatrix} \frac{1}{\sqrt{2}}a_{11}^2 & a_{11}a_{12} & \cdots & a_{11}a_{1d} & \frac{1}{\sqrt{2}}a_{12}^2 & a_{12}a_{13} & \cdots & \frac{1}{\sqrt{2}}a_{1d}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}a_{n1}^2 & a_{n1}a_{n2} & \cdots & a_{n1}a_{nd} & \frac{1}{\sqrt{2}}a_{n2}^2 & a_{n2}a_{n3} & \cdots & \frac{1}{\sqrt{2}}a_{nd}^2 \end{pmatrix} \end{aligned} \tag{2.2}$$

A is an $n \times d$ matrix, B is an $n \times \binom{d+1}{2}$ matrix, and c_ℓ is a $\binom{d+1}{2}$-dimensional vector. They yield:

$$\begin{pmatrix} F_\ell(\mathsf{a}_1) \\ \vdots \\ F_\ell(\mathsf{a}_n) \end{pmatrix} \;=\; (\mathsf{A}\;\; \mathsf{B}) \cdot \begin{pmatrix} \mathsf{g}_\ell \\ \mathsf{c}_\ell \end{pmatrix} + \cdots . \tag{2.3}$$

The coefficient $1/\sqrt{2}$ of a_{pi}^2 in the definition of B is needed so that the eigenvalues of $\mathsf{B}^t\mathsf{B}$ are independent of rotations in $\mathbb{R}^d$ that keep the origin fixed. We will establish this fact in Lemma 6.2. The vectors g_ℓ and c_ℓ are unknowns in (2.3), and c_ℓ is a linearization of the matrix Q_ℓ. Since $F_\ell^{-1}(0)$ contains a_p for $p \in [1,n]$, the left hand side of (2.3) is a zero vector.

To approximate the F_ℓ's using quadratic functions, the first attempt is to retain just $(\mathsf{A}\;\; \mathsf{B}) \cdot (\mathsf{g}_\ell^t\;\; \mathsf{c}_\ell^t)^t$ in the right hand side of (2.3) because this keeps only the linear and quadratic terms. The subsequent analysis in Section 7 demands $\|\mathsf{B}^\dagger\|$ to be comparable to the reciprocal of the $\left(\binom{m+1}{2}+1\right)$-th largest singular value of B. Unfortunately, $\|\mathsf{B}^\dagger\|$ is determined by the possibly much larger reciprocal of the smallest singular value of B. As a result, we modify B by changing its $n - \binom{m+1}{2}$ smallest singular values as follows. Recall that $\binom{m+1}{2} + 1 \le n \le \binom{d+1}{2}$ by assumption. Define:

$$\begin{aligned} m_0 &\stackrel{\text{def}}{=} \tbinom{m+1}{2} \\ \mathsf{L}\Lambda\mathsf{R}^t &\stackrel{\text{def}}{=} \text{thin SVD of } \mathsf{B}\text{, where } \Lambda = \mathrm{diag}_n(\lambda_1, \lambda_2, \cdots, \lambda_n) \text{ and} \\ &\qquad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0 \\ \widehat{\Lambda} &\stackrel{\text{def}}{=} \mathrm{diag}_n\big(\lambda_1, \lambda_2, \cdots, \lambda_{m_0}, \underbrace{\lambda_{m_0+1}, \cdots, \lambda_{m_0+1}}_{n-m_0 \text{ copies}}\big) \\ \widehat{\mathsf{B}} &\stackrel{\text{def}}{=} \mathsf{L}\widehat{\Lambda}\mathsf{R}^t \end{aligned}$$

L is an $n \times n$ matrix, Λ and $\widehat{\Lambda}$ are $n \times n$ diagonal matrices, and R is a $\binom{d+1}{2} \times n$ matrix. $\widehat{\mathsf{B}}$ is the replacement of B.

Lemma 2.1 below shows that the case of $\lambda_{m_0+1} = 0$ can be dealt with separately in the algorithm. (The proof of Lemma 2.1 is given in Section 8.) Thus, we can assume that λ_{m_0+1} is positive. For ease of presentation, we also assume that $\lambda_n > 0$; otherwise, for each $\lambda_i = 0$ (such an i can range from $m_0 + 2$ to n), we set the corresponding diagonal entry of $\widehat{\Lambda}$ to zero. The proof of Lemma 7.5 requires that if a diagonal entry of Λ is zero, the corresponding entry in $\widehat{\Lambda}$ is also zero.

Lemma 2.1 *There exists a constant c that is a polynomial in m such that if $n \geq c$, ε is sufficiently small, and $\lambda_{m_0+1} = 0$, then with probability $1 - O(n^{-1/3})$, $\mathcal{T}$ is equal to the space spanned by the eigenvectors corresponding to the m largest eigenvalues of $\mathsf{A}^t\mathsf{A}$.*

By keeping only $(\mathsf{A} \;\; \mathsf{B}) \cdot (\mathsf{g}_\ell^t \;\; \mathsf{c}_\ell^t)^t$ in the right hand side of (2.3) and replacing B by $\widehat{\mathsf{B}}$, we allude to some nonlinear functions $\widehat{F}_\ell : \mathbb{R}^d \to \mathbb{R}$, $\ell \in [m+1, d]$, that approximate the F_ℓ's and satisfy the following system:

$$\forall\, \ell \in [m+1, d], \qquad \begin{pmatrix} \widehat{F}_\ell(\mathsf{a}_1) \\ \vdots \\ \widehat{F}_\ell(\mathsf{a}_n) \end{pmatrix} \;=\; (\mathsf{A} \;\; \widehat{\mathsf{B}}) \begin{pmatrix} \widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell \end{pmatrix}, \tag{2.4}$$

where $\widehat{\mathsf{g}}_\ell$ and $\widehat{\mathsf{c}}_\ell$ are the new unknowns. The vector $\widehat{\mathsf{g}}_\ell$ is the unknown gradient of $\widehat{F}_\ell$. Since the unknown gradients are supposed to span the approximate normal space at the origin, we require them to satisfy the following constraints:

$$\begin{aligned} \forall\, \ell \in [m+1, d], \qquad & \|\widehat{\mathsf{g}}_\ell\| = 1 \\ \forall\, \ell_1 \neq \ell_2, \qquad & \widehat{\mathsf{g}}_{\ell_1} \perp \widehat{\mathsf{g}}_{\ell_2} \end{aligned} \tag{2.5}$$

2.2 Algorithm and overview

Our algorithm solves an eigenvalue problem derived from $\widehat{\mathsf{B}}$. Define the following matrix:

$$\mathsf{H} = (\widehat{\mathsf{B}}^\dagger \mathsf{A})^t.$$

We find the eigenvectors corresponding to the m largest eigenvalues of the following matrix:

$$\mathsf{H}\mathsf{H}^t = \mathsf{A}^t(\widehat{\mathsf{B}}^\dagger)^t \cdot \widehat{\mathsf{B}}^\dagger \mathsf{A} = \mathsf{A}^t \mathsf{L} \widehat{\Lambda}^\dagger \widehat{\Lambda}^\dagger \mathsf{L}^t \mathsf{A}.$$

These m eigenvectors span the approximate tangent space. We compute the full SVD of the $n \times d$ matrix $\widehat{\Lambda}^\dagger \mathsf{L}^t \mathsf{A}$ to obtain the eigenvectors of $\mathsf{H}\mathsf{H}^t$. The pseudocode is given below.

TANGENT(A)

1. Compute the thin SVD $\mathsf{L}\Lambda\mathsf{R}^t$ of B.
2. If $\lambda_{m_0+1} = 0$, then return the eigenvectors corresponding to the m largest eigenvalues of $\mathsf{A}^t\mathsf{A}$ as an orthonormal basis of the estimated tangent space.
3. Compute $\widehat{\Lambda}^\dagger$ and the full SVD $\mathsf{C}\mathsf{D}\mathsf{E}^t$ of $\widehat{\Lambda}^\dagger \mathsf{L}^t \mathsf{A}$. Assume that the diagonal entries of D are in descending order and that $\mathsf{E} = (\widehat{\mathsf{g}}_1, \ldots, \widehat{\mathsf{g}}_d)$, where $\widehat{\mathsf{g}}_\ell$ corresponds to the ℓth largest diagonal entry in D.
4. Return $(\widehat{\mathsf{g}}_1, \ldots, \widehat{\mathsf{g}}_m)$ as an orthonormal basis of the approximate tangent space.

The running time is $O\big(T_{\text{tsvd}}\big(n, \binom{d+1}{2}\big) + T_{\text{fsvd}}(n, d) + d^2 n\big)$, which is $O(d^2 n^2)$ in practice [7, 16, 17].

Section 3 describes our experiments that demonstrate the accuracy of our tangent estimation in both noiseless and noisy cases.

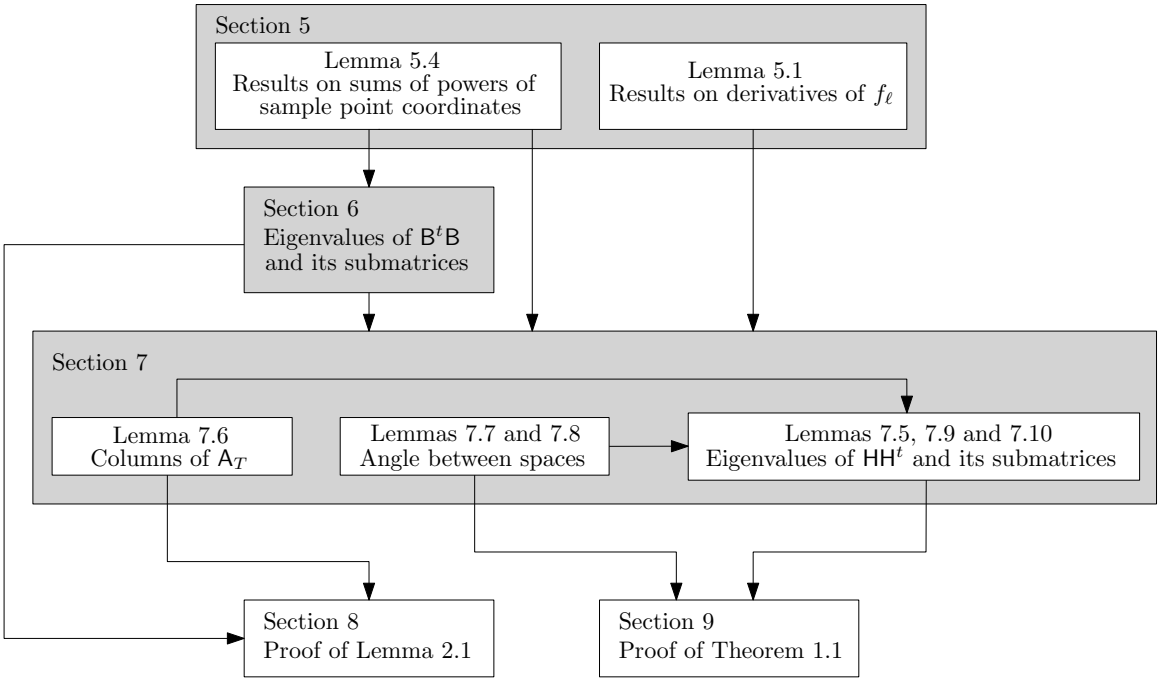


Figure 2: Proof overview.

Section 4 turns (2.4) and (2.5) into an eigenvalue problem. The solution minimizes the quantity $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2 + \sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$, that is, the sum of squared interpolation error with the "curviness" penalized by the term $\sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$. Tuning the parameter α is time-consuming though. We push α to zero in the limit to derive the eigenvalue problem for $\mathsf{H}\mathsf{H}^t$.

Sections 5–9 contain the analysis that leads to an $O(\varepsilon^2)$ bound on the angular error. Figure 2 shows the dependence of results in different sections. Assume that the coordinate axes $x_1, \ldots, x_m$ span the tangent space of $\mathcal{M}$ at the origin. So the coordinate axes $x_{m+1}, \ldots, x_d$ span the normal space at the origin.

Section 5 presents several results on the derivatives of the coordinate functions and on the sums of powers of sample point coordinates. These results may be of independent interest. Lemma 5.1(i) shows that $\left\|\begin{pmatrix} \mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) & \cdots & \mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) \end{pmatrix}\right\| = O(1/\varrho)$ for any unit vector $\mathsf{u} \in \mathbb{R}^m$, independent of the dimension and other factors. Similarly, Lemma 5.1(ii) shows that $(f_{m+1}(\mathsf{v}), \ldots, f_d(\mathsf{v}))$ is approximated well by $(\frac{1}{2}\mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{v},\mathsf{v}), \ldots, \frac{1}{2}\mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{v},\mathsf{v}))$ for any $\mathsf{v} \in \mathbb{R}^m$ such that $\|\mathsf{v}\| \leq \varrho\varepsilon$, that is, $\left\|(f_{m+1}(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{v},\mathsf{v}), \ldots, f_d(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{v},\mathsf{v}))\right\| = O(\varrho\varepsilon^3)$. Lemma 5.4 gives concentration bounds for $\left|\sum_{p=1}^{n} a_{pi}a_{pj}a_{pk}a_{pl}\right|$, $\left|\sum_{p=1}^{n} a_{pi}a_{pj}a_{pk}\right|$, and $\left|\sum_{p=1}^{n} a_{pi}a_{pj}\right|$ for $i, j, k, l \in [1, m]$.

Since the approximate tangent space is spanned by some eigenvectors of $\mathsf{H}\mathsf{H}^t$, we need to analyze the eigenvalues of $\mathsf{H}\mathsf{H}^t = \mathsf{A}^t(\widehat{\mathsf{B}}^\dagger)^t \cdot \widehat{\mathsf{B}}^\dagger\mathsf{A}$, which requires us to bound the eigenvalues of $\widehat{\mathsf{B}}^\dagger$. These results are presented in Section 6. Rearrange the columns of B so that $\mathsf{B} = (\mathsf{B}_{TT}\ \mathsf{B}_{TN}\ \mathsf{B}_{NN})$, where B_{TT} consists of columns in $a_{pi}a_{pj}$ for possibly non-distinct $i, j \in [1, m]$, B_{TN} consists of columns in $a_{pi}a_{pj}$ for $i \in [1, m]$ and $j \in [m+1, d]$, and B_{NN} consists of columns

in $a_{pi}a_{pj}$ for possibly non-distinct $i, j \in [m+1, d]$. We divide $\mathsf{B}^t\mathsf{B}$ into blocks as follows.

$$\mathsf{B}^t\mathsf{B} \;=\; \begin{pmatrix} \mathsf{B}_{TT}^t\mathsf{B}_{TT} & \mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN} \\ \mathsf{B}_{TN}^t\mathsf{B}_{TT} & \mathsf{B}_{TN}^t\mathsf{B}_{TN} & \mathsf{B}_{TN}^t\mathsf{B}_{NN} \\ \mathsf{B}_{NN}^t\mathsf{B}_{TT} & \mathsf{B}_{NN}^t\mathsf{B}_{TN} & \mathsf{B}_{NN}^t\mathsf{B}_{NN} \end{pmatrix}$$

We apply the concentration bounds in Lemma 5.4. Then, the Gershgorin Circle Theorem [13, 17] says that the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ are dominated by the diagonal entries which are $\Theta(n\varrho^4\varepsilon^4)$ (Lemma 6.3). Using known bounds on $\sum_{i=1}^m a_{pi}^2$ and $\sum_{i=m+1}^d a_{pi}^2$, one can easily show that $\|\mathsf{B}_{TN}\| = O(\sqrt{n}\varrho^2\varepsilon^3)$ and $\|\mathsf{B}_{NN}\| = O(\sqrt{n}\varrho^2\varepsilon^4)$ (Lemma 6.4). Then, applying the Gershgorin Circle Theorem to the division of $\mathsf{B}^t\mathsf{B}$ above shows that the m_0 largest eigenvalues of $\mathsf{B}^t\mathsf{B}$ are dominated by those of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ and hence are $\Theta(n\varrho^4\varepsilon^4)$ (Lemma 6.5). A finer analysis then shows that the (m_0+1)-th largest eigenvalue $\lambda_{m_0+1}^2$ of $\mathsf{B}^t\mathsf{B}$ is $O(n\varrho^4\varepsilon^6)$ (Lemma 6.6). Note that $\|\widehat{\mathsf{B}}^\dagger\| = 1/\lambda_{m_0+1}$.

We write $\mathsf{A} = (\mathsf{A}_T\ \mathsf{A}_N)$, where A_T consists of the columns $(a_{1i}, \ldots, a_{ni})^t$ for $i \in [1, m]$ and A_N consists of the columns $(a_{1i}, \ldots, a_{ni})^t$ for $i \in [m+1, d]$. Since $\mathsf{H} = (\widehat{\mathsf{B}}^\dagger\mathsf{A})^t$, we can write

$$\mathsf{H}_T = (\widehat{\mathsf{B}}^\dagger\mathsf{A}_T)^t, \quad \mathsf{H}_N = (\widehat{\mathsf{B}}^\dagger\mathsf{A}_N)^t, \quad \mathsf{H} = \begin{pmatrix}\mathsf{H}_T \\ \mathsf{H}_N\end{pmatrix}, \quad \mathsf{H}\mathsf{H}^t = \begin{pmatrix}\mathsf{H}_T\mathsf{H}_T^t & \mathsf{H}_T\mathsf{H}_N^t \\ \mathsf{H}_N\mathsf{H}_T^t & \mathsf{H}_N\mathsf{H}_N^t\end{pmatrix}.$$

Note that H_T is an $m \times \binom{d+1}{2}$ submatrix and H_N is a $(d-m) \times \binom{d+1}{2}$ submatrix. In Section 7, we prove that $\|\mathsf{H}_N\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$ and every singular value of H_T is $\Theta(\sqrt{n}\varrho\varepsilon/\lambda_{m_0+1})$. Then, the Gershgorin Circle Theorem implies that the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ are dominated by those of $\mathsf{H}_T\mathsf{H}_T^t$ and hence they are $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$, and the $d-m$ smallest eigenvalues of $\mathsf{H}\mathsf{H}^t$ are at most $\|\mathsf{H}_N\|^2 + \|\mathsf{H}_T\|\|\mathsf{H}_N\| = O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$ (Lemma 7.10).

In the analysis of $\mathsf{H}_N^t = \widehat{\mathsf{B}}^\dagger\mathsf{A}_N$ (Lemma 7.5), our definition of $\widehat{\mathsf{B}}$ gives $\|\widehat{\mathsf{B}}^\dagger\mathsf{A}_N\| \le \|\mathsf{B}^\dagger\mathsf{A}_N\|$. Notice that $\mathsf{B}(\mathsf{B}^\dagger\mathsf{A}_N) = \mathsf{A}_N$. By the Taylor expansion, we can approximate each entry $a_{p\ell}$ in A_N by $\frac{1}{2}(a_{p1}\ \cdots\ a_{pm}) \cdot \mathrm{D}^2 f_\ell|_{\mathsf{0}} \cdot (a_{p1}\ \cdots\ a_{pm})^t$. The definition of B allows us to write these Taylor expansions as $\mathsf{B}\mathsf{Z} \approx \mathsf{A}_N$, where Z is a $\binom{d+1}{2} \times (d-m)$ matrix such that the $(\ell-m)$-th column is

$$\left(\tfrac{1}{\sqrt{2}}q_{\ell,11}\quad q_{\ell,12}\quad \cdots\quad q_{\ell,1d}\quad \tfrac{1}{\sqrt{2}}q_{\ell,22}\quad q_{\ell,23}\quad \cdots\quad q_{\ell,2d}\quad \cdots\quad \tfrac{1}{\sqrt{2}}q_{\ell,dd}\right)^t,$$

where $q_{\ell,ij}$ is the (i,j) entry of the Hessian matrix $\mathrm{D}^2 f_\ell|_{\mathsf{0}}$ if $i, j \in [1, m]$, and $q_{\ell,ij} = 0$ otherwise. Then, Lemma 5.1(i) allows us to conclude that $\|\mathsf{Z}\| = O(1/\varrho)$, and by the property of pseudoinverse, $\|\mathsf{B}^\dagger\mathsf{A}_N\| \lesssim \|\mathsf{Z}\| = O(1/\varrho) = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$ as $\lambda_{m_0+1} = O(\sqrt{n}\varrho^2\varepsilon^3)$ by the result in Section 6.

The analysis of H_T makes use of the thin SVD $\mathsf{L}\widehat{\Lambda}\mathsf{R}^t$ of $\widehat{\mathsf{B}}$ (Lemma 7.9). Recall that the m_0 largest singular values of $\widehat{\mathsf{B}}$ are those of B. The other $n-m_0$ singular values are equal to λ_{m_0+1}. If we group the largest m_0 singular values in an $m_0 \times m_0$ diagonal submatrix $\widehat{\Lambda}_0$, then we can write $\widehat{\Lambda}$, L, and R as follows.

$$\widehat{\Lambda} = \begin{pmatrix}\widehat{\Lambda}_0 & \mathsf{0}_{m_0,n-m_0} \\ \mathsf{0}_{n-m_0,m_0} & \lambda_{m_0+1}\mathsf{I}_{n-m_0}\end{pmatrix}, \mathsf{L} = \Big(\underbrace{\mathsf{L}_0}_{m_0\text{ columns}} \quad \underbrace{\mathsf{L}_1}_{n-m_0\text{ columns}}\Big), \mathsf{R} = \Big(\underbrace{\mathsf{R}_0}_{m_0\text{ columns}} \quad \underbrace{\mathsf{R}_1}_{n-m_0\text{ columns}}\Big).$$

Combining the above with the relation $\mathsf{H}_T^t = \widehat{\mathsf{B}}^\dagger\mathsf{A}_T$, we obtain

$$\mathsf{H}_T\mathsf{H}_T^t = \mathsf{A}_T^t\mathsf{L}_0\left(\widehat{\Lambda}_0^\dagger\right)^2\mathsf{L}_0^t\mathsf{A}_T \;+\; \tfrac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T.$$

Thus, the maximum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is at most $\|\mathsf{A}_T^t\mathsf{L}_0(\widehat{\Lambda}_0^\dagger)^2\mathsf{L}_0^t\mathsf{A}_T\| + \frac{1}{\lambda_{m_0+1}^2}\|\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T\|$, which can be verified to be $O(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ using the facts that $\|\mathsf{L}_0\| = \|\mathsf{L}_1\| = 1$, $\|\mathsf{A}_T\| =$

$O(\sqrt{n}\varrho\varepsilon)$ as $\|\mathsf{a}_p\| = O(\varrho\varepsilon)$, $\|\widehat{\Lambda}_0^\dagger\| = \Theta(1/(\sqrt{n}\varrho^2\varepsilon^2))$, and $\lambda_{m_0+1} = O(\sqrt{n}\varrho^2\varepsilon^3)$. The minimum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is at least the minimum eigenvalue of $\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$. Using the concentration bounds in Lemma 5.4, we can show that every column vector $(a_{1i},\ldots,a_{ni})^t$ in A_T has a 2-norm of $\Theta(\sqrt{n}\varrho\varepsilon)$, and that the angle between two distinct column vectors $(a_{1i},\ldots,a_{ni})^t$ and $(a_{1j},\ldots,a_{nj})^t$ in A_T is large (Lemma 7.6). Then, we show that every column vector in A_T makes a small angle with the column space of L_1 (Lemmas 7.7–7.8). It follows that $\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T \approx \|\mathsf{A}_T\|^2 = \sqrt{m}\cdot\Omega(n\varrho^2\varepsilon^2)$, and so the minimum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is at least $\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T = \Omega(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ as desired.

We analyze the angular error (Theorem 1.1) in Section 9 using the bounds on the singular values of H_T and $\|\mathsf{H}_N\|$. We take an arbitrary unit eigenvector e corresponding to any of the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$, say σ. Let v be the vector consisting of the first m coordinates of e. Let w be the vector consisting of the other $d-m$ coordinates of e. We check the angle that e makes the true tangent space spanned by the coordinate axes $x_1,\ldots,x_m$. This is done by examining the equation:

$$\mathsf{H}\mathsf{H}^t\mathsf{e} \;=\; \mathsf{H}\mathsf{H}^t\begin{pmatrix}\mathsf{v}\\ \mathsf{w}\end{pmatrix} \;=\; \begin{pmatrix}\mathsf{H}_T\mathsf{H}_T^t & \mathsf{H}_T\mathsf{H}_N^t\\ \mathsf{H}_N\mathsf{H}_T^t & \mathsf{H}_N\mathsf{H}_N^t\end{pmatrix}\begin{pmatrix}\mathsf{v}\\ \mathsf{w}\end{pmatrix} \;=\; \sigma\begin{pmatrix}\mathsf{v}\\ \mathsf{w}\end{pmatrix}.$$

Then, $\mathsf{w} = \left(\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t\right)^{-1}\mathsf{H}_N\mathsf{H}_T^t\mathsf{v}$. Therefore, the angle between e and the true tangent space is $\arctan(\|\mathsf{w}\|/\|\mathsf{v}\|) \le \|\left(\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t\right)^{-1}\|\cdot\|\mathsf{H}_N\|\cdot\|\mathsf{H}_T^t\|$. Since $\sigma = \Theta(\|\mathsf{H}_T\|^2) = \Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ and $\|\mathsf{H}_N\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$ by the results in Section 7, we conclude that $\arctan(\|\mathsf{w}\|/\|\mathsf{v}\|) = O(\varepsilon^2)$. It follows that the space spanned by the eigenvectors corresponding to the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ makes an $O(\varepsilon^2)$ angle with the true tangent space, completing the analysis of the angular error.

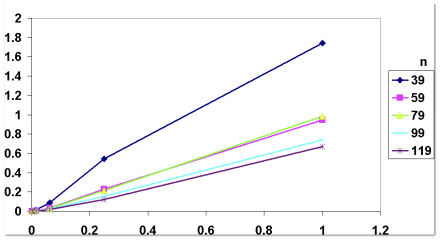
3 Experimental results

We carried out some experiments to estimate tangent spaces using sample points drawn from different manifolds, including spheres, manifolds with saddles, and sinusoidal curves in high dimensions. Since our algorithm works locally, sample points are only needed in a local neighborhood. We do not know how to locally sample a manifold uniformly or according to a Poisson distribution in general, so some adhoc heuristics are used for each class of manifolds. We will describe the sampling heuristics in each case. For each manifold tested, we fix a point p in the manifold and generate point samples in its neighborhood. Since we know the true tangent space, we can compute the angular error for each trial, which allows us to report the mean angular error over all trials. We conducted 25 trials for every manifold, every value of n, and every neighborhood radius. The smallest value of n is $\binom{m+1}{2} + m + 30$ and we increase n by adding multiples of 20.

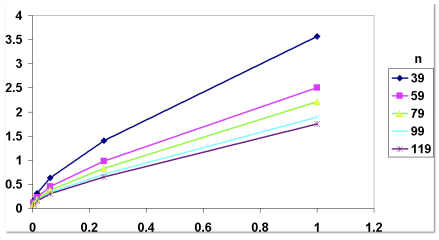
3.1 Sphere

We tried unit spheres $\mathbb{S}^m$ for $3 \le m \le 9$ in $\mathbb{R}^{200}$. We choose p to be the north pole and put an m-ball D of radius r tangent to $\mathbb{S}^m$ at p. We generate n sample points in D uniformly at random and then project them towards the origin onto $\mathbb{S}^m$.

