Journal of Machine Learning Research 22 (2021) 1-92 Submitted 10/20; Revised 4/21; Published 6/21

Kernel Smoothing, Mean Shift, and Their Learning Theory with Directional Data

Yikun Zhang YIKUN@UW.EDU
Yen-Chi Chen YENCHIC@UW.EDU
Department of Statistics
University of Washington
Seattle, WA 98195, USA

Editor: Sayan Mukherjee

Abstract

Directional data consist of observations distributed on a (hyper)sphere, and appear in many applied fields, such as astronomy, ecology, and environmental science. This paper studies both statistical and computational problems of kernel smoothing for directional data. We generalize the classical mean shift algorithm to directional data, which allows us to identify local modes of the directional kernel density estimator (KDE). The statistical convergence rates of the directional KDE and its derivatives are derived, and the problem of mode estimation is examined. We also prove the ascending property of the directional mean shift algorithm and investigate a general problem of gradient ascent on the unit hypersphere. To demonstrate the applicability of the algorithm, we evaluate it as a mode clustering method on both simulated and real-world data sets.

Keywords: Directional data, mean shift algorithm, kernel smoothing, mode clustering, optimization on a manifold

1. Introduction

A directional data set (or simply directional data) is the collection of observations on a (hyper)sphere. It occurs in many scientific problems when measurements are taken on the surface of a spherical object, such as Earth or other planets. For instance, the locations of earthquakes are often represented by their longitudes and latitudes (Taylor and Yin, 2009; Craig et al., 2011); thus, the locations can be viewed as random variables on a two-dimensional (2D) sphere. In astronomical surveys, the locations of galaxies are usually recorded by their angular positions (right ascensions and declinations) in the sky, leading to observations on a 2D sphere (York et al., 2000; Skrutskie et al., 2006; Abbott et al., 2016). In planetary science, observations often comprise locations on a planet, such as Mars, and can also be considered as random variables on a 2D sphere (Cabrol and Grin, 2010; Barlow, 2015; García-Portugués et al., 2020).

These observations on a sphere can be regarded as independently and identically distributed random variables from a density function supported on the sphere (called a directional density function). The local modes of a density function are often of research interest because they signal high density areas (Scott, 2012) and can be used to cluster data (Sasaki et al., 2018; Chacón, 2020). However, identifying the local modes of a directional density

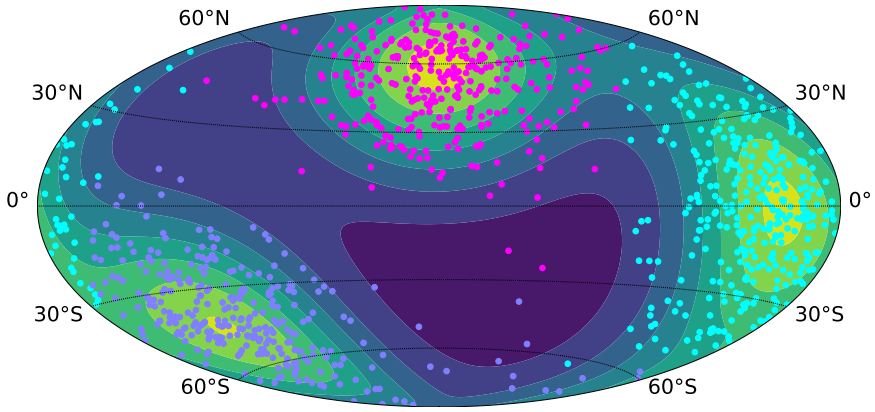


Figure 1: Clustering of directional data using the proposed directional mean shift algorithm (Algorithm 1). Additional details of the simulated data can be found in Section 6.1.2.

function is a nontrivial task that involves both statistical and computational challenges. From a statistical perspective, it is necessary to obtain an accurate estimator of the underlying directional density (as well as its derivatives). From a computational perspective, it is needful to design an algorithm to efficiently compute the local modes of the density estimator.

To address the aforementioned challenges, we consider the idea of kernel smoothing because the kernel density estimator (KDE; Rosenblatt 1956; Parzen 1962) in the Euclidean data setting is highly successful. Its statistical properties have been well-studied (Wasserman, 2006; Scott, 2015; Chen, 2017), and the local modes of a Euclidean KDE are often good estimators of the local modes of the underlying density function (Parzen, 1962; Romano, 1988; Vieu, 1996; Chen et al., 2016). Moreover, in Euclidean KDEs, there is an elegant algorithm known as the *mean shift* algorithm (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002; Carreira-Perpiñán, 2015) that allows us to numerically obtain the local modes at a low cost.

Although kernel smoothing has been applied to directional data since the seminal work of Hall et al. (1987) and other studies have been conducted on analyzing its performance as a density estimator (Bai et al., 1988; Zhao and Wu, 2001; García-Portugués, 2013; Ley and Verdebout, 2018), little is known about the behavior of the derivatives of a directional KDE. To the best of our knowledge, Klemelä (2000) was the only work to examine the derivatives of a particular type of directional KDE; however, their estimators are rarely used in practice. Thus, the statistical properties of the gradient system induced by a general directional KDE and the resulting local modes are still open problems.

Computationally, the standard mean shift algorithm was first generalized to directional data setting by Oba et al. (2005). Using the directional mean shift algorithm, we are able

to determine the local modes of the directional KDE and perform mode clustering (mean shift clustering) of spherical data. Figure 1 presents an example of mode clustering with our proposed algorithm. However, the algorithmic rate of convergence of the mean shift algorithm with directional data remains unclear. We address this problem by viewing the directional mean shift algorithm as a special case of gradient ascent methods on the q-dimensional unit sphere $\Omega_q = \left\{\boldsymbol{x} \in \mathbb{R}^{q+1} : ||\boldsymbol{x}||_2^2 = x_1^2 + \cdots + x_{q+1}^2 = 1\right\}$ and develop some linear convergence results for the general gradient ascent method on Ω_q.

Notation. Bold-faced variables (e.g., $\boldsymbol{x}, \boldsymbol{\mu}$) represent vectors, while capitalized (bold-faced) variables (e.g., $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$) denote random variables (or random vectors). The set of real numbers is denoted by $\mathbb{R}$, while the unit q-dimensional sphere embedded in $\mathbb{R}^{q+1}$ is denoted by Ω_q. The norm $||\cdot||_2$ is the usual Euclidean norm (or so-called L_2-norm) in $\mathbb{R}^d$ for some positive integer d. The directional density is denoted by f unless otherwise specified, and the probability of a set of events is denoted by $\mathbb{P}$. If a random vector $\boldsymbol{X}$ is distributed as $f(\cdot)$, the expectations of functions of $\boldsymbol{X}$ are denoted by $\mathbb{E}_f$ or $\mathbb{E}$ when the underlying distribution function is clear. We use the big-O notation $h(x) = O(g(x))$ if the absolute value of $h(x)$ is upper bounded by a positive constant multiple of $g(x)$ for all sufficiently large x. In contrast, $h(x) = o(g(x))$ when $\lim_{x\to\infty} \frac{|h(x)|}{g(x)} = 0$. For random vectors, the notation $o_P(1)$ is short for a sequence of random vectors that converges to zero in probability. The expression $O_P(1)$ denotes a sequence that is bounded in probability. Additional details of stochastic o and O symbols can be found in Section 2.2 of van der Vaart (1998). The notation $a_n \asymp b_n$ indicates that $\frac{a_n}{b_n}$ has lower and upper bounds away from zero and infinity, respectively.

Main results.

1. We revisit the mean shift algorithm with directional data (Algorithm 1) and provide some new insights on its iterative formula, which can be expressed in terms of the total gradient of the directional KDE (Sections 3 and 4.1).

2. From the perspective of statistical learning theory, we establish uniform convergence rates of the gradient and Hessian of the directional KDE (Theorem 2 and 4).

3. Moreover, we derive the asymptotic properties of estimated local modes around the true (population) local modes (Theorem 6).

4. With regard to computational learning theory, we prove the ascending and converging properties of the directional mean shift algorithm (Theorems 8 and 11).

5. In addition, we prove that the directional mean shift algorithm converges linearly to an estimated local mode under suitable initialization (Theorem 12).

6. We demonstrate the applicability of the directional mean shift algorithm by using it as a clustering method on both simulated and real-world data sets (Section 6).

Related work. The directional KDE has a long history in statistics since the work of Hall et al. (1987). Its statistical convergence rates and asymptotic distributions have been studied by Bai et al. (1988); Zhao and Wu (2001). In addition, Hall et al. (1987); Bai et al. (1988); García-Portugués (2013); García-Portugués et al. (2013) considered the problem of

selecting the smoothing bandwidth of directional KDEs. A study by Klemelä (2000) was the first to estimate the derivatives of a directional density. More generally, Hendriks (1990); Pelletier (2005); Berry and Sauer (2017) considered the nonparametric density estimation on (Riemannian) manifolds (with boundary). The uniform convergence rate and asymptotic results of the KDE on Riemannian manifolds have also been investigated in Henry and Rodriguez (2009); Jiang (2017); Kim et al. (2019). As the unit hypersphere Ω_q is a q-dimensional manifold with constant curvature and positive reach (Federer, 1959), their analyses and results are applicable to the directional KDE.

The standard mean shift algorithm with Euclidean data is a popular approach to various tasks such as clustering (Fukunaga and Hostetler, 1975), image segmentation (Comaniciu and Meer, 2002), and object tracking (Comaniciu et al., 2003); see a comprehensive review in Carreira-Perpiñán (2015). Its convergence properties have been well-studied in Cheng (1995); Li et al. (2007); Aliyari Ghassabeh (2013, 2015); Arias-Castro et al. (2016); Wang et al. (2016). The algorithmic convergence rates of mean shift algorithms with Gaussian and Epanechnikov kernels are generally linear, except for some extreme values of the bandwidth (Carreira-Perpiñán, 2007; Huang et al., 2018). It can be improved to be superlinear by dynamically updating the data set for estimating the density (Zhang et al., 2006). There are other methods to accelerate the mean shift algorithm by combining stochastic optimization with blurring or random sampling (Carreira-Perpiñán, 2006, 2008; Yuan and Li, 2009; Hyrien and Baran, 2016). The mean shift algorithm with directional data was studied by Oba et al. (2005); Kafai et al. (2010); Kobayashi and Otsu (2010); Shou-Jen Chang-Chien et al. (2010); Chang-Chien et al. (2012); Yang et al. (2014) in the last two decades. More generally, Tuzel et al. (2005); Subbarao and Meer (2006, 2009); Cetingul and Vidal (2009); Caseiro et al. (2012); Ashizawa et al. (2017) proposed their mean shift algorithms on manifolds using logarithmic and exponential maps, heat kernel, or direct log-density estimation via least squares. These mean shift algorithms on general manifolds are applicable to directional data, though they are more complicated than our interested method.

Outline. The remainder of the paper is organized as follows. Section 2 reviews some background knowledge on directional KDEs and differential geometry, while Section 3 provides a detailed derivation of the mean shift algorithm with directional data. Section 4 focuses on the statistical learning theory of the directional KDE; we formulate the gradient and Hessian estimators of directional KDEs and establish their pointwise and uniform consistency results as well as a mode consistency theory. Section 5 considers the computational learning theory of the directional mean shift algorithm; we study the ascending and converging properties of the algorithm. Simulation studies and applications to real-world data sets are unfolded in Section 6. Proofs of theorems and technical lemmas are deferred to Appendix D. All the code for our experiments is available at `https://github.com/zhangyk8/DirMS`.

2. Preliminaries

This section is devoted to a brief review of the directional KDE and some technical concepts of differential geometry on Ω_q.

2.1 Kernel Density Estimation with Directional Data

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q \subset \mathbb{R}^{q+1}$ be a random sample generated from the underlying directional density function f on Ω_q with $\int_{\Omega_q} f(\boldsymbol{x})\, \omega_q(d\boldsymbol{x}) = 1$, where ω_q is the Lebesgue measure on Ω_q. A well-known fact about the surface area of Ω_q is that

$$\bar{\omega}_q \equiv \omega_q(\Omega_q) = \frac{2\pi^{\frac{q+1}{2}}}{\Gamma(\frac{q+1}{2})} \quad \text{for any integer } q \geq 1, \tag{1}$$

where Γ is the Gamma function defined as $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ with the real part of the complex integration variable z (if applicable) being positive. The directional KDE at point $\boldsymbol{x} \in \Omega_q$ is often written as (Hall et al., 1987; Bai et al., 1988; García-Portugués, 2013):

$$\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1 - \boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right), \tag{2}$$

where L is a directional kernel (a rapidly decaying function with nonnegative values and defined on $(-\delta_L, \infty) \subset \mathbb{R}$ for some constant $\delta_L > 0$)[1], $h > 0$ is the bandwidth parameter, and $c_{h,q}(L)$ is a normalizing constant satisfying

$$c_{h,q}(L)^{-1} = \int_{\Omega_q} L\left(\frac{1 - \boldsymbol{x}^T\boldsymbol{y}}{h^2}\right) \omega_q(d\boldsymbol{y}) = h^q \lambda_{h,q}(L) \asymp h^q \lambda_q(L) \tag{3}$$

with $\lambda_{h,q}(L) = \bar{\omega}_{q-1} \int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1}(2 - rh^2)^{\frac{q}{2}-1} dr$ and $\lambda_q(L) = 2^{\frac{q}{2}-1}\bar{\omega}_{q-1} \int_0^\infty L(r) r^{\frac{q}{2}-1} dr$; see (a) of Lemma 21 in Appendix D.2 for details.