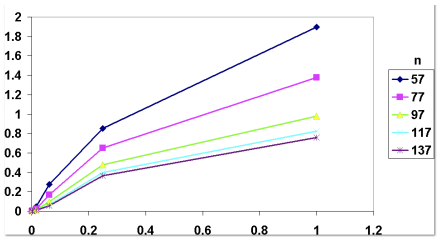
The projection of D towards the origin onto $\mathbb{S}^m$ is the spherical cap that sample points may occupy. Projecting this spherical cap orthogonally onto D gives a concentric, smaller m-ball D'. We can view D' as part of the domain of the coordinate functions $f_{m+1},\cdots,f_d$, which map a point in D' to a point on $\mathbb{S}^m$. Therefore, $\max_{\mathsf{z}\in D'}\|f_{m+1}(\mathsf{z}),\cdots,f_d(\mathsf{z})\|$ is the maximum distance between a point in D' and $\mathbb{S}^m$. By elementary trigonometry, this maximum distance



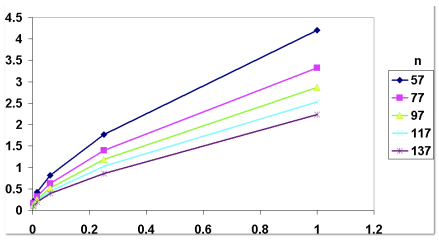
(a) $\mathbb{S}^3$: noiseless



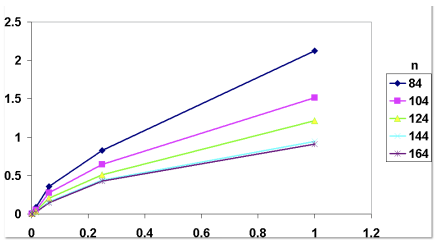
(b) $\mathbb{S}^3$: noisy



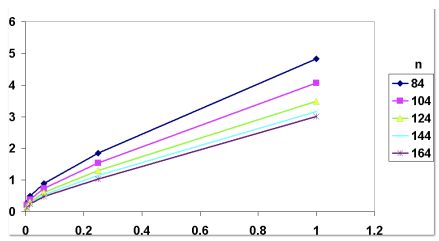
(c) $\mathbb{S}^6$: noiseless



(d) $\mathbb{S}^6$: noisy



(e) $\mathbb{S}^9$: noiseless



(f) $\mathbb{S}^9$: noisy

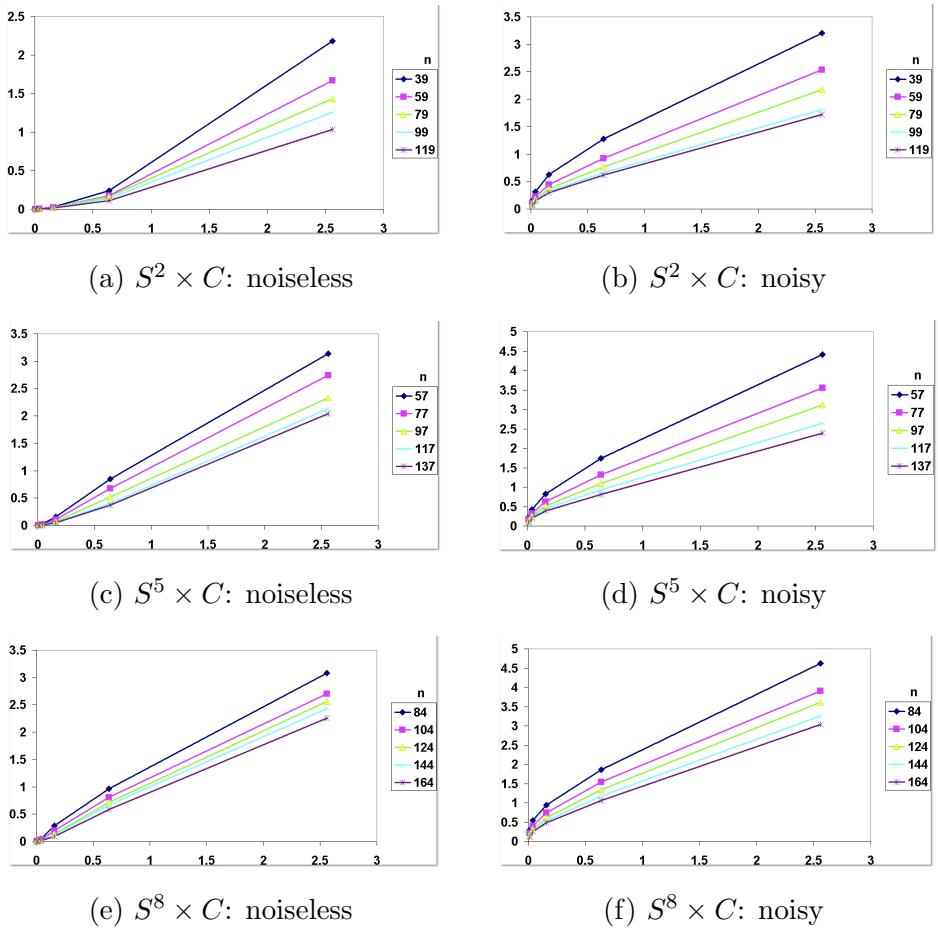
Figure 3: Plots of mean angular error for $\mathbb{S}^3$, $\mathbb{S}^6$ and $\mathbb{S}^9$. The vertical axes measure the mean angular errors. The horizontal axes measure r^2. The maximum noise level is at least 16%.

is $r^2/(1+r^2+\sqrt{1+r^2})$. We vary r between 0 and 1, so $r^2/(1+r^2+\sqrt{1+r^2})$ lies between $r^2/(2+\sqrt{2})$ and $r^2/2$. Although the neighborhood radius is not exactly r, it is roughly cr for some constant c.

We plot the mean angular error against r^2 for $\mathbb{S}^3$ in Figure 3(a). A mean angular error of 1.8° can be achieved with $n = 40$ even when r is as large as 1. We experimented with noisy data by adding random noise. We perturb each sample point by a random displacement chosen from $[0, 0.08r^2]$ in a random direction in $\mathbb{R}^d$. Therefore, the maximum noise level is at least 16% of $\max_{\mathsf{z}\in D'} \|f_{m+1}(\mathsf{z}), \cdots, f_d(\mathsf{z})\|$. We plot the mean angular error against r^2 for $\mathbb{S}^3$ in the noisy case in Figure 3(b). A mean angular error of roughly 3.5° can still be achieved with $n = 40$ even when r is as large as 1. Both plots in the noiseless and noisy cases demonstrate that the angular error is roughly proportional to r^2. Figures 3(c)–(f) show the plots of the mean angular errors for $\mathbb{S}^6$ and $\mathbb{S}^9$ in the noiseless and noisy cases.

3.2 Manifold with saddles

Let S^{m-1} be the $(m-1)$-sphere centered at the origin with radius 4 in the subspace spanned by the $x_1, \ldots, x_m$ axes. For each point $\mathsf{q} \in S^{m-1}$, construct the circle centered at q with radius 2 embedded in the plane spanned by the vector q and the x_{m+1}-axis. The union of all such



(a) $S^2 \times C$: noiseless (b) $S^2 \times C$: noisy

(c) $S^5 \times C$: noiseless (d) $S^5 \times C$: noisy

(e) $S^8 \times C$: noiseless (f) $S^8 \times C$: noisy

Figure 4: Plots of mean angular error for $S^2 \times C$, $S^5 \times C$ and $S^8 \times C$. The vertical axes measure the mean angular errors. The horizontal axes measure r^2. The maximum noise level is at least 16%.

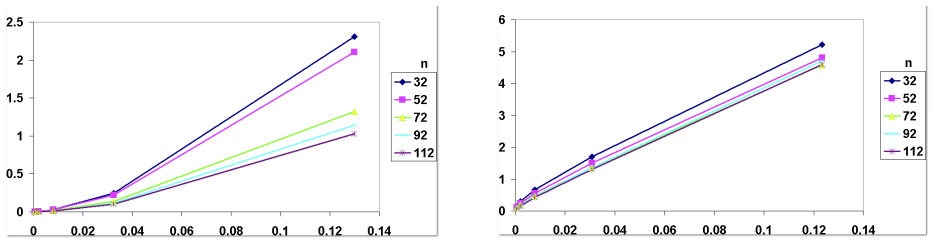
circles is an m-dimensional manifold, which we denote by $S^{m-1} \times C$. We choose the point p on the x_1-axis at distance 2 from the origin, which is a saddle of $S^{m-1} \times C$. We sample n points in an m-ball D of radius r and tangent at p. Then, we lift the points orthogonally away from D and onto $S^{m-1} \times C$.

Since we lift sampled points in D orthogonally onto $S^{m-1} \times C$, D is part of the domain of the coordinate functions $f_{m+1}, \cdots, f_d$. The maximum distance between D and $S^{m-1} \times C$ is $\max_{\mathsf{z} \in D} \|f_{m+1}(\mathsf{z}), \cdots, f_d(\mathsf{z})\|$, which by elementary trigonometry is $r^2/(2 + \sqrt{4 - r^2})$. We vary r between 0 and 1.6, so $r^2/(2 + \sqrt{4 - r^2})$ lies between $0.25r^2$ and $0.3125r^2$. Although the neighborhood radius is not exactly r, it is roughly cr for some constant c.

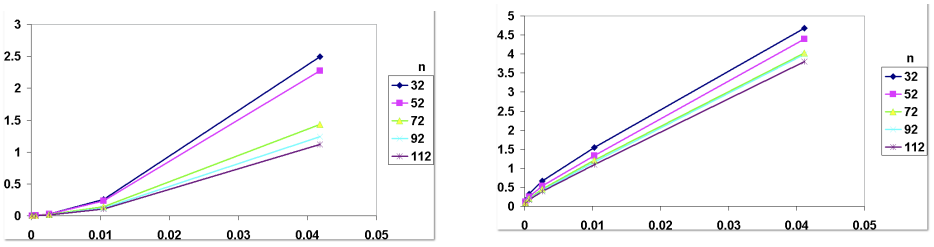
We plot the mean angular error against r^2 for $S^2 \times C$ in Figure 4(a). Random noise is added in the same way as before. We perturb each sample point by a random displacement chosen from $[0, 0.05r^2]$ in a random direction in $\mathbb{R}^d$. Therefore, the maximum noise level is at least 16% of $\max_{\mathsf{z} \in D} \|f_{m+1}(\mathsf{z}), \cdots, f_d(\mathsf{z})\|$. Figure 4(b) shows the plot of the mean angular error against r^2 for $S^2 \times C$ in the noisy case. Figures 4(c)–(f) show the plots for $S^5 \times C$ and $S^8 \times C$.

3.3 Curve

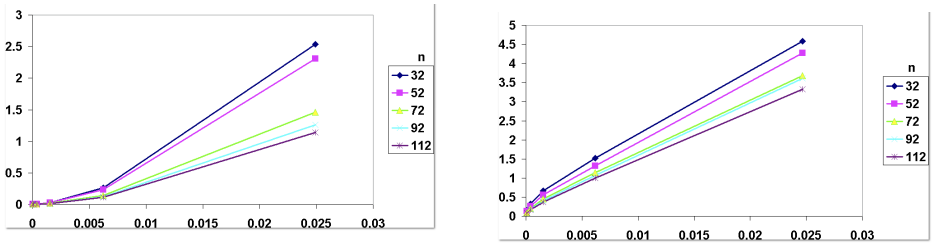
We experimented with the curve $\varphi : [0, \pi] \to \mathbb{R}^d$ such that $\varphi(\theta) = (\theta, \sin\theta, \ldots, \sin\theta)$ for $d =$ 20, 60 and 100. We pick the point $\mathsf{p} = (\pi/2, 1, \ldots, 1)$, where the curve twists a lot. We vary



(a) $\mathbb{R}^{20}$: noiseless (b) $\mathbb{R}^{20}$: noisy



(c) $\mathbb{R}^{60}$: noiseless (d) $\mathbb{R}^{60}$: noisy



(a) $\mathbb{R}^{100}$: noiseless (b) $\mathbb{R}^{100}$: noisy

Figure 5: Plots of mean angular error for the sinusoidal curve in $\mathbb{R}^{20}$, $\mathbb{R}^{60}$ and $\mathbb{R}^{100}$. The vertical axes measure the mean angular error. The horizontal axes measure r^2/d. The maximum noise level is roughly 16%.

r between 0 and $\pi/2$. For each r, the sample points are generated by sampling n values of θ from $\left[\pi/2 - r/\sqrt{d},\ \pi/2 + r/\sqrt{d}\right]$. The minimum value of $\sin\theta$ over $\left[\pi/2 - r/\sqrt{d},\ \pi/2 + r/\sqrt{d}\right]$ is equal to $\cos(r/\sqrt{d}) \approx 1 - r^2/(2d)$. Therefore, the neighborhood radius is roughly $(r^2/d + (d-1)r^4/(4d^2))^{1/2} \approx r/\sqrt{d}$. We plot the mean angular error against $r^2/20$ for $\mathbb{R}^{20}$ in Figure 5(a).

Random noise is added by perturbing each sample point by a random displacement chosen from $[0, 0.08r^2/\sqrt{d}]$ in a random direction in $\mathbb{R}^d$. As explained before, the minimum value of $\sin\theta$ over $[\pi/2 - r/\sqrt{d},\ \pi/2 + r/\sqrt{d}]$ is roughly $1 - r^2/(2d)$. For every $\ell \in [2, d]$, the coordinate function $f_\ell(\theta)$ is $1 - \sin\theta \approx r^2/(2d)$. So the maximum value of $\|f_2(\theta), \cdots, f_d(\theta)\|$ is roughly $\sqrt{(d-1)r^4/(4d^2)} \approx r^2/(2\sqrt{d})$. It follows that the maximum noise magnitude of $0.08r^2/\sqrt{d}$ is roughly 16% of the maximum value of $\|f_2(\theta), \cdots, f_d(\theta)\|$. Figure 5(b) shows the plot for the noisy cases in $\mathbb{R}^{20}$. Figures 5(c)–(f) show the plots for $\mathbb{R}^{60}$ and $\mathbb{R}^{100}$.

4 Transformation to an eigenvalue problem

In this section, we show how to reduce our tangent estimation problem to finding the eigenvalues of the matrix $\mathsf{H}\mathsf{H}^t$ as described in Section 2.2.

For every d-dimensional vector $\widehat{\mathsf{g}}_\ell$ that satisfies (2.5) and every $\binom{d+1}{2}$-dimensional vector $\widehat{\mathsf{c}}_\ell$,

we can apply (2.4) to evaluate the corresponding $\widehat{F}_\ell$ at a_p for $p \in [1,n]$. Since $F_\ell(\mathsf{a}_p) = 0$ and $\widehat{F}_\ell$ is supposed to approximate F_ℓ, we should choose $\widehat{\mathsf{g}}_\ell$ and $\widehat{\mathsf{c}}_\ell$ to minimize some fitting error. Lemma 4.1 below shows that solving the eigenvalue problem for the matrix $\mathsf{H}\mathsf{H}^t$ gives the best, less "curvy" fit.

Since $F_\ell(\mathsf{a}_p) = 0$, the mean squared error $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2$ is a natural error measure. However, since $n \le \binom{d+1}{2}$, the system is under-determined and $\widehat{F}_\ell(\mathsf{a}_p)$ can be made zero for all $p \in [1,n]$ for many choices of $\widehat{\mathsf{g}}_\ell$ and $\widehat{\mathsf{c}}_\ell$. We change the objective function to $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2 + \sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$ for some positive parameter α. Intuitively, $\widehat{\mathsf{c}}_\ell$ is an "approximation" of the linearization c_ℓ of Q_ℓ, so the penalty $\sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$ favors a less "curvy" fit. We convert this optimization problem to a matrix problem as follows. Define:

$$\begin{aligned} d_0 &\stackrel{\text{def}}{=} \tbinom{d+1}{2} \\ \mathsf{U}_\alpha &\stackrel{\text{def}}{=} \tfrac{1}{n}\widehat{\mathsf{B}}^t\widehat{\mathsf{B}} + \alpha\mathsf{I}_{d_0} \\ \mathsf{H}_\alpha &\stackrel{\text{def}}{=} \tfrac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}}\mathsf{U}_\alpha^{-1} \\ \mathsf{W}_\alpha &\stackrel{\text{def}}{=} \tfrac{1}{n}\mathsf{A}^t\mathsf{A} - \mathsf{H}_\alpha\mathsf{U}_\alpha\mathsf{H}_\alpha^t \end{aligned} \tag{4.1}$$

U_α is a $d_0 \times d_0$ matrix, H_α is an $d \times d_0$ matrix, and W_α is a $d \times d$ matrix. U_α and W_α are square and symmetric, and U_α is invertible because $\alpha > 0$. Lemma 4.1(i) below shows that the minimization of $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2 + \sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$ subject to (2.5) is equivalent to finding the smallest $d-m$ eigenvalues of W_α. Lemma 4.1(i) also gives the optimal setting of $\widehat{\mathsf{c}}_\ell$.

A typical experimental approach is to solve the above optimization problem on some training data in order to tune the parameter α. It may thus be time-consuming to set α appropriately. For a small enough α, the quantity $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2$ will be made zero as the system is under-determined, which means that the objective function value is effectively $\sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$. Thus, a curvy fit is penalized no matter how small α is. This motivates us to push α to zero in the limit. Lemma 4.1(ii) and (iii) show that $\lim_{\alpha\to 0}\frac{1}{\alpha}\mathsf{W}_\alpha$ gives another eigenvalue problem that does not require any parameter, which is the problem solved by our tangent estimation algorithm in Section 2.2.

Lemma 4.1

(i) *For all $\alpha > 0$ and for all mutually orthogonal unit vectors $\widehat{\mathsf{g}}_{m+1}, \ldots, \widehat{\mathsf{g}}_d$, the value of $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2 + \sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$ is minimized with respect to α and $\widehat{\mathsf{g}}_{m+1}, \ldots, \widehat{\mathsf{g}}_d$ when $\widehat{\mathsf{c}}_\ell = -\mathsf{H}_\alpha^t\,\widehat{\mathsf{g}}_\ell$ for every $\ell \in [m+1, d]$, and this minimum is equal to $\sum_{\ell=m+1}^{d}\widehat{\mathsf{g}}_\ell^t\,\mathsf{W}_\alpha\,\widehat{\mathsf{g}}_\ell$.*

(ii) $\frac{1}{\alpha}\mathsf{W}_\alpha = \mathsf{A}^t\mathsf{L}\Sigma_\alpha\mathsf{L}^t\mathsf{A}$, *where* $\Sigma_\alpha = \mathrm{diag}_n\Big(\frac{1}{\alpha n+\lambda_1^2}, \ldots, \frac{1}{\alpha n+\lambda_{m_0}^2}, \frac{1}{\alpha n+\lambda_{m_0+1}^2}, \ldots, \frac{1}{\alpha n+\lambda_{m_0+1}^2}\Big)$.

(iii) *Let* $\mathsf{H} = \lim_{\alpha\to 0}\mathsf{H}_\alpha$. *Let* $\Sigma = \lim_{\alpha\to 0}\Sigma_\alpha$. *Then,* $\mathsf{H} = (\widehat{\mathsf{B}}^\dagger\mathsf{A})^t$ *and* $\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A} = \mathsf{H}\mathsf{H}^t$. *For every $\ell \in [m+1, d]$, if we set $\widehat{\mathsf{g}}_\ell$ to be the unit eigenvector corresponding to the ℓth largest eigenvalue of $\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A}$ and set $\widehat{\mathsf{c}}_\ell = -\mathsf{H}^t\widehat{\mathsf{g}}_\ell$, then $\widehat{F}_\ell(\mathsf{a}_p) = 0$ for every $p \in [1,n]$.*

Proof. We follow the argument of Taubin who proved a result similar to (i) for reconstructing algebraic curves [32] in the presence of enough sample points so that the system is not under-

determined and the penalty $\sum_{\ell=m+1}^{d} \alpha \|\widehat{\mathsf{c}}_\ell\|^2$ is not needed. By (2.4),

$$\begin{aligned}\frac{1}{n}\sum_{p=1}^{n} \widehat{F}_\ell(\mathsf{a}_p)^2 &= \frac{1}{n}\begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\begin{pmatrix}\mathsf{A}^t \\ \widehat{\mathsf{B}}^t\end{pmatrix}\begin{pmatrix}\mathsf{A} & \widehat{\mathsf{B}}\end{pmatrix}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix} \\ &= \begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\begin{pmatrix}\frac{1}{n}\mathsf{A}^t\mathsf{A} & \frac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}} \\ \frac{1}{n}\widehat{\mathsf{B}}^t\mathsf{A} & \frac{1}{n}\widehat{\mathsf{B}}^t\widehat{\mathsf{B}}\end{pmatrix}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix}.\end{aligned}$$

Also, $\alpha\|\widehat{\mathsf{c}}_\ell\|^2 = \begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\begin{pmatrix}\mathsf{0}_{d,d} & \mathsf{0}_{d,d_0} \\ \mathsf{0}_{d_0,d} & \alpha\mathsf{I}_{d_0}\end{pmatrix}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix}$. Adding it to the above gives:

$$\frac{1}{n}\sum_{p=1}^{n} \widehat{F}_\ell(\mathsf{a}_p)^2 + \alpha\|\widehat{\mathsf{c}}_\ell\|^2 = \begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\mathsf{Z}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix},$$

where

$$\mathsf{Z} = \begin{pmatrix}\frac{1}{n}\mathsf{A}^t\mathsf{A} & \frac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}} \\ \frac{1}{n}\widehat{\mathsf{B}}^t\mathsf{A} & \frac{1}{n}\widehat{\mathsf{B}}^t\widehat{\mathsf{B}} + \alpha\mathsf{I}_{d_0}\end{pmatrix} = \begin{pmatrix}\mathsf{W}_\alpha + \mathsf{H}_\alpha\mathsf{U}_\alpha\mathsf{H}_\alpha^t & \mathsf{H}_\alpha\mathsf{U}_\alpha \\ \mathsf{U}_\alpha\mathsf{H}_\alpha^t & \mathsf{U}_\alpha\end{pmatrix}.$$

We use the fact that U_α is symmetric in deriving $\mathsf{U}_\alpha\mathsf{H}_\alpha^t$ in the lower left quadrant of Z. We write Z as the product of three matrices:

$$\mathsf{Z} = \begin{pmatrix}\mathsf{I}_d & \mathsf{H}_\alpha \\ \mathsf{0}_{d_0,d} & \mathsf{I}_{d_0}\end{pmatrix}\begin{pmatrix}\mathsf{W}_\alpha & \mathsf{0}_{d,d_0} \\ \mathsf{0}_{d_0,d} & \mathsf{U}_\alpha\end{pmatrix}\begin{pmatrix}\mathsf{I}_d & \mathsf{0}_{d,d_0} \\ \mathsf{H}_\alpha^t & \mathsf{I}_{d_0}\end{pmatrix}.$$

Then,

$$\begin{aligned}\begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\mathsf{Z}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix} &= \begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t\end{pmatrix}\begin{pmatrix}\mathsf{I}_d & \mathsf{H}_\alpha \\ \mathsf{0}_{d_0,d} & \mathsf{I}_{d_0}\end{pmatrix}\begin{pmatrix}\mathsf{W}_\alpha & \mathsf{0}_{d,d_0} \\ \mathsf{0}_{d_0,d} & \mathsf{U}_\alpha\end{pmatrix}\begin{pmatrix}\mathsf{I}_d & \mathsf{0}_{d,d_0} \\ \mathsf{H}_\alpha^t & \mathsf{I}_{d_0}\end{pmatrix}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell\end{pmatrix} \\ &= \begin{pmatrix}\widehat{\mathsf{g}}_\ell^t & \widehat{\mathsf{c}}_\ell^t + \widehat{\mathsf{g}}_\ell^t\mathsf{H}_\alpha\end{pmatrix}\begin{pmatrix}\mathsf{W}_\alpha & \mathsf{0}_{d,d_0} \\ \mathsf{0}_{d_0,d} & \mathsf{U}_\alpha\end{pmatrix}\begin{pmatrix}\widehat{\mathsf{g}}_\ell \\ \widehat{\mathsf{c}}_\ell + \mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell\end{pmatrix} \\ &= \widehat{\mathsf{g}}_\ell^t\mathsf{W}_\alpha\widehat{\mathsf{g}}_\ell + \left(\widehat{\mathsf{c}}_\ell + \mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell\right)^t\mathsf{U}_\alpha\left(\widehat{\mathsf{c}}_\ell + \mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell\right).\end{aligned}$$

Observe that U_α is positive semidefinite. It implies that $\left(\widehat{\mathsf{c}}_\ell + \mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell\right)^t\mathsf{U}_\alpha\left(\widehat{\mathsf{c}}_\ell + \mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell\right) \geq 0$. The minimum value of $\frac{1}{n}\sum_{\ell=m+1}^{d}\sum_{p=1}^{n}\widehat{F}_\ell(\mathsf{a}_p)^2 + \sum_{\ell=m+1}^{d}\alpha\|\widehat{\mathsf{c}}_\ell\|^2$ is thus achieved when $\widehat{\mathsf{c}}_\ell = -\mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell$ for every $\ell \in [m+1, d]$, and this minimum equals $\sum_{\ell=m+1}^{d}\widehat{\mathsf{g}}_\ell^t\mathsf{W}_\alpha\widehat{\mathsf{g}}_\ell$.

Consider (ii). Plugging the definition $\mathsf{H}_\alpha = \frac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}}\mathsf{U}_\alpha^{-1}$ into the definition of W_α gives:

$$\begin{aligned}\mathsf{W}_\alpha &= \tfrac{1}{n}\mathsf{A}^t\mathsf{A} - \mathsf{H}_\alpha\mathsf{U}_\alpha\mathsf{H}_\alpha^t \\ &= \tfrac{1}{n}\mathsf{A}^t\mathsf{A} - \tfrac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}}\mathsf{H}_\alpha^t \\ &= \tfrac{1}{n}\mathsf{A}^t\mathsf{A} - \tfrac{1}{n^2}\mathsf{A}^t\widehat{\mathsf{B}}\left(\mathsf{U}_\alpha^{-1}\right)^t\widehat{\mathsf{B}}^t\mathsf{A} \\ &= \tfrac{1}{n}\mathsf{A}^t\Big(\mathsf{I}_n - \tfrac{1}{n}\widehat{\mathsf{B}}\big(\mathsf{U}_\alpha^{-1}\big)^t\widehat{\mathsf{B}}^t\Big)\mathsf{A}. \qquad (4.2)\end{aligned}$$

Recall that $m_0 < n \leq d_0$ by assumption and $\mathsf{L}\widehat{\Lambda}\mathsf{R}^t$ is the thin SVD of $\widehat{\mathsf{B}}$. Let $\mathsf{L}\begin{pmatrix}\widehat{\Lambda} & \mathsf{0}_{n,d_0-n}\end{pmatrix}\bar{\mathsf{R}}^t$ be the full SVD of $\widehat{\mathsf{B}}$. So L is an $n \times n$ orthogonal matrix and $\bar{\mathsf{R}}$ is a $d_0 \times d_0$ orthogonal matrix,

which means $\mathsf{L}^t = \mathsf{L}^{-1}$ and $\bar{\mathsf{R}}^t = \bar{\mathsf{R}}^{-1}$. Moreover, R is the leftmost $d_0 \times n$ submatrix of $\bar{\mathsf{R}}$.

$$\begin{aligned}
\mathsf{U}_\alpha &= \tfrac{1}{n}\widehat{\mathsf{B}}^t\widehat{\mathsf{B}} + \alpha \mathsf{I}_{d_0} \\
&= \tfrac{1}{n}\bar{\mathsf{R}}\Big(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n}\Big)^t \Big(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n}\Big)\bar{\mathsf{R}}^t + \alpha \mathsf{I}_{d_0} \\
&= \bar{\mathsf{R}} \cdot \mathrm{diag}_{d_0}\Big(\tfrac{\lambda_1^2}{n}, \cdots, \tfrac{\lambda_{m_0}^2}{n}, \tfrac{\lambda_{m_0+1}^2}{n}, \cdots, \tfrac{\lambda_{m_0+1}^2}{n}, \underbrace{0, \cdots, 0}_{d_0-n \text{ copies}}\Big) \cdot \bar{\mathsf{R}}^t + \bar{\mathsf{R}}(\alpha \mathsf{I}_{d_0})\bar{\mathsf{R}}^t \\
&= \bar{\mathsf{R}} \cdot \mathrm{diag}_{d_0}\Big(\tfrac{\alpha n+\lambda_1^2}{n}, \cdots, \tfrac{\alpha n+\lambda_{m_0}^2}{n}, \tfrac{\alpha n+\lambda_{m_0+1}^2}{n}, \cdots, \tfrac{\alpha n+\lambda_{m_0+1}^2}{n}, \underbrace{\alpha, \cdots, \alpha}_{d_0-n \text{ copies}}\Big) \cdot \bar{\mathsf{R}}^t.
\end{aligned}$$

It implies that

$$\mathsf{U}_\alpha^{-1} = \bar{\mathsf{R}} \cdot \mathrm{diag}_{d_0}\Big(\tfrac{n}{\alpha n+\lambda_1^2}, \cdots, \tfrac{n}{\alpha n+\lambda_{m_0}^2}, \tfrac{n}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{n}{\alpha n+\lambda_{m_0+1}^2}, \tfrac{1}{\alpha}, \cdots, \tfrac{1}{\alpha}\Big) \cdot \bar{\mathsf{R}}^t. \quad (4.3)$$

Therefore,

$$\begin{aligned}
\tfrac{1}{n}\widehat{\mathsf{B}}\left(\mathsf{U}_\alpha^{-1}\right)^t\widehat{\mathsf{B}}^t &= \mathsf{L}\Big(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n}\Big) \cdot \\
&\quad \mathrm{diag}_{d_0}\Big(\tfrac{1}{\alpha n+\lambda_1^2}, \cdots, \tfrac{1}{\alpha n+\lambda_{m_0}^2}, \tfrac{1}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{1}{\alpha n+\lambda_{m_0+1}^2}, \tfrac{1}{\alpha n}, \cdots, \tfrac{1}{\alpha n}\Big) \cdot \\
&\quad \Big(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n}\Big)^t \mathsf{L}^t \\
&= \mathsf{L} \cdot \mathrm{diag}_n\Big(\tfrac{\lambda_1^2}{\alpha n+\lambda_1^2}, \cdots, \tfrac{\lambda_{m_0}^2}{\alpha n+\lambda_{m_0}^2}, \tfrac{\lambda_{m_0+1}^2}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{\lambda_{m_0+1}^2}{\alpha n+\lambda_{m_0+1}^2}\Big) \cdot \mathsf{L}^t.
\end{aligned}$$

Plugging this into (4.2) gives:

$$\begin{aligned}
\mathsf{W}_\alpha &= \tfrac{1}{n}\mathsf{A}^t\mathsf{L}\Big(\mathsf{I}_n - \mathrm{diag}_n\Big(\tfrac{\lambda_1^2}{\alpha n+\lambda_1^2}, \cdots, \tfrac{\lambda_{m_0}^2}{\alpha n+\lambda_{m_0}^2}, \tfrac{\lambda_{m_0+1}^2}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{\lambda_{m_0+1}^2}{\alpha n+\lambda_{m_0+1}^2}\Big)\Big)\mathsf{L}^t\mathsf{A} \\
&= \tfrac{1}{n}\mathsf{A}^t\mathsf{L} \cdot \mathrm{diag}_n\Big(\tfrac{\alpha n}{\alpha n+\lambda_1^2}, \cdots, \tfrac{\alpha n}{\alpha n+\lambda_{m_0}^2}, \tfrac{\alpha n}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{\alpha n}{\alpha n+\lambda_{m_0+1}^2}\Big) \cdot \mathsf{L}^t\mathsf{A} \\
&= \alpha \mathsf{A}^t\mathsf{L}\Sigma_\alpha\mathsf{L}^t\mathsf{A}.
\end{aligned}$$

Consider (iii). Using the definition of H_α, (4.3), and the full SVD of $\widehat{\mathsf{B}}$, we obtain

$$\begin{aligned}
\mathsf{H}_\alpha &= \tfrac{1}{n}\mathsf{A}^t\widehat{\mathsf{B}}\mathsf{U}_\alpha^{-1} \\
&= \mathsf{A}^t\mathsf{L}\big(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n}\big) \cdot \mathrm{diag}_{d_0}\Big(\tfrac{1}{\alpha n+\lambda_1^2}, \cdots, \tfrac{1}{\alpha n+\lambda_{m_0}^2}, \tfrac{1}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{1}{\alpha n+\lambda_{m_0+1}^2}, \tfrac{1}{\alpha n}, \cdots, \tfrac{1}{\alpha n}\Big) \cdot \bar{\mathsf{R}}^t \\
&= \mathsf{A}^t\mathsf{L} \cdot \mathrm{diag}_n\Big(\tfrac{\lambda_1}{\alpha n+\lambda_1^2}, \cdots, \tfrac{\lambda_{m_0}}{\alpha n+\lambda_{m_0}^2}, \tfrac{\lambda_{m_0+1}}{\alpha n+\lambda_{m_0+1}^2}, \cdots, \tfrac{\lambda_{m_0+1}}{\alpha n+\lambda_{m_0+1}^2}\Big) \cdot \mathsf{R}^t.
\end{aligned}$$

In the last step, we use the property that R is the leftmost $d_0 \times n$ submatrix of $\bar{\mathsf{R}}$. Define $\mathsf{H} = \lim_{\alpha\to 0}\mathsf{H}_\alpha$ and $\Sigma = \lim_{\alpha\to 0}\Sigma_\alpha$. As a result,

$$\mathsf{H} = \lim_{\alpha\to 0}\mathsf{H}_\alpha = \mathsf{A}^t\mathsf{L} \cdot \mathrm{diag}_n\Big(\tfrac{1}{\lambda_1}, \cdots, \tfrac{1}{\lambda_{m_0}}, \tfrac{1}{\lambda_{m_0+1}}, \cdots, \tfrac{1}{\lambda_{m_0+1}}\Big) \cdot \mathsf{R}^t.$$

Observe that $\widehat{\mathsf{B}}^\dagger = \mathsf{R} \cdot \mathrm{diag}_n(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_{m_0}}, \frac{1}{\lambda_{m_0+1}}, \cdots, \frac{1}{\lambda_{m_0+1}}) \cdot \mathsf{L}^t$. Therefore, $\mathsf{H} = (\widehat{\mathsf{B}}^\dagger\mathsf{A})^t$ and $\mathsf{H}\mathsf{H}^t = \mathsf{A}^t\mathsf{L} \cdot \mathrm{diag}_n(\frac{1}{\lambda_1^2}, \cdots, \frac{1}{\lambda_{m_0}^2}, \frac{1}{\lambda_{m_0+1}^2}, \cdots \frac{1}{\lambda_{m_0+1}^2}) \cdot \mathsf{L}^t\mathsf{A} = \mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A}$.

Take any $\ell \in [m+1, d]$. Let e_α be the unit eigenvector corresponding to the ℓth largest eigenvalue of $\mathsf{A}^t\mathsf{L}\Sigma_\alpha\mathsf{L}^t\mathsf{A}$ for a positive but arbitrarily small α. By (i), $\widehat{\mathsf{c}}_\ell$ should be set to

$-\mathsf{H}_\alpha^t\mathsf{e}_\alpha$ to minimize the fitting error, which makes $\mathsf{e}_\alpha^t\mathsf{W}_\alpha\mathsf{e}_\alpha = \frac{1}{n}\sum_{p=1}^n \widehat{F}_\ell(\mathsf{a}_p)^2 + \alpha\|\widehat{\mathsf{c}}_\ell\|^2 = \frac{1}{n}\sum_{p=1}^n \widehat{F}_\ell(\mathsf{a}_p)^2 + \alpha\mathsf{e}_\alpha^t(\mathsf{H}_\alpha\mathsf{H}_\alpha^t)\mathsf{e}_\alpha$. By (ii),

$$\mathsf{e}_\alpha^t(\mathsf{A}^t\mathsf{L}\Sigma_\alpha\mathsf{L}^t\mathsf{A})\mathsf{e}_\alpha \;=\; \mathsf{e}_\alpha^t(\tfrac{1}{\alpha}\mathsf{W}_\alpha)\mathsf{e}_\alpha \;=\; \frac{1}{\alpha n}\sum_{p=1}^n \widehat{F}_\ell(\mathsf{a}_p)^2 + \mathsf{e}_\alpha^t(\mathsf{H}_\alpha\mathsf{H}_\alpha^t)\mathsf{e}_\alpha.$$

As $\alpha \to 0$, the left hand side and the second term on the right hand side converge to the same value. Therefore, if we set $\widehat{\mathsf{g}}_\ell = \mathsf{e}_\alpha$ and $\widehat{\mathsf{c}}_\ell = -\mathsf{H}_\alpha^t\widehat{\mathsf{g}}_\ell$, then $\lim_{\alpha\to 0}\frac{1}{\alpha n}\sum_{p=1}^n \widehat{F}_\ell(\mathsf{a}_p)^2 = 0$, which makes $\lim_{\alpha\to 0}\widehat{F}_\ell(\mathsf{a}_p) = 0$ for every $p \in [1, n]$. This proves (iii). ◻

Remark: The condition $n \leq \binom{d+1}{2}$ is needed for the proofs of Lemma 4.1(ii) and the identity $\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A} = \mathsf{H}\mathsf{H}^t$ in Lemma 4.1(iii) to go through. The subsequent error analysis studies the eigenvalues of $\mathsf{H}\mathsf{H}^t$ and use the identity $\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A} = \mathsf{H}\mathsf{H}^t$ to bound the angular error.

5 Derivatives and Sums

5.1 Derivatives of coordinate functions

We derive two results in this subsection. Lemma 5.1(i) puts an upper bound on the 2-norm of the second derivatives of the coordinate functions. Lemma 5.1(ii) bounds the error of approximating the coordinate functions using their second derivatives. We use Taylor expansions [23] heavily in the proofs. Given a smooth function $h : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is an open subset of $\mathbb{R}^k$ for some $k \geq 1$, if two points $\mathsf{a}, \mathsf{x} \in \mathcal{X}$ are connected by a segment in $\mathcal{X}$, then for every integer $j \geq 1$, there exists a point z in the interior of the segment connecting a and x such that

$$\begin{aligned} h(\mathsf{x}) \;=\; & h(\mathsf{a}) + \mathrm{D}h|_{\mathsf{a}}(\mathsf{x}-\mathsf{a}) + \cdots + \frac{1}{(j-1)!}\mathrm{D}^{j-1}h|_{\mathsf{a}}(\mathsf{x}-\mathsf{a},\ldots,\mathsf{x}-\mathsf{a}) + \\ & \frac{1}{j!}\mathrm{D}^j h|_{\mathsf{z}}(\mathsf{x}-\mathsf{a},\ldots,\mathsf{x}-\mathsf{a}). \end{aligned}$$

If we let u be the unit vector $(\mathsf{x}-\mathsf{a})/\|\mathsf{x}-\mathsf{a}\|$ and let $\varepsilon = \|\mathsf{x}-\mathsf{a}\|$, then

$$h(\mathsf{x}) = h(\mathsf{a}) + \varepsilon\,\mathrm{D}h|_{\mathsf{a}}(\mathsf{u}) + \cdots + \frac{\varepsilon^{j-1}}{(j-1)!}\mathrm{D}^{j-1}h|_{\mathsf{a}}(\mathsf{u},\ldots,\mathsf{u}) + \frac{\varepsilon^j}{j!}\mathrm{D}^j h|_{\mathsf{z}}(\mathsf{u},\ldots,\mathsf{u}).$$

As $\varepsilon = \|\mathsf{x}-\mathsf{a}\|$ approaches zero, the magnitude of the remainder $\frac{\varepsilon^j}{j!}\mathrm{D}^j h|_{\mathsf{z}}(\mathsf{u},\ldots,\mathsf{u})$ approaches zero faster than the magnitude of every preceding term in the expansion. In our case, we will use the Taylor expansion of a coordinate function f_ℓ around the origin, and there are no constant and linear term in the expansion.

Lemma 5.1 *There exists a value $\varepsilon_0 \in (0,1)$ such that for every $\varepsilon \in (0, \varepsilon_0]$ and every unit vector $\mathsf{u} \in \mathbb{R}^m$, the following properties hold.*

(i) $\left\|\begin{pmatrix} \mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) & \cdots & \mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) \end{pmatrix}\right\| = O(1/\varrho)$.

(ii) *Let* $\mathsf{v} = \varrho\varepsilon\mathsf{u}$. $\left\|\begin{pmatrix} f_{m+1}(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) & \cdots & f_d(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \end{pmatrix}\right\| = O(\varrho\varepsilon^3)$.

Proof. Rotate space such that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$ and the coordinate axes $x_{m+1}, \ldots, x_d$ span the normal space of $\mathcal{M}$ at the origin. Let $\mathsf{v} = \varrho\varepsilon\mathsf{u}$.

Consider (i). Take any $\ell \in [m+1, d]$. We claim that if ε_0 is sufficiently small, then $|f_\ell(\mathsf{v})| \geq \frac{1}{4}\varrho^2\varepsilon^2\left|\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})\right|$ for all $\varepsilon \in (0, \varepsilon_0]$. If $\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) = 0$, our claim is trivially true. Assume that

$\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) \neq 0$. The Taylor expansion of $f_\ell(\mathsf{v})$ is $\frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) + \cdots = \frac{1}{2}\varrho^2\varepsilon^2\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) + \cdots$. As ε_0 approaches 0, the remainder approaches zero faster than $\frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v})$ does. Therefore, there exists $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0]$ and $\mathsf{v} = \varrho\varepsilon\mathsf{u}$,

$$\left| f_\ell(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \right| \leq \frac{1}{4}\left|\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v})\right| = \frac{1}{4}\varrho^2\varepsilon^2\left|\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})\right|.$$

Thus, $|f_\ell(\mathsf{v})| \geq \frac{1}{4}\varrho^2\varepsilon^2\left|\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})\right|$, proving our claim.

Any point in $\mathcal{M}$ within a distance $\varrho\varepsilon$ from the origin is at distance $O(\varrho\varepsilon^2)$ from $\mathcal{T}$ [15, Lemma 6], which implies that $\| \begin{pmatrix} f_{m+1}(\mathsf{v}) & \cdots & f_d(\mathsf{v}) \end{pmatrix} \| = O(\varrho\varepsilon^2)$. Therefore, by our claim,

$$\frac{1}{4}\varrho^2\varepsilon^2 \left\| \begin{pmatrix} \mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) & \cdots & \mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} f_{m+1}(\mathsf{v}) & \cdots & f_d(\mathsf{v}) \end{pmatrix} \right\| = O(\varrho\varepsilon^2).$$

Dividing both sides by $\varrho^2\varepsilon^2/4$ gives $\left\| \begin{pmatrix} \mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) & \cdots & \mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) \end{pmatrix} \right\| = O(1/\varrho)$, establishing the correctness of (i).

Consider (ii). Let B_ε denote the d-ball centered at the origin with radius $\varrho\varepsilon$. Take any $\ell \in [m{+}1, d]$. Define the plane $\mathcal{L} = \{a\mathsf{u}{+}\mathsf{z} : a \in \mathbb{R} \wedge \mathsf{z} \text{ is a vector parallel to the } x_\ell \text{ axis}\}$. Since u is a vector in the tangent space of $\mathcal{M}$ at the origin, for a sufficiently small ε_0, $\mathcal{M} \cap B_\varepsilon \cap \mathcal{L}$ is a one-dimensional curve in the plane $\mathcal{L}$ and $\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})$ is the rate of change of the ℓ-th coordinate of the unit tangent along $\mathcal{M} \cap B_\varepsilon \cap \mathcal{L}$ at the origin [28, Chapter 5]. That is, $\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})$ is the reciprocal of the radius of curvature of $\mathcal{M} \cap B_\varepsilon \cap \mathcal{L}$ at the origin.

If we scale down the unit length linearly, all lengths increase linearly; in particular, the radius of curvature of $\mathcal{M} \cap B_\varepsilon \cap \mathcal{L}$ at the origin, the local feature size ϱ at the origin, and the value $|f_\ell(\mathsf{v})|$. The linear increase in the radius of curvature means that $\left|\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})\right|$ decreases linearly, which implies that $\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) = \varrho^2\varepsilon^2\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u})$ increases linearly.

A Taylor expansion of $f_\ell(\mathsf{v})$ is $\frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) + \frac{1}{6}\mathrm{D}^3 f_\ell|_{\mathsf{p}_{\ell,\mathsf{u},\varepsilon}}(\mathsf{v},\mathsf{v},\mathsf{v})$, where $\frac{1}{6}\mathrm{D}^3 f_\ell|_{\mathsf{p}_{\ell,\mathsf{u},\varepsilon}}(\mathsf{v},\mathsf{v},\mathsf{v})$ is the remainder and $\mathsf{p}_{\ell,\mathsf{u},\varepsilon} = \varrho\omega_{\ell,\mathsf{u},\varepsilon}\mathsf{u}$ for some $\omega_{\ell,\mathsf{u},\varepsilon} \in (0,\varepsilon)$. Notice that $\mathsf{p}_{\ell,\mathsf{u},\varepsilon}$ and $\omega_{\ell,\mathsf{u},\varepsilon}$ depend on ε and u. The remainder can be rewritten as:

$$f_\ell(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \;=\; \frac{1}{6}\mathrm{D}^3 f_\ell|_{\mathsf{p}_{\ell,\mathsf{u},\varepsilon}}(\mathsf{v},\mathsf{v},\mathsf{v}) \;=\; \varrho\varepsilon^3 \cdot \frac{1}{6}\varrho^2\mathrm{D}^3 f_\ell|_{\mathsf{p}_{\ell,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u}) \tag{5.1}$$

Either $f_\ell(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v})$ is zero or it changes linearly when we scale down the unit length because $f_\ell(\mathsf{v})$ and $\frac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{v},\mathsf{v})$ do. The leading factor $\rho\varepsilon^3$ on the right hand side of (5.1) changes linearly too as we scale down the unit length. Thus, $\frac{1}{6}\varrho^2\mathrm{D}^3 f_\ell|_{\mathsf{p}_{\ell,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u})$ is scale independent.

We have the following equation:

$$\left\| \begin{pmatrix} f_{m+1}(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \\ \vdots \\ f_d(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \end{pmatrix} \right\| \;=\; \varrho\varepsilon^3 \cdot \left\| \begin{pmatrix} \frac{1}{6}\varrho^2\mathrm{D}^3 f_{m+1}|_{\mathsf{p}_{m+1,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u}) \\ \vdots \\ \frac{1}{6}\varrho^2\mathrm{D}^3 f_d|_{\mathsf{p}_{d,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u}) \end{pmatrix} \right\| \tag{5.2}$$

We maximize the norm of the vector on the right hand side of (5.2) by going over all possible unit vectors u in the tangent space at the origin, but v on the left hand side of (5.2) is kept fixed independent of the variation of u. For every u, we also vary $\mathsf{p}_{\ell,\mathsf{u},\varepsilon}$ by varying ε over $(0,\varepsilon_0]$ to maximize the norm of the vector on the right hand side of (5.2). Therefore,

$$\left\| \begin{pmatrix} f_{m+1}(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_{m+1}|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \\ \vdots \\ f_d(\mathsf{v}) - \frac{1}{2}\mathrm{D}^2 f_d|_{\mathsf{0}}(\mathsf{v},\mathsf{v}) \end{pmatrix} \right\| \;\leq\; \varrho\varepsilon^3 \cdot \sup_{\substack{\|\mathsf{u}\|=1 \\ \varepsilon\in(0,\varepsilon_0]}} \left\| \begin{pmatrix} \frac{1}{6}\varrho^2\mathrm{D}^3 f_{m+1}|_{\mathsf{p}_{m+1,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u}) \\ \vdots \\ \frac{1}{6}\varrho^2\mathrm{D}^3 f_d|_{\mathsf{p}_{d,\mathsf{u},\varepsilon}}(\mathsf{u},\mathsf{u},\mathsf{u}) \end{pmatrix} \right\|.$$

In the inequality above, the rightmost factor on the right hand side is a constant that depends on $\mathcal{M}$ but not on scale. This establishes the correctness of (ii). □

5.2 Sums of powers of coordinates

The main result in this subsection is Lemma 5.4 which bounds certain sums of powers of sample point coordinates. These bounds are proved using integration, the Chebyshev's inequality, and the technical result stated in Lemma 5.3 below. We need some notation. Let p and q denote two non-negative integers.

$$\beta(q) \stackrel{\text{def}}{=} \begin{cases} \dfrac{\pi(q-1)(q-3)\cdots 1}{q(q-2)\cdots 2}, & \text{if } q \text{ is even and positive,} \\ \dfrac{2(q-1)(q-3)\cdots 2}{q(q-2)\cdots 1}, & \text{if } q \text{ is odd.} \end{cases}$$

$$V_q \stackrel{\text{def}}{=} \text{volume of a unit } q\text{-ball.}$$

Note that $\beta(1) = 2$ and $V_0 = 1$. It can be verified that β and V_q satisfy the following recurrences:

$$\forall\, q \geq 2, \qquad \beta(q-1)\beta(q+2) = 2\pi(q+1)/(q(q+2)) \tag{5.3}$$

$$\forall\, q \geq 1, \qquad V_q = \beta(q)V_{q-1} \tag{5.4}$$

$$\forall\, q \geq 2, \qquad V_q = 2\pi V_{q-2}/q \tag{5.5}$$

We will need the following two technical results. The Chebyshev's inequality bounds the probability of a random variable deviating from the mean by a multiple of the standard deviation.

Lemma 5.2 (Chebyshev's Inequality) *Let Y be a random variable with finite expected value μ and finite positive variance σ^2. For any positive real number a, $\Pr(|Y - \mu| \geq a\sigma) \leq 1/a^2$.*

The next result gives the value of a particular integral that will be used often in the proof of Lemma 5.4.

Lemma 5.3 *Let p and q be two non-negative integers. Let r_i, x_i and r_{i+1} be three variables such that $x_i \in [-r_i, r_i]$ and $r_{i+1}^2 = r_i^2 - x_i^2$. If p is non-negative and even, then*

$$\int_{-r_i}^{r_i} x_i^p r_{i+1}^q \,\mathrm{d}x_i = \frac{(p-1)(p-3)\cdots 1}{(p+q+1)(p+q-1)\cdots(q+3)} \cdot \beta(q+1) \cdot r_i^{p+q+1},$$

where $\dfrac{(p-1)(p-3)\cdots 1}{(p+q+1)(p+q-1)\cdots(q+3)}$ *is interpreted as* 1 *when* $p = 0$.

Proof. Perform a change of variables: $x_i = r_i \sin\theta$ and $r_{i+1} = r_i \cos\theta$ and we obtain

$$\int x_i^p r_{i+1}^q \,\mathrm{d}x_i = r_i^{p+q+1} \int \sin^p\theta \cos^{q+1}\theta \,\mathrm{d}\theta.$$

The limits of the integrals also change: $[-r_i, r_i]$ becomes $[-\pi/2, \pi/2]$. The following two recursive formulae are from [19]:

$$\forall\, p \geq 2, q \geq 0, \qquad \int \sin^p\theta \cos^{q+1}\theta \,\mathrm{d}\theta =$$
$$-\frac{\sin^{p-1}\theta \cos^{q+2}\theta}{p+q+1} + \frac{p-1}{p+q+1} \cdot \int \sin^{p-2}\theta \cos^{q+1}\theta \,\mathrm{d}\theta \tag{5.6}$$

$$\forall\, p \geq 0, q \geq 1, \qquad \int \sin^p\theta \cos^{q+1}\theta \,\mathrm{d}\theta =$$
$$\frac{\sin^{p+1}\theta \cos^q\theta}{p+q+1} + \frac{q}{p+q+1} \cdot \int \sin^p\theta \cos^{q-1}\theta \,\mathrm{d}\theta. \tag{5.7}$$

In our case, since the limits of the integrals are $[-\pi/2, \pi/2]$, the leading additive terms in both (5.6) and (5.7) vanish. Then, we first repeatedly apply (5.6) to decrease the exponent p. Since p is even, it will eventually become zero and we then repeatedly apply (5.7) to decrease the exponent q. One can then verify that we obtain the result stated in the lemma. □

Lemma 5.4 below gives upper and lower bounds on certain sums of powers of sample point coordinates. These upper and lower bounds are obtained by calculating the expected values and variances of the sums of powers. Then, the Chebyshev's inequality is applied to obtain the high probability bound. Recall that the sample points in $\mathcal{M}$ are generated by a Poisson process and the neighborhood of the origin being examined contains exactly n sample points (excluding the origin). It follows that these n sample points are uniformly distributed in that neighborhood [5]. As a result, the probability of a sample point falling into a region within the neighborhood is the ratio of the volume of that region to the neighborhood volume. This observation allows us to calculate the expected values and variances of the sums of powers by integration and applying Lemma 5.3.

Lemma 5.4 *Assume that the coordinate axes $x_1, \ldots, x_m$ span the tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin. Let i, j, k, and l be four distinct integers from $[1, m]$. If ε is sufficiently small, then for every constant $c > 0$, the following properties hold simultaneously with probability $1 - O(n^{-1/3})$, where the hidden constant in the probability bound depends on c.*

(i) $\left|\sum_{p=1}^{n} a_{pi}^4 - \frac{3}{(m+2)(m+4)} n\varrho^4\varepsilon^4\right| < cn^{2/3}\varrho^4\varepsilon^4 + O(n\varrho^4\varepsilon^6)$.

(ii) $\left|\sum_{p=1}^{n} a_{pi}^2 a_{pj}^2 - \frac{1}{(m+2)(m+4)} n\varrho^4\varepsilon^4\right| < cn^{2/3}\varrho^4\varepsilon^4 + O(n\varrho^4\varepsilon^6)$.

(iii) $\left|\sum_{p=1}^{n} a_{pi}^2 - \frac{1}{m+2} n\varrho^2\varepsilon^2\right| < cn^{2/3}\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)$.

(iv) $\left|\sum_{p=1}^{n} a_{pi}^3 a_{pj}\right|$, $\left|\sum_{p=1}^{n} a_{pi}^2 a_{pj} a_{pk}\right|$ *and* $\left|\sum_{p=1}^{n} a_{pi} a_{pj} a_{pk} a_{pl}\right|$ *are less than* $cn^{2/3}\varrho^4\varepsilon^4 + O(n\varrho^4\varepsilon^6)$.

(v) $\left|\sum_{p=1}^{n} a_{pi}^3\right|$, $\left|\sum_{p=1}^{n} a_{pi}^2 a_{pj}\right|$ *and* $\left|\sum_{p=1}^{n} a_{pi} a_{pj} a_{pk}\right|$ *are less than* $cn^{2/3}\varrho^3\varepsilon^3 + O(n\varrho^3\varepsilon^5)$.

(vi) $\left|\sum_{p=1}^{n} a_{pi} a_{pj}\right| < cn^{2/3}\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)$.

Proof. Let B_ε be the d-ball centered at the origin with radius $\varrho\varepsilon$. Since $\varrho\varepsilon$ is defined to be the distance from the origin to the $(n+1)$-th sample point, there are exactly n sample points in the interior of B_ε, excluding the origin. As the sample points are generated by a Poisson process, these n sample points are uniformly distributed in the interior of $\mathcal{M} \cap B_\varepsilon$.