As in Euclidean kernel smoothing, bandwidth selection is a critical component in determining the performance of directional KDEs. There is extensive literature (Hall et al., 1987; Bai et al., 1988; Taylor, 2008; Marzio et al., 2011; Oliveira et al., 2012; García-Portugués, 2013; Saavedra-Nieves and Crujeiras, 2020) that investigates various reliable bandwidth selection mechanisms. On the contrary, kernel selection is less crucial, and a popular candidate is the so-called von Mises kernel $L(r) = e^{-r}$. Its name originates from the famous q-von Mises-Fisher (vMF) distribution on Ω_q, which is denoted by $\text{vMF}(\boldsymbol{\mu}, \nu)$ and has the density

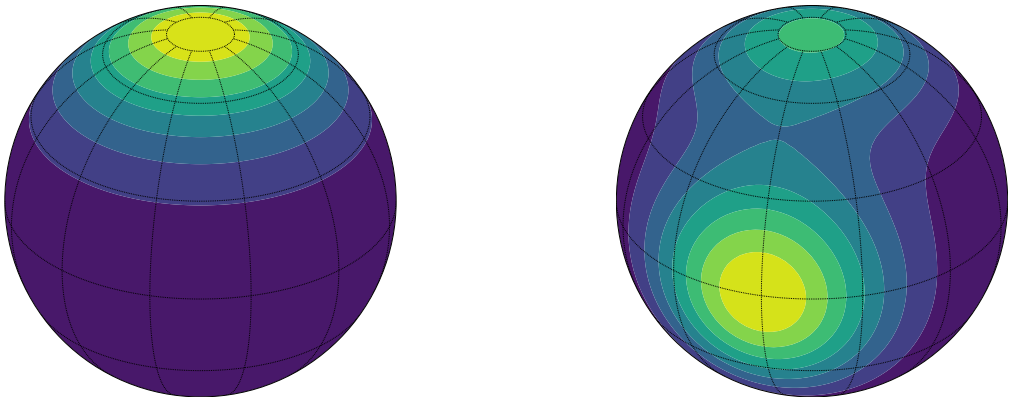
$$f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \nu) = C_q(\nu) \cdot \exp(\nu \boldsymbol{\mu}^T \boldsymbol{x}) \quad \text{with} \quad C_q(\nu) = \frac{\nu^{\frac{q-1}{2}}}{(2\pi)^{\frac{q+1}{2}} \mathcal{I}_{\frac{q-1}{2}}(\nu)}, \tag{4}$$

where $\boldsymbol{\mu} \in \Omega_q$ is the directional mean, $\nu \geq 0$ is the concentration parameter, and

$$\mathcal{I}_\alpha(\nu) = \frac{\left(\frac{\nu}{2}\right)^\alpha}{\pi^{\frac{1}{2}}\Gamma\left(\alpha + \frac{1}{2}\right)} \int_{-1}^{1} (1 - t^2)^{\alpha - \frac{1}{2}} \cdot e^{\nu t} dt$$

is the modified Bessel function of the first kind of order ν. See Figure 2 for contour plots of a von Mises-Fisher density and a mixture of von Mises-Fisher densities on Ω_2, respectively.

1. Normally, the kernel L is only required to be defined on $[0, \infty)$. We extend its domain to $(-\delta_L, \infty) \subset \mathbb{R}$ so that the usual derivatives of $\widehat{f}_h$ can be defined in $\mathbb{R}^{q+1}$ or at least a small neighborhood around Ω_q in $\mathbb{R}^{q+1}$ under some mild conditions on L. See Section 2.2 and condition (D2') in Section 4.2 for details.



(a) $f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}, \nu)$ with $\boldsymbol{\mu} = (0, 0, 1)$ and $\nu = 4.0$

(b) $\frac{2}{5} \cdot f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + \frac{3}{5} \cdot f_{\text{vMF},2}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2)$ with $\boldsymbol{\mu}_1 = (0, 0, 1), \boldsymbol{\mu}_2 = (1, 0, 0)$, and $\nu_1 = \nu_2 = 5.0$

Figure 2: Contour plots of a 2-von Mises-Fisher density and a mixture of 2-vMF densities

Using the von-Mises kernel, the directional KDE in (2) becomes a mixture of q-von Mises-Fisher densities as follows:

$$\widehat{f}_h(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} f_{\text{vMF}}\left(\boldsymbol{x}; \boldsymbol{X}_i, \frac{1}{h^2}\right) = \frac{1}{n(2\pi)^{\frac{q+1}{2}} \mathcal{I}_{\frac{q-1}{2}}(1/h^2) h^{q-1}} \sum_{i=1}^{n} \exp\left(\frac{\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right).$$

For a more detailed discussion of the statistical properties of the von Mises-Fisher distribution and directional KDE, we refer the interested reader to Mardia and Jupp (2000); Banerjee et al. (2005); Pewsey and García-Portugués (2021).

2.2 Gradient and Hessian on a Sphere

For a function defined on a manifold, its gradient and Hessian are defined through the tangent space of the manifold. Whereas the formal definitions of the gradient and Hessian on a general manifold are often involved (see Appendix B), their representations are simple when the manifold is a (hyper)sphere Ω_q.

Let $T_{\boldsymbol{x}} \equiv T_{\boldsymbol{x}}(\Omega_q)$ be the tangent space of the sphere Ω_q at point $\boldsymbol{x} \in \Omega_q$. For the sphere Ω_q, the tangent space has a simple representation in the ambient space $\mathbb{R}^{q+1}$ as follows:

$$T_{\boldsymbol{x}} \simeq \left\{\boldsymbol{v} \in \mathbb{R}^{q+1} : \boldsymbol{x}^T \boldsymbol{v} = 0\right\}, \tag{5}$$

where $V_1 \simeq V_2$ signifies that the two vector spaces are isomorphic. In what follows, the expression $\boldsymbol{v} \in T_{\boldsymbol{x}}$ indicates that $\boldsymbol{v}$ is a vector tangent to Ω_q at $\boldsymbol{x}$.

A *geodesic* on Ω_q is a non-constant, parametrized curve $\gamma : [0, 1] \to \Omega_q$ of constant speed and (locally) minimum length between two points on Ω_q. It can be represented by part of a great circle on the sphere Ω_q. For a smooth function $f : \Omega_q \to \mathbb{R}$, its differential in the

(tangent) direction $\boldsymbol{v} \in T_{\boldsymbol{x}}$ with $||\boldsymbol{v}||_2 = 1$ at point $\boldsymbol{x} \in \Omega_q$ is defined as follows. We first define a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$. Then the differential (at $\boldsymbol{x}$) $df_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \mathbb{R}$ is given by

$$df_{\boldsymbol{x}}(\boldsymbol{v}) = \frac{d}{dt} f(\alpha(t))\Big|_{t=0}. \tag{6}$$

With this, the *Riemannian gradient* $\texttt{grad}\, f(\boldsymbol{x}) \in T_{\boldsymbol{x}} \subset \mathbb{R}^{q+1}$ is defined as

$$df_{\boldsymbol{x}}(\boldsymbol{v}) = \langle \texttt{grad}\, f(\boldsymbol{x}), \boldsymbol{v} \rangle = \boldsymbol{v}^T \texttt{grad}\, f(\boldsymbol{x}). \tag{7}$$

The Riemannian Hessian $\mathcal{H}f(\boldsymbol{x}) \in T_{\boldsymbol{x}} \times T_{\boldsymbol{x}}$ is the second derivative of f within the tangent space $T_{\boldsymbol{x}}$. We characterize its matrix representation as follows. Let $\boldsymbol{v}, \boldsymbol{u} \in T_{\boldsymbol{x}} \subset \mathbb{R}^{q+1}$ be two unit vectors inside the tangent space $T_{\boldsymbol{x}}$. We consider two geodesic curves $\alpha, \beta : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \beta(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$ and $\beta'(0) = \boldsymbol{u}$. We define a second-order differential as

$$d^2 f_{\boldsymbol{x}}(\boldsymbol{v}, \boldsymbol{u}) = \frac{d}{dt} df_{\beta(t)} \left(\alpha'(t)\right) \Big|_{t=0}$$

and the *Riemannian Hessian* $\mathcal{H}f(\boldsymbol{x})$ is a $(q+1) \times (q+1)$ matrix satisfying

$$d^2 f_{\boldsymbol{x}}(\boldsymbol{v}, \boldsymbol{u}) = \langle \texttt{grad} \langle \texttt{grad}\, f, \boldsymbol{v} \rangle (\boldsymbol{x}), \boldsymbol{u} \rangle = \boldsymbol{v}^T \mathcal{H}f(\boldsymbol{x}) \boldsymbol{u} \tag{8}$$

and belongs to $T_{\boldsymbol{x}} \times T_{\boldsymbol{x}}$. To ensure that $\mathcal{H}f(\boldsymbol{x})$ belongs to $T_{\boldsymbol{x}} \times T_{\boldsymbol{x}}$, it has to satisfy

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\mathcal{H}f(\boldsymbol{x}) = \mathcal{H}f(\boldsymbol{x})(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T), \tag{9}$$

where I_{q+1} is the $(q+1) \times (q+1)$ identity matrix and $(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)$ is a projection matrix onto the tangent space $T_{\boldsymbol{x}}$. Note that $d^2 f_{\boldsymbol{x}}(\boldsymbol{v}, \boldsymbol{u}) = d^2 f_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v})$ can be easily verified.

Although (7) and (8) define the Riemannian gradient and Riemannian Hessian on a sphere, it is unclear how they are related to the total gradient operator ∇, where $\nabla g(\boldsymbol{x}) \in \mathbb{R}^{q+1}$ and the ℓ-th component is

$$[\nabla g(\boldsymbol{x})]_\ell = \frac{dg(\boldsymbol{x})}{dx_\ell}$$

for any differentiable function $g : \mathbb{R}^{q+1} \to \mathbb{R}$. Whereas the total gradient ∇ cannot be applied to a directional density (because it is only supported on Ω_q), the directional KDE $\widehat{f}_h$ is well-defined outside of Ω_q (after smoothly extending the domain of the kernel L from $[0, \infty)$ to $\mathbb{R}$), and its total gradient $\nabla \widehat{f}_h(\boldsymbol{x}) \in \mathbb{R}^{q+1}$ can be defined for any point $\boldsymbol{x} \in \mathbb{R}^{q+1}$.

To associate the total gradient with the Riemannian gradient, we consider the following construction. Assume tentatively that f is well-defined and smooth in $\mathbb{R}^{q+1} \backslash \{\boldsymbol{0}\}$, not limited to Ω_q. In this case, $\nabla f(\boldsymbol{x})$ is well-defined $\mathbb{R}^{q+1} \backslash \{\boldsymbol{0}\}$ and all subsequent derivations can also be applied to the directional KDE $\widehat{f}_h$. For any point $\boldsymbol{x} \in \Omega_q$ and unit vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$, we define a geodesic curve $\alpha : (-\epsilon, \epsilon) \to \Omega_q$ with $\alpha(0) = \boldsymbol{x}$ and $\alpha'(0) = \boldsymbol{v}$. Then, a differential of f at $\boldsymbol{x} \in \Omega_q$ is a linear map characterized by

$$df_{\boldsymbol{x}}(\boldsymbol{v}) = \frac{d}{dt} f(\alpha(t))\Big|_{t=0} = \nabla f(\alpha(t))^T \alpha'(t)\Big|_{t=0} = \nabla f(\boldsymbol{x})^T \alpha'(0) = \nabla f(\boldsymbol{x})^T \boldsymbol{v}$$

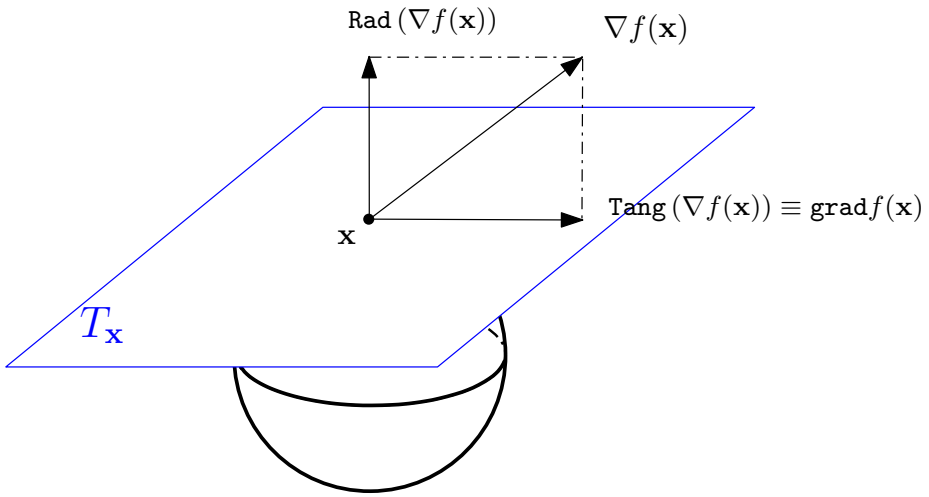


Figure 3: Visualization of a differential of the directional density f on the unit sphere and its gradient

for any given $\boldsymbol{v} \in T_{\boldsymbol{x}}$. Thus, by the definition of the Riemannian gradient in (7),

$$df_{\boldsymbol{x}}(\boldsymbol{v}) = \boldsymbol{v}^T \mathtt{grad}\, f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^T \boldsymbol{v} = \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right)^T \boldsymbol{v},$$

and we conclude that

$$\mathtt{grad}\, f(\boldsymbol{x}) \equiv \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right) = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}), \tag{10}$$

which is the tangent component of the total gradient $\nabla f(\boldsymbol{x})$. That is, the Riemannian gradient is the same as the tangent component of the total gradient. In addition, we can define the radial component of the total gradient as

$$\mathtt{Rad}(\nabla f(\boldsymbol{x})) = \nabla f(\boldsymbol{x}) - \mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right) = \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}). \tag{11}$$

See Figure 3 for a graphical illustration.

In the same context, we use the fact that $\alpha''(0) = -\boldsymbol{x}$ for the geodesic curve α and deduce that for any unit vector $\boldsymbol{v} \in T_{\boldsymbol{x}} \subset \mathbb{R}^{q+1}$,

$$\begin{aligned}
\boldsymbol{v}^T \mathcal{H} f(\boldsymbol{x}) \boldsymbol{v} &= \frac{d^2}{dt^2} f(\alpha(t))\Big|_{t=0} \\
&= \frac{d}{dt}\left[\nabla f(\alpha(t))^T \alpha'(t)\right]\Big|_{t=0} \\
&\left(= \left[\sum_{i=1}^{q+1}\sum_{j=1}^{q+1} \frac{\partial^2}{\partial x_i \partial x_j} f(\alpha(t)) \cdot \alpha_i'(t)\alpha_j'(t) + \sum_{i=1}^{q+1} \frac{\partial}{\partial x_i} f(\alpha(t)) \cdot \alpha_i''(t)\right]\Bigg|_{t=0}\right) \\
&= \alpha'(0)^T \nabla\nabla f(\alpha(0))\alpha'(0) + \nabla f(\alpha(0))^T \alpha''(0) \\
&= \boldsymbol{v}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{v} + \nabla f(\boldsymbol{x})^T \alpha''(0) \\
&= \boldsymbol{v}^T (\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1})\boldsymbol{v}.
\end{aligned} \tag{12}$$

One may conjecture that $(\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \alpha''(0))$ is the Riemannian Hessian matrix. However, it does not satisfy the projection condition in Equation (9). To this end, we select

$$\mathcal{H}f(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T). \tag{13}$$

One can verify that the Hessian matrix in (13) satisfies both (8) and (9); thus, it characterizes the relationship between the Riemannian Hessian and total gradient operator. More importantly, the Hessian matrix in (13) is indeed the Riemannian Hessian on Ω_q. Detailed definitions of Riemannian Hessians can be found in Section 2 and 4.2 of Absil et al. (2013).

3. Mean Shift Algorithm with Directional Data

In this section, we present a detailed derivation of the mean shift algorithm with directional data. Given the directional KDE $\widehat{f}_h(\boldsymbol{x})$ in (2), Kobayashi and Otsu (2010); Yang et al. (2014) introduced a Lagrangian multiplier to maximize $\widehat{f}_h(\boldsymbol{x})$ under the constraint $\boldsymbol{x}^T\boldsymbol{x} = 1$ and derived the directional mean shift algorithm. To make a better comparison with the standard mean shift algorithm with Euclidean data, we provide an alternative derivation.

Given a Euclidean KDE of the form $\widehat{p}_n(\boldsymbol{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)$ with a differentiable kernel profile $k : [0,\infty) \to [0,\infty)$, its (total) gradient has the following decomposition:

$$\nabla\widehat{p}_n(\boldsymbol{x}) = \underbrace{\frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]}_{\text{term 1}} \underbrace{\left[\frac{\sum_{i=1}^{n}\boldsymbol{X}_i g\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n} g\left(\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)} - \boldsymbol{x}\right]}_{\text{term 2}}, \tag{14}$$

where $g(x) = -k'(x)$ is the derivative of the selected kernel profile. As noted by Comaniciu and Meer (2002), the first term is proportional to the density estimate at $\boldsymbol{x}$ with the "kernel" $G(\boldsymbol{x}) = c_{g,d} \cdot g(||\boldsymbol{x}||_2^2)$, and the second term is the so-called *mean shift* vector, which points toward the direction of maximum increase in the density estimator $\widehat{p}_n$. Thus, the standard mean shift algorithm with Euclidean data translates each query point according to the corresponding mean shift vector, which leads to a converging path to a local mode of $\widehat{p}_n$ under some conditions (Li et al., 2007; Aliyari Ghassabeh, 2015; Arias-Castro et al., 2016).

The key insight in our derivation of the directional mean shift algorithm is the following alternative representation of the directional KDE as:

$$\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n}\sum_{i=1}^{n} L\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right), \tag{15}$$

given a directional random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \Omega_q$. Recall that the original directional KDE in (2) is $\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{n}\sum_{i=1}^{n} L\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$. Both $\widehat{f}_h$ and $\widetilde{f}_h$ can be defined on any point in $\mathbb{R}^{q+1}\backslash\{\boldsymbol{0}\}$. Although $\widehat{f}_h(\boldsymbol{x}) \neq \widetilde{f}_h(\boldsymbol{x})$ for $\boldsymbol{x} \notin \Omega_q$, their function values are identical on the sphere; that is,

$$\widehat{f}_h(\boldsymbol{x}) = \widetilde{f}_h(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \Omega_q \tag{16}$$

due to the fact that $\frac{1}{2}\left|\left|\boldsymbol{x}-\boldsymbol{X}_i\right|\right|_2^2 = 1-\boldsymbol{x}^T\boldsymbol{X}_i$ for any $\boldsymbol{x}\in\Omega_q$.

Since the two directional KDEs are the same on Ω_q, either of them can be used to express our density estimator. The power of the expression $\widetilde{f}_h$ is that its total gradient has a similar decomposition as the total gradient of the Euclidean KDE (cf. (14)):

$$\begin{aligned}\nabla\widetilde{f}_h(\boldsymbol{x}) &= \frac{c_{h,q}(L)}{nh^2}\sum_{i=1}^{n}(\boldsymbol{x}-\boldsymbol{X}_i)\cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\\ &= \underbrace{\frac{c_{h,q}(L)}{nh^2}\left[\sum_{i=1}^{n}-L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)\right]}_{\text{term 1}}\cdot\underbrace{\left[\frac{\sum_{i=1}^{n}\boldsymbol{X}_i\cdot L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n}L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}-\boldsymbol{x}\right]}_{\text{term 2}}.\end{aligned}\tag{17}$$

Similar to the density gradient estimator with Euclidean data (cf. Equation (14)), the first term of the product in (17) can be viewed as a proportional form of the directional density estimate at $\boldsymbol{x}$ with "kernel" $G(r) = -L'(r)$:

$$\widetilde{f}_{h,G}(\boldsymbol{x}) = \frac{c_{h,q}(G)}{n}\sum_{i=1}^{n}-L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right) = \frac{c_{h,q}(G)}{n}\sum_{i=1}^{n}-L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)\tag{18}$$

given that $-L'(r)$ is non-negative on $[0,\infty)$. Some commonly used kernel functions, such as the von-Mises kernel $L(r) = e^{-r}$, easily satisfy this condition. The second term of the product in (17) is indeed the *directional mean shift* vector

$$\Xi_h(\boldsymbol{x}) = \frac{\sum_{i=1}^{n}\boldsymbol{X}_iL'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}{\sum_{i=1}^{n}L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\right|\right|_2^2\right)}-\boldsymbol{x} = \frac{\sum_{i=1}^{n}\boldsymbol{X}_iL'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^{n}L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}-\boldsymbol{x},\tag{19}$$

which is the difference between a weighted sample mean with weights $\frac{L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^{n}L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}$, $i = 1,...,n$, and $\boldsymbol{x}$, the current query point of the directional density estimation. It is worth mentioning that these weights are strictly positive when the von-Mises kernel $L(r) = e^{-r}$ is applied. From Equations (18) and (19), the total gradient estimator at $\boldsymbol{x}$ becomes

$$\nabla\widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{c_{h,q}(G)h^2}\cdot\widetilde{f}_{h,G}(\boldsymbol{x})\cdot\Xi_h(\boldsymbol{x}),$$

yielding

$$\Xi_h(\boldsymbol{x}) = \frac{c_{h,q}(G)h^2}{c_{h,q}(L)}\cdot\frac{\nabla\widetilde{f}_h(\boldsymbol{x})}{\widetilde{f}_{h,G}(\boldsymbol{x})}.$$

As is illustrated in (10), the total gradient of the directional KDE at $\boldsymbol{x}$, $\nabla\widetilde{f}_h(\boldsymbol{x})$, becomes the Riemannian gradient of $\widetilde{f}_h(\boldsymbol{x}) = \widehat{f}_h(\boldsymbol{x})$ on Ω_q after being projected onto the tangent space $T_{\boldsymbol{x}}$. This suggests that the directional mean shift vector $\Xi_h(\boldsymbol{x})$, which is parallel to

Algorithm 1 Mean Shift Algorithm with Directional Data

Input:

- Directional data sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim f(\boldsymbol{x})$ on Ω_q.
- The smoothing bandwidth h.
- An initial point $\widehat{\boldsymbol{y}}_0 \in \Omega_q$ and the precision threshold $\epsilon > 0$.

while $1 - \widehat{\boldsymbol{y}}_{s+1}^T \widehat{\boldsymbol{y}}_s > \epsilon$ **do**

$$\widehat{\boldsymbol{y}}_{s+1} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2} \tag{20}$$

end while

Output: A candidate local mode of directional KDE, $\widehat{\boldsymbol{y}}_s$.

the total gradient of $\widetilde{f_h}$ at $\boldsymbol{x}$, points in the direction of maximum increase in the estimated density $\widetilde{f_h}$ after being projected onto the tangent space $T_{\boldsymbol{x}}$. However, due to the manifold structure of Ω_q, translating a point $\boldsymbol{x} \in \Omega_q$ in the mean shift direction $\Xi_h(\boldsymbol{x})$ deviates the point from Ω_q. We thus project the translated point $\boldsymbol{x} + \Xi_h(\boldsymbol{x})$ onto Ω_q by a simple standardization: $\boldsymbol{x} + \Xi_h(\boldsymbol{x}) \mapsto \frac{\boldsymbol{x}+\Xi_h(\boldsymbol{x})}{||\boldsymbol{x}+\Xi_h(\boldsymbol{x})||_2}$. In summary, starting at point $\boldsymbol{x}$, the directional mean shift algorithm moves this point to a new location $\frac{\boldsymbol{x}+\Xi_h(\boldsymbol{x})}{||\boldsymbol{x}+\Xi_h(\boldsymbol{x})||_2}$. This movement creates a path leading to a local mode of the estimated directional density under suitable conditions (Theorems 8 and 11).