Consider an m-dimensional triangulated hyperrectangle R in $\mathcal{T} \cap B_\varepsilon$ with infinitesimal side lengths $\mathrm{d}x_1$, …, $\mathrm{d}x_m$. Suppose that R lies well inside B_ε so that its 2^m vertices are the orthogonal projections of 2^m points in $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$. Connect these 2^m points in $\mathcal{M} \cap B_\varepsilon$ as in the triangulation of R to produce some m-simplices. The union of these m-simplices is an infinitesimal volume U which is a deformed version of R. Let $\mathrm{d}V$ denote the volume of U. Since U projects orthogonally onto R, the volume of U is at least the volume of R, and therefore, $\mathrm{d}V \geq \mathrm{d}x_1 \cdots \mathrm{d}x_m$. For every segment s' in an m-simplex in U, it projects to a segment s in an m-simplex in the triangulation of R such that $\mathrm{length}(s') = \sec\theta \cdot \mathrm{length}(s)$, where θ is the acute angle between the support lines of s' and s. The angle θ is no more than the angle between $\mathcal{T}$ and the tangent space at an endpoint of s', which is known to be $O(\varepsilon)$ [9, Lemma 15]. Therefore, $\mathrm{d}V \leq (1 + O(\varepsilon^2))^m \,\mathrm{d}x_1 \cdots \mathrm{d}x_m = (1 + O(\varepsilon^2))\,\mathrm{d}x_1 \cdots \mathrm{d}x_m$. In the last step, we use the fact that m is a constant, so $m \cdot O(\varepsilon^2) + \binom{m}{2} \cdot O(\varepsilon^4) + \cdots = O(\varepsilon^2)$. It follows that:

$$\mathrm{d}x_1 \ldots \mathrm{d}x_m \;\leq\; \mathrm{d}V \;\leq\; (1 + O(\varepsilon^2)) \cdot \mathrm{d}x_1 \ldots \mathrm{d}x_m \tag{5.8}$$

Define the following symbols:

$$\begin{aligned} r_0 &\stackrel{\text{def}}{=} \varrho\varepsilon \\ r_1 &\stackrel{\text{def}}{=} \kappa r_0 \text{ for some parameter } \kappa \text{ to be specified later} \\ r_{i+1}^2 &\stackrel{\text{def}}{=} r_i^2 - x_i^2, \text{ for every } i \in [1, m-1] \end{aligned}$$

Since the tangent space at any point in $\mathcal{M} \cap B_\varepsilon$ makes an $O(\varepsilon)$ angle with $\mathcal{T}$ [9], the projection of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$ covers an m-ball centered at the origin with radius $r_0 \cos(O(\varepsilon)) = (1 - O(\varepsilon^2))r_0$. The volume of this m-ball is thus $(1-O(\varepsilon^2))^m V_m r_0^m \geq (1-O(\varepsilon^2))V_m r_0^m$. Combining with (5.8), we obtain:

$$(1 - O(\varepsilon^2))V_m r_0^m \;\leq\; \text{vol}(\mathcal{M} \cap B_\varepsilon) \;\leq\; (1 + O(\varepsilon^2))V_m r_0^m. \tag{5.9}$$

We will prove that each of (i)–(vi) holds with probability $1 - O(n^{-1/3})$. Therefore, they hold simultaneously with probability at least $1 - 6 \cdot O(n^{-1/3}) = 1 - O(n^{-1/3})$ as well. In the following, $\mathtt{z} = (z_1, z_2, \ldots, z_d)^t$ denotes a random point in $\mathcal{M} \cap B_\varepsilon$.

Analysis of (i): The variance of $\sum_{p=1}^n a_{pi}^4$ equals $n\mathtt{E}[z_i^8] - n\big(\mathtt{E}[z_i^4]\big)^2 \leq n\mathtt{E}[z_i^8] \leq n\varrho^8\varepsilon^8$ because $|z_i| \leq \varrho\varepsilon$. Lemma 5.2 implies that $\left|\sum_{p=1}^n a_{pi}^4 - \mathtt{E}\left[\sum_{p=1}^n a_{pi}^4\right]\right| < cn^{2/3}\varrho^4\varepsilon^4$ with probability $1 - c^{-2}n^{-1/3}$. It remains to bound $\mathtt{E}\left[\sum_{p=1}^n a_{pi}^4\right]$.

Since $\mathtt{E}[\sum_{p=1}^n a_{pi}^4] = \frac{n}{\text{vol}(\mathcal{M}\cap B_\varepsilon)} \int_{\mathcal{M}\cap B_\varepsilon} z_i^4 \, \mathrm{d}V$, it follows from (5.9) that:

$$\frac{(1 - O(\varepsilon^2))n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^4 \, \mathrm{d}V \;\leq\; \mathtt{E}\left[\sum_{p=1}^n a_{pi}^4\right] \;\leq\; \frac{(1 + O(\varepsilon^2))n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^4 \, \mathrm{d}V \tag{5.10}$$

We first calculate $\frac{n}{V_m r_0^m} \int_D x_i^4 \, \mathrm{d}x_m \cdots \mathrm{d}x_1$, where D is the m-ball in $\mathcal{T}$ centered at the origin with radius r_1. Consider an $(m - i + 1)$-ball with radius r_i centered at the origin. Its volume, which is $V_{m-i+1} r_i^{m-i+1}$, can also be written as the integration of the volume of its $(m - i)$-dimensional cross-section—an $(m - i)$-ball—that is perpendicular to the x_i axis and at distance x_i from the origin as the value x_i varies from $-r_i$ to r_i. Figure 6 gives an illustration. The radius of the $(m-i)$-dimensional cross-section is $r_{i+1} = \sqrt{r_i^2 - x_i^2}$. Therefore, $V_{m-i+1} r_i^{m-i+1} = \int_{-r_i}^{r_i} V_{m-i} r_{i+1}^{m-i} \, \mathrm{d}x_i$, where $V_{m-1} r_{i+1}^{m-i}$ denotes the volume function of an $(m-i)$-ball with radius r_{i+1}. Inductively, we obtain

$$V_{m-i+1} r_i^{m-i+1} = \int_{-r_i}^{r_i} \int_{-r_{i+1}}^{r_{i+1}} \cdots \int_{-r_m}^{r_m} \mathrm{d}x_m \cdots \mathrm{d}x_i.$$

We now return to calculating $\frac{n}{V_m r_0^m} \int_D x_i^4 \, \mathrm{d}x_m \cdots \mathrm{d}x_1$. By symmetry, we can assume that $i = 1$.

$$\begin{aligned} \frac{n}{V_m r_0^m} \int_D x_i^4 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 &= \frac{n}{V_m r_0^m} \int_{-r_1}^{r_1} \cdots \int_{-r_m}^{r_m} x_1^4 \, \mathrm{d}x_m \ldots \mathrm{d}x_1 \\ &= \frac{nV_{m-1}}{V_m r_0^m} \int_{-r_1}^{r_1} x_1^4 r_2^{m-1} \, \mathrm{d}x_1 \\ &= \frac{nV_{m-1}}{V_m r_0^m} \cdot \frac{3\beta(m) r_1^{m+4}}{(m+2)(m+4)} && (\because \text{Lemma 5.3}) \\ &= \frac{3n r_1^{m+4}}{(m+2)(m+4) r_0^m} && (\because (5.4)) \end{aligned}$$

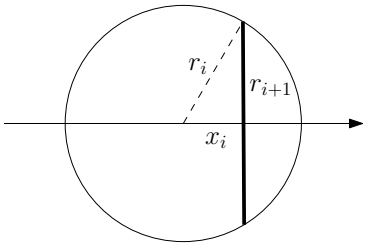


Figure 6: The circle represents an $(m-i+1)$-ball centered at the origin with radius r_i. The bold segment represents an $(m-i)$-dimensional cross-section of the $(m-i+1)$-ball, and its radius r_{i+1} is equal to $\sqrt{r_i^2 - x_i^2}$.

The orthogonal projection of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$ is contained in the m-ball centered at the origin with radius r_0. By (5.8), if we set $r_1 = r_0$, then

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^4 \, \mathrm{d}V &\leq (1 + O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_D x_i^4 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 \\
&= (1 + O(\varepsilon^2)) \cdot \frac{3nr_0^4}{(m+2)(m+4)} \\
&= \frac{3n\varrho^4\varepsilon^4}{(m+2)(m+4)} + O(n\varrho^4\varepsilon^6).
\end{aligned}$$

The orthogonal projection of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$ covers an m-ball centered at the origin with radius $(1 - O(\varepsilon^2))r_0$. Thus, if we set r_1 to be this radius, then

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^4 \, \mathrm{d}V &\geq \frac{n}{V_m r_0^m} \int_D x_i^4 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 \\
&= (1 - O(\varepsilon^2))^{m+4} \cdot \frac{3nr_0^4}{(m+2)(m+4)} \\
&\geq (1 - O(\varepsilon^2)) \cdot \frac{3nr_0^4}{(m+2)(m+4)} \\
&= \frac{3n\varrho^4\varepsilon^4}{(m+2)(m+4)} - O(n\varrho^4\varepsilon^6).
\end{aligned}$$

Then, it follows from (5.10) that $\left| \mathtt{E}\left[\sum_{p=1}^n a_{pi}^4\right] - \frac{3}{(m+2)(m+4)} n\varrho^4\varepsilon^4 \right| = O(n\varrho^4\varepsilon^6)$, establishing the correctness of (i).

Analysis of (ii): The variance of $\sum_{p=1}^n a_{pi}^2 a_{pj}^2$ equals $n\mathtt{E}[z_i^4 z_j^4] - n(\mathtt{E}[z_i^2 z_j^2])^2 \leq n\mathtt{E}[z_i^4 z_j^4] \leq n\varrho^8\varepsilon^8$ because $|z_i| \leq \varrho\varepsilon$. Lemma 5.2 implies that $\left| \sum_{p=1}^n a_{pi}^2 a_{pj}^2 - \mathrm{E}\left[\sum_{p=1}^n a_{pi}^2 a_{pj}^2\right] \right| < cn^{2/3}\varrho^4\varepsilon^4$ with probability $1 - c^{-2}n^{-1/3}$. It remains to bound $\mathrm{E}\left[\sum_{p=1}^n a_{pi}^2 a_{pj}^2\right]$.

Since $\mathtt{E}\left[\sum_{p=1}^n a_{pi}^2 a_{pj}^2\right] = \frac{n}{\mathrm{vol}(\mathcal{M}\cap B_\varepsilon)} \int_{\mathcal{M}\cap B_\varepsilon} z_i^2 z_j^2 \, \mathrm{d}V$, we can derive as in (i) the relation $\frac{(1-O(\varepsilon^2))n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^2 z_j^2 \, \mathrm{d}V \leq \mathtt{E}\left[\sum_{p=1}^n a_{pi}^2 a_{pj}^2\right] \leq \frac{(1+O(\varepsilon^2))n}{V_m r_0^m} \int_{\mathcal{M}\cap B_\varepsilon} z_i^2 z_j^2 \, \mathrm{d}V$. Let D be the m-ball in $\mathcal{T}$ centered at the origin with radius r_1. By symmetry, we can assume that $i = 1$

and $j = 2$.

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_D x_i^2 x_j^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 &= \frac{nV_{m-2}}{V_m r_0^m} \int_{-r_1}^{r_1} \int_{-r_2}^{r_2} x_1^2 x_2^2 r_3^{m-2} \, \mathrm{d}x_2 \mathrm{d}x_1 \\
&= \frac{n\beta(m-1)\beta(m+2)V_{m-2} r_1^{m+4}}{(m+1)(m+4)V_m r_0^m} \quad (\because \text{Lemma 5.3}) \\
&= \frac{n r_1^{m+4}}{(m+2)(m+4) r_0^m} \quad (\because (5.3) \,\&\, (5.5))
\end{aligned}$$

As in (i), if we set $r_1 = r_0$, then

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 z_j^2 \, \mathrm{d}V &\le (1 + O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_D x_i^2 x_j^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 \\
&= (1 + O(\varepsilon^2)) \cdot \frac{n r_0^4}{(m+2)(m+4)} \\
&= \frac{n\varrho^4 \varepsilon^4}{(m+2)(m+4)} + O(n\varrho^4 \varepsilon^6).
\end{aligned}$$

The orthogonal projection of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$ covers an m-ball centered at the origin with radius $(1 - O(\varepsilon^2))r_0$. Thus, if we set r_1 to be this radius, then

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 z_j^2 \, \mathrm{d}V &\ge \frac{n}{V_m r_0^m} \int_D x_i^2 x_j^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 \\
&= (1 - O(\varepsilon^2))^{m+4} \cdot \frac{n r_0^4}{(m+2)(m+4)} \\
&\ge (1 - O(\varepsilon^2)) \cdot \frac{n r_0^4}{(m+2)(m+4)} \\
&= \frac{n\varrho^4 \varepsilon^4}{(m+2)(m+4)} - O(n\varrho^4 \varepsilon^6).
\end{aligned}$$

Therefore, $\left| \mathrm{E}\Big[\sum_{p=1}^n a_{pi}^2 a_{pj}^2\Big] - \frac{n\varrho^4\varepsilon^4}{(m+2)(m+4)} \right| = O(n\varrho^4\varepsilon^6)$, establishing the correctness of (ii).

Analysis of (iii): The variance of $\sum_{p=1}^n a_{pi}^2$ is at most $n\mathrm{E}\big[z_i^4\big] \le n\varrho^4\varepsilon^4$. Lemma 5.2 implies that $\left| \sum_{p=1}^n a_{pi}^2 - \mathrm{E}\big[\sum_{p=1}^n a_{pi}^2\big] \right| < cn^{2/3}\varrho^2\varepsilon^2$ with probability $1 - c^{-2}n^{-1/3}$. It remains to bound $\mathrm{E}\big[\sum_{p=1}^n a_{pi}^2\big]$. Since $\mathrm{E}\Big[\sum_{p=1}^n a_{pi}^2\Big] = \frac{n}{\mathrm{vol}(\mathcal{M} \cap B_\varepsilon)} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 \, \mathrm{d}V$, by (5.9), $\mathrm{E}\Big[\sum_{p=1}^n a_{pi}^2\Big]$ lies between $(1 - O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 \, \mathrm{d}V$ and $(1 + O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 \, \mathrm{d}V$. Let D be the m-ball in $\mathcal{T}$ centered at the origin with radius r_1.

$$\begin{aligned}
\frac{n}{V_m r_0^m} \int_D x_i^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 &= \frac{n}{V_m r_0^m} \int_{-r_1}^{r_1} \cdots \int_{-r_m}^{r_m} x_1^2 \, \mathrm{d}x_m \ldots \mathrm{d}x_1 \\
&= \frac{nV_{m-1}}{V_m r_0^m} \int_{-r_1}^{r_1} x_1^2 r_2^{m-1} \, \mathrm{d}x_1 \\
&= \frac{n r_1^{m+2}}{(m+2) r_0^m} \quad (\because \text{Lemma 5.3 and (5.4)})
\end{aligned}$$

If we set $r_1 = r_0$, then $\mathrm{E}\Big[\sum_{p=1}^n a_{pi}^2\Big] \le (1 + O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 \, \mathrm{d}V \le (1 + O(\varepsilon^2))^2 \cdot \frac{n}{V_m r_0^m} \int_D x_i^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 = \frac{1 + O(\varepsilon^2)}{m+2} \cdot n r_0^2 = \frac{1}{m+2} \cdot n\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)$. The orthogonal projection

of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$ covers an m-ball centered at the origin with radius $(1 - O(\varepsilon^2))r_0$. Thus, if we set r_1 to be this radius, then $\mathrm{E}\Big[\sum_{p=1}^n a_{pi}^2\Big] \geq (1 - O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_{\mathcal{M} \cap B_\varepsilon} z_i^2 \, \mathrm{d}V \geq (1 - O(\varepsilon^2)) \cdot \frac{n}{V_m r_0^m} \int_D x_i^2 \, \mathrm{d}x_m \cdots \mathrm{d}x_1 = \frac{(1-O(\varepsilon^2))^{m+3}}{m+2} \cdot n r_0^2 = \frac{1}{m+2} \cdot n\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4)$. Therefore, $\Big|\mathrm{E}\big[\sum_{p=1}^n a_{pi}^2\big] - \frac{1}{m+2} n\varrho^2\varepsilon^2\Big| = O(n\varrho^2\varepsilon^4)$. This establishes the correctness of (iii).

Analysis of (iv), (v) *and* (vi): We prove a more general statement. Consider $\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}$, where $k \geq 1$, $i_j \in [1, m]$ for every $j \in [0, k]$, e_0 is an even (possibly zero) integer, and e_j is an odd integer for every $j \in [1, k]$. We show that $\Big|\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}\Big| < cn^{2/3}\varrho^e\varepsilon^e + O(n\varrho^e\varepsilon^{e+2})$ with probability $1 - c^{-2}n^{-1/3}$, where $e = \sum_{j=0}^k e_j$.

The variance of $\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}$ is at most $n\mathrm{E}\Big[\prod_{j=0}^k z_{i_j}^{2e_j}\Big] \leq n\varrho^{2e}\varepsilon^{2e}$ because $z_i \leq \varrho\varepsilon$. By Lemma 5.2, it holds with probability $1 - c^{-2}n^{-1/3}$ that $\Big|\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}\Big| < \Big|\mathrm{E}\Big[\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}\Big]\Big| + cn^{2/3}\varrho^e\varepsilon^e$. Thus, it suffices to bound $\Big|\mathrm{E}\Big[\sum_{p=1}^n \prod_{j=0}^k a_{pi_j}^{e_j}\Big]\Big|$.

Let D be the m-ball in $\mathcal{T}$ centered at the origin with radius $r_1 = (1 - O(\varepsilon^2))r_0$ covered by the orthogonal projection of $\mathcal{M} \cap B_\varepsilon$ onto $\mathcal{T}$. Let M be the portion of $\mathcal{M} \cap B_\varepsilon$ that projects onto D. We deal with M and $(\mathcal{M} \cap B_\varepsilon) \setminus M$ separately.

Divide $\mathbb{R}^d$ into 2^k subsets so that for every subset and every $j \in [1, k]$, the sign of $z_{i_j}^{e_j}$ does not flip within the subset. (Note that $z_{i_0}^{e_0} \geq 0$ as e_0 is even.) Consider two such subsets $H_k^+ = \{\mathsf{x} \in \mathbb{R}^d : x_{i_k}^{e_k} \geq 0 \ \wedge \ \forall\, j \in [1, k-1], x_{i_j}^{e_j} \geq 0\}$ and $H_k^- = \{\mathsf{x} \in \mathbb{R}^d : x_{i_k}^{e_k} \leq 0 \ \wedge \ \forall\, j \in [1, k-1], x_{i_j}^{e_j} \geq 0\}$. By (5.8),

$$\int_{M \cap H_k^+} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V \leq (1 + O(\varepsilon^2)) \cdot \int_{D \cap H_k^+} \prod_{j=0}^k x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1.$$

Since $z_{i_k}^{e_k} \leq 0$ in H_k^- and $\int_{M \cap H_k^-} |z_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} z_{i_j}^{e_j} \, \mathrm{d}V \geq \int_{D \cap H_k^-} |x_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1$, we get $\int_{M \cap H_k^-} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V = -\int_{M \cap H_k^-} |z_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} z_{i_j}^{e_j} \, \mathrm{d}V \leq -\int_{D \cap H_k^-} |x_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1$. By symmetry, $\int_{D \cap H_k^-} |x_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1 = \int_{D \cap H_k^+} \prod_{j=0}^k x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1$. Therefore,

$$\int_{M \cap H_k^-} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V \leq -\int_{D \cap H_k^+} \prod_{j=0}^k x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1.$$

Let $H = \{\mathsf{x} \in \mathbb{R}^d : \forall\, j \in [1, k-1], x_{i_j}^{e_j} \geq 0\}$. It follows that

$$\begin{aligned}
\int_{M \cap H} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V &= \int_{M \cap H_k^+} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V + \int_{M \cap H_k^-} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V \\
&\leq O(\varepsilon^2) \cdot \int_{D \cap H_k^+} \prod_{j=0}^k x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1 \\
&\leq O(\varepsilon^2) \cdot V_m r_1^{m+e}/2^k \qquad (\because x_{i_j} \leq r_1) \\
&\leq O(\varepsilon^2) \cdot V_m r_0^{m+e}/2^k \qquad (\because r_1 \leq r_0)
\end{aligned}$$

Conversely, $\int_{M \cap H_k^+} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V \geq \int_{D \cap H_k^+} \prod_{j=0}^k x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1$ and $\int_{M \cap H_k^-} \prod_{j=0}^k z_{i_j}^{e_j} \, \mathrm{d}V = -\int_{M \cap H_k^-} |z_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} z_{i_j}^{e_j} \, \mathrm{d}V \geq -(1 + O(\varepsilon)^2) \cdot \int_{D \cap H_k^-} |x_{i_k}^{e_k}| \cdot \prod_{j=0}^{k-1} x_{i_j}^{e_j} \, \mathrm{d}x_m \cdots \mathrm{d}x_1$, which is equal

to $-(1+O(\varepsilon)^2)\cdot\int_{D\cap H_k^+}\prod_{j=0}^k x_{i_j}^{e_j}\,\mathrm{d}x_m\cdots\mathrm{d}x_1$, Therefore,

$$\int_{M\cap H}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V \;\;\geq\;\; -O(\varepsilon^2)\cdot\int_{D\cap H_k^+}\prod_{j=0}^k x_{i_j}^{e_j}\,\mathrm{d}x_m\cdots\mathrm{d}x_1 \;\;\geq\;\; -O(\varepsilon^2)\cdot V_m r_0^{m+e}/2^k.$$

As a result, $\left|\int_{M\cap H}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \leq O(\varepsilon^2)\cdot V_m r_0^{m+e}/2^k$.

There are $2^{k-1}-1$ other combinations of signs of $z_{i_j}^{e_j}$ for $j\in[1,k-1]$. Each combination gives rise to a subset G of $\mathbb{R}^d$ and one can derive as in the above that $\left|\int_{M\cap G}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \leq O(\varepsilon^2)\cdot V_m r_0^{m+e}/2^k$. Consequently, $\left|\int_M\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \leq 2^{k-1}\cdot O(\varepsilon^2)\cdot V_m r_0^{m+e}/2^k = O(\varepsilon^2)\cdot V_m r_0^{m+e}$.

Let D_0 be the m-ball in $\mathcal{T}$ centered at the origin with radius r_0. Since $\left|\prod_{j=0}^k z_{i_j}^{e_j}\right| \leq r_0^e$, by (5.8), $\left|\int_{(\mathcal{M}\cap B_\varepsilon)\setminus M}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \leq (1+O(\varepsilon^2))\cdot\int_{D_0\setminus D} r_0^e\,\mathrm{d}x_m\cdots\mathrm{d}x_1 = (1+O(\varepsilon^2))\cdot V_m r_0^e(r_0^m - r_1^m) = O(\varepsilon^2)\cdot V_m r_0^{m+e}$.

We conclude that $\left|\int_{\mathcal{M}\cap B_\varepsilon}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \leq \left|\int_M\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| + \left|\int_{(\mathcal{M}\cap B_\varepsilon)\setminus M}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| = O(\varepsilon^2)\cdot V_m r_0^{m+e}$. Then,

$$\begin{aligned}\left|\mathrm{E}\left[\sum_{p=1}^n\prod_{j=0}^k a_{pi_j}^{e_j}\right]\right| &\leq \frac{n}{\mathrm{vol}(\mathcal{M}\cap B_\varepsilon)}\left|\int_{\mathcal{M}\cap B_\varepsilon}\prod_{j=0}^k z_{i_j}^{e_j}\,\mathrm{d}V\right| \\ &\leq \frac{n}{\mathrm{vol}(D)}\cdot O(\varepsilon^2)\cdot V_m r_0^{m+e} \\ &= O(n\varrho^e\varepsilon^{e+2}).\end{aligned}$$

This establishes the correctness of (iv), (v) and (vi). □

6 Eigenvalues of $\mathsf{B}^t\mathsf{B}$

We show in this section that the largest m_0 eigenvalues of $\mathsf{B}^t\mathsf{B}$ are $\Theta(n\varrho^4\varepsilon^4)$ and the largest (m_0+1)-th eigenvalue of $\mathsf{B}^t\mathsf{B}$ is $O(n\varrho^4\varepsilon^6)$. The bounds on the eigenvalues of $\mathsf{B}^t\mathsf{B}$ are obtained by proving a series of lemmas using the Gershgorin Circle Theorem [17] and its generalization [13].

6.1 Preliminaries

Let C be a square matrix. The Gershgorin Circle Theorem states that for each eigenvalue σ of C, there exists a row i of C such that $|\sigma - c_{ii}| \leq \sum_{j\neq i}|c_{ij}|$. It follows that $|c_{ii}| - \sum_{j\neq i}|c_{ij}| \leq \sigma \leq \sum_j |c_{ij}|$. By applying the Gershgorin Circle Theorem to C^t, there also exists a column j of C such that $|\sigma - c_{jj}| \leq \sum_{i\neq j}|c_{ij}|$, which implies that $|c_{jj}| - \sum_{i\neq j}|c_{ij}| \leq \sigma \leq \sum_i |c_{ij}|$. This result has been generalized to the case when C is partitioned into blocks [13]. Consider the following partition of C:

$$\begin{pmatrix}\mathsf{C}_{11} & \cdots & \mathsf{C}_{1r}\\ \vdots & \ddots & \vdots\\ \mathsf{C}_{r1} & \cdots & \mathsf{C}_{rr}\end{pmatrix} \tag{6.1}$$

That is, there exist integers n_i such that C_{ij} is an $n_i\times n_j$ matrix. Note that the matrices C_{ii}'s are square, but the other C_{ij}'s may not be square. Each row of blocks $\begin{pmatrix}\mathsf{C}_{i1} & \cdots & \mathsf{C}_{ii} & \cdots & \mathsf{C}_{ir}\end{pmatrix}$ defines a *generalized gershgorin set* G_i which contains all real numbers μ such that $\|(\mathsf{C}_{ii} -$

$\mu \mathsf{I}_{n_i})^{-1}\|^{-1} \leq \sum_{j \neq i} \|\mathsf{C}_{ij}\|$. The eigenvalues of C_{ii} are in G_i by a continuity argument [13]. The definition of G_i implies that:

$$\min\{\mu \in G_i\} \;\geq\; \text{smallest eigenvalue of } \mathsf{C}_{ii} - \sum_{i \neq j} \|\mathsf{C}_{ij}\| \tag{6.2}$$

$$\max\{\mu \in G_i\} \;\leq\; \text{largest eigenvalue of } \mathsf{C}_{ii} + \sum_{i \neq j} \|\mathsf{C}_{ij}\| \tag{6.3}$$

Equations (6.2) and (6.3) help to bound the eigenvalues of C because G_i contains some eigenvalues of C under certain conditions as stated in the following result.

Lemma 6.1 ([13]) *Consider any partition of a square matrix C into blocks as in* (6.1)*. Every eigenvalue of C lies in some generalized gershgorin set G_i with respect to this partition. Moreover, if a generalized gershgorin set G_i is disjoint from the union of the other generalized gershgorin sets, then G_i contains exactly n_i eigenvalues of C, where n_i is the dimension of C_{ii}.*

In addition to the Gershgorin Circle Theorem, there are also some easy bounds on the 2-norm and eigenvalues of a matrix [17]. For any $r \times k$ matrix U,

$$\|\mathsf{U}\| \;=\; \|\mathsf{U}^t\| \quad \text{and} \quad \|\mathsf{U}^t\mathsf{U}\| \;=\; \|\mathsf{U}\|^2.$$

Moreover,

$$\max_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{U}^t\mathsf{U}\mathsf{e} = \sigma_{\max}^2 \quad \text{and} \quad \min_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{U}^t\mathsf{U}\mathsf{e} = \sigma_{\min}^2,$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values of U. Since $\|\cdot\|$ satisfies triangle inequality, if $\mathsf{U} = \mathsf{V} + \mathsf{W}$, then

$$\|\mathsf{U}\| \;\leq\; \|\mathsf{V}\| + \|\mathsf{W}\|.$$

If $\mathsf{U} = (\mathsf{V}\ \mathsf{W})$, where the row dimension of U is r and the column dimensions of V and W are i and j, respectively, then since we can write $\mathsf{U} = (\mathsf{V}\ \mathsf{0}_{r,j}) + (\mathsf{0}_{r,i}\ \mathsf{W})$, we get

$$\|\mathsf{U}\| \;\leq\; \|\mathsf{V}\| + \|\mathsf{W}\|.$$

If $\mathsf{U} = \mathsf{V}\,\mathsf{W}$, then

$$\|\mathsf{U}\| \;\leq\; \|\mathsf{V}\| \cdot \|\mathsf{W}\|.$$

Suppose that the row dimension of U is r. Then, $\|\mathsf{U}\| \leq \|\mathsf{U}\|_F \leq \sqrt{r}\,\|\mathsf{U}\|$. Note that $\|\mathsf{U}\|_F = (\sum_{ij} u_{ij}^2)^{1/2} = \left(\sum_{i=1}^r \|\mathsf{u}_{i*}\|^2\right)^{1/2}$. Therefore,

$$\|\mathsf{U}\| \;\leq\; \sqrt{r} \max_{i \in [1,r]} \|\mathsf{u}_{i*}\|.$$

Suppose that $\mathsf{U} = \mathsf{V} + \mathsf{W}$ and all three matrices U, V and W are symmetric (of dimension k) and positive semi-definite. In this case, the minimum eigenvalue of U is $\min_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{U}\mathsf{e} = \min_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{V}\mathsf{e} + \mathsf{e}^t\mathsf{W}\mathsf{e}$, which is greater than or equal to both $\min_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{V}\mathsf{e}$ and $\min_{\mathsf{e} \in \mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{W}\mathsf{e}$. Therefore, if $\mathsf{U} = \mathsf{V} + \mathsf{W}$ for some symmetric and positive semi-definite matrices U, V and W, then

$$\text{min. eigenvalue of } \mathsf{U} \geq \max\{\text{min. eigenvalue of } \mathsf{V},\ \text{min. eigenvalue of } \mathsf{W}\}.$$

To facilitate the analysis of the eigenvalues of $\mathsf{B}^t\mathsf{B}$, it is convenient to assume that the coordinate axes $x_1, \ldots, x_m$ span the tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin. That is, for every $p \in [1, n]$ and every $i \in [1, m]$, a_{pi} is the coordinate of a_p on the x_i-axis. The following result shows that the eigenvalues of $\mathsf{B}^t\mathsf{B}$ are preserved by rotations in $\mathbb{R}^d$ that keep the origin fixed.