We can encapsulate the directional mean shift algorithm into a single fixed-point equation. Let $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty} \subset \Omega_q$ denote the path of successive points defined by the directional mean shift algorithm, where $\widehat{\boldsymbol{y}}_0$ is the initial point of the iteration. Translating the query point $\widehat{\boldsymbol{y}}_s$ by the directional mean shift vector (19) at step s leads to

$$\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s = \frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^n L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}.$$

When $L(r)$ is decreasing, $L'(r)$ is non-positive on $[0, \infty)$ and

$$||\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s||_2 = \frac{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2}{\left|\sum_{i=1}^n L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|} = -\frac{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2}{\sum_{i=1}^n L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}$$

given that $\sum_{i=1}^n L'\left(\frac{1-\boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right) \neq 0$. (Here $L'(r)$ can be replaced by subgradients at non-differentiable points of L. See also Remark 9.) Again, many commonly used kernel functions, such as the von-Mises kernel $L(r) = e^{-r}$, have nonzero derivatives on $[0, \infty)$ and satisfy this mild condition. Therefore,

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s}{||\Xi_h\left(\widehat{\boldsymbol{y}}_s\right) + \widehat{\boldsymbol{y}}_s||_2} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)}{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2}$$

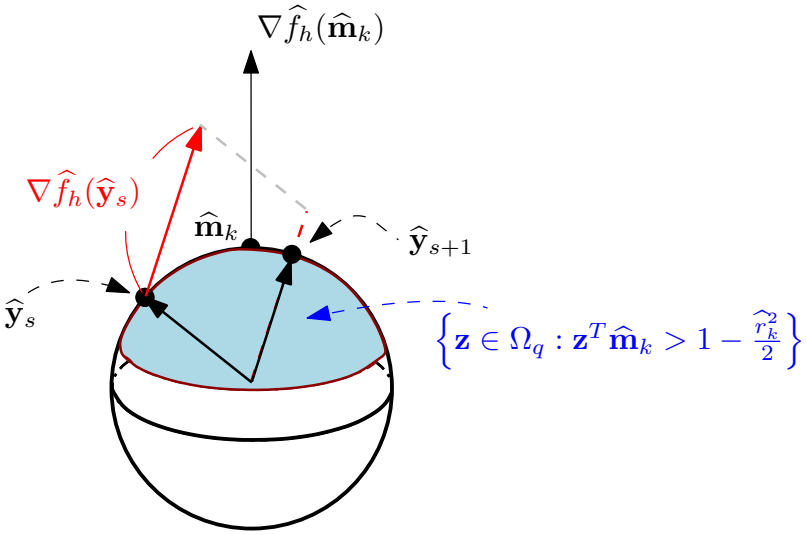


Figure 4: Illustration of one-step iteration of Algorithm 1

is the resulting fixed-point equation for $s = 0, 1, ...$, whose right-hand side is a standardized weighted sample mean at $\widehat{\boldsymbol{y}}_s$ computed with "kernel" $G(r) = -L'(r)$. The entire mean shift algorithm with directional data is summarized in Algorithm 1 (see also Figure 4 for a graphical illustration).

Analogous to the mean shift algorithm with Euclidean data, Algorithm 1 can be leveraged for mode seeking and clustering with directional data. We derive statistical and computational learning theory for mode seeking in Sections 4 and 5. For clustering, we demonstrate with both simulated and real-world data sets that the algorithm can be used to cluster directional data in Section 6. It should also be noted that the directional mean shift algorithm can be viewed as a gradient ascent method on Ω_q with an adaptive step size; see Section 5.2 for details.

More importantly, similar to the standard mean shift algorithm with Euclidean data, the directional mean shift algorithm has several advantages over a regular gradient ascent method. First, the directional mean shift algorithm requires no tuning of the step size parameter, yet exhibits mathematical simplicity when it is written as the fixed-point iteration (20). Second, the algorithm does not need to estimate the normalizing constant $c_{h,q}(L)$ of the directional KDE in its application. Specifically, in order to identify local modes of the directional KDE using our algorithm, it is only necessary to specify the directional kernel L up to a constant. This avoids additional computational cost in estimating the normalizing constant $c_{h,q}(L)$ for the kernel, because the constant $c_{h,q}(L)$ often involves complicated functions for high dimensional directional data. For instance, estimating the normalizing constant of the von Mises kernel involves an approximation of a modified Bessel function of the first kind, though several efficient algorithms have been developed; see, for instance, Sra (2012).

4. Statistical Learning Theory of Directional KDE and its Derivatives

Because the (directional) mean shift algorithm is inspired by a gradient ascent method, we study the gradient and Hessian systems of the two estimators $\widehat{f}_h$ and $\widetilde{f}_h$.

4.1 Gradient and Hessian of Directional KDEs

We have demonstrated that it is valid to deduce two mathematically equivalent directional KDEs (2) and (15) for estimating the true directional density f. Somewhat surprisingly, the corresponding total gradients are different in general. The total gradient of $\widetilde{f}_h$ is

$$\nabla \widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot L' \left(\frac{1}{2} \left|\left| \frac{\boldsymbol{x} - \boldsymbol{X}_i}{h} \right|\right|_2^2 \right), \tag{21}$$

while the total gradient of $\widehat{f}_h$ is

$$\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{X}_i L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right). \tag{22}$$

Although the total gradients $\nabla \widetilde{f}_h$ and $\nabla \widehat{f}_h$ have different values even on Ω_q, they both play a vital role in the directional mean shift algorithm (Algorithm 1). On the one hand, we have argued in Section 3 that $\nabla \widetilde{f}_h(\boldsymbol{x})$ has a similar decomposition as the total gradient of the Euclidean KDE, and derived Algorithm 1 based on $\nabla \widetilde{f}_h(\boldsymbol{x})$. On the other hand, given the form of $\nabla \widehat{f}_h(\boldsymbol{x})$ in (22), the fixed-point equation (20) in Algorithm 1 can be written as

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left| \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s) \right|\right|_2}. \tag{23}$$

As argued in Section 2.2 and (11), any total gradient at $\boldsymbol{x} \in \Omega_q$ can be decomposed into radial and tangent components. Therefore, the total gradient $\nabla \widetilde{f}_h(\boldsymbol{x})$ is decomposed as

$$\begin{aligned}
\nabla \widetilde{f}_h(\boldsymbol{x}) &= \boldsymbol{x}\boldsymbol{x}^T \nabla \widetilde{f}_h(\boldsymbol{x}) + \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla \widetilde{f}_h(\boldsymbol{x}) \\
&= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{x} \left(1 - \boldsymbol{x}^T \boldsymbol{X}_i\right) L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \\
&\quad + \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\boldsymbol{x} \cdot \boldsymbol{x}^T \boldsymbol{X}_i - \boldsymbol{X}_i\right) L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \\
&\equiv \texttt{Rad}\left(\nabla \widetilde{f}_h(\boldsymbol{x}) \right) + \texttt{Tang}\left(\nabla \widetilde{f}_h(\boldsymbol{x}) \right),
\end{aligned}$$

where `Rad` and `Tang` are the radial and tangent components of the total gradient, as in (11) and (10). Similarly, we decompose $\nabla \widehat{f}_h(\boldsymbol{x})$ as

$$\begin{aligned}
\nabla \widehat{f}_h(\boldsymbol{x}) &= \boldsymbol{x}\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x}) + \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla \widehat{f}_h(\boldsymbol{x}) \\
&= -\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{X}_i \cdot L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\boldsymbol{x} \cdot \boldsymbol{x}^T \boldsymbol{X}_i - \boldsymbol{X}_i\right) \cdot L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) \\
&\equiv \mathtt{Rad}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) + \mathtt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right).
\end{aligned}
$$

Therefore, the difference between the two total gradients $\nabla \widetilde{f}_h(\boldsymbol{x})$ and $\nabla \widehat{f}_h(\boldsymbol{x})$ is

$$
\nabla \widetilde{f}_h(\boldsymbol{x}) - \nabla \widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) \cdot \boldsymbol{x}, \tag{24}
$$

which is parallel to the radial direction $\boldsymbol{x}$. This implies that given kernel L, the Riemannian gradients of the two estimators are the same, that is,

$$
\begin{aligned}
\mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) \equiv \mathtt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) &= \nabla \widehat{f}_h(\boldsymbol{x}) - \left[\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x})\right] \cdot \boldsymbol{x} \\
&= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\boldsymbol{x}^T \boldsymbol{X}_i \cdot \boldsymbol{x} - \boldsymbol{X}_i\right) \cdot L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) \\
&= \mathtt{grad}\, \widetilde{f}_h(\boldsymbol{x}) \equiv \mathtt{Tang}\left(\nabla \widetilde{f}_h(\boldsymbol{x})\right).
\end{aligned} \tag{25}
$$

Later, we demonstrate in Theorems 2 and 4 that the Riemannian gradients of $\widehat{f}_h$ and $\widetilde{f}_h$ are consistent estimators of the Riemannian gradient of the underlying density f that generates directional data. One can also deduce the same fixed-point equation (20) (or equivalently (23)) from the Riemannian/tangent gradient estimator $\mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) \equiv \mathtt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right)$, although the assumption on the directional estimated density $\widehat{f}_h$ is stricter. See Appendix C for detailed derivations.

Having demonstrated that the Riemannian gradients of $\widetilde{f}_h$ and $\widehat{f}_h$ are identical, we now study the Riemannian Hessians of $\widetilde{f}_h$ and $\widehat{f}_h$. By (13), the Riemannian Hessian of $\widehat{f}_h$ is associated with the total gradient operator ∇ via

$$
\mathcal{H}\widehat{f}_h(\boldsymbol{x}) = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\left[\nabla\nabla \widehat{f}_h(\boldsymbol{x}) - \nabla \widehat{f}_h(\boldsymbol{x})^T \boldsymbol{x} I_{q+1}\right](I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \tag{26}
$$

and similarly for $\mathcal{H}\widetilde{f}_h(\boldsymbol{x})$. The following lemma shows that when a directional kernel L is smooth, the two Riemannian Hessians are identical.

Lemma 1 *Assume that kernel L is twice continuously differentiable. Then,*

$$
\mathcal{H}\widetilde{f}_h(\boldsymbol{x}) = \mathcal{H}\widehat{f}_h(\boldsymbol{x})
$$

for any point $\boldsymbol{x} \in \Omega_q$.

The proof of Lemma 1 can be found in Appendix D.1. As a result, we can compute the Riemannian Hessian estimator at point $\boldsymbol{x} \in \Omega_q$ based on either $\widetilde{f}_h(\boldsymbol{x})$ or $\widehat{f}_h(\boldsymbol{x})$, which will produce the same expression. Later, in Theorems 2 and 4, we demonstrate that $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ is a (uniformly) consistent estimator of $\mathcal{H}f(\boldsymbol{x})$ defined in (13).

4.2 Assumptions

To apply the total gradient operator ∇ to a directional density f that generates data, we extend it from Ω_q to $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ by defining $f(\boldsymbol{x}) \equiv f\left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}\right)$ for all $\boldsymbol{x} \in \mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$. In this extension, we assume that the total gradient $\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, ..., \frac{\partial f(\boldsymbol{x})}{\partial x_{q+1}}\right)^T$ and total Hessian matrix $\nabla\nabla f(\boldsymbol{x}) = \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right)_{1 \le i,j \le q+1}$ in $\mathbb{R}^{q+1}$ exist, and are continuous on $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ and square integrable on Ω_q. This extension has also been used by Zhao and Wu (2001); García-Portugués et al. (2013); García-Portugués (2013). Note that the Riemannian gradient and Hessian are invariant under this extension.

To establish the consistency results of gradient and Hessian estimators (cf. (25) and (26) or (38) in its explicit form), we consider the following assumptions.

- **(D1)** Assume that the extended density function f is at least three times continuously differentiable on $\mathbb{R}^{q+1} \setminus \{\mathbf{0}\}$ and that its derivatives are square integrable on Ω_q.
- **(D2)** Assume that $L : [0, \infty) \to [0, \infty)$ is a bounded and Riemann integrable function such that
$$0 < \int_0^\infty L^k(r) r^{\frac{q}{2}-1} dr < \infty$$
for all $q \ge 1$ and $k = 1, 2$.
- **(D2')** Under (D2), we further assume that L is a twice continuously differentiable function on $(-\delta_L, \infty) \subset \mathbb{R}$ for some constant $\delta_L > 0$ such that
$$0 < \int_0^\infty L'(r)^k r^{\frac{q}{2}-1} dr < \infty, \quad 0 < \int_0^\infty L''(r)^k r^{\frac{q}{2}-1} dr < \infty$$
for all $q \ge 1$ and $k = 1, 2$.

Here, conditions (D1) and (D2) are required for the consistency of the directional KDE (Hall et al., 1987; Klemelä, 2000; Zhao and Wu, 2001; García-Portugués et al., 2013; García-Portugués, 2013). The stronger condition (D2') is imposed for the consistency of Riemannian gradient estimator $\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) \equiv \texttt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right)$ and Hessian estimator $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$. The differentiability condition in (D2') can be relaxed so that L, after being smoothly extrapolated from $[0, \infty)$ to $(-\delta_L, \infty)$ for some constant $\delta_L > 0$, is (twice) continuously differentiable except for a set of points with Lebesgue measure 0 on $(-\delta_L, \infty)$. One can justify via integration by parts that many commonly used kernels, such as the von-Mises kernel $L(r) = e^{-r}$ or compactly supported kernels, satisfy condition (D2').

Under conditions (D1) and (D2), the pointwise convergence rate of $\widehat{f}_h$ is

$$\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^q}}\right);$$

see, for instance, Hall et al. (1987); Zhao and Wu (2001); García-Portugués (2013); García-Portugués et al. (2013). Moreover, Bai et al. (1988) used a piecewise constant kernel function

to approximate the given kernel L and derived the uniform convergence rate as

$$\|\widehat{f}_h - f\|_\infty = \sup_{\boldsymbol{x}\in\Omega_q} \left|\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})\right| = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^q}}\right). \tag{27}$$

One can also prove the uniform consistency of the directional KDE by slightly modifying the technique in Giné and Guillou (2002) and Einmahl and Mason (2005) for the consistency results of the usual Euclidean KDE. We will leverage such technique in our proof for the uniform convergence rates of the Riemannian gradient and Hessian estimators.

4.3 Pointwise Consistency

Our derivations of the pointwise convergence rates of the (Riemannian) gradient and Hessian estimators of the directional KDE $\widehat{f}_h$ are analogous to the arguments for the usual Euclidean KDE (Silverman, 1986; Scott, 2015), which rely on the Taylor's expansion. The difference in the directional KDE case is that the integrals are taken over the Lebesgue measure ω_q on Ω_q when we compute the expectations $\mathbb{E}\left[\texttt{grad}\,\widehat{f}_h(\boldsymbol{x})\right]$ and $\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]$. The key argument for evaluating directional integrals is the following change-of-variable formula

$$\omega_q(d\boldsymbol{x}) = (1-t^2)^{\frac{q}{2}-1} dt\, \omega_{q-1}(d\boldsymbol{\xi}),$$

where $t = \boldsymbol{x}^T\boldsymbol{y}$ for a fixed point $\boldsymbol{y} \in \Omega_q$ and $\boldsymbol{\xi} \in \Omega_q$ is a unit vector orthogonal to $\boldsymbol{y}$. The formula is proved in Lemma 2 of García-Portugués et al. (2013) and on pages 91-93 in Efthimiou and Frye (2014) in two different ways. The surface area of Ω_q in (1) easily follows from this formula. With this formula, we have the following convergence results.

Theorem 2 *Assume conditions (D1) and (D2'). For any fixed $\boldsymbol{x} \in \Omega_q$, we have*

$$\texttt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\,f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$$

as $h \to 0$ and $nh^{q+2} \to \infty$.
Under the same condition, for any fixed $\boldsymbol{x} \in \Omega_q$, we have

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$$

as $h \to 0$ and $nh^{q+4} \to \infty$.

The proof of Theorem 2 is lengthy and deferred to Appendix D.2. Theorem 2 demonstrates that the Riemannian gradient of a directional KDE is a consistent estimator of the Riemannian gradient of the directional density that generates data. A similar result holds for the Riemannian Hessian. It cannot be claimed that the total gradients $\nabla\widehat{f}_h$ or $\nabla\widetilde{f}_h$ converge to ∇f since the radial component of f depends on how f is extended to points outside Ω_q. Lemma 10 below and the proof of Theorem 2 demonstrate that the limiting behaviors of $\texttt{Rad}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)$ and $\texttt{Rad}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right)$ are different; the former one is of the order

$O\left(h^{-2}\right) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$ while the latter one is of the order $O(1) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$. Note that Klemelä (2000) also derived similar convergence rates of the derivatives of a directional KDE, while the definitions of directional KDE and its derivatives in Klemelä (2000) are different from ours and the results are more complex.

Remark 3 *Under some smoothness conditions (Chacón et al., 2011), the pointwise convergence rates of gradient and Hessian estimators defined by the usual Euclidean KDE are*

$$O(h^2) + O_P\left(\frac{1}{nh^{d+2}}\right) \quad \text{and} \quad O(h^2) + O_P\left(\frac{1}{nh^{d+4}}\right),$$

where d represents the dimension of the Euclidean data. Therefore, our pointwise consistency results for the Riemannian gradient and Hessian of the directional KDE in Theorem 2 align with the pointwise convergence rates of the usual Euclidean KDE, in the sense that the dimension d is replaced by the (intrinsic) manifold dimension q of directional data.

4.4 Uniform Consistency

We now strengthen the convergence results in Theorem 2 to uniform convergence rates with the assumptions and techniques developed by Giné and Guillou (2002) and Einmahl and Mason (2005).

Let $[\tau] = (\tau_1, ..., \tau_{q+1})$ be a multi-index (that is, $\tau_1, ..., \tau_{q+1}$ are non-negative integers and $|[\tau]| = \sum_{j=1}^{q+1} \tau_j$). Define $D^{[\tau]} = \frac{\partial^{\tau_1}}{\partial x_1^{\tau_1}} \cdots \frac{\partial^{\tau_{q+1}}}{\partial x_1^{\tau_{q+1}}}$ as the $|[\tau]|$-th order partial derivative operator. Given the directional KDE in (15), we define the following function class of the kernel function L and its partial derivatives as

$$\mathcal{K} = \left\{ \boldsymbol{u} \mapsto K\left(\frac{\boldsymbol{z}-\boldsymbol{u}}{h}\right) : \boldsymbol{u}, \boldsymbol{z} \in \Omega_q, h > 0, K(\boldsymbol{x}) = D^{[\tau]} L\left(\frac{1}{2}||\boldsymbol{x}||_2^2\right), |[\tau]| = 0, 1, 2 \right\}.$$

Under condition (D2'), $\mathcal{K}$ is a collection of bounded measurable functions on Ω_q. To guarantee the uniform consistency of the directional KDE itself as well as its (Riemannian) gradient and Hessian, we assume the following:

- **(K1)** $\mathcal{K}$ is a bounded VC (subgraph) class of measurable functions on Ω_q, that is, there exist constants $A, \vartheta > 0$ such that for any $0 < \epsilon < 1$,

$$\sup_Q N\left(\mathcal{K}, L_2(Q), \epsilon||F||_{L_2(Q)}\right) \leq \left(\frac{A}{\epsilon}\right)^{\vartheta},$$

  where $N(T, d_T, \epsilon)$ is the ϵ-covering number of the pseudometric space (T, d_T), Q is any probability measure on Ω_q, and F is an envelope function of $\mathcal{K}$. The constants A and ϑ are usually called the VC (Vapnik-Chervonenkis) characteristics of $\mathcal{K}$ and the norm $||F||_{L_2(Q)}$ is defined as $\left[\int_{\Omega_q} |F(\boldsymbol{x})|^2 dQ(\boldsymbol{x})\right]^{\frac{1}{2}}$.

Given the differentiability of kernel L guaranteed by (D2'), we can take F as a constant envelope function

$$C_{\mathcal{K}} = \sup_{\boldsymbol{x} \in \mathbb{R}^{q+1}, |[\tau]|=0,1,2} \left| D^{[\tau]} L\left(\frac{1}{2}||\boldsymbol{x}||_2^2\right) \right|$$

when it is finite. Condition (K1) is not stringent in practice and can be satisfied by many kernel functions, such as the von-Mises kernel $L(r) = e^{-r}$ and many compactly supported kernels on $[0, \infty)$. For these kernel options, the resulting function class $\mathcal{K}$ comprises only functions of the form $\boldsymbol{u} \mapsto \Phi(|\boldsymbol{A}\boldsymbol{u} + \boldsymbol{b}|)$, where Φ is a real-valued function of bounded variation on $[0, \infty)$, $\boldsymbol{A}$ ranges over matrices in $\mathbb{R}^{(q+1)\times(q+1)}$, and $\boldsymbol{b}$ ranges over $\mathbb{R}^{q+1}$. Thus, $\mathcal{K}$ is of VC (subgraph) class by Lemma 22 in Nolan and Pollard (1987).

Under conditions (D1), (D2'), and (K1), the uniform consistency results of the directional KDE (restated) as well as its Riemannian gradient and Hessian estimators are summarized as the following theorem, whose proof can be founded in Appendix D.3.

Theorem 4 *Assume (D1), (D2'), and (K1). The uniform convergence rate of $\widehat{f}_h$ is given by*

$$\sup_{\boldsymbol{x}\in\Omega_q} |\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})| = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^q}}\right)$$

as $h \to 0$ and $\frac{nh^q}{|\log h|} \to \infty$.
Furthermore, the uniform convergence rate of $\texttt{grad}\, \widehat{f}_h(\boldsymbol{x})$ on Ω_q is

$$\sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right),$$

as $h \to 0$ and $\frac{nh^{q+2}}{|\log h|} \to \infty$. Finally, the uniform convergence rate of $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ on Ω_q is

$$\sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right),$$

as $h \to 0$ and $\frac{nh^{q+4}}{|\log h|} \to \infty$, where $||\cdot||_{\max}$ is the elementwise maximum norm for a vector in $\mathbb{R}^{q+1}$ or a matrix in $\mathbb{R}^{(q+1)\times(q+1)}$.

Remark 5 *Theorem 4 can also be generalized to higher-order derivatives. All that is necessary is to modify the assumptions (D2') and (K1) to higher-order derivatives (projected on the tangent direction) as well as strengthen the differentiable assumptions on f in (D1). The elementwise maximum norm between the derivative estimator and the true quantity will embrace the rate*

$$O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2m}}}\right),$$

where m is the highest order of derivatives desired.

4.5 Mode Consistency

Consistency of estimating local modes has been established for the usual Euclidean KDE by Chen et al. (2016), where the authors demonstrated that with probability tending to 1, the number of estimated local modes is the same as the number of true local modes under appropriate assumptions. Moreover, the convergence rate of the Hausdorff distance (a

common distance between two sets) between the collection of local modes and its estimator is elucidated. Here, we reproduce the consistency of estimating local modes of a directional density f supported on Ω_q by the local modes of the directional KDE $\widehat{f}_h$.

Given two sets $A, B \subset \Omega_q$, their Hausdorff distance is

$$\texttt{Haus}(A, B) = \inf\{r > 0 : A \subset B \oplus r, B \subset A \oplus r\}, \tag{28}$$

where $A \oplus r = \{\boldsymbol{y} \in \Omega_q : \inf_{\boldsymbol{x}\in A} ||\boldsymbol{x} - \boldsymbol{y}||_2 \leq r\} = \left\{\boldsymbol{y} \in \Omega_q : \sup_{\boldsymbol{x}\in A} \boldsymbol{x}^T\boldsymbol{y} \geq 1 - \frac{r^2}{2}\right\}$. The equality follows from the fact that $||\boldsymbol{x}||_2^2 = 1$ for any $\boldsymbol{x} \in \Omega_q$.

Let C_3 be the upper bound for the partial derivatives of the directional density f on the compact manifold Ω_q up to the third order. Such constant exists under condition (D1). Let $\widehat{\mathcal{M}}_n = \left\{\widehat{\boldsymbol{m}}_1, ..., \widehat{\boldsymbol{m}}_{\widehat{K}_n}\right\}$ be the collection of local modes of $\widehat{f}_h$ and $\mathcal{M} = \{\boldsymbol{m}_1, ..., \boldsymbol{m}_K\}$ be the collection of local modes of f. Here, $\widehat{K}_n$ is the number of estimated local modes and K is the number of true local modes. We consider the following assumptions.

- **(M1)** There exists $\lambda_* > 0$ such that

  $$0 < \lambda_* \leq |\lambda_1(\boldsymbol{m}_j)|, \quad \text{for all } j = 1, ..., K,$$

  where $0 > \lambda_1(\boldsymbol{x}) \geq \cdots \geq \lambda_q(\boldsymbol{x})$ are the q smallest (negatively-largest) eigenvalues of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$.

- **(M2)** There exist $\Theta_1, \rho_* > 0$ such that

  $$\left\{\boldsymbol{x} \in \Omega_q : ||\texttt{Tang}(\nabla f(\boldsymbol{x}))||_{\max} \leq \Theta_1, \lambda_1(\boldsymbol{x}) \leq -\frac{\lambda_*}{2} < 0\right\} \subset \mathcal{M} \oplus \rho_*,$$

  where λ_* is defined in (M1) and $0 < \rho_* < \min\left\{\sqrt{2 - 2\cos\left(\frac{3\lambda_*}{2C_3}\right)}, 2\right\}$.

Condition (M1) is imposed so that every local mode of f is isolated from other critical points; see Lemma 3.2 in Banyaga and Hurtubise (2004). The condition also guarantees that the number of local modes of f supported on the compact manifold Ω_q is finite. As noted by Chen et al. (2016), condition (M1) always holds when f is a Morse function on Ω_q. The second condition (M2) regularizes the behavior of f so that points with near 0 (Riemannian) gradients and negative eigenvalues of $\mathcal{H}f(\boldsymbol{x})$ within the tangent space $T_{\boldsymbol{x}}$ must be close to local modes. See the paper by Chen et al. (2016) for detailed discussion. The constant $\sqrt{2 - 2\cos\left(\frac{3\lambda_*}{2C_3}\right)}$ is selected so that the great-circle distance from $\boldsymbol{m}_k$ to the boundary of $\boldsymbol{m}_k \oplus \rho_*$, that is, $\arccos(\boldsymbol{m}_k^T\boldsymbol{x})$ with $\boldsymbol{x} \in \partial S_k$, is less than $\frac{3\lambda_*}{2C_3}$ for any $\boldsymbol{m}_k \in \mathcal{M}$, where $S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \boldsymbol{m}_k||_2 \leq \rho_*\}$ and $\partial S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \boldsymbol{m}_k||_2 = \rho_*\}$.