Lemma 6.2 *The eigenvalues of* $\mathsf{B}^t\mathsf{B}$ *are preserved by rotations in* $\mathbb{R}^d$ *that keep the origin fixed.*

Proof. Recall that $d_0 = \binom{d+1}{2}$. Each row of B is a vector in $\mathbb{R}^{d_0}$ and it is the image of the function $h : \begin{pmatrix} y_1 & \cdots & y_d \end{pmatrix}^t \mapsto \begin{pmatrix} \frac{1}{\sqrt{2}}y_1^2 & y_1y_2 & \cdots & y_1y_d & \frac{1}{\sqrt{2}}y_2^2 & y_2y_3 & \cdots & \frac{1}{\sqrt{2}}y_d^2 \end{pmatrix}^t$. The linear space spanned by $h(\mathbb{R}^d)$ is $\mathbb{R}^{d_0}$. To see this, consider the vectors e_{ij}, where $i \in [1,d]$ and $j \in [i,d]$, such that the i-th and j-th entries of e_{ij} are ones and all other entries of e_{ij} are zeros. There are d_0 such e_{ij}'s and the vectors $h(\mathsf{e}_{ij})$'s are linearly independent because for any $k < l$, $h(\mathsf{e}_{kl})$ contains a 1 in the position of y_ky_l and no other $h(\mathsf{e}_{ij})$'s do.

Take any rotation R in $\mathbb{R}^d$ that keeps the origin fixed. Define the transformation $\varphi : h(\mathsf{v}) \mapsto h \circ R(\mathsf{v})$. For every pair of vectors $\mathsf{u}, \mathsf{v} \in \mathbb{R}^d$, $h(\mathsf{u})^t \cdot h(\mathsf{v}) = \frac{1}{2}\sum_{i=1}^d u_i^2v_i^2 + \sum_{i=1}^d \sum_{j=i+1}^d u_iu_jv_iv_j = \frac{1}{2}(\sum_{i=1}^d u_iv_i)^2 = \frac{1}{2}(\mathsf{u}^t \cdot \mathsf{v})^2$. Since R preserves distances and angles in $\mathbb{R}^d$, $\mathsf{u}^t \cdot \mathsf{v} = R(\mathsf{u})^t \cdot R(\mathsf{v})$, which implies that $(\varphi \circ h(\mathsf{u}))^t \cdot \varphi \circ h(\mathsf{v}) = h(\mathsf{u})^t \cdot h(\mathsf{v})$. That is, φ preserves distances and angles in $h(\mathbb{R}^d)$.

Since the $h(\mathsf{e}_{ij})$'s form a basis of $\mathbb{R}^{d_0}$, we can define a linear transformation ψ in $\mathbb{R}^{d_0}$ such that $\psi \circ h(\mathsf{e}_{ij}) = \varphi \circ h(\mathsf{e}_{ij})$ for every $i \in [1,d]$ and every $j \in [i,d]$. For any $i,k \in [1,d]$, any $j \in [i,d]$ and any $l \in [k,d]$, $(\psi \circ h(\mathsf{e}_{ij}))^t \cdot \psi \circ h(\mathsf{e}_{kl}) = (\varphi \circ h(\mathsf{e}_{ij}))^t \cdot \varphi \circ h(\mathsf{e}_{kl}) = h(\mathsf{e}_{ij})^t \cdot h(\mathsf{e}_{kl})$, which implies that ψ preserves distances and angles in $\mathbb{R}^{d_0}$ and hence ψ is an isometry in $\mathbb{R}^{d_0}$. Since ψ and φ agree on the $h(\mathsf{e}_{ij})$'s by definition and both ψ and φ preserve distances in $h(\mathbb{R}^d)$, $\psi(\mathsf{z})$ must be equal to $\varphi(\mathsf{z})$ for every vector $\mathsf{z} \in h(\mathbb{R}^d)$.

We conclude that the effect on $h(\mathbb{R}^d)$ caused by the rotation R in $\mathbb{R}^d$ is produced by the isometry ψ in $\mathbb{R}^{d_0}$. Since the eigenvalues of $\mathsf{B}^t\mathsf{B}$ are invariant under isometries in $\mathbb{R}^{d_0}$, they are not changed by the rotation R. □

6.2 Analysis

Suppose that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. There are two kinds of columns in B, namely, the "double" columns $\begin{pmatrix} \frac{1}{\sqrt{2}}a_{1i}^2 & \cdots & \frac{1}{\sqrt{2}}a_{ni}^2 \end{pmatrix}^t$ for $i \in [1,d]$ and the "cross" columns $\begin{pmatrix} a_{1i}a_{1j} & \cdots & a_{ni}a_{nj} \end{pmatrix}^t$ for $i \in [1,d]$ and $j \in [i+1,d]$. Rearranging the columns in B does not change the eigenvalues of $\mathsf{B}^t\mathsf{B}$. For convenience, we rearrange the columns of B so that $\mathsf{B} = (\mathsf{B}_{TT} \;\; \mathsf{B}_{TN} \;\; \mathsf{B}_{NN})$, where B_{TT} consists of the "double" columns for $i \in [1,m]$ and the "cross" columns for $i \in [1,m]$ and $j \in [i+1,m]$, B_{TN} consists of the "cross" columns for $i \in [1,m]$ and $j \in [m+1,d]$, and B_{NN} consists of the "double" columns for $i \in [m+1,d]$ and the "cross" columns for $i \in [m+1,d]$ and $j \in [i+1,d]$.

Recall that $d_0 = \binom{d+1}{2}$ and $m_0 = \binom{m+1}{2}$. B_{TT} has m_0 columns, B_{TN} has $m(d-m)$ columns, and B_{NN} contains $d_0 - m_0 - md + m^2$ columns. The matrix $\mathsf{B}^t\mathsf{B}$ can be divided into blocks as follows.

$$\mathsf{B}^t\mathsf{B} \;=\; \begin{pmatrix} \mathsf{B}_{TT}^t\mathsf{B}_{TT} & \mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN} \\ \mathsf{B}_{TN}^t\mathsf{B}_{TT} & \mathsf{B}_{TN}^t\mathsf{B}_{TN} & \mathsf{B}_{TN}^t\mathsf{B}_{NN} \\ \mathsf{B}_{NN}^t\mathsf{B}_{TT} & \mathsf{B}_{NN}^t\mathsf{B}_{TN} & \mathsf{B}_{NN}^t\mathsf{B}_{NN} \end{pmatrix} \tag{6.4}$$

We first analyze the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ and the singular values of B_{TN} and B_{NN}.

Lemma 6.3 *Assume that the coordinate axes* $x_1, \ldots, x_m$ *span* $\mathcal{T}$*. If* ε *is sufficiently small, then with probability* $1 - O(n^{-1/3})$*, the eigenvalues of* $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ *are* $\Theta(n\varrho^4\varepsilon^4)$ *and so* $\|\mathsf{B}_{TT}\| = \Theta(\sqrt{n}\varrho^2\varepsilon^2)$*.*

Proof. For every $p \in [1,n]$ and every $i \in [1,m]$, $|a_{pi}| \leq \varrho\varepsilon$. Thus, for every $p \in [1,n]$ and every quadruple of possibly non-distinct $i,j,k,l \in [1,m]$, $\sum_{p=1}^n |a_{pi}a_{pj}a_{pk}a_{pl}| \leq n\varrho^4\varepsilon^4$. It follows that

the maximum absolute row sum of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is at most $m_0 n\varrho^4\varepsilon^4 = O(n\varrho^4\varepsilon^4)$, which is an upper bound on the largest eigenvalue by the Gershgorin Circle Theorem.

To prove the lower bound, rearrange the columns of B_{TT} such that its leftmost $n \times m$ submatrix is $\begin{pmatrix} \frac{1}{\sqrt{2}}a_{11}^2 & \cdots & \frac{1}{\sqrt{2}}a_{1m}^2 \\ \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{2}}a_{n1}^2 & \cdots & \frac{1}{\sqrt{2}}a_{nm}^2 \end{pmatrix}$. This rearrangement does not change the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$. Let V be the trailing $(m_0 - m) \times (m_0 - m)$ submatrix of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$. Then,

$$\mathsf{B}_{TT}^t\mathsf{B}_{TT} = \left(\begin{array}{cccc|c} \frac{1}{2}\sum_{p=1}^n a_{p1}^4 & \frac{1}{2}\sum_{p=1}^n a_{p1}^2a_{p2}^2 & \cdots & \frac{1}{2}\sum_{p=1}^n a_{p1}^2a_{pm}^2 & * \\ \frac{1}{2}\sum_{p=1}^n a_{p1}^2a_{p2}^2 & \frac{1}{2}\sum_{p=1}^n a_{p2}^4 & \cdots & \frac{1}{2}\sum_{p=1}^n a_{p2}^2a_{pm}^2 & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2}\sum_{p=1}^n a_{p1}^2a_{pm}^2 & \frac{1}{2}\sum_{p=1}^n a_{p2}^2a_{pm}^2 & \cdots & \frac{1}{2}\sum_{p=1}^n a_{pm}^4 & * \\ \hline * & * & \cdots & * & \mathsf{V} \end{array}\right) \tag{6.5}$$

Define an $m \times m$ matrix W whose entries are identical and equal to $\frac{1}{2(m+2)(m+4)}n\varrho^4\varepsilon^4$. Define another $m \times m$ matrix U such that u_{ij} is equal to the (i,j) entry of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ minus $\frac{1}{2(m+2)(m+4)}n\varrho^4\varepsilon^4$. We split $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ into the following sum of matrices.

$$\mathsf{B}_{TT}^t\mathsf{B}_{TT} = \begin{pmatrix} \mathsf{U} & * \\ * & \mathsf{V} \end{pmatrix} + \begin{pmatrix} \mathsf{W} & \mathsf{0}_{m,m_0-m} \\ \mathsf{0}_{m_0-m,m} & \mathsf{0}_{m_0-m,m_0-m} \end{pmatrix} \tag{6.6}$$

All matrices in (6.6) are symmetric, and $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is clearly positive semi-definite. We show that the two matrices on the right hand side of (6.6) are also positive semi-definite by bounding their eigenvalues from below. Then, we can conclude that the minimum eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is at least the maximum of the minimum eigenvalues of the two matrices on the right hand side of (6.6). Since the entries of W are identical, the matrix $\begin{pmatrix} \mathsf{W} & \mathsf{0}_{m,m_0-m} \\ \mathsf{0}_{m_0-m,m} & \mathsf{0}_{m_0-m,m_0-m} \end{pmatrix}$ has rank one. One can verify that $\begin{pmatrix} 1/\sqrt{m} & \cdots & 1/\sqrt{m} & \mathsf{0}_{m_0-m} \end{pmatrix}^t$ is a unit eigenvector and the only non-zero eigenvalue is $\frac{m}{2(m+2)(m+4)}n\varrho^4\varepsilon^4$. So the minimum eigenvalue of $\begin{pmatrix} \mathsf{W} & \mathsf{0}_{m,m_0-m} \\ \mathsf{0}_{m_0-m,m} & \mathsf{0}_{m_0-m,m_0-m} \end{pmatrix}$ is zero. It remains to bound the minimum eigenvalue of $\begin{pmatrix} \mathsf{U} & * \\ * & \mathsf{V} \end{pmatrix}$ from below.

Apply Lemma 5.4(i) with the constant $c = \frac{2}{3(m+2)(m+4)}$. Then, it follows from (6.5) that, with probability $1 - O(n^{-1/3})$, for every $i \in [1,m]$, the (i,i) entry of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is at least $\frac{3}{2(m+2)(m+4)}n\varrho^4\varepsilon^4 - \frac{1}{3(m+2)(m+4)}n^{2/3}\varrho^4\varepsilon^4 - O(n\varrho^4\varepsilon^6)$, which implies that

$$\begin{aligned} |u_{ii}| &\geq \frac{3n\varrho^4\varepsilon^4}{2(m+2)(m+4)} - \frac{n\varrho^4\varepsilon^4}{2(m+2)(m+4)} - \frac{n^{2/3}\varrho^4\varepsilon^4}{3(m+2)(m+4)} - O(n\varrho^4\varepsilon^6) \\ &= \frac{n\varrho^4\varepsilon^4}{(m+2)(m+4)} - \frac{n^{2/3}\varrho^4\varepsilon^4}{3(m+2)(m+4)} - O(n\varrho^4\varepsilon^6). \end{aligned} \tag{6.7}$$

Apply Lemma 5.4(ii) with the constant $c = \frac{2}{3m_0(m+2)(m+4)}$. Then, it follows from (6.5) that, with probability $1 - O(n^{-1/3})$, for every pair of distinct $i, j \in [1,m]$, the (i,j) entry of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is at most $\frac{1}{2(m+2)(m+4)}n\varrho^4\varepsilon^4 + \frac{1}{3m_0(m+2)(m+4)}n^{2/3}\varrho^4\varepsilon^4 + O(n\varrho^4\varepsilon^6)$, which implies that

$$\begin{aligned} |u_{ij}| &\leq \frac{n\varrho^4\varepsilon^4}{2(m+2)(m+4)} - \frac{n\varrho^4\varepsilon^4}{2(m+2)(m+4)} + \frac{n^{2/3}\varrho^4\varepsilon^4}{3m_0(m+2)(m+4)} + O(n\varrho^4\varepsilon^6) \\ &= \frac{n^{2/3}\varrho^4\varepsilon^4}{3m_0(m+2)(m+4)} + O(n\varrho^4\varepsilon^6). \end{aligned} \tag{6.8}$$

By the Gershgorin Circle Theorem, every eigenvalue of $\begin{pmatrix} \mathsf{U} & * \\ * & \mathsf{V} \end{pmatrix}$ is at least the absolute value of the diagonal entry minus the absolute values of the off-diagonal entries in the same row for some row.

Consider a row that contains the entry u_{ii} of U for some $i \in [1, m]$. By (6.8), the absolute value of each off-diagonal entry in U is at most $\frac{1}{3m_0(m+2)(m+4)} n^{2/3} \varrho^4 \varepsilon^4 + O(n\varrho^4\varepsilon^6)$. The other off-diagonal entries are $\frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^3 a_{pj}$ and $\frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^2 a_{pj} a_{pk}$ for some distinct $i, j, k \in [1, m]$. Apply Lemma 5.4(iv) with the constant $c = \frac{\sqrt{2}}{3m_0(m+2)(m+4)}$. It implies that $\left| \frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^3 a_{pj} \right|$ and $\left| \frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^2 a_{pj} a_{pk} \right|$ are at most $\frac{1}{3m_0(m+2)(m+4)} n^{2/3} \varrho^4 \varepsilon^4 + O(n\varrho^4\varepsilon^6)$ with probability $1 - O(n^{-1/3})$. By (6.7), $|u_{ii}|$ minus the absolute values of the off-diagonal entries in the ith row is greater than $\frac{1}{(m+2)(m+4)} n\varrho^4\varepsilon^4 - \frac{2}{3(m+2)(m+4)} n^{2/3}\varrho^4\varepsilon^4 - O(n\varrho^4\varepsilon^6)$, which is $\Omega(n\varrho^4\varepsilon^4)$.

Consider a row that contains a diagonal entry of V. This diagonal entry of V is $\sum_{p=1}^{n} a_{pi}^2 a_{pj}^2$ for some distinct $i, j \in [1, m]$. Apply Lemma 5.4(ii) with the constant $c = \frac{1}{3(m+2)(m+4)}$. It implies that $\sum_{p=1}^{n} a_{pi}^2 a_{pj}^2 \geq \frac{1}{(m+2)(m+4)} n\varrho^4\varepsilon^4 - \frac{1}{3(m+2)(m+4)} n^{2/3}\varrho^4\varepsilon^4 - O(n\varrho^4\varepsilon^6)$ with probability $1 - O(n^{-1/3})$. The off-diagonal entries are $\frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^3 a_{pj}$, $\frac{1}{\sqrt{2}} \sum_{p=1}^{n} a_{pi}^2 a_{pj} a_{pk}$ and $\sum_{p=1}^{n} a_{pi} a_{pj} a_{pk} a_{pl}$ for some distinct $i, j, k, l \in [1, m]$. A similar analysis as in the previous paragraph shows that $\sum_{p=1}^{n} a_{pi}^2 a_{pj}^2$ minus the off-diagonal entries in the same row is $\Omega(n\varrho^4\varepsilon^4)$. □

Lemma 6.4 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. $\|\mathsf{B}_{TN}\| = O(\sqrt{n}\varrho^2\varepsilon^3)$ and $\|\mathsf{B}_{NN}\| = O(\sqrt{n}\varrho^2\varepsilon^4)$.*

Proof. For every $p \in [1, n]$, $\left(\sum_{i=1}^{m} a_{pi}^2\right)^{1/2} \leq \|\mathsf{a}_p\| \leq \varrho\varepsilon$. The 2-norm of a row of B_{TN} is $\left(\sum_{i=1}^{m} a_{pi}^2 \sum_{j=m+1}^{d} a_{pj}^2\right)^{1/2} = \left(\sum_{i=1}^{m} a_{pi}^2\right)^{1/2} \cdot \left(\sum_{j=m+1}^{d} a_{pj}^2\right)^{1/2}$. The distance from a_p to $\mathcal{T}$ is known to be $O(\varrho\varepsilon^2)$ [15, Lemma 6], so $\left(\sum_{j=m+1}^{d} a_{pj}^2\right)^{1/2} = O(\varrho\varepsilon^2)$. It follows that the 2-norm of a row of B_{TN} is $O(\varrho^2\varepsilon^3)$. There are n rows in B_{TN}, implying that $\|\mathsf{B}_{TN}\| = O(\sqrt{n}\varrho^2\varepsilon^3)$. The 2-norm of a row of B_{NN} is no more than $\left(\sum_{i=m+1}^{d} a_{pi}^2 \sum_{j=i}^{d} a_{pj}^2\right)^{1/2} \leq \sum_{j=m+1}^{d} a_{pj}^2 = O(\varrho^2\varepsilon^4)$. Summing over the n rows in B_{NN} shows that $\|\mathsf{B}_{NN}\| = O(\sqrt{n}\varrho^2\varepsilon^4)$. □

We are now ready to give bounds on the eigenvalues of $\mathsf{B}^t\mathsf{B}$. The bound on the $d_0 - m_0$ smallest eigenvalues is not the best possible yet. We boostrap a better bound from it later.

Lemma 6.5 *If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, the m_0 largest eigenvalues of $\mathsf{B}^t\mathsf{B}$ are $\Theta(n\varrho^4\varepsilon^4)$ and the $d_0 - m_0$ smallest eigenvalues of $\mathsf{B}^t\mathsf{B}$ are $O(n\varrho^4\varepsilon^5)$.*

Proof. By Lemma 6.2, we can rotate $\mathbb{R}^d$ so that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$ because the eigenvalues of $\mathsf{B}^t\mathsf{B}$ are not affected. Then, we can partition $\mathsf{B}^t\mathsf{B}$ as shown in (6.4). We define three generalized gershgorin sets [13] as follows.

The first set G_1 is for the row of blocks $\begin{pmatrix} \mathsf{B}_{TT}^t\mathsf{B}_{TT} & \mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN} \end{pmatrix}$. By (6.2) and (6.3), the real numbers in G_1 are at least the minimum eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ minus $\|\mathsf{B}_{TT}^t\mathsf{B}_{TN}\| + \|\mathsf{B}_{TT}^t\mathsf{B}_{NN}\|$ and at most the maximum eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ plus $\|\mathsf{B}_{TT}^t\mathsf{B}_{TN}\| + \|\mathsf{B}_{TT}^t\mathsf{B}_{NN}\|$. By Lemma 6.3, with probability $1 - O(n^{-1/3})$, $\|\mathsf{B}_{TT}^t\mathsf{B}_{TT}\| = \Theta(n\varrho^4\varepsilon^4)$ and $\|\mathsf{B}_{TT}\| = \Theta(\sqrt{n}\varrho^2\varepsilon^2)$. Then, it follows from Lemma 6.4 that, with probability $1 - O(n^{-1/3})$, $\|\mathsf{B}_{TT}^t\mathsf{B}_{TN}\| \leq \|\mathsf{B}_{TT}\| \cdot \|\mathsf{B}_{TN}\| = O(n\varrho^4\varepsilon^5)$ and $\|\mathsf{B}_{TT}^t\mathsf{B}_{NN}\| \leq \|\mathsf{B}_{TT}\| \cdot \|\mathsf{B}_{NN}\| = O(n\varrho^4\varepsilon^6)$. Thus, the numbers in G_1 are $\Theta(n\varrho^4\varepsilon^4)$ with probability $1 - O(n^{-1/3})$.

The second set G_2 is for the row of blocks $\begin{pmatrix} \mathsf{B}_{TN}^t\mathsf{B}_{TT} & \mathsf{B}_{TN}^t\mathsf{B}_{TN} & \mathsf{B}_{TN}^t\mathsf{B}_{NN} \end{pmatrix}$. By (6.3), the real numbers in G_2 are at most the maximum eigenvalue of $\mathsf{B}_{TN}^t\mathsf{B}_{TN}$ plus $\|\mathsf{B}_{TN}^t\mathsf{B}_{TT}\| +$

$\|\mathsf{B}_{TN}^t\mathsf{B}_{NN}\|$. By Lemmas 6.3 and 6.4, with probability $1-O(n^{-1/3})$, $\|\mathsf{B}_{TN}^t\mathsf{B}_{TN}\| = \|\mathsf{B}_{TN}\|^2 = O(n\varrho^4\varepsilon^6)$ and $\|\mathsf{B}_{TN}^t\mathsf{B}_{TT}\| + \|\mathsf{B}_{TN}^t\mathsf{B}_{NN}\| \le \|\mathsf{B}_{TN}\|\cdot\|\mathsf{B}_{TT}\| + \|\mathsf{B}_{TN}\|\cdot\|\mathsf{B}_{NN}\| = O(n\varrho^4\varepsilon^5)$. Thus, the numbers in G_2 are $O(n\varrho^4\varepsilon^5)$ with probability $1-O(n^{-1/3})$.

The third set G_3 is for the row of blocks $(\mathsf{B}_{NN}^t\mathsf{B}_{TT} \;\; \mathsf{B}_{NN}^t\mathsf{B}_{TN} \;\; \mathsf{B}_{NN}^t\mathsf{B}_{NN})$. By (6.3), the real numbers in G_3 are at most the maximum eigenvalue of $\mathsf{B}_{NN}^t\mathsf{B}_{NN}$ plus $\|\mathsf{B}_{NN}^t\mathsf{B}_{TT}\| + \|\mathsf{B}_{NN}^t\mathsf{B}_{TN}\|$. By Lemmas 6.3 and 6.4, with probability $1-O(n^{-1/3})$, $\|\mathsf{B}_{NN}^t\mathsf{B}_{NN}\| = \|\mathsf{B}_{NN}\|^2 = O(n\varrho^4\varepsilon^8)$ and $\|\mathsf{B}_{NN}^t\mathsf{B}_{TT}\| + \|\mathsf{B}_{NN}^t\mathsf{B}_{TN}\| \le \|\mathsf{B}_{NN}\|\cdot\|\mathsf{B}_{TT}\| + \|\mathsf{B}_{NN}\|\cdot\|\mathsf{B}_{TN}\| = O(n\varrho^4\varepsilon^6)$. Thus, the numbers in G_3 are $O(n\varrho^4\varepsilon^6)$ with probability $1-O(n^{-1/3})$.

If ε is sufficiently small, the numbers in G_1 are much bigger than those in G_2 and G_3, implying that $G_1 \cap (G_2 \cup G_3)$ is empty. By Lemma 6.1, the disjointness of G_1 from $G_2 \cup G_3$ implies that G_1 and $G_2 \cup G_3$ contain exactly m_0 and $d_0 - m_0$ eigenvalues of $\mathsf{B}^t\mathsf{B}$, respectively. Hence, G_1 contains the m_0 largest eigenvalues and $G_2 \cup G_3$ contains the $d_0 - m_0$ smallest eigenvalues. □

Lemma 6.5 allows us to show a tighter bound $O(n\varrho^4\varepsilon^6)$ on the (m_0+1)-th largest eigenvalue of $\mathsf{B}^t\mathsf{B}$. It will be important later that this bound is smaller than the bound on the m_0 largest eigenvalues by a factor ε^2.

Lemma 6.6 *If ε is sufficiently small, then with probability $1-O(n^{-1/3})$, the (m_0+1)-th largest eigenvalue of $\mathsf{B}^t\mathsf{B}$ is $O(n\varrho^4\varepsilon^6)$.*

Proof. By Lemma 6.2, we can rotate $\mathbb{R}^d$ so that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$ and then partition $\mathsf{B}^t\mathsf{B}$ as shown in (6.4). Let σ be an eigenvalue of $\mathsf{B}^t\mathsf{B}$ among the $d_0 - m_0$ smallest ones. By Lemma 6.5, σ is $O(n\varrho^4\varepsilon^5)$ with probability $1-O(n^{-1/3})$. Let e be an eigenvector of $\mathsf{B}^t\mathsf{B}$ corresponding to σ. Divide e into two parts $\begin{pmatrix}\mathsf{v}^t & \mathsf{w}^t\end{pmatrix}^t$, where v consists of the first m_0 coordinates of e and w consists of the last $d_0 - m_0$ coordinates of e.

We claim that $\mathsf{w} \neq \mathsf{0}_{d_0-m_0,1}$ with probability $1-O(n^{-1/3})$. If $\mathsf{w} = \mathsf{0}_{d_0-m_0,1}$, then the following relation holds as e is an eigenvector of $\mathsf{B}^t\mathsf{B}$.

$$\mathsf{B}^t\mathsf{B}\cdot\mathsf{e} \;=\; \begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TT}\cdot\mathsf{v}\\ *\end{pmatrix} \;=\; \sigma\mathsf{e} \;=\; \begin{pmatrix}\sigma\mathsf{v}\\ \mathsf{0}_{d_0-m_0,1}\end{pmatrix}.$$

This implies that σ is an eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$. Then, either σ is not $O(n\varrho^4\varepsilon^5)$ which occurs with probability $O(n^{-1/3})$ by Lemma 6.5, or $\sigma = O(n\varrho^4\varepsilon^5)$ is an eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ which occurs with probability $O(n^{-1/3})$ by Lemma 6.3. This proves our claim.