It should be emphasized that condition (M1) is a weak condition that can be satisfied by the local modes of common directional densities. We take the von-Mises-Fisher density as an example. With the formula (4), we naturally extend f_{vMF} to $\mathbb{R}^{q+1}$ and deduce that

$$\nabla f_{\text{vMF}}(\boldsymbol{x}) = \nu\boldsymbol{\mu}C_q(\nu) \cdot \exp\left(\nu\boldsymbol{\mu}^T\boldsymbol{x}\right) \quad \text{and} \quad \nabla\nabla f_{\text{vMF}}(\boldsymbol{x}) = \nu^2\boldsymbol{\mu}\boldsymbol{\mu}^T C_q(\nu) \cdot \exp\left(\nu\boldsymbol{\mu}^T\boldsymbol{x}\right),$$

which in turn indicates that at the mode $\boldsymbol{\mu} \in \Omega_q$,

$$\begin{aligned}\mathcal{H} f_{\mathrm{vMF}}(\boldsymbol{\mu}) &= \left(I_{q+1} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right) \nabla\nabla f_{\mathrm{vMF}}(\boldsymbol{\mu}) \left(I_{q+1} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right) - \boldsymbol{\mu}^T \nabla f_{\mathrm{vMF}}(\boldsymbol{\mu}) \left(I_{q+1} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right) \\ &= -\nu C_q(\nu) \cdot e^{\nu} \left(I_{q+1} - \boldsymbol{\mu}\boldsymbol{\mu}^T\right).\end{aligned}$$

By Brauer's theorem (Example 1.2.8 in Horn and Johnson 2012), we conclude that the eigenvalues of $\mathcal{H} f_{\mathrm{vMF}}(\boldsymbol{\mu})$ are 0 with (algebraic) multiplicity 1, which is associated with the eigenvector $\boldsymbol{\mu}$, and $-\nu C_q(\nu) \cdot e^{\nu}$ with multiplicity q, which are associated with the eigenvectors in $T_{\boldsymbol{\mu}}$.

Given these assumptions, the mode consistency of the directional KDE is as follows.

Theorem 6 *Assume (D1), (D2'), (K1), and (M1-2). For any $\delta \in (0,1)$, when h is sufficiently small and n is sufficiently large,*

(a) there must be at least one estimated local mode $\widehat{\boldsymbol{m}}_k$ within $S_k = \boldsymbol{m}_k \oplus \rho_*$ for every $\boldsymbol{m}_k \in \mathcal{M}$, and

(b) the collection of estimated modes satisfies $\widehat{\mathcal{M}}_n \subset \mathcal{M} \oplus \rho_*$ and there is a unique estimated local mode $\widehat{\boldsymbol{m}}_k$ within $S_k = \boldsymbol{m}_k \oplus \rho_*$

with probability at least $1 - \delta$. In total, when h is sufficiently small and n is sufficiently large, there exist some constants $A_3, B_3 > 0$ such that

$$\mathbb{P}\left(\widehat{K}_n \neq K\right) \leq B_3 e^{-A_3 n h^{q+4}}.$$

(c) The Hausdorff distance between the collection of local modes and its estimator satisfies

$$\texttt{Haus}\left(\mathcal{M}, \widehat{\mathcal{M}}_n\right) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right),$$

as $h \to 0$ and $nh^{q+2} \to \infty$.

The proof of Theorem 6 is in Appendix D.4. It states that asymptotically, the set of estimated local modes are close to the set of true local modes and there exists a $1-1$ mapping between pairs of estimated and true local modes. Thus, the local modes of the directional KDE are good estimators of the local modes of the population directional density.

Remark 7 *Unlike the statement of Theorem 1 by Chen et al. (2016), the radius ρ_* in (M2) for $\mathcal{M}$ to contain $\widehat{\mathcal{M}}_n$ can be selected to be independent of the dimension of the data. The reason lies in the fact that the proof of statement (a) in Theorem 6 performs a Taylor's expansion to the third order and leverages the constant upper bound for the third-order partial derivatives. The same technique can be used to improve the original proof in Theorem 1 of Chen et al. (2016) to obtain a dimension-free radius for mode consistency.*

5. Computational Learning Theory of Directional Mean Shift Algorithm

In this section, we study the algorithmic convergence of Algorithm 1. We start with the ascending property and convergence of Algorithm 1, and then prove the linear convergence of gradient ascent algorithms on the sphere Ω_q. By shrinking the bandwidth parameter, the adaptive step size of Algorithm 1 as a gradient ascent iteration on Ω_q can be sufficiently small so that the algorithm converges linearly to the estimated local modes around their neighborhoods. Finally, we discuss on the computational complexity of Algorithm 1.

5.1 Ascending Property and Convergence of Algorithm 1

Let $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$ be the path of successive points generated by Algorithm 1. The corresponding sequence of directional density estimates is given by

$$\widehat{f}_h(\widehat{\boldsymbol{y}}_s) = \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} L\left(\frac{1-\widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right) \quad \text{for } s = 0, 1, \ldots.$$

Theorem 8 (Ascending Property) *If kernel $L : [0,\infty) \to [0,\infty)$ is monotonically decreasing, differentiable, and convex with $L(0) < \infty$, then the sequence $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0}^{\infty}$ is monotonically increasing and thus converges.*

At a high level, the proof of Theorem 8 follows from the inequality

$$L(x_2) - L(x_1) \geq L'(x_1) \cdot (x_2 - x_1), \tag{29}$$

which is guaranteed by the convexity and differentiability of the kernel function L; see Appendix D.5 for details.

Remark 9 *Note that the differentiability of kernel L in Theorem 8 can be relaxed. The monotonicity and convexity of L already imply that L is differentiable except for a countable set of points $\mathcal{N}$ (see Section 6.2 and 6.6 in Royden and Fitzpatrick 2010). Moreover, the left and right derivatives of L on $\mathcal{N}$ exist and are finite. Therefore, for any $x_1 \in \mathcal{N}$, we can replace $L'(x_1)$ in (29) by any subgradient g_{x_1} without impacting other parts of the inequality. Furthermore, as the left or right derivatives of the convex function L are non-decreasing, any subgradient g_{x_1} at point x_1 satisfies $L'(x_1^-) \leq g_{x_1} \leq L'(x_1^+)$; thus, (29) holds.*

The ascending property of $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0}^{\infty}$ under the directional mean shift algorithm is not sufficient to guarantee the convergence of its iterative sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$. To derive the convergence of $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$, we make the following assumptions on the directional KDE $\widehat{f}_h$.

- **(C1)** The number of local modes of $\widehat{f}_h$ on Ω_q is finite, and the modes are isolated from other critical points.
- **(C2)** Given the current values of n and $h > 0$, we assume that $\widehat{\boldsymbol{m}}_k^T \nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k) \neq 0$ for all $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, that is, $\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i \cdot L'\left(\frac{1-\widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i}{h^2}\right) \neq 0$.

Condition (C1) is a weak condition when the uniform consistency (Theorem 4) and mode consistency (Theorem 6) are established. In reality, condition (C1) is implied by conditions (D1) and (M1-2) on f as well as (D2') and (K1) on the kernel function L with a probability tending to 1 as the sample size increases and the bandwidth parameter decreases accordingly.

Condition (C2) may look strange at first glance; however, it is a reasonable and common assumption. In practice, it will be valid with those commonly chosen kernel functions, a reasonable sample size n, and a properly tuned bandwidth parameter $h > 0$. More importantly, because the directional density f is always positive around its local mode, the following lemma demonstrates that condition (C2) holds with probability tending to 1 as the sample size increases to infinity and the bandwidth parameter tends to 0 accordingly.

Lemma 10 *Assume conditions (D1) and (D2'). For any fixed $\boldsymbol{x} \in \Omega_q$, we have*

$$h^2 \cdot \mathtt{Rad}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) \asymp h^2 \cdot \nabla \widehat{f}_h(\boldsymbol{x}) = \boldsymbol{x} f(\boldsymbol{x}) C_{L,q} + o(1) + O_P\left(\sqrt{\frac{1}{nh^q}}\right)$$

as $nh^q \to \infty$ and $h \to 0$, where $C_{L,q} = -\frac{\int_0^\infty L'(r) r^{\frac{q}{2}-1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} > 0$ is a constant depending only on kernel L and dimension q and "$\asymp$" stands for an asymptotic equivalence.

The proof of Lemma 10 can be found in Appendix D.5. With Lemma 10, we know that while the tangent component of $\nabla \widehat{f}_h$ at each local mode is 0, its radial component is diverging; thus, condition (C2) holds asymptotically. This is not a surprising result, because observations in a directional data sample are supported on the sphere and the directional KDE $\widehat{f}_h$ would thus decrease rapidly when moving away from the sphere. In addition, the limiting behavior of $\nabla \widehat{f}_h$ determines the adaptive step size of the directional mean shift algorithm when it approaches the estimated local modes (see Section 5.2 for details). A similar asymptotic behavior of the step size of the mean shift algorithm in the Euclidean setting has been noticed by Cheng (1995) and restated by Arias-Castro et al. (2016).

We now state the convergence of Algorithm 1 under conditions (C1) and (C2).

Theorem 11 *Assume (C1) and (C2) and the conditions on kernel L in Theorem 8. We further assume that L is continuously differentiable. Then, for each local mode $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, there exists a $\widehat{r}_k > 0$ such that the sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^\infty$ converges to $\widehat{\boldsymbol{m}}_k$ whenever the initial point $\widehat{\boldsymbol{y}}_0 \in \Omega_q$ satisfies $||\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{m}}_k||_2 \leq \widehat{r}_k$. Moreover, under conditions (D1) and (D2'), there exists a fixed constant $r^* > 0$ such that $\mathbb{P}(\widehat{r}_k \geq r^*) \to 1$ as $h \to 0$ and $nh^q \to \infty$.*

The proof of Theorem 11 is in Appendix D.5. The theorem implies that when we initialize the directional mean shift algorithm (Algorithm 1) sufficiently close to an estimated local mode, it will converge to this mode.

5.2 Linear Convergence of Gradient Ascent Algorithms on Ω_q

We now discuss the linear convergence of gradient ascent algorithms on Ω_q. Because the sphere Ω_q is not a conventional Euclidean space but a Riemannian manifold, the definition

of a gradient ascent update is more complex. We first provide a brief introduction to some useful concepts from differential geometry. The interested readers can consult Appendix B for additional details.

An *exponential map* at $\boldsymbol{x} \in \Omega_q$ is a mapping $\texttt{Exp}_{\boldsymbol{x}} : T_{\boldsymbol{x}} \to \Omega_q$ such that a vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$ is mapped to point $\boldsymbol{y} := \texttt{Exp}_{\boldsymbol{x}}(\boldsymbol{v}) \in \Omega_q$ with $\gamma(0) = \boldsymbol{x}, \gamma(1) = \boldsymbol{y}$ and $\gamma'(0) = \boldsymbol{v}$, where $\gamma : [0,1] \to \Omega_q$ is a geodesic. An intuitive way of thinking of the exponential map evaluated at $\boldsymbol{v}$ on the sphere Ω_q is that starting at point $\boldsymbol{x}$, we identify another point $\boldsymbol{y}$ on Ω_q along the great circle in the direction of $\boldsymbol{v}$ so that the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is $||\boldsymbol{v}||_2$.

The inverse of the exponential map is a mapping $\texttt{Exp}_{\boldsymbol{x}}^{-1} : U \subset \Omega_q \to T_{\boldsymbol{x}}$ such that $\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$ represents the vector in $T_{\boldsymbol{x}}$ starting at $\boldsymbol{x}$, pointing to $\boldsymbol{y}$, and with its length equal to the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. $\texttt{Exp}_{\boldsymbol{x}}^{-1}$ is sometimes called the logarithmic map.

On Ω_q, the notion of *parallel transport* provides a sensible way to transport a vector along a geodesic (Zhang and Sra, 2016). Intuitively, a tangent vector $\boldsymbol{v} \in T_{\boldsymbol{x}}$ at $\boldsymbol{x} \in \Omega_q$ of a geodesic γ is still a tangent vector $\Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{v}) \in T_{\boldsymbol{y}}$ of γ after being transported to point $\boldsymbol{y}$ along γ. Furthermore, parallel transport preserves inner products, that is, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{x}} = \langle \Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{u}), \Gamma_{\boldsymbol{x}}^{\boldsymbol{y}}(\boldsymbol{v}) \rangle_{\boldsymbol{y}}$. The above concepts can be defined on a general Riemannian manifold; however, for our purposes, it suffices to focus on the case of Ω_q.

Adopting the notation of Zhang and Sra (2016), a gradient ascent algorithm applied to an objective function f on Ω_q (a Riemannian manifold) is written as

$$\boldsymbol{y}_{s+1} = \texttt{Exp}_{\boldsymbol{y}_s} \left(\eta \cdot \texttt{grad}\, f(\boldsymbol{y}_s) \right). \tag{30}$$

Recall that given a sequence $\{\boldsymbol{y}_s\}_{s=0}^{\infty}$ converging to $\boldsymbol{m}_k \in \mathcal{M}$, the convergence is said to be linear if there exists a positive constant $\Upsilon < 1$ (rate of convergence) such that $||\boldsymbol{y}_{s+1} - \boldsymbol{m}_k|| \leq \Upsilon \, ||\boldsymbol{y}_s - \boldsymbol{m}_k||$ when s is sufficiently large (Boyd and Vandenberghe, 2004). In our context, the norm $||\cdot||$ refers to the geodesic (or great-circle) distance $d(\boldsymbol{x}, \boldsymbol{y}) = \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2$ between two points $\boldsymbol{x}, \boldsymbol{y} \in \Omega_q$. An equivalent statement of linear convergence is that the algorithm takes $O(\log(1/\epsilon))$ iterations to converge to an ϵ-error of $\widehat{\boldsymbol{m}}_k$.

Here, we first prove the linear convergence results for the gradient ascent algorithm with f and $\widehat{f}_h$ on Ω_q under a feasible range of step size η. We then demonstrate that the directional mean shift algorithm is an exemplification of the gradient ascent algorithm on Ω_q with an adaptive step size, and that its step size eventually falls into the feasible range with a properly tuned bandwidth parameter. Using the notation in Zhang and Sra (2016), we let $\zeta(1,c) \equiv \frac{c}{\tanh(c)}$. One can show by differentiating $\zeta(1,c)$ that $\zeta(1,c)$ is strictly increasing and $\zeta(1,c) > 1$ for any $c > 0$.

Theorem 12 *Assume (D1) and (M1).*

(a) ***Linear convergence of gradient ascent with*** f*: Given a convergence radius* r_0 *with* $0 < r_0 \leq \sqrt{2 - 2\cos\left[\frac{3\lambda_*}{2(q+1)^{\frac{3}{2}} C_3}\right]}$*, the iterative sequence* $\{\boldsymbol{y}_s\}_{s=0}^{\infty}$ *defined by the population-level gradient ascent algorithm* (30) *satisfies*

$$d(\boldsymbol{y}_s, \boldsymbol{m}_k) \leq \Upsilon^s \cdot d(\boldsymbol{y}_0, \boldsymbol{m}_k) \quad \textit{with} \quad \Upsilon = \sqrt{1 - \frac{\eta \lambda_*}{2}},$$

whenever $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)C_3\zeta(1,r_0)}\right\}$ *and the initial point* $\boldsymbol{y}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ *for some* $\boldsymbol{m}_k \in \mathcal{M}$. *We recall from Section 4.5 that* C_3 *is an upper bound for the derivatives of the directional density* f *up to the third order,* $\lambda_* > 0$ *is defined in (M1), and* $\mathcal{M}$ *is the set of local modes of the directional density* f.

We further assume (D2') and (K1) in the sequel.

(b) ***Linear convergence of gradient ascent with*** $\widehat{f}_h$: *Let the sample-based gradient ascent update on* Ω_q *be* $\widehat{\boldsymbol{y}}_{s+1} = \texttt{Exp}_{\boldsymbol{y}_s}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$. *With the same choice of the convergence radius* $r_0 > 0$ *and* $\Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}}$ *as in (a), if* $h \to 0$ *and* $\frac{nh^{q+2}}{|\log h|} \to \infty$, *then for any* $\delta \in (0,1)$,

$$d\left(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k\right) \leq \Upsilon^s \cdot d\left(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k\right) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$$

with probability at least $1 - \delta$, *whenever* $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)C_3 \cdot \zeta(1,r_0)}\right\}$ *and the initial point* $\widehat{\boldsymbol{y}}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ *for some* $\boldsymbol{m}_k \in \mathcal{M}$.

The proof of Theorem 12 is in Appendix D.6. As shown in (a) of Theorem 12, the linear convergence radius of gradient ascent algorithm (30) on Ω_q generally depends on the lower bound λ_* on absolute eigenvalues of the Riemannian Hessian $\mathcal{H}f(\boldsymbol{x})$ (within the tangent space $T_{\boldsymbol{x}}$), the upper bound C_3 for the (partial) derivatives of f up to the third order, and manifold dimension q.

In practice, we are more interested in the algorithmic convergence rate of sample-based gradient ascent algorithms with directional KDEs to the estimated local modes $\widehat{M}_n$. As indicated by Theorem 4, the Hessian matrices of $\widehat{f}_h$ at its local modes have only negative eigenvalues within the corresponding tangent spaces given (M1), sufficiently small h, and sufficiently large $\frac{nh^{q+4}}{|\log h|}$. In reality, unless the data configuration is highly abnormal, the local modes of directional KDEs are non-degenerate and $\widehat{f}_h$ is geodesically strongly concave (see Appendix B for a precise definition) around small neighborhoods of the estimated local modes. Together with an application of smooth kernels, says the von Mises kernel, $\widehat{f}_h$ is β-smooth on Ω_q and, consequently, a sample-based gradient ascent algorithm with the directional KDE $\widehat{f}_h$ converges linearly to the estimated local modes around their small neighborhoods, given a proper step size.

With respect to the directional mean shift algorithm, we recall from the fixed-point equation (23) that the geodesic distance between $\widehat{\boldsymbol{y}}_{s+1}$ and $\widehat{\boldsymbol{y}}_s$ (one-step iteration) is

$$\arccos\left(\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T \widehat{\boldsymbol{y}}_s}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right).$$

To derive the adaptive step size $\widehat{\eta}_s$ of the directional mean shift algorithm as a sample-based gradient ascent iteration $\widehat{\boldsymbol{y}}_{s+1} = \texttt{Exp}_{\widehat{\boldsymbol{y}}_s}\left(\widehat{\eta}_s \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$ on Ω_q, we notice the following

geodesic distance equation:

$$\left|\left|\widehat{\eta}_s \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2 = \arccos\left(\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T \widehat{\boldsymbol{y}}_s}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right).$$

This shows that the directional mean shift algorithm is a gradient ascent method on Ω_q with an adaptive step size

$$\widehat{\eta}_s = \arccos\left(\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)^T \widehat{\boldsymbol{y}}_s}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}\right) \cdot \frac{1}{\left|\left|\texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}$$

for $s = 0, 1, ...$ We denote the angle between the total gradient estimator $\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)$ and $\widehat{\boldsymbol{y}}_s$ by $\widehat{\theta}_s$. By the orthogonality of $\widehat{\boldsymbol{y}}_s$ and $\texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s) \equiv \texttt{Tang}\left(\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$, the expression for the adaptive step size $\widehat{\eta}_s$ becomes

$$\widehat{\eta}_s = \frac{\widehat{\theta}_s}{\left(\sin \widehat{\theta}_s\right) \cdot \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}$$

for $s = 0, 1, ...$ As the directional mean shift algorithm approaches a local mode of $\widehat{f}_h$, $\widehat{\theta}_s$ tends to 0 and $\frac{\widehat{\theta}_s}{\sin \widehat{\theta}_s}$ is approximately equal to 1. Thus, the step size $\widehat{\eta}_s$ is essentially controlled by the (Euclidean) norm of the total gradient estimator, that is, $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$. The larger the norm of $\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)$ at step s, the shorter the step size of Algorithm 1. Because the tangent component of $\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)$ is small around the estimated local modes, its radial component $\texttt{Rad}\left(\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$ dominates the norm $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$. Lemma 10 suggests that $\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2$ can be sufficiently large as the sample size increases to infinity and the bandwidth parameter decreases to 0 accordingly; therefore, one can always select a small bandwidth parameter h such that the step size $\widehat{\eta}_s$ lies within the feasible range for linear convergence. Algorithm 1 thus converges (at least) linearly around the local modes of the directional KDE $\widehat{f}_h$.

5.3 Computational Complexity

Given a fixed data set $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\} \subset \Omega_q$, the time complexity of Algorithm 1 is $O(n \times q)$ for one iteration of the algorithm on a single query point. When Algorithm 1 is applied to the entire data set as the set of query points, each iteration exhibits $O(n^2 \times q)$ time complexity. Assuming that the algorithm converges linearly, the total time complexity for reaching an ϵ-error is $O\left(n^2 \times q \times \log(1/\epsilon)\right)$. The space complexity of mode clustering with Algorithm 1 is, in general, $O(n \times q)$ if mode clustering is performed on the entire data set to estimate the directional density and only the current set of iteration points are stored in memory. Algorithm 1 inevitably faces a computational bottleneck or even inferior performance when the (intrinsic) dimension q is large. This drawback of the algorithm results not only from its time and space complexity, but also from its original dependency on nonparametric density estimation, which is known to suffer from *the curse of dimensionality*.

6. Experiments

In this section, we present our experimental results of the directional mean shift algorithm on both simulated and real-world data sets. Unless stated otherwise, we use the von Mises kernel $L(r) = e^{-r}$ in the directional KDE (2) to estimate the directional densities and their derivatives. Given the data sample $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$, the default bandwidth parameter is selected via the rule of thumb in Proposition 2 in García-Portugués (2013):

$$h_{\mathrm{ROT}} = \left[\frac{4\pi^{\frac{1}{2}} \mathcal{I}_{\frac{q-1}{2}}(\widehat{\nu})^2}{\widehat{\nu}^{\frac{q+1}{2}} \left[2q \cdot \mathcal{I}_{\frac{q+1}{2}}(2\widehat{\nu}) + (q+2)\widehat{\nu} \cdot \mathcal{I}_{\frac{q+3}{2}}(2\widehat{\nu}) \right] n} \right]^{\frac{1}{q+4}} \tag{31}$$

for $q \geq 1$, which is the optimal bandwidth for the directional KDE that minimizes the asymptotic mean integrated squared error when the underlying density is a von Mises-Fisher density and the von Mises kernel is applied. The estimated concentration parameter $\widehat{\nu}$ is given by (4.4) in Banerjee et al. (2005) as

$$\widehat{\nu} = \frac{\bar{R}(q + 1 - \bar{R}^2)}{1 - \bar{R}^2},$$

where $\bar{R} = \frac{\left|\left|\sum_{i=1}^n \boldsymbol{X}_i\right|\right|_2}{n}$ (see also Sra (2012) for a detailed discussion and experiments on the numerical approximation of the concentration parameter for von Mises-Fisher distributions). We also perform mode clustering (Chen et al., 2016) (sometimes called mean shift clustering in Fukunaga and Hostetler 1975; Cheng 1995) on the original data sets in our simulation studies, in which data points are assigned to the same cluster if they converge to the same (estimated) local mode. When such procedure is carried out on the entire data space, it partitions the space into different regions called *basins of attraction* of the (directional) KDE. As the true density component from which a data point is generated is known a priori in our simulation studies (i.e., we know the label of each observation), we also provide the misclassification rates or confusion matrices of mode clustering with the directional mean shift algorithm, though one should be aware that mode clustering, by its nature, embraces non-overlapping basins of attraction (Banyaga and Hurtubise, 2004; Chacón, 2015). Figures 6, 8, and 9 in this section as well as Figures 10 and 11 in Appendix A are plotted via the Matplotlib Basemap Toolkit (`https://matplotlib.org/basemap/`).