From now on, assume that $\mathsf{w} \neq \mathsf{0}_{d_0-m_0,1}$ and e is scaled such that $\|\mathsf{w}\| = 1$.

Next, we show that $\|\mathsf{v}\| = O(\varepsilon)$. Refer to the partition of $\mathsf{B}^t\mathsf{B}$ in (6.4). Expanding the equation $\mathsf{B}^t\mathsf{B}\cdot\mathsf{e} = \sigma\mathsf{e}$ gives the equation $\mathsf{B}_{TT}^t\mathsf{B}_{TT}\cdot\mathsf{v} + \begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\cdot\mathsf{w} = \sigma\mathsf{v}$. It implies that $\mathsf{v} = -(\mathsf{B}_{TT}^t\mathsf{B}_{TT} - \sigma\mathsf{I}_{m_0})^{-1}\cdot\begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\cdot\mathsf{w}$. Therefore,

$$\|\mathsf{v}\| \;\le\; \left\|(\mathsf{B}_{TT}^t\mathsf{B}_{TT} - \sigma\mathsf{I}_{m_0})^{-1}\right\| \cdot \left\|\begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\right\|. \tag{6.9}$$

Note that $\left\|\begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\right\| \le \|\mathsf{B}_{TT}^t\mathsf{B}_{TN}\| + \|\mathsf{B}_{TT}^t\mathsf{B}_{NN}\| \le \|\mathsf{B}_{TT}\|\cdot\|\mathsf{B}_{TN}\| + \|\mathsf{B}_{TT}\|\cdot\|\mathsf{B}_{NN}\|$. Then, Lemmas 6.3 and 6.4 imply that $\left\|\begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\right\| = O(n\varrho^4\varepsilon^5)$ with probability $1-O(n^{-1/3})$. By Lemmas 6.3 and 6.5, with probability $1-O(n^{-1/3})$, the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT} - \sigma\mathsf{I}_{m_0}$ are $\Theta(n\varrho^4\varepsilon^4) - O(n\varrho^4\varepsilon^5) = \Theta(n\varrho^4\varepsilon^4)$, implying that $\|(\mathsf{B}_{TT}^t\mathsf{B}_{TT} - \sigma\mathsf{I}_{m_0})^{-1}\| = \Theta(1/(n\varrho^4\varepsilon^4))$. Plugging the bounds on $\|\begin{pmatrix}\mathsf{B}_{TT}^t\mathsf{B}_{TN} & \mathsf{B}_{TT}^t\mathsf{B}_{NN}\end{pmatrix}\|$ and $\|(\mathsf{B}_{TT}^t\mathsf{B}_{TT} - \sigma\mathsf{I}_{m_0})^{-1}\|$ into (6.9) gives $\|\mathsf{v}\| = O(\varepsilon)$.

By the definition of eigenvalues, if we project each row vector of B onto the support line of e, the sum of the squared lengths of the projections is σ. We show that this sum is $O(n\varrho^4\varepsilon^6)$ as follows.

Take the pth row b_{p*} of B. Divide b_{p*} into two parts $\begin{pmatrix}\bar{\mathsf{b}}_{p*} & \tilde{\mathsf{b}}_{p*}\end{pmatrix}$, where $\bar{\mathsf{b}}_{p*}$ consists of the first m_0 entries and $\tilde{\mathsf{b}}_{p*}$ consists of the last $d_0 - m_0$ entries. Note that $\|\bar{\mathsf{b}}_{p*}\| = \sqrt{\frac{1}{2}\sum_{i=1}^{m} a_{pi}^4 + \sum_{i=1}^{m}\sum_{j=i+1}^{m} a_i^2 a_j^2} = \sqrt{\frac{1}{2}\left(\sum_{i=1}^{m} a_{pi}^2\right)^2} < \|\mathsf{a}_p\|^2 \le \varrho^2\varepsilon^2$. Therefore,

$$\mathsf{v}^t \cdot \bar{\mathsf{b}}_{p*} \;\le\; \|\mathsf{v}\| \cdot \|\bar{\mathsf{b}}_{p*}\| \;\le\; O(\varepsilon)\cdot\varrho^2\varepsilon^2 \;=\; O(\varrho^2\varepsilon^3).$$

By grouping terms in $\|\tilde{\mathsf{b}}_{p*}\|^2$, we get

$$\begin{aligned}\|\tilde{\mathsf{b}}_{p*}\|^2 &= \sum_{i=1}^{m}\sum_{j=m+1}^{d} a_{pi}^2 a_{pj}^2 + \sum_{i=m+1}^{d}\frac{1}{2}a_{pi}^4 + \sum_{i=m+1}^{d}\sum_{j=i+1}^{d} a_{pi}^2 a_{pj}^2 \\ &= \left(\sum_{i=1}^{m} a_{pi}^2\right)\left(\sum_{j=m+1}^{d} a_{pj}^2\right) + \frac{1}{2}\left(\sum_{j=m+1}^{d} a_{pj}^2\right)^2.\end{aligned}$$

The distance from a_p to $\mathcal{T}$ is $O(\varrho\varepsilon^2)$ [15, Lemma 6], so $\sum_{j=m+1}^{d} a_{pj}^2 = O(\varrho^2\varepsilon^4)$. It follows that $\|\tilde{\mathsf{b}}_{p*}\| \le \sqrt{\varrho^2\varepsilon^2\cdot O(\varrho^2\varepsilon^4) + O(\varrho^4\varepsilon^8)} = O(\varrho^2\varepsilon^3)$. Since $\|\mathsf{w}\| = 1$,

$$\mathsf{w}^t \cdot \tilde{\mathsf{b}}_{p*} \;\le\; \|\mathsf{w}\| \cdot \|\tilde{\mathsf{b}}_{p*}\| \;=\; O(\varrho^2\varepsilon^3).$$

The squared length of the projection of b_{p*} onto the support line of e is

$$\left(\frac{\mathsf{e}^t \cdot \mathsf{b}_{p*}}{\|\mathsf{e}\|}\right)^2 \;=\; \left(\frac{\mathsf{v}^t \cdot \bar{\mathsf{b}}_{p*} + \mathsf{w}^t \cdot \tilde{\mathsf{b}}_{p*}}{\sqrt{\|\mathsf{v}\|^2 + \|\mathsf{w}\|^2}}\right)^2 \;=\; \left(\frac{O(\varrho^2\varepsilon^3)}{\sqrt{1 + O(\varepsilon^2)}}\right)^2 \;=\; O(\varrho^4\varepsilon^6).$$

Summing this bound over the n rows of B gives $O(n\varrho^4\varepsilon^6)$, which is an upper bound of σ. □

7 Eigenvalues of $\mathsf{H}\mathsf{H}^t$

7.1 Preliminaries

Given a $k \times l$ matrix U, what happens to $\|\mathsf{U}\|$ if we multiply each row u_{i*} by a factor which is less than or equal to one? Let V be the matrix obtained after changing U. Then

$$\begin{aligned}\|\mathsf{V}\| &= \left(\max_{\mathsf{e}\in\mathbb{R}^l, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{V}^t\mathsf{V}\mathsf{e}\right)^{1/2} = \left(\max_{\mathsf{e}\in\mathbb{R}^l, \|\mathsf{e}\|=1} \sum_{i=1}^{k} (\mathsf{v}_{i*}\mathsf{e})^2\right)^{1/2} \\ &\le \left(\max_{\mathsf{e}\in\mathbb{R}^l, \|\mathsf{e}\|=1} \sum_{i=1}^{k} (\mathsf{u}_{i*}\mathsf{e})^2\right)^{1/2} = \left(\max_{\mathsf{e}\in\mathbb{R}^l, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{U}^t\mathsf{U}\mathsf{e}\right)^{1/2} \\ &= \|\mathsf{U}\|.\end{aligned}$$

Also, for every orthonormal $k \times k$ matrix R, $\mathsf{R}^t\mathsf{R} = \mathsf{I}_k$, and therefore,

$$\|\mathsf{R}\mathsf{U}\| = \left(\max_{\mathsf{e}\in\mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t(\mathsf{R}\mathsf{U})^t(\mathsf{R}\mathsf{U})\mathsf{e}\right)^{1/2} = \left(\max_{\mathsf{e}\in\mathbb{R}^k, \|\mathsf{e}\|=1} \mathsf{e}^t\mathsf{U}^t\mathsf{U}\mathsf{e}\right)^{1/2} = \|\mathsf{U}\|.$$

We need three more technical results concerning the angles between vectors and spaces. The first one is from [12]. The other two are folklore and we include their proofs for completeness.

Lemma 7.1 ([12, Lemma 1.1]) *Let* M *be an* $s \times s$ *real symmetric matrix with eigenvalues* $\mu_1, \ldots, \mu_s$ *in an arbitrary order. Let* v_i *denote a unit eigenvector of* M *corresponding to* μ_i*. If* $\mathsf{M} + \mathsf{M}'$ *is a real symmetric matrix,* μ' *is an eigenvalue of* $\mathsf{M} + \mathsf{M}'$*, and* z *is a unit eigenvector of* $\mathsf{M} + \mathsf{M}'$ *corresponding to* μ'*, then for every* $r \in [1, s-1]$*, the angle between* z *and the space spanned by* $\{\mathsf{v}_1, \ldots, \mathsf{v}_r\}$ *is at most* $\arcsin\left(\frac{\|\mathsf{M}'\|}{\min_{i \in [r+1,s]} |\mu_i - \mu'|}\right)$.

Lemma 7.2 *Let* $\begin{pmatrix}\mathsf{U} & \mathsf{V}\end{pmatrix}$ *be an* $s \times s$ *orthonormal matrix such that* U *is* $s \times (s-r)$ *and* V *is* $s \times r$ *for some* $r \in [1, s-1]$*. Let* $\mathsf{Z} = \begin{pmatrix}\mathsf{z}_{*1} & \cdots & \mathsf{z}_{*r}\end{pmatrix}$ *be an* $s \times r$ *orthonormal matrix. The angle between the column spaces of* V *and* Z *is* $\arcsin\left(\|\mathsf{U}^t\mathsf{Z}\|\right) \leq \arcsin\left(\sqrt{r}\max_{i \in [1,r]} \|\mathsf{U}^t\mathsf{z}_{*i}\|\right)$.

Proof. Let θ denote the angle between the column spaces of V and Z. For every unit vector z in the column space of Z, the sine of the angle between z and the column space of V is $\left\|\mathsf{U}^t\mathsf{z}\right\|/\|\mathsf{z}\| = \left\|\mathsf{U}^t\mathsf{z}\right\|$. Each such vector z is a linear combination of the columns of Z, i.e., $\mathsf{z} = \mathsf{Z}\mathsf{e}$ for some $r \times 1$ unit vector e. It follows that $\sin\theta = \max_{\|\mathsf{e}\|=1}\left\|\mathsf{U}^t\mathsf{Z}\mathsf{e}\right\| = \left\|\mathsf{U}^t\mathsf{Z}\right\|$. Moreover, $\left\|\mathsf{U}^t\mathsf{Z}\right\| \leq \left\|\mathsf{U}^t\mathsf{Z}\right\|_F = \sqrt{\sum_{i=1}^r \left\|\mathsf{U}^t\mathsf{z}_{*i}\right\|^2} \leq \sqrt{r} \cdot \max_{i \in [1,r]}\left\|\mathsf{U}^t\mathsf{z}_{*i}\right\|$. ◻

Lemma 7.3 *Let* v *be a* r*-dimensional vector. Let* $\{\mathsf{e}_i : 1 \leq i \leq k\}$ *be an orthonormal basis of the column space of a* $r \times l$ *matrix* U*. If* θ_i *denotes the acute angle between the support lines of* v *and* e_i*, then the acute angle between* v *and the column space of* U *is at least* $\arccos(\sum_{i=1}^k \cos\theta_i) = \arccos(\frac{1}{\|\mathsf{v}\|}\sum_{i=1}^k |\mathsf{v}^t\mathsf{e}_i|)$*, provided that* $\frac{1}{\|\mathsf{v}\|}\sum_{i=1}^k |\mathsf{v}^t\mathsf{e}_i| \leq 1$.

Proof. Take the projection of v onto the column space of U, and normalize the projection to a unit vector w. So $\mathsf{w} = \sum_{i=1}^k \alpha_i \mathsf{e}_i$ for some $\alpha_i \in [-1, 1]$. The cosine of the angle between v and w is equal to $\frac{1}{\|\mathsf{v}\|}\mathsf{v}^t\mathsf{w} = \frac{1}{\|\mathsf{v}\|}\sum_{i=1}^k \alpha_i \mathsf{v}^t\mathsf{e}_i \leq \frac{1}{\|\mathsf{v}\|}\sum_{i=1}^k |\alpha_i|\,|\mathsf{v}^t\mathsf{e}_i| \leq \frac{1}{\|\mathsf{v}\|}\sum_{i=1}^k |\mathsf{v}^t\mathsf{e}_i|$. ◻

The main result of this section is Lemma 7.10 which gives bounds on the eigenvalues of $\mathsf{H}\mathsf{H}^t$. The proof of Lemma 7.10 will be facilitated by a rotation of $\mathbb{R}^d$ so that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. We prove below that such a rotation does not change the eigenvalues of $\mathsf{H}\mathsf{H}^t$.

Lemma 7.4 *If we apply a rotation to* $\mathbb{R}^d$ *that keeps the origin fixed, the eigenvalues of* $\mathsf{H}\mathsf{H}^t$ *are preserved and the eigenvectors of* $\mathsf{H}\mathsf{H}^t$ *are rotated correspondingly.*

Proof. Recall that $\mathsf{L}\Lambda\mathsf{R}^t$ denotes the thin SVD of B and that $\mathsf{H}\mathsf{H}^t = \mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A}$ by Lemma 4.1(iii). Let M be a $d \times d$ rotation matrix. The proof of Lemma 6.2 reveals that the effect of applying M is produced by an isometry in $\mathbb{R}^{d_0}$. It follows that the application of M only changes the matrix R but not L or Λ in the thin SVD of B. Therefore, when we apply M, the middle part $\mathsf{L}\Sigma\mathsf{L}^t$ of $\mathsf{H}\mathsf{H}^t$ remains fixed and A is changed to $\mathsf{A}\mathsf{M}^t$ by the rotation. This changes $\mathsf{H}\mathsf{H}^t$ to $\mathsf{M}\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A}\mathsf{M}^t = \mathsf{M}\mathsf{H}\mathsf{H}^t\mathsf{M}^t$. Since M is a rotation matrix, multiplying M on the left and M^t on the right does not change the eigenvalues of $\mathsf{H}\mathsf{H}^t$, but it does rotate the eigenvectors of $\mathsf{H}\mathsf{H}^t$ correspondingly. ◻

7.2 Analysis

Suppose that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. Then, A can be divided into two submatrices. Let A_T be the leftmost $n \times m$ submatrix of A. Let A_N be the rightmost $n \times (d-m)$ submatrix

of A. That is, $\mathsf{A} = \begin{pmatrix}\mathsf{A}_T & \mathsf{A}_N\end{pmatrix}$. Lemma 4.1(iii) implies that $\mathsf{H} = (\widehat{\mathsf{B}}^\dagger\mathsf{A})^t = \begin{pmatrix}\widehat{\mathsf{B}}^\dagger\mathsf{A}_T & \widehat{\mathsf{B}}^\dagger\mathsf{A}_N\end{pmatrix}^t$. We define two submatrices of H as follows.

$$\mathsf{H} = \begin{pmatrix}\mathsf{H}_T \\ \mathsf{H}_N\end{pmatrix}, \text{ where } \mathsf{H}_T \stackrel{\text{def}}{=} (\widehat{\mathsf{B}}^\dagger\mathsf{A}_T)^t \text{ and } \mathsf{H}_N \stackrel{\text{def}}{=} (\widehat{\mathsf{B}}^\dagger\mathsf{A}_N)^t.$$

Note that H_T is an $m \times d_0$ matrix and H_N is a $(d-m) \times d_0$ matrix. Our analysis begins with bounding $\|\mathsf{H}_N\|$.

Lemma 7.5 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, $\|\mathsf{H}_N\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$.*

Proof. Recall that f_ℓ, $\ell \in [m+1, d]$, denotes a coordinate function of $\mathcal{M}$ at the origin. Define:

$$\begin{aligned}
\tilde{a}_{p\ell} &\stackrel{\text{def}}{=} \tfrac{1}{2}\mathrm{D}^2 f_\ell|_{\mathsf{0}}\big(\begin{pmatrix}a_{p1} & \cdots & a_{pm}\end{pmatrix}^t, \begin{pmatrix}a_{p1} & \cdots & a_{pm}\end{pmatrix}^t\big) \\
\tilde{\mathsf{A}}_N &\stackrel{\text{def}}{=} \begin{pmatrix}\tilde{a}_{1,m+1} & \cdots & \tilde{a}_{1d} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{n,m+1} & \cdots & \tilde{a}_{nd}\end{pmatrix} \\
\tilde{\mathsf{H}}_N &\stackrel{\text{def}}{=} (\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N)^t
\end{aligned}$$

In the Taylor expansion of f_ℓ, there is no constant term or first order term because $\mathcal{M}$ contains the origin and the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. As a result, if ε is sufficiently small, then $\tilde{a}_{p\ell}$ is close to $a_{p\ell}$, and therefore, $\tilde{\mathsf{A}}_N$ and $\tilde{\mathsf{H}}_N$ are approximations of A_N and H_N, respectively. Substituting $\mathsf{A}_N = \tilde{\mathsf{A}}_N + (\mathsf{A}_N - \tilde{\mathsf{A}}_N)$ into $\mathsf{H}_N = (\widehat{\mathsf{B}}^\dagger\mathsf{A}_N)^t$, we obtain $\mathsf{H}_N = \tilde{\mathsf{H}}_N + (\widehat{\mathsf{B}}^\dagger(\mathsf{A}_N - \tilde{\mathsf{A}}_N))^t$. Therefore, $\big\|\mathsf{H}_N\big\| \le \big\|\tilde{\mathsf{H}}_N\big\| + \big\|\widehat{\mathsf{B}}^\dagger(\mathsf{A}_N - \tilde{\mathsf{A}}_N)\big\|$.

We first bound $\big\|\widehat{\mathsf{B}}^\dagger(\mathsf{A}_N - \tilde{\mathsf{A}}_N)\big\|$. For every $p \in [1, n]$, the pth row of $\mathsf{A}_N - \tilde{\mathsf{A}}_N$ contains the third and higher order terms in the Taylor expansions of $f_\ell((a_{p1} \ \cdots \ a_{pm})^t)$ for $\ell \in [m+1, d]$. By Lemma 5.1(ii), if ε is small enough, the 2-norm of each row of $\mathsf{A}_N - \tilde{\mathsf{A}}_N$ is $O(\varrho\varepsilon^3)$. Thus, $\big\|\mathsf{A}_N - \tilde{\mathsf{A}}_N\big\| \le \big\|\mathsf{A}_N - \tilde{\mathsf{A}}_N\big\|_F = O(\sqrt{n}\varrho\varepsilon^3)$. Since λ_{m_0+1} is the smallest singular value of $\widehat{\mathsf{B}}$, $\|\widehat{\mathsf{B}}^\dagger\| = 1/\lambda_{m_0+1}$. Therefore, $\big\|\widehat{\mathsf{B}}^\dagger(\mathsf{A}_N - \tilde{\mathsf{A}}_N)\big\| \le \big\|\widehat{\mathsf{B}}^\dagger\big\| \cdot \big\|\mathsf{A}_N - \tilde{\mathsf{A}}_N\big\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$.

It remains to show that $\big\|\tilde{\mathsf{H}}_N\big\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$. Observe that $\big\|\tilde{\mathsf{H}}_N\big\| = \big\|\tilde{\mathsf{H}}_N^t\big\| = \big\|\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N\big\|$. The smallest singular value of B is smaller than or equal to the smallest singular value of $\widehat{\mathsf{B}}$ by construction. Intuitively, we would expect $\big\|\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N\big\| \le \big\|\mathsf{B}^\dagger\tilde{\mathsf{A}}_N\big\|$. Therefore, we can bound $\big\|\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N\big\|$ if we can bound $\big\|\mathsf{B}^\dagger\tilde{\mathsf{A}}_N\big\|$.

Recall that $\mathrm{D}^2 f_\ell|_{\mathsf{0}}$ can be viewed as the $m \times m$ symmetric matrix $\big(\partial^2 f_\ell/\partial x_i \partial x_j\big)_{i,j\in[1,m]}$ with every entry evaluated at the origin. Let $q_{\ell,ij}$ denote the (i,j) entry of $\mathrm{D}^2 f_\ell|_{\mathsf{0}}$. By definition, $\tilde{a}_{p\ell} = \frac{1}{2}(a_{p1} \ \cdots \ a_{pm}) \cdot \mathrm{D}^2 f_\ell|_{\mathsf{0}} \cdot (a_{p1} \ \cdots \ a_{pm})^t$. Expanding this equation gives:

$$\tilde{a}_{p\ell} \ = \ \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{2}q_{\ell,ij}a_{pi}a_{pj} \ = \ \sum_{i=1}^{m}\frac{1}{2}q_{\ell,ii}a_{pi}^2 + \sum_{i=1}^{m}\sum_{j=i+1}^{m}q_{\ell,ij}a_{pi}a_{pj}.$$

We extend the range of i and j in $q_{\ell,ij}$ to $[1, d]$ by letting $q_{\ell,ij} = 0$ whenever $i \in [m+1, d]$ or $j \in [m+1, d]$. Then, define Z to be the $d_0 \times (d-m)$ matrix $\begin{pmatrix}\mathsf{z}_{*1} & \cdots & \mathsf{z}_{*d-m}\end{pmatrix}$, where

$$\mathsf{z}_{*\ell-m} = \begin{pmatrix}\frac{1}{\sqrt{2}}q_{\ell,11} & q_{\ell,12} & \cdots & q_{\ell,1d} & \frac{1}{\sqrt{2}}q_{\ell,22} & q_{\ell,23} & \cdots & q_{\ell,2d} & \cdots & \frac{1}{\sqrt{2}}q_{\ell,dd}\end{pmatrix}^t.$$

The definition of $\mathsf{z}_{*\ell-m}$ is crafted so that $\mathsf{B}\mathsf{z}_{*\ell-m}$ equals $\begin{pmatrix}\tilde{a}_{1\ell} & \cdots & \tilde{a}_{n\ell}\end{pmatrix}^t$. Therefore,

$$\mathsf{B}\mathsf{Z} = \tilde{\mathsf{A}}_N. \tag{7.1}$$

Consider the square matrix obtained by appending $d_0 - d + m$ zero columns to the right of Z. Then, $\|\mathsf{Z}\|$ is equal to the 2-norm of this square matrix, which by the Gershgorin Circle Theorem is at most $\max_{j\in[1,d-m]} \sum_{i=1}^{d_0} |z_{ij}|$. Among the entries in z_{*j}, m_0 entries are from the upper triangular portion of $\mathrm{D}^2 f_{j+m}|_{\mathsf{0}}$ and the other $d_0 - m_0$ entries are zeros by definition. Thus, $\sum_{i=1}^{d_0} |z_{ij}| \le \sqrt{m_0} \cdot \|\mathsf{z}_{*j}\|$. By Lemma 5.1(i), $\left\|\mathrm{D}^2 f_{j+m}|_{\mathsf{0}}\right\| = O(1/\varrho)$ and therefore, $\|\mathsf{z}_{*j}\| \le \left\|\mathrm{D}^2 f_{j+m}|_{\mathsf{0}}\right\|_F \le \sqrt{m} \cdot \left\|\mathrm{D}^2 f_{j+m}|_{\mathsf{0}}\right\| = O(1/\varrho)$. By the Gershgorin Circle Theorem, $\|\mathsf{Z}\| \le \max_{j\in[1,d-m]} \sum_{i=1}^{d_0} |z_{ij}| \le \max_{j\in[1,d-m]} \sqrt{m_0} \cdot \|\mathsf{z}_{*j}\| = O(1/\varrho)$. Since $d_0 - m_0$ rows of Z contain only zeros by construction, $\|\mathsf{Z}\|_F \le \sqrt{m_0} \cdot \|\mathsf{Z}\| = O(1/\varrho)$.

$\mathsf{B}^\dagger \tilde{\mathsf{A}}_N$ also satisfies (7.1), i.e., $\mathsf{B}(\mathsf{B}^\dagger \tilde{\mathsf{A}}_N) = \tilde{\mathsf{A}}_N$. By the property of pseudoinverse, $\left\|\mathsf{B}^\dagger \tilde{\mathsf{A}}_N\right\|_F$ is no more than $\|\mathsf{Z}\|_F$ or the Frobenius norm of any matrix that satisfies (7.1). As a result,

$$\left\|\mathsf{B}^\dagger \tilde{\mathsf{A}}_N\right\| \le \left\|\mathsf{B}^\dagger \tilde{\mathsf{A}}_N\right\|_F \le \|\mathsf{Z}\|_F = O(1/\varrho). \tag{7.2}$$

We relate $\left\|\widehat{\mathsf{B}}^\dagger \tilde{\mathsf{A}}_N\right\|$ to $\left\|\mathsf{B}^\dagger \tilde{\mathsf{A}}_N\right\|$ as follows. Recall that $\mathsf{L}\Lambda\mathsf{R}^t$ and $\mathsf{L}\widehat{\Lambda}\mathsf{R}^t$ are the thin SVDs of B and $\widehat{\mathsf{B}}$, respectively, and $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ are the singular values of B. Let σ_i and $\widehat{\sigma}_i$ denote the diagonal entries of $\Lambda^\dagger$ and $\widehat{\Lambda}^\dagger$, respectively. By Lemmas 6.5 and 6.6, it holds with probability $1 - O(n^{-1/3})$ that $\lambda_{m_0} = \Theta(\sqrt{n}\varrho^2\varepsilon^2)$ and $\lambda_{m_0+1} = O(\sqrt{n}\varrho^2\varepsilon^3)$. Therefore,

$$\forall\, i \in [1, m_0], \quad \widehat{\sigma}_i = \sigma_i = 1/\lambda_i,$$
$$\forall\, i \in [m_0 + 1, n], \quad \widehat{\sigma}_i = \tfrac{1}{\lambda_{m_0+1}} \le \tfrac{1}{\lambda_i} = \sigma_i.$$

(For simplicity, we assume that the smallest singular value λ_n of Λ is positive, and therefore, $\widehat{\Lambda}$ has n positive diagonal entries. Otherwise, if $\lambda_i = 0$, both σ_i and $\widehat{\sigma}_i$ are zero by our definition of $\widehat{\Lambda}$.)

The full SVD of B is $\mathsf{L}\,(\Lambda \;\; \mathsf{0}_{n,d_0-n})\,\bar{\mathsf{R}}^t$, where $\bar{\mathsf{R}}$ consists of the d_0 unit eigenvectors of $\mathsf{B}^t\mathsf{B}$ and R is the leftmost $d_0 \times n$ submatrix of $\bar{\mathsf{R}}$. It follows that $\mathsf{L}\,(\widehat{\Lambda} \;\; \mathsf{0}_{n,d_0-n})\,\bar{\mathsf{R}}^t$ is the full SVD of $\widehat{\mathsf{B}}$, and therefore, $\mathsf{B}^\dagger = \bar{\mathsf{R}}\,(\Lambda^\dagger \;\; \mathsf{0}_{n,d_0-n})^t\,\mathsf{L}^t$ and $\widehat{\mathsf{B}}^\dagger = \bar{\mathsf{R}}\,(\widehat{\Lambda}^\dagger \;\; \mathsf{0}_{n,d_0-n})^t\,\mathsf{L}^t$. Observe that:

- For $i \in [1, n]$, the ith row in $\bar{\mathsf{R}}^t\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N$ equals the ith row in $\bar{\mathsf{R}}^t\mathsf{B}^\dagger\tilde{\mathsf{A}}_N$ multiplied by $\widehat{\sigma}_i/\sigma_i$.
- The bottom $d_0 - n$ rows in both $\bar{\mathsf{R}}^t\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N$ and $\bar{\mathsf{R}}^t\mathsf{B}^\dagger\tilde{\mathsf{A}}_N$ contain only zeros.

Therefore, $\left\|\bar{\mathsf{R}}^t\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N\right\| \le \left\|\bar{\mathsf{R}}^t\mathsf{B}^\dagger\tilde{\mathsf{A}}_N\right\|$ as $\widehat{\sigma}_i \le \sigma_i$ for $i \in [1, n]$.

Multiplying $\tilde{\mathsf{H}}_{\mathsf{N}}^t$, $\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N$ and $\mathsf{B}^\dagger\tilde{\mathsf{A}}_N$ on the left by $\bar{\mathsf{R}}^t$ does not change their 2-norms. Therefore,

$$\left\|\tilde{\mathsf{H}}_N\right\| = \left\|\tilde{\mathsf{H}}_N^t\right\| = \left\|\bar{\mathsf{R}}^t\widehat{\mathsf{B}}^\dagger\tilde{\mathsf{A}}_N\right\| \le \left\|\bar{\mathsf{R}}^t\mathsf{B}^\dagger\tilde{\mathsf{A}}_N\right\| = \left\|\mathsf{B}^\dagger\tilde{\mathsf{A}}_N\right\| \overset{(7.2)}{=} O(1/\varrho).$$

Thus, $\left\|\tilde{\mathsf{H}}_N\right\| = O(\sqrt{n}\varrho\varepsilon^3/\lambda_{m_0+1})$ because $\lambda_{m_0+1} = O(\sqrt{n}\varrho^2\varepsilon^3)$ by Lemma 6.6. □

Before we analyze the singular values of H_T, we prove two technical results about the column vectors of A_T and the column space of B_{TT}. First, we show that the acute angle between the support lines of any two distinct columns of A_T is large and A_T has rank m.

Lemma 7.6 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, for every distinct $i, j \in [1, m]$, the acute angle between the support lines of a_{*i} and a_{*j} is at least* $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$*, and A_T has rank m.*

Proof. Take an arbitrary pair of distinct columns a_{*i} and a_{*j} of A_T. By Lemma 5.4(iii), it holds with probability $1 - O(n^{-1/3})$ that $\|\mathsf{a}_{*i}\|$ and $\|\mathsf{a}_{*j}\|$ are at least $c_1\sqrt{n}\varrho\varepsilon$ for some constant $c_1 > 0$. The inner product $\mathsf{a}_{*i}^t \cdot \mathsf{a}_{*j}$ equals $\sum_{p=1}^{n} a_{pi}a_{pj}$. Apply Lemma 5.4(vi) with the

constant $c = \frac{1}{18}c_1^2 m^{-1}$. It implies that if ε is small enough, then with probability $1 - O(n^{-1/3})$, $\left|\mathsf{a}_{*i}^t \cdot \mathsf{a}_{*j}\right| \le cn^{2/3}\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4) \le 2cn^{2/3}\varrho^2\varepsilon^2 = \frac{c_1^2}{9}m^{-1}n^{2/3}\varrho^2\varepsilon^2$. Thus, $\left|\mathsf{a}_{*i}^t \cdot \mathsf{a}_{*j}\right| / (\|\mathsf{a}_{*i}\|\|\mathsf{a}_{*j}\|) \le \frac{1}{9}m^{-1}n^{-1/3}$, which implies that the acute angle between the support lines of a_{*i} and a_{*j} is at least $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$.

We prove that A_T has rank m by bounding the minimum eigenvalue of $\mathsf{A}_T^t\mathsf{A}_T$ away from zero. Observe that:

$$\mathsf{A}_T^t\mathsf{A}_T = \begin{pmatrix} \sum_{p=1}^n a_{p1}^2 & \sum_{p=1}^n a_{p1}a_{p2} & \cdots & \sum_{p=1}^n a_{p1}a_{pm} \\ \sum_{p=1}^n a_{p2}a_{p1} & \sum_{p=1}^n a_{p2}^2 & \cdots & \sum_{p=1}^n a_{p2}a_{pm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{p=1}^n a_{pm}a_{p1} & \sum_{p=1}^n a_{pm}a_{p2} & \cdots & \sum_{p=1}^n a_{pm}^2 \end{pmatrix}$$

By the Gershgorin Circle Theorem, the minimum eigenvalue of $\mathsf{A}_T^t\mathsf{A}_T$ is greater than or equal to $\min_{i\in[1,m]}\left\{\sum_{p=1}^n a_{pi}^2 - \sum_{j\in[1,m]\setminus\{i\}}\left|\sum_{p=1}^n a_{pi}a_{pj}\right|\right\}$. Apply Lemma 5.4(iii) with $c = \frac{1}{3(m+2)}$ to obtain $\sum_{p=1}^n a_{pi}^2 \ge \frac{1}{m+2}n\varrho^2\varepsilon^2 - \frac{1}{3(m+2)}n^{2/3}\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4)$ with probability $1 - O(n^{-1/3})$. Apply Lemma 5.4(vi) with $c = \frac{1}{3(m-1)(m+2)}$ to obtain $\sum_{j\in[1,m]\setminus\{i\}}\left|\sum_{p=1}^n a_{pi}a_{pj}\right| \le \frac{1}{3(m+2)}n^{2/3}\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)$ with probability $1 - O(n^{-1/3})$. Then, with probability $1 - O(n^{-1/3})$, the minimum eigenvalue of $\mathsf{A}_T^t\mathsf{A}_T$ is at least $\frac{1}{3(m+2)}n\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4)$, which is positive when ε is small enough. □

Next, we show a lower bound on the angle between any column vector of A_T and the column space of B_{TT}.

Lemma 7.7 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, every column vector of A_T makes an angle* $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$ *or more with the column space of B_{TT}.*

Proof. We first introduce three constants c_1, c_2 and c_3. It holds with probability $1 - O(n^{-1/3})$ that all eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ are greater than $c_1 n\varrho^4\varepsilon^4$ for some constant $c_1 > 0$ (Lemma 6.3), the 2-norm of each column of B_{TT} is greater than $c_2\sqrt{n}\varrho^2\varepsilon^2$ for some constant $c_2 > 0$ (Lemma 5.4(i) and (ii)), and the 2-norm of every column of A_T is greater than $c_3\sqrt{n}\varrho\varepsilon$ for some constant $c_3 > 0$ (Lemma 5.4(iii)).

Let $\mathsf{e}_1, \ldots, \mathsf{e}_{m_0}$ be m_0 unit eigenvectors corresponding to the eigenvalues of $\mathsf{B}_{TT}\mathsf{B}_{TT}^t$. Let $\mathsf{b}_{*1}, \mathsf{b}_{*2}, \ldots, \mathsf{b}_{*m_0}$ denote the columns of B_{TT}.

We first show that for every $j \in [1, m_0]$, $\mathsf{e}_j = \sum_{i=1}^{m_0}\beta_{ij}\mathsf{b}_{*i}/\|\mathsf{b}_{*i}\|$ for some coefficients β_{ij}'s such that $|\beta_{ij}| \le 1/\sqrt{c_1}$. By Lemma 6.3, $\mathsf{B}_{TT}\mathsf{B}_{TT}^t$ has rank m_0, implying that $\mathsf{e}_j = \sum_{i=1}^{m_0}\beta_{ij}\mathsf{b}_{*i}/\|\mathsf{b}_{*i}\|$ for some coefficients β_{ij}'s. It remains to bound $|\beta_{ij}|$. We put the equations $\mathsf{e}_j = \sum_{i=1}^{m_0}\beta_{ij}\mathsf{b}_{*i}/\|\mathsf{b}_{*i}\|$ in matrix form as follows.

$$\mathsf{E} = \begin{pmatrix} \mathsf{e}_1 & \cdots & \mathsf{e}_{m_0} \end{pmatrix} = \mathsf{B}_{TT}\begin{pmatrix} \frac{1}{\|\mathsf{b}_{*1}\|}\beta_{11} & \cdots & \frac{1}{\|\mathsf{b}_{*1}\|}\beta_{1m_0} \\ \vdots & \ddots & \vdots \\ \frac{1}{\|\mathsf{b}_{*m_0}\|}\beta_{m_0 1} & \cdots & \frac{1}{\|\mathsf{b}_{*m_0}\|}\beta_{m_0 m_0} \end{pmatrix} \tag{7.3}$$

The thin SVD of B_{TT}^t is $\mathsf{V}\mathsf{D}\mathsf{E}^t$, where D is an $m_0 \times m_0$ diagonal matrix whose (j, j) entry is the square root μ_j of the j-th largest eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$, and V is an $m_0 \times m_0$ matrix whose j-th column v_{*j} is the unit eigenvector of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ corresponding to μ_j^2. Multiplying both sides

of (7.3) by $\mathsf{V}\mathsf{D}^{-1}\mathsf{E}^t$ gives:

$$\mathsf{V}\mathsf{D}^{-1} = \begin{pmatrix} \frac{1}{\|\mathsf{b}_{*1}\|}\beta_{11} & \cdots & \frac{1}{\|\mathsf{b}_{*1}\|}\beta_{1m_0} \\ \vdots & \ddots & \vdots \\ \frac{1}{\|\mathsf{b}_{*m_0}\|}\beta_{m_01} & \cdots & \frac{1}{\|\mathsf{b}_{*m_0}\|}\beta_{m_0m_0} \end{pmatrix}.$$

Comparing the matrices on the two sides term by term shows that $\beta_{ij} = \|\mathsf{b}_{*i}\|v_{ij}/\mu_j$ for $i, j \in [1, m_0]$. By assumption, $\mu_j \geq \sqrt{c_1}\sqrt{n}\varrho^2\varepsilon^2$. Since $|a_{pr}a_{ps}| \leq \varrho^2\varepsilon^2$ for $r, s \in [1, d]$, $\|\mathsf{b}_{*i}\| \leq \sqrt{n}\varrho^2\varepsilon^2$. Also, $|v_{ij}| \leq 1$ as v_{*j} is a unit vector. It follows that $|\beta_{ij}| \leq 1/\sqrt{c_1}$.

We bound $\mathsf{a}_{*k}^t \cdot \mathsf{b}_{*i}$ for $k \in [1, m]$ and $i \in [1, m_0]$ as follows. The inner product $\mathsf{a}_{*k}^t \cdot \mathsf{b}_{*i}$ is equal to $\sum_{p=1}^n \frac{1}{\sqrt{2}}a_{pk}a_{pr}^2$ or $\sum_{p=1}^n a_{pk}a_{pr}a_{ps}$ for some possibly non-distinct $k, r, s \in [1, m]$. Apply Lemma 5.4(v) with the constant $c = \frac{\sqrt{c_1}c_2c_3}{18}m^{-1}m_0^{-2}$. It implies that, with probability $1 - O(n^{-1/3})$, $\left|\mathsf{a}_{*k}^t \cdot \mathsf{b}_{*i}\right| \leq cn^{2/3}\varrho^3\varepsilon^3 + O(n\varrho^3\varepsilon^5)$, which is at most $2cn^{2/3}\varrho^3\varepsilon^3 = \frac{\sqrt{c_1}c_2c_3}{9}m^{-1}m_0^{-2}n^{2/3}\varrho^3\varepsilon^3$ when ε is sufficiently small.

For $k \in [1, m]$ and $j \in [1, m_0]$, $\frac{1}{\|\mathsf{a}_{*k}\|}\left|\mathsf{a}_{*k}^t \cdot \mathsf{e}_j\right| \leq \sum_{i=1}^{m_0} \frac{|\beta_{ij}|}{\|\mathsf{a}_{*k}\|\,\|\mathsf{b}_{*i}\|}\left|\mathsf{a}_{*k}^t \cdot \mathsf{b}_{*i}\right| \leq \sum_{i=1}^{m_0} \frac{1}{\sqrt{c_1}c_2c_3n\varrho^3\varepsilon^3} \cdot \frac{\sqrt{c_1}c_2c_3}{9}m^{-1}m_0^{-2}n^{2/3}\varrho^3\varepsilon^3 = \frac{1}{9}m^{-1}m_0^{-1}n^{-1/3}$. Therefore, $\frac{1}{\|\mathsf{a}_{*k}\|}\sum_{j=1}^{m_0}\left|\mathsf{a}_{*k}^t \cdot \mathsf{e}_j\right| \leq \frac{1}{9}m^{-1}n^{-1/3}$. Then, Lemma 7.3 implies that the acute angle between a_{*k} and the column space of B_{TT} is $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$ or more. □

The matrix $\widehat{\Lambda}$ in the thin SVD $\mathsf{L}\widehat{\Lambda}\mathsf{R}^t$ of $\widehat{\mathsf{B}}$ can be partitioned into blocks to separate the m_0 largest singular values of B from the λ_{m_0+1}'s. The matrices L and R can then be partitioned correspondingly. Specifically, we obtain

$$\begin{aligned} \widehat{\Lambda} &= \begin{pmatrix} \widehat{\Lambda}_0 & \mathsf{0}_{m_0,n-m_0} \\ \mathsf{0}_{n-m_0,m_0} & \lambda_{m_0+1}\mathsf{I}_{n-m_0} \end{pmatrix}, \\ \mathsf{L} &= \Big(\underbrace{\mathsf{L}_0}_{m_0 \text{ columns}} \quad \underbrace{\mathsf{L}_1}_{n-m_0 \text{ columns}} \Big), \\ \mathsf{R} &= \Big(\underbrace{\mathsf{R}_0}_{m_0 \text{ columns}} \quad \underbrace{\mathsf{R}_1}_{n-m_0 \text{ columns}} \Big). \end{aligned} \tag{7.4}$$

Note that $\|\mathsf{L}_0\| = \|\mathsf{L}_1\| = 1$.

Our subsequent analysis of the eigenvalues of H_T requires an upper bound on the angle between the column spaces of L_0 and B_{TT}, which is given in the following result.

Lemma 7.8 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, the angle between the column spaces of L_0 and B_{TT} is $O(\varepsilon^2)$.*

Proof. Recall that $\mathsf{B} = \begin{pmatrix}\mathsf{B}_{TT} & \mathsf{B}_{TN} & \mathsf{B}_{NN}\end{pmatrix}$. Therefore, $\mathsf{B}\mathsf{B}^t = \mathsf{B}_{TT}\mathsf{B}_{TT}^t + \mathsf{B}_{TN}\mathsf{B}_{TN}^t + \mathsf{B}_{NN}\mathsf{B}_{NN}^t$. Let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_{m_0}$ be the m_0 largest eigenvalues of $\mathsf{B}_{TT}\mathsf{B}_{TT}^t$. Let V be an $n \times m_0$ matrix whose columns are the unit eigenvectors of $\mathsf{B}_{TT}\mathsf{B}_{TT}^t$ corresponding to $\mu_1, \ldots, \mu_{m_0}$. The diagonalization of $\mathsf{B}_{TT}\mathsf{B}_{TT}^t$ is:

$$\mathsf{B}_{TT}\mathsf{B}_{TT}^t = \begin{pmatrix}\mathsf{V} & *\end{pmatrix} \cdot \mathrm{diag}_n(\mu_1, \ldots, \mu_{m_0}, 0, \ldots, 0) \cdot \begin{pmatrix}\mathsf{V} & *\end{pmatrix}^t. \tag{7.5}$$

Take any column vector z of L_0. Thus, $\mathsf{B}\mathsf{B}^t\mathsf{z} = \mu'\mathsf{z}$ where μ' is one of the largest m_0 eigenvalues of $\mathsf{B}\mathsf{B}^t$. We expand $\mathsf{B}\mathsf{B}^t$ to obtain $(\mathsf{B}_{TT}\mathsf{B}_{TT}^t + \mathsf{B}_{TN}\mathsf{B}_{TN}^t + \mathsf{B}_{NN}\mathsf{B}_{NN}^t)\mathsf{z} = \mu'\mathsf{z}$. We invoke Lemma 7.1 with $\mathsf{M} = \mathsf{B}_{TT}\mathsf{B}_{TT}^t$ and $\mathsf{M}' = \mathsf{B}_{TN}\mathsf{B}_{TN}^t + \mathsf{B}_{NN}\mathsf{B}_{NN}^t$. Lemma 7.1 and (7.5) imply that the angle between z and the column space of V is at most $\arcsin(\|\mathsf{M}'\|/|\mu'|)$. By Lemma 6.4,

$\|\mathsf{M}'\| \le \|B_{TN}\|^2 + \|B_{NN}\|^2 = O(n\varrho^4\varepsilon^6)$. Lemma 6.5 implies that the largest m_0 eigenvalues of $\mathsf{B}\mathsf{B}^t$, including μ', are $\Theta(n\varrho^4\varepsilon^4)$ with probability $1-O(n^{-1/3})$. It follows that, with probability $1-O(n^{-1/3})$, for every column z of L_0, the angle between z and the column space of V is at most $\arcsin(O(\varepsilon^2)) = O(\varepsilon^2)$.

Let U be an $n \times (n-m_0)$ matrix such that the columns of U and V form an orthonormal basis of the column space of B. For every column z of L_0, since z makes an $O(\varepsilon^2)$ angle with the column space of V, the angle between z and the column space of U is $\pi/2 - O(\varepsilon^2)$. That is, $\left\|\mathsf{U}^t\mathsf{z}\right\| = \cos(\pi/2 - O(\varepsilon^2)) = O(\varepsilon^2)$. Then, Lemma 7.2 implies that the angle between the column spaces of L_0 and V is $\arcsin(\sqrt{m_0}\cdot O(\varepsilon^2)) = O(\varepsilon^2)$. Since the columns in V form an orthonormal basis of the column space of B_{TT}, the angle between the column spaces of L_0 and B_{TT} is $O(\varepsilon^2)$. □

We are ready to bound the eigenvalues of H_T. The analysis uses the tools that we have just developed, namely Lemmas 7.6, 7.7, and 7.8.

Lemma 7.9 *Assume that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. If ε is sufficiently small, then with probability $1-O(n^{-1/3})$, the singular values of H_T are $\Theta(\sqrt{n}\varrho\varepsilon/\lambda_{m_0+1})$ and so $\|\mathsf{H}_T\| = \Theta(\sqrt{n}\varrho\varepsilon/\lambda_{m_0+1})$.*

Proof. Refer to the partitions of $\widehat{\Lambda}$, L and R in (7.4). Note that $\widehat{\Lambda}_0 = \mathrm{diag}_{m_0}(\lambda_1, \cdots, \lambda_{m_0})$, where $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{m_0}$ are the m_0 largest singular values of B. Then,

$$\begin{aligned}
\mathsf{H}_T^t &= \widehat{\mathsf{B}}^\dagger \mathsf{A}_T = (\mathsf{R}_0 \;\; \mathsf{R}_1) \begin{pmatrix} \widehat{\Lambda}_0^\dagger & \mathsf{0}_{m_0,n-m_0} \\ \mathsf{0}_{n-m_0,m_0} & \frac{1}{\lambda_{m_0+1}}\mathsf{I}_{n-m_0} \end{pmatrix} \begin{pmatrix} \mathsf{L}_0^t \\ \mathsf{L}_1^t \end{pmatrix} \mathsf{A}_T \\
&= \mathsf{R}_0\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T + \tfrac{1}{\lambda_{m_0+1}}\mathsf{R}_1\mathsf{L}_1^t\mathsf{A}_T.
\end{aligned}$$

Since every column vector of R_0 is orthogonal to any column vector of R_1 (i.e., $\mathsf{R}_0^t\mathsf{R}_1 = \mathsf{0}_{m_0,n-m_0}$ and $\mathsf{R}_1^t\mathsf{R}_0 = \mathsf{0}_{n-m_0,m_0}$), we obtain

$$\begin{aligned}
\mathsf{H}_T\mathsf{H}_T^t &= \left(\mathsf{R}_0\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T + \tfrac{1}{\lambda_{m_0+1}}\mathsf{R}_1\mathsf{L}_1^t\mathsf{A}_T\right)^t \left(\mathsf{R}_0\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T + \tfrac{1}{\lambda_{m_0+1}}\mathsf{R}_1\mathsf{L}_1^t\mathsf{A}_T\right) \\
&= \left(\mathsf{R}_0\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T\right)^t \left(\mathsf{R}_0\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T\right) + \tfrac{1}{\lambda_{m_0+1}^2}\left(\mathsf{R}_1\mathsf{L}_1^t\mathsf{A}_T\right)^t\left(\mathsf{R}_1\mathsf{L}_1^t\mathsf{A}_T\right) \\
&= \mathsf{A}_T^t\mathsf{L}_0\left(\widehat{\Lambda}_0^\dagger\right)^2\mathsf{L}_0^t\mathsf{A}_T + \tfrac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T.
\end{aligned}$$

The three matrices $\mathsf{H}_T\mathsf{H}_T^t$, $\mathsf{A}_T^t\mathsf{L}_0(\widehat{\Lambda}_0^\dagger)^2\mathsf{L}_0^t\mathsf{A}_T$, and $\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$ are symmetric and positive semi-definite. It follows that the maximum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is at most the sum of the maximum eigenvalues of $\mathsf{A}_T^t\mathsf{L}_0(\widehat{\Lambda}_0^\dagger)^2\mathsf{L}_0^t\mathsf{A}_T$ and $\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$, and the minimum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is at least the maximum of the minimum eigenvalues of $\mathsf{A}_T^t\mathsf{L}_0(\widehat{\Lambda}_0^\dagger)^2\mathsf{L}_0^t\mathsf{A}_T$ and $\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$.

For every $i \in [1,m]$, $\|\mathsf{a}_{*i}\| = \sqrt{\sum_{p=1}^n a_{pi}^2} \le \sqrt{n}\varrho\varepsilon$ as $|a_{pi}| \le \varrho\varepsilon$. Therefore, $\|\mathsf{A}_T\| \le \sqrt{mn}\varrho\varepsilon$. Then, $\left\|\frac{1}{\lambda_{m_0+1}^2}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T\right\| \le \|\mathsf{A}_T\|^2\|\mathsf{L}_1\|^2/\lambda_{m_0+1}^2 = \|\mathsf{A}_T\|^2/\lambda_{m_0+1}^2 = O(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$.

By Lemma 6.5, it holds with probability $1-O(n^{-1/3})$ that the diagonal entries of $\widehat{\Lambda}_0^\dagger$ are $\Theta(1/(\sqrt{n}\varrho^2\varepsilon^2))$. Thus, it holds with probability $1-O(n^{-1/3})$ that $\left\|\mathsf{A}_T^t\mathsf{L}_0(\widehat{\Lambda}_0^\dagger)^2\mathsf{L}_0^t\mathsf{A}_T\right\| = \left\|\widehat{\Lambda}_0^\dagger\mathsf{L}_0^t\mathsf{A}_T\right\|^2 \le \left\|\widehat{\Lambda}_0^\dagger\right\|^2\|\mathsf{L}_0\|^2\|\mathsf{A}_T\|^2 = \|\mathsf{A}_T\|^2/\Theta(n\varrho^4\varepsilon^4) = O(1/(\varrho^2\varepsilon^2))$, which is $O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$ as $\lambda_{m_0+1} = O(\sqrt{n}\varrho^2\varepsilon^3)$ with probability $1-O(n^{-1/3})$ by Lemma 6.6.

We conclude from the previous two paragraphs that $\left\|\mathsf{H}_T\mathsf{H}_T^t\right\| = O(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ or, equivalently, $\|\mathsf{H}_T\| = O(\sqrt{n}\varrho\varepsilon/\lambda_{m_0+1})$. In the rest of the proof, we show that every eigenvalue of $\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$ is $\Omega(n\varrho^2\varepsilon^2)$. This implies that the minimum eigenvalue of $\mathsf{H}_T\mathsf{H}_T^t$ is $\Omega(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ or, equivalently, the minimum singular value of H_T is $\Omega(\sqrt{n}\varrho\varepsilon/\lambda_{m_0+1})$.

Lemmas 7.7 and 7.8 imply that, with probability $1 - O(n^{-1/3})$, every column vector a_{*i} of A_T makes an angle at least $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right) - O(\varepsilon^2)$ with the column space of L_0. Therefore, with probability $1 - O(n^{-1/3})$, every column vector a_{*i} of A_T makes an angle at most $O(\varepsilon^2) + \arcsin\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$ with the column space of L_1. Let $\bar{\mathsf{a}}_{*i}$ denote the projection of a_{*i} in the column space of L_1.

Take an arbitrary pair of distinct columns a_{*i} and a_{*j} of A_T. By Lemma 7.6, the acute angle between the support lines of a_{*i} and a_{*j} is at least $\arccos\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$. We have argued in the previous paragraph that each of a_{*i} and a_{*j} makes an angle at most $O(\varepsilon^2) + \arcsin\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$ with the column space of L_1. Therefore, the acute angle between the support lines of $\bar{\mathsf{a}}_{*i}$ and $\bar{\mathsf{a}}_{*j}$ is at least $\pi/2 - O(\varepsilon^2) - 3\arcsin\left(\frac{1}{9}m^{-1}n^{-1/3}\right)$.

For every $i \in [1, m]$, $\bar{\mathsf{a}}_{*i} = \mathsf{L}_1\mathsf{L}_1^t\mathsf{a}_{*i}$. Thus, $\begin{pmatrix}\bar{\mathsf{a}}_{*1} & \cdots & \bar{\mathsf{a}}_{*m}\end{pmatrix} = \mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$. It follows that:

$$\begin{aligned}\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T &= \left(\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T\right)^t \cdot \left(\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T\right) \\ &= \begin{pmatrix} \sum_{p=1}^n \bar{a}_{p1}^2 & \sum_{p=1}^n \bar{a}_{p1}\bar{a}_{p2} & \cdots & \sum_{p=1}^n \bar{a}_{p1}\bar{a}_{pm} \\ \sum_{p=1}^n \bar{a}_{p2}\bar{a}_{p1} & \sum_{p=1}^n \bar{a}_{p2}^2 & \cdots & \sum_{p=1}^n \bar{a}_{p2}\bar{a}_{pm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{p=1}^n \bar{a}_{pm}\bar{a}_{p1} & \sum_{p=1}^n \bar{a}_{pm}\bar{a}_{p2} & \cdots & \sum_{p=1}^n \bar{a}_{pm}^2 \end{pmatrix}\end{aligned}$$

By the Gershgorin Circle Theorem, the minimum eigenvalue of $\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$ is greater than or equal to $\min_{i\in[1,m]}\left\{\sum_{p=1}^n \bar{a}_{pi}^2 - \sum_{j\in[1,m]\setminus\{i\}}\left|\sum_{p=1}^n \bar{a}_{pi}\bar{a}_{pj}\right|\right\}$.

We have shown earlier that $\|\bar{\mathsf{a}}_{*i}\| \geq \|\mathsf{a}_{*i}\|\cos\left(O(\varepsilon^2) + \arcsin\left(\frac{1}{9}m^{-1}n^{-1/3}\right)\right)$, which is at least $\|\mathsf{a}_{*i}\|\left(1 - \frac{1}{8}m^{-1}n^{-1/3}\right)$ for a small enough ε. Note that $\sum_{p=1}^n \bar{a}_{pi}^2 = \|\bar{\mathsf{a}}_{*i}\|^2$. We have also shown that $\left|\bar{\mathsf{a}}_{*i}^t \cdot \bar{\mathsf{a}}_{*j}\right| \leq \|\bar{\mathsf{a}}_{*i}\|\,\|\bar{\mathsf{a}}_{*j}\|\cos\left(\pi/2 - O(\varepsilon^2) - 3\arcsin\left(\frac{1}{9}m^{-1}n^{-1/3}\right)\right)$, which is at most $\|\bar{\mathsf{a}}_{*i}\|\,\|\bar{\mathsf{a}}_{*j}\| \cdot \frac{1}{2}m^{-1}n^{-1/3} \leq \|\mathsf{a}_{*i}\|\,\|\mathsf{a}_{*j}\| \cdot \frac{1}{2}m^{-1}n^{-1/3}$ for a small enough ε. Note that $\sum_{p=1}^n \bar{a}_{pi}\bar{a}_{pj} = \bar{\mathsf{a}}_{*i}^t \cdot \bar{\mathsf{a}}_{*j}$.