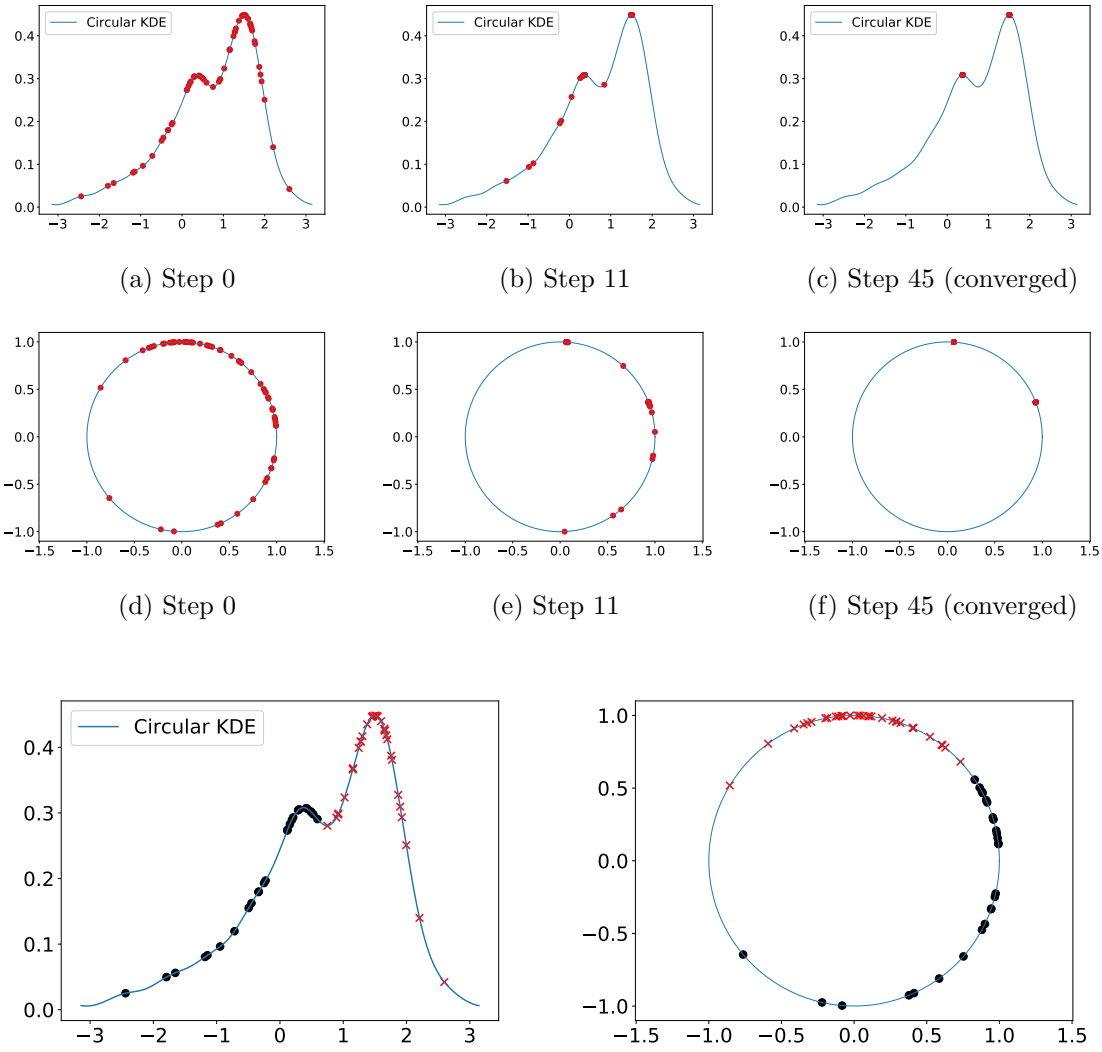
6.1 Simulation Studies

6.1.1 Circular Case

To evaluate the effectiveness of Algorithm 1, we first randomly generate 60 data points from a circular density

$$f_1(x) = \frac{e^{-|x|}}{4(1 - e^{-\pi})} \cdot \mathbb{1}_{[-\pi,\pi]}(x) + \frac{1}{4\pi \mathcal{I}_0(6)} \exp\left[6 \cos\left(x - \frac{\pi}{2}\right)\right],$$

which is a mixture of a Laplace density with mean 0 and scale 1 truncated to $[-\pi, \pi]$ and a von Mises density with mean $\frac{\pi}{2}$ and concentration parameter $\nu = 6$. The von Mises(-Fisher) distributed samples are generated via rejection sampling with the uniform distribution as



(a) Step 0 (b) Step 11 (c) Step 45 (converged)

(d) Step 0 (e) Step 11 (f) Step 45 (converged)

(g) Mode clustering (viewed on function values) (h) Mode clustering (viewed on Ω_1)

Figure 5: Directional mean shift algorithm performed on simulated data on Ω_1. **Panel (a)-(c):** Outcomes under different iterations of the algorithm. **Panel (d)-(f):** Corresponding locations of points in panels (a)-(c) on a unit circle. **Panel (g) and (h):** Visualization of the affiliations of data points after mode clustering.

the proposal density. The true local modes are 0 and $\frac{\pi}{2}$ in terms of angular representations or $(0, 0)$ and $(0, 1)$ in Cartesian coordinates. The directional KDE on the simulated data and directional mean shift iterations are presented in Figure 5. The bandwidth parameter here is selected as $h = 0.3 < h_{\text{ROT}} \approx 0.4181$ because the aforementioned rule of thumb h_{ROT} tends to be oversmoothing when the underlying density is not von Mises distributed. In addition, the tolerance level for terminating the algorithm is set to $\epsilon = 10^{-7}$.

Figure 5 empirically demonstrates the validity of Algorithm 1 on the unit circle Ω_1, in which all the simulated data points converge to the local modes of the circular density estimator. In addition, the misclassification rate in this simulation study is 0.1.

6.1.2 Spherical Case

We simulate 1000 data points from the following density

$$f_3(\boldsymbol{x}) = 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_1, \nu_1) + 0.3 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_2, \nu_2) + 0.4 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_3, \nu_3)$$

with $\boldsymbol{\mu}_1 \approx (-0.35, -0.61, -0.71)$, $\boldsymbol{\mu}_2 \approx (0.5, 0, 0.87)$, $\boldsymbol{\mu}_3 = (-0.87, 0.5, 0)$ (or $[-120°, -45°]$, $[0°, 60°]$, $[150°, 0°]$ in their precise spherical [longitude, latitude] coordinates), and $\nu_1 = \nu_2 = 8$, $\nu_3 = 5$. The bandwidth parameter is selected using (31), and the tolerance level for terminating the algorithm is again set to $\epsilon = 10^{-7}$. The results are presented in Figure 6.

In Figure 6, all simulated data points converge to the local modes of the estimated directional density under the application of Algorithm 1; therefore, all the original data points are clustered according to where they converge. The confusion matrix in this simulation study is $\begin{bmatrix} \mathbf{278} & 0 & 9 \\ 0 & \mathbf{323} & 1 \\ 20 & 8 & \mathbf{361} \end{bmatrix}$ and the misclassification rate is thus 0.038. Moreover, the total number of iterative steps is much lower than the case with a single mode in Appendix A.1 (Figure 10). We also observe that most of data points already converge to the local modes of the directional KDE after a few initial steps, while most of the subsequent iterations handle those points that are geodesically far away from an estimated local mode and have small estimated (tangent/Riemannian) gradients on their iterative paths.

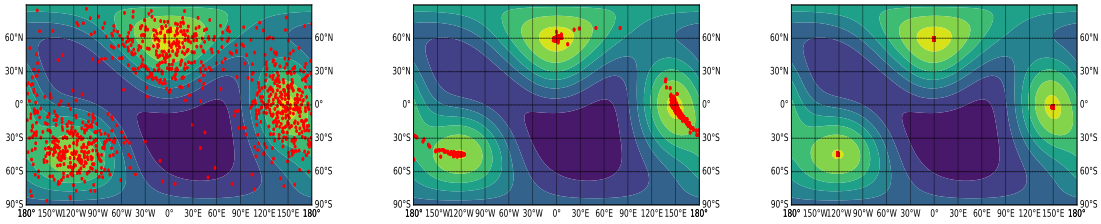
6.1.3 q-Directional Case with $q > 2$

Our algorithmic formulation of the directional mean shift algorithm (Algorithm 1) and its associated learning theory are valid on any general (intrinsic) dimension q of Ω_q. For this reason, we are also interested in how the algorithm behaves as the dimension q of directional data increases. We randomly simulate 1000 data points from each of the following densities repeatedly,

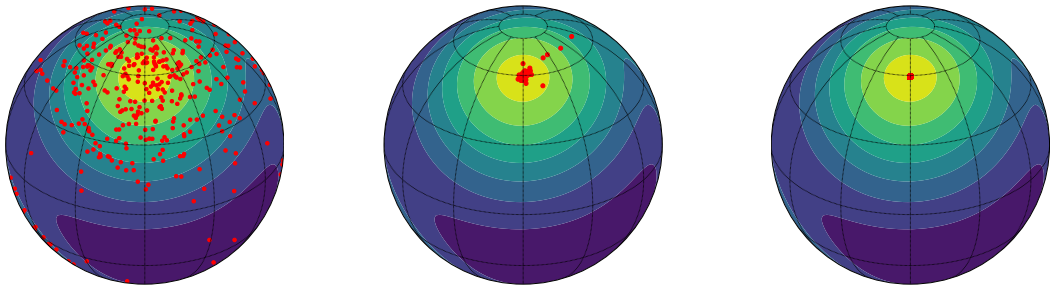
$$\sum_{i=1}^{4} 0.25 \cdot f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}_{i,q}, \nu')$$

with $\boldsymbol{\mu}_{i,q} = \boldsymbol{e}_{i,q+1} \in \Omega_q \subset \mathbb{R}^{q+1}$ for $q = 3, 4, ..., 12$ and $i = 1, ..., 4$, where the concentration parameter $\nu' = 10$ and the mixture weight of each density component are constant. Here, $\{\boldsymbol{e}_{i,q+1}\}_{i=1}^{q+1} \subset \Omega_q$ is the standard basis of the ambient Euclidean space $\mathbb{R}^{q+1}$. For each value of dimension q, we repeat the data simulation process 20 times and compute the average misclassification rate of mode clustering with Algorithm 1 on each simulated data set accordingly. Figure 7 shows the boxplots of misclassification rates.

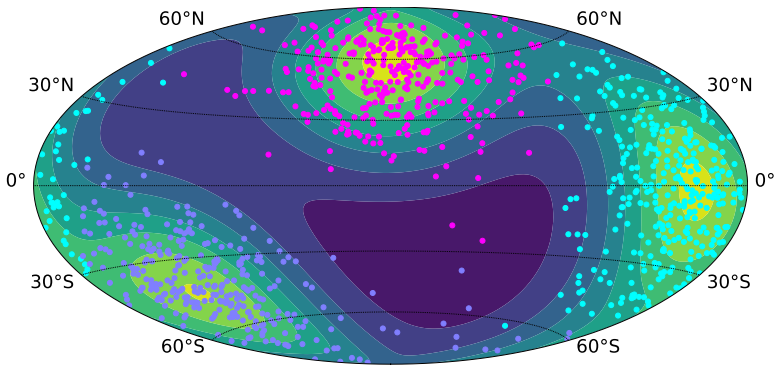
As the dimension q of directional data becomes larger, the misclassification rates of mode clustering with Algorithm 1 also gradually increase to 1 (the worst case), which in turn implies that the ability of Algorithm 1 to identify the density component from which a data point is simulated tends to deteriorate with respect to the dimension. Such inferior performances of the directional mean shift algorithm on higher-dimensional hyperspheres



(a) Step 0 (b) Step 5 (c) Step 28 (converged)



(d) Step 0 (e) Step 5 (f) Step 28 (converged)



(g) Mode clustering (Hammer projection view)

Figure 6: Directional mean shift algorithm performed on simulated data with three local modes on Ω_2. **Panel (a)-(c):** Outcomes under different iterations of the algorithm displayed in a cylindrical equidistant view. **Panel (d)-(f):** Corresponding locations of points in panels (a)-(c) in an orthographic view. Note: two local modes are at the back of the sphere; thus, we cannot directly see them. **Panel (g):** Clustering result under the Hammer projection (page 160 in Snyder et al. 1989).

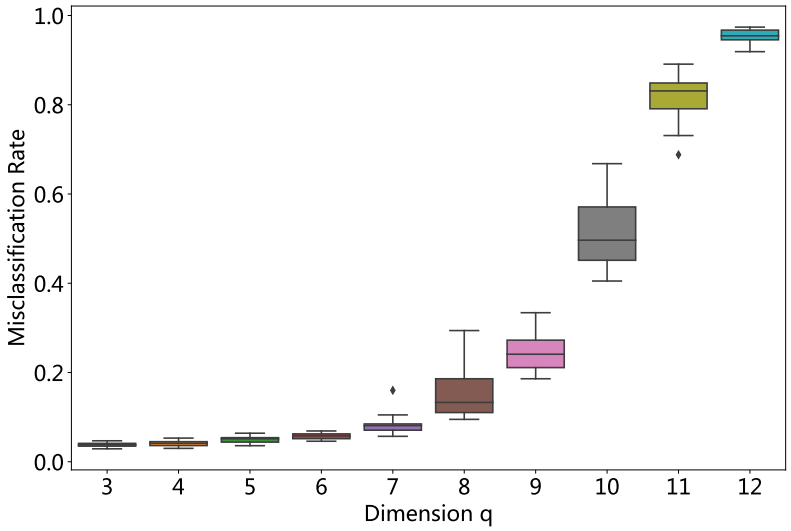


Figure 7: Boxplots of misclassification rates of mode clustering under different values of dimension q

are not surprising because (i) we do not fine-tune the bandwidth parameter