Apply Lemma 5.4(iii) with $c = \frac{1}{10(m+2)}$ to show that the relation $\left|\|\mathsf{a}_{*i}\|^2 - \frac{1}{m+2}n\varrho^2\varepsilon^2\right| \leq \frac{1}{10(m+2)}n^{2/3}\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)$ holds with probability $1 - O(n^{-1/3})$. Thus, $\sum_{p=1}^n \bar{a}_{pi}^2 = \|\bar{\mathsf{a}}_{*i}\|^2 \geq \|\mathsf{a}_{*i}\|^2\left(1 - \frac{1}{4}n^{-1/3}\right) \geq \frac{1}{m+2}n\varrho^2\varepsilon^2 - \frac{1}{10(m+2)}n^{2/3}\varrho^2\varepsilon^2 - \frac{1}{4(m+2)}n^{2/3}\varrho^2\varepsilon^2 - \frac{1}{40(m+2)}n^{1/3}\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4)$. Also, $\sum_{j\in[1,m]\setminus\{i\}}\left|\sum_{p=1}^n \bar{a}_{pi}\bar{a}_{pj}\right| = \sum_{j\in[1,m]\setminus\{i\}}\left|\mathsf{a}_{*i}^t \cdot \mathsf{a}_{*j}\right| \leq \sum_{j\in[1,m]\setminus\{i\}}\|\mathsf{a}_{*i}\| \cdot \|\mathsf{a}_{*j}\| \cdot \frac{1}{2}m^{-1}n^{-1/3} \leq (m-1)\cdot\left(\frac{11}{10(m+2)}n\varrho^2\varepsilon^2 + O(n\varrho^2\varepsilon^4)\right)\cdot\frac{1}{2}m^{-1}n^{-1/3} < \frac{11}{20(m+2)}n^{2/3}\varrho^2\varepsilon^2 + O(n^{2/3}\varrho^2\varepsilon^4)$. Hence, with probability $1 - O(n^{-1/3})$, the minimum eigenvalue of $\mathsf{A}_T^t\mathsf{L}_1\mathsf{L}_1^t\mathsf{A}_T$ is greater than or equal to $\frac{1}{m+2}n\varrho^2\varepsilon^2 - \frac{1}{m+2}\left(\frac{1}{10} + \frac{1}{4} + \frac{1}{40} + \frac{11}{20}\right)n\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4) = \frac{3}{40}n\varrho^2\varepsilon^2 - O(n\varrho^2\varepsilon^4)$ which is $\Omega(n\varrho^2\varepsilon^2)$ for a small enough ε. □

We are now ready to bound the eigenvalues of $\mathsf{H}\mathsf{H}^t$—the main result of this section.

Lemma 7.10 *If ε is sufficiently small, then with probability $1 - O(n^{-1/3})$, the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ are $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ and the $d-m$ smallest eigenvalues of $\mathsf{H}\mathsf{H}^t$ are $O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$.*

Proof. By Lemma 7.4, we can rotate $\mathbb{R}^d$ so that the coordinate axes $x_1, \ldots, x_m$ span $\mathcal{T}$. Partition $\mathsf{H}\mathsf{H}^t$ into blocks using H_T and H_N:

$$\mathsf{H}\mathsf{H}^t = \begin{pmatrix}\mathsf{H}_T\mathsf{H}_T^t & \mathsf{H}_T\mathsf{H}_N^t \\ \mathsf{H}_N\mathsf{H}_T^t & \mathsf{H}_N\mathsf{H}_N^t\end{pmatrix}$$

We apply the generalization of the Gershgorin circle theorem [13]. Let G_1 be the set of all real numbers μ such that $(\|(\mathsf{H}_T\mathsf{H}_T^t - \mu\mathsf{I}_m)^{-1}\|)^{-1} \leq \|\mathsf{H}_T\mathsf{H}_N^t\|$, and let G_2 be the set of all real numbers μ such that $(\|(\mathsf{H}_N\mathsf{H}_N^t - \mu\mathsf{I}_{d-m})^{-1}\|)^{-1} \leq \|\mathsf{H}_N\mathsf{H}_T^t\|$.

By (6.2) and (6.3), the numbers in G_1 are at least the minimum eigenvalue value of $\mathsf{H}_T\mathsf{H}_T^t$ minus $\|\mathsf{H}_T\mathsf{H}_N^t\|$ and at most the maximum eigenvalue value of $\mathsf{H}_T\mathsf{H}_T^t$ plus $\|\mathsf{H}_T\mathsf{H}_N^t\|$. If ε is small enough, the numbers in G_1 are $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ with probability $1 - O(n^{-1/3})$ because Lemmas 7.5 and 7.9 imply that, with probability $1 - O(n^{-1/3})$, the eigenvalues of $\mathsf{H}_T\mathsf{H}_T^t$ are $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$, which dominates $\|\mathsf{H}_T\mathsf{H}_N^t\| \leq \|\mathsf{H}_T\|\,\|\mathsf{H}_N\| = O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$.

By (6.3), the numbers in G_2 are at most the maximum eigenvalue value of $\mathsf{H}_N\mathsf{H}_N^t$ plus $\|\mathsf{H}_N\mathsf{H}_T^t\|$. If ε is small enough, the numbers in G_2 are $O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$ with probability $1 - O(n^{-1/3})$ because Lemmas 7.5 and 7.9 imply that, with probability $1 - O(n^{-1/3})$, $\|\mathsf{H}_N\mathsf{H}_N^t\| = \|\mathsf{H}_N\|^2 = O(n\varrho^2\varepsilon^6/\lambda_{m_0+1}^2)$ and $\|\mathsf{H}_N\mathsf{H}_T^t\| \leq \|\mathsf{H}_N\|\,\|\mathsf{H}_T\| = O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$.

As a result, if ε is small enough, then with probability $1 - O(n^{-1/3})$, all numbers in G_2 are smaller than those in G_1. By Lemma 6.1, the disjointness of G_1 and G_2 implies that G_1 and G_2 contain exactly m and $d-m$ eigenvalues of $\mathsf{H}\mathsf{H}^t$, respectively. Hence, the lemma follows. □

8 Proof of Lemma 2.1

Let S be a set of sample points in $\mathcal{M}$. Without loss of generality, we assume that the origin is a sample point in S. We require S to satisfy the condition that the distance between the origin and its $(n+1)$-th nearest sample point is at most $\varrho\varepsilon$ for some sufficiently small $\varepsilon \in (0, \varepsilon_0]$ where ε_0 is the constant in Lemma 5.1. We rotate $\mathbb{R}^d$ so that the coordinate axes $x_1, \ldots, x_m$ span the tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin. By Lemma 6.2, this does not change the singular values of B.

Let E denote the event that $\lambda_{m_0+1} = 0$ and there exists $q \in [1, n]$ such that $a_{q,k} \neq 0$ for some $k \in [m+1, d]$. Our goal is to prove that $\Pr(E) = O(n^{-1/3})$. Then, it holds with probability $1 - O(n^{-1/3})$ that if $\lambda_{m_0+1} = 0$, we have $a_{p,i} = 0$ for all $p \in [1, n]$ and for all $i \in [m+1, d]$, and so Lemma 2.1 is true.

By Lemma 6.3, it holds with probability $1 - O(n^{-1/3})$ that the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ are at least $c_0 n\varrho^4\varepsilon^4$ and at most $c_1 n\varrho^4\varepsilon^4$ for some constants c_0 and c_1. The proof of Lemma 6.3 reveals that c_0 and c_1 are polynomials in m. Let F denote the event that the eigenvalues of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ lie between $c_0 n\varrho^4\varepsilon^4$ and $c_1 n\varrho^4\varepsilon^4$. The probability $\Pr(E)$ can be split up into the following sum:

$$\Pr(E|F) \cdot \Pr(F) + \Pr(E|\neg F) \cdot \Pr(\neg F).$$

The second term is $O(n^{-1/3})$ because $\Pr(\neg F) = O(n^{-1/3})$ by Lemma 6.3. We show that $\Pr(E|F) = 0$ below. From now on, we assume that the condition F holds.

When the event E happens, $\lambda_{m_0+1} = 0$ and there exists index $q \in [1, n]$ such that $a_{q,k} \neq 0$ for some $k \in [m+1, d]$. By swapping coordinate axes if necessary, we can further assume that $a_{q,m+1} \neq 0$. Let $\mathsf{C}_{q,TT}$ denote the matrix obtained by deleting the row in B_{TT} for a_q. Since $n \geq m_0 + 1$, there are at least m_0 rows in $\mathsf{C}_{q,TT}$. We claim that $\mathsf{C}_{q,TT}$ has rank m_0, provided that $n > c_0 m_0$. Otherwise, the smallest eigenvalue of $\mathsf{C}_{q,TT}^t\mathsf{C}_{q,TT}$ is zero, which implies that there is a unit direction $\mathsf{u} \in \mathbb{R}^{m_0}$ that is orthogonal to every row vector of $\mathsf{C}_{q,TT}$. Since every coordinate of a_q has magnitude at most $\varrho\varepsilon$, the projection of the row for a_q in B_{TT} onto u has a squared length at most $m_0\varrho^4\varepsilon^4$. Therefore, if we project the row vectors of B_{TT} onto u, the sum of the squared lengths of the projections is at most $m_0\varrho^4\varepsilon^4$, which implies that the smallest eigenvalue of $\mathsf{B}_{TT}^t\mathsf{B}_{TT}$ is at most $m_0\varrho^4\varepsilon^4 < c_0 n\varrho^4\varepsilon^4$, a contradiction to the condition F.

Let C_q denote the matrix obtained from B by removing the row for a_q. Since $\mathsf{C}_{q,TT}$ is a

submatrix of C_q which is in turn a submatrix of B, we have

$$m_0 = \operatorname{rank}(\mathsf{C}_{q,TT}) \;\le\; \operatorname{rank}(\mathsf{C}_q) \;\le\; \operatorname{rank}(\mathsf{B}).$$

Since $\lambda_{m_0+1} = 0$, the rank of B is at most m_0, and so $\operatorname{rank}(\mathsf{B}) \ge \operatorname{rank}(\mathsf{C}_{q,TT}) = m_0$ implies that $\operatorname{rank}(\mathsf{B}) = m_0$. This allows us to conclude that C_q has rank m_0.

Since $\operatorname{rank}(\mathsf{B}) = m_0$, the column space of B has rank m_0, which implies that every column in B is a linear combination of the columns of B_{TT}. Therefore, there exist coefficients g_{ij}'s such that

$$\forall\, p \in [1,n], \quad \frac{1}{\sqrt{2}} a_{p,m+1}^2 = \sum_{i \le j \in [1,m]} g_{ij} a_{pi} a_{pj}. \tag{8.1}$$

Since $n \ge m_0 + 1$, there are m_0 coefficients g_{ij}'s, and C_q has rank m_0, the coefficients g_{ij}'s are completely determined by the following smaller system:

$$\forall\, p \in [1,n] \setminus \{q\}, \quad \frac{1}{\sqrt{2}} a_{p,m+1}^2 = \sum_{i \le j \in [1,m]} g_{ij} a_{pi} a_{pj}. \tag{8.2}$$

In other words, when E happens under condition F, the coefficients g_{ij}'s are determined irrespective of the coordinates of a_q and yet a_q must satisfy (8.1). We can interpret (8.1) as saying that the sample points a_p, $p \in [1,n]$, lie in the following hypersurface:

$$H(\mathsf{x}) = \frac{1}{\sqrt{2}} x_{m+1}^2 - \sum_{i \le j \in [1,m]} g_{ij} x_i x_j = 0. \tag{8.3}$$

Both the hypersurface $H(\mathsf{x}) = 0$ and $\mathcal{M}$ contain the origin. Let L denote the linear subspace spanned by axes $x_1, \ldots, x_{m+1}$. The intersection of L and the hypersurface $H(\mathsf{x}) = 0$ is a conic surface, and this can be seen as follows. Recall that the axes $x_1, \ldots, x_m$ span the true tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin. Take any unit vector $\mathsf{u} \in \mathcal{T}$ that makes an angle θ_i with the axis x_i for $i \in [1,m]$. For any $c \in \mathbb{R}$, the x_{m+1} coordinate of the point in the hypersurface $H(\mathsf{x}) = 0$ that projects to the point $c\mathsf{u} \in \mathcal{T}$ can be written as

$$\begin{aligned} x_{m+1}^2 &= \left(\sum_{i \le j \in [1,m]} \sqrt{2} g_{ij} \cos\theta_i \cos\theta_j \right) c^2 \\ \Rightarrow \quad x_{m+1} &= \pm \left(\sum_{i \le j \in [1,m]} \sqrt{2} g_{ij} \cos\theta_i \cos\theta_j \right)^{1/2} c. \end{aligned} \tag{8.4}$$

Therefore, the cross-section of $H(\mathsf{x}) = 0$ in the plane spanned by u and the axis x_{m+1} consists of two lines through the origin with slopes of the same magnitude but opposite signs.

Recall the coordinate function $f_{m+1} : \mathbb{R}^m \to \mathbb{R}$ for $\mathcal{M}$ such that, given a point $(x_1 \;\ldots\; x_d)^t \in \mathcal{M}$, we have $x_{m+1} = f_{m+1}((x_1 \;\ldots\; x_m)^t)$ in a local neighborhood of the origin. We can choose the constant ε_0 in Lemma 5.1 so that $\varrho\varepsilon_0$ is at most the radius of this local neighborhood. Also, recall that the Taylor expansion of f_{m+1} does not have a constant or a linear term, that is,

$$\forall\, \mathsf{u} \in \mathbb{R}^m,\ \forall\, c \in \mathbb{R}, \quad f_{m+1}(c\mathsf{u}) = \frac{c^2}{2} \mathrm{D}^2 f_\ell|_{\mathsf{0}}(\mathsf{u},\mathsf{u}) + \cdots. \tag{8.5}$$

We claim that the set of points $\mathcal{K} = \{\mathsf{x} \in \mathcal{M} : x_{m+1} \ne 0 \wedge H(\mathsf{x}) = 0\}$ has measure zero. Suppose not. Then, there exists a unit vector $\mathsf{v} = (v_1 \;\ldots\; v_m)^t \in \mathcal{T}$ such that the intersection between $\mathcal{K}$ and the plane spanned by v and the coordinate axis x_{m+1} consists of some curve

segment(s) of positive length. By the definition of $\mathcal{K}$, f_{m+1} is not identically zero along the directions v and $-\mathsf{v}$. Then, since (8.4) is a linear equation, it can only agree with (8.5) at isolated values of c after we substitute u by v. This is a contradiction to the existence of curve segment(s) of positive length in the intersection of $\mathcal{K}$ and the plane spanned by v and the axis x_{m+1}. We conclude that $\mathcal{K}$ has measure zero. It follows that the probability of drawing the sample point a_q from $\mathcal{K}$ is zero, which implies that $\Pr(E|F) = 0$.

In summary, $\Pr(E|F) \cdot \Pr(F) + \Pr(E|\neg F) \cdot \Pr(\neg F) = O(n^{-1/3})$, completing the proof of Lemma 2.1.

9 Proof of Theorem 1.1

By Lemma 7.4, we can assume that $\mathbb{R}^d$ has been rotated so that the coordinate axes $x_1, \ldots, x_m$ span the tangent space $\mathcal{T}$ of $\mathcal{M}$ at the origin. Then, we can partition $\mathsf{H} = \begin{pmatrix} \mathsf{H}_T \\ \mathsf{H}_N \end{pmatrix}$.

We call TANGENT(A). If $\lambda_{m_0+1} = 0$, then Lemma 2.1 implies that, with probability $1 - O(n^{-1/3})$, the estimated tangent space is equal to $\mathcal{T}$, and therefore, there is no angular error.

Assume that $\lambda_{m_0+1} > 0$ for the rest of the proof. Then, the estimated tangent space is spanned by the unit eigenvectors corresponding to the m largest eigenvalues of $\mathsf{A}^t\mathsf{L}\Sigma\mathsf{L}^t\mathsf{A} = \mathsf{H}\mathsf{H}^t$. Let e be one of these m unit eigenvectors. Divide e into two parts $\begin{pmatrix} \mathsf{v}^t & \mathsf{w}^t \end{pmatrix}^t$, where v consists of the first m coordinates and w consists of the last $d-m$ coordinates. Let σ be the eigenvalue of $\mathsf{H}\mathsf{H}^t$ corresponding to e. Then,

$$\mathsf{H}\mathsf{H}^t\mathsf{e} \;=\; \begin{pmatrix} \mathsf{H}_T\mathsf{H}_T^t & \mathsf{H}_T\mathsf{H}_N^t \\ \mathsf{H}_N\mathsf{H}_T^t & \mathsf{H}_N\mathsf{H}_N^t \end{pmatrix} \begin{pmatrix} \mathsf{v} \\ \mathsf{w} \end{pmatrix} \;=\; \sigma \begin{pmatrix} \mathsf{v} \\ \mathsf{w} \end{pmatrix} \tag{9.1}$$

We want to bound $\arctan(\|\mathsf{w}\|/\|\mathsf{v}\|)$, the angle between e and $\mathcal{T}$.

We first show that $\mathsf{v} \neq \mathsf{0}_{m,1}$ with probability $1 - O(n^{-1/3})$. Suppose that $\mathsf{v} = \mathsf{0}_{m,1}$. By (9.1), we get $\mathsf{H}_N\mathsf{H}_N^t\mathsf{w} = \sigma\mathsf{w}$, meaning that σ is also an eigenvalue of $\mathsf{H}_N\mathsf{H}_N^t$ in addition to being one of the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$. This occurs with probability $O(n^{-1/3})$ because, with probability $1 - O(n^{-1/3})$, the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ are $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$ by Lemma 7.10, but the eigenvalues of $\mathsf{H}_N\mathsf{H}_N^t$ are $O(n\varrho^2\varepsilon^6/\lambda_{m_0+1}^2)$ by Lemma 7.5.

Assume that v is non-zero. By (9.1), $\mathsf{w} = \left(\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t\right)^{-1}\mathsf{H}_N\mathsf{H}_T^t\mathsf{v}$. By Lemmas 7.5 and 7.9, it holds with probability $1 - O(n^{-1/3})$ that $\|\mathsf{H}_N\mathsf{H}_T^t\| \leq \|\mathsf{H}_N\|\,\|\mathsf{H}_T\| = O(n\varrho^2\varepsilon^4/\lambda_{m_0+1}^2)$. By Lemmas 7.5 and 7.10, it holds with probability $1 - O(n^{-1/3})$ that the eigenvalues of $\mathsf{H}_N\mathsf{H}_N^t$ are $O(n\varrho^2\varepsilon^6/\lambda_{m_0+1}^2)$ and $\sigma = \Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$. The smallest eigenvalue of $\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t$ is thus $\Theta(n\varrho^2\varepsilon^2/\lambda_{m_0+1}^2)$, implying that $\|(\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t)^{-1}\| = O(\lambda_{m_0+1}^2/(n\varrho^2\varepsilon^2))$. Hence, $\|\mathsf{w}\| \leq \left\|\left(\sigma\mathsf{I}_{d-m} - \mathsf{H}_N\mathsf{H}_N^t\right)^{-1}\right\| \cdot \left\|\mathsf{H}_N\mathsf{H}_T^t\right\| \cdot \|\mathsf{v}\| = O(\varepsilon^2) \cdot \|\mathsf{v}\|$. The angle between e and $\mathcal{T}$ is $\arctan(\|\mathsf{w}\|/\|\mathsf{v}\|) = O(\varepsilon^2)$.

We conclude that, with probability at least $1 - m \cdot O(n^{-1/3}) = 1 - O(n^{-1/3})$, all eigenvectors corresponding to the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ make an $O(\varepsilon^2)$ angle with $\mathcal{T}$. Let $\begin{pmatrix} \mathsf{U} & \mathsf{V} \end{pmatrix}$ be a $d \times d$ orthonormal matrix such that the columns of V form an orthonormal basis of $\mathcal{T}$. This implies that, with probability $1 - O(n^{-1/3})$, all eigenvectors corresponding to the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$ make an $O(\varepsilon^2)$ angle with the column space of V and hence an $\pi/2 - O(\varepsilon^2)$ angle with the column space of U. Then, Lemma 7.2 implies that $\sqrt{m} \cdot O(\varepsilon^2) = O(\varepsilon^2)$ is an upper bound on the angle between $\mathcal{T}$ and the space spanned by the eigenvectors corresponding to the m largest eigenvalues of $\mathsf{H}\mathsf{H}^t$.

10 Conclusion

We present an algorithm for estimating tangent spaces from a given set of sample points in an unknown manifold. The algorithm works locally and uses the n sample points nearest to p. The distance from p to the $(n+1)$-th nearest sample point can be expressed as $\varrho\varepsilon$, where $\varepsilon \in (0,1)$ and ϱ is the local feature size at p. (The algorithm does not need to know ϱ though.) When we fix n, the value ε decreases as sampling density increases. Assuming that the sample points are distributed according to a Poisson process with an unknown parameter, our algorithm guarantees an $O(\varepsilon^2)$ bound on the angular error with high probability. The quadratic angular error convergence has been confirmed in our experiments.

The angular error bounds in [2, 10, 11, 15, 26] hold for all sample points. Our $O(\varepsilon^2)$ angular error bound applies to the center of a local neighborhood in which the sample points are used to estimate the tangent space at the center. One should be able to extend our result so that it applies simultaneously to centers of disjoint neighborhoods by restricting the range of ε further and estimating the failure probability using the union bound. Further work is needed to see if our angular error bound can be guaranteed at all sample points.

Our algorithm assumes that the manifold dimension m is known to us. The results in Section 6 show that the largest $\binom{m+1}{2}$ eigenvalues of $\mathsf{B}^t\mathsf{B}$ are $\Theta(n\varrho^4\varepsilon^4)$ and the other eigenvalues are $O(n\varrho^4\varepsilon^6)$. Therefore, one should be able to determine the manifold dimension automatically by detecting this $\Omega(\varepsilon^2)$ factor gap in the eigenvalues. This approach of finding gaps in the spectrum of eigenvalues has been used in several previous work on detecting manifold dimension.

One may wonder what happens if we omit the conversion of B to $\widehat{\mathsf{B}}$ in practice. This is equivalent to defining Σ as $\mathrm{diag}_n\left(\frac{1}{\lambda_1^2}, \frac{1}{\lambda_2^2}, \ldots, \frac{1}{\lambda_n^2}\right)$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the singular values of B. We experimented with this alternative method. While this method performs reasonably in the noiseless case, it fails badly in the noisy cases. So the conversion of B to $\widehat{\mathsf{B}}$ makes a real difference. There is also the possibility that λ_{m_0+1} is positive but close to zero. In that case, although the theoretical analysis holds, there will be numerical issues in forming $\Sigma = \mathrm{diag}_n\left(\frac{1}{\lambda_1^2}, \cdots, \frac{1}{\lambda_{m_0}^2}, \frac{1}{\lambda_{m_0+1}^2}, \cdots, \frac{1}{\lambda_{m_0+1}^2}\right)$. Therefore, some thresholding of the singular values may be necessary for numerical stability.

Acknowledgment

We thank the anonymous referees for their helpful comments and useful advice.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in Neural Information Processing Systems*, 14 (2002).

[2] M. Belkin, J. Sun, and Y. Wang. Constructing Laplace operator from point clouds in R^d. *Proceedings of the 20th Annual ACM–SIAM Symposium on Discrete Algorithms*, 2009, 1021–1040.

[3] J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. *Proceedings of the 26th Annual Symposium on Computational Geometry*, 2010, 324–333.

[4] J.-D. Boissonnat, L.J. Guibas and S.Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Discrete and Computational Geometry*, 42 (2009), 37–70.

[5] O.J. Boxma and U. Yechiali. Poisson Processes, Ordinary and Compound. *Encyclopedia of Statistics in Quality and Reliability*, F. Ruggeri, R.S. Kenett and F.W. Faltin (eds.), Wiley, New York, 2007.

[6] K.M. Carter, R. Raich, and A.O. Hero. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58 (2010), 650–663.

[7] T.F. Chan. An improved algorithm for computing the singular value decomposition. *ACM Transactions on Mathematical Software*, 8 (1982), 72–83.

[8] S.-W. Cheng and M-K. Chiu. Dimension detection via slivers. *Proceedings of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009, 1001-1010.

[9] S.-W. Cheng, T.K. Dey and E.A. Ramos. Manifold reconstruction from point samples. *Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms*, 2005, 1018–1027.

[10] S.-W. Cheng, Y. Wang, and Z. Wu. Provable Dimension Detection using Principal Component Analysis. *International Journal of Computational Geometry and Applications*, 18 (2008), 414–440.

[11] T. K. Dey, J. Giesen, S. Goswami and W. Zhao. Shape dimension and approximation from samples. *Discrete and Computational Geometry*, 29 (2003), 419–434.

[12] S.C. Eisenstat and I.C.F. Ipsen. Relative perturbation bounds for eigenspaces and singular vector subspaces. *Proceedings of the 5th SIAM Conference on Applied Linear Algebra*, 1994, 62–66.

[13] D.G. Feingold, R.S. Varga, Block diagonally dominant matrices and generalizations of the Gershgorin circle theorem, *Pacific Journal of Mathematics*, 12 (1962), 1241–1250.

[14] M. Gashler and T. Martinez. Tangent space guided intelligent neighbor finding. *Proceedings of International Joint Conference on Neural Networks*, 2011, 2617–2624.

[15] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds with high codimension. *Discrete and Computational Geometry*, 32 (2004), 245–267.

[16] G.H. Golub and C. Reinsch. Singular value decomposition and least square solutions. In *Handbook for Automatic Computation*, II, *Linear Algebra*, J.H. Wilkinson and C. Reinsch (eds.), Springer-Verlag, New York, 1971.

[17] G.H. Golub and C.F. van Loan. *Matrix Computations*, Johns Hopkins University Press, 1996.

[18] D. Gong, X. Zhao, and G. Medioni. Robust multiple manifolds structure learning. *Proceedings of the International Conference on Machine Learning*, 2012,

[19] I.S. Gradshtein and I.M. Ryzhik. *Table of Integrals, Series, and Products*, Academic Press, 1994.

[20] P.S. Heckbert and M. Garland. Optimal triangulation and quadric-based surface simplification. *Computational Geometry: Theory and Applications*, 14 (1999), 49–65.

[21] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in Euclidean space. *Proceedings of the 22nd International Conference on Machine Learning*, 2005, 289–296.

[22] B. Kégl. Intrinsic dimension estimation using packing numbers. *Advances in Neural Information Processing Systems*, 14 (2003), 681–688.

[23] S. Lang. *Calculus of Several Variables*, Springer-Verlag, 1987.

[24] F. Le Gall. Faster Algorithms for Rectangular Matrix Multiplication. *Proceedings of the 53rd IEEE Annual Symposium on Foundations of Computer Science*, 2012, 514–523.

[25] E. Levina and P.J. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17 (2005), 777—784.

[26] A.V. Little, M. Maggioni, and L. Rosasco. Multiscale geometric methods for data sets I: multiscale SVD, noise and curvature. *Computer Science and Artificial Intelligence Laboratory Technical Report*, MIT-CSAIL-TR-2012-029, CBCL-310, September 8, 2012.

[27] A. massoud Farahmand, C. Szepesvári and J.-Y. Audibert. Manifold-adaptive dimension estimation. *Proceedings of the 24th International Conference on Machine Learning*, 2007, 265–272.

[28] F. Morgan. *Riemannian Geometry: a beginner's guide*, A. K. Peters, 1998.

[29] J.C. Nascimento and J.G. Silva. Manifold learning for object tracking with multiple motion dynamics. *Proceedings of ECCV 2010*, Part III, LNCS 6313, 172–185.

[30] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (2000), 2323–2326.

[31] B. Schölkopf, A. Smola, and K.-R¿ Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 (1998), 1299–1319.

[32] G. Taubin. Estimation of planar curves, surfaces and nonplanar space curves defined by implicit equations, with applications to edge and range image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 (1991), 1115–1138.

[33] J.B. Tenenbaum, V. de Silva and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (2000), 2319–2323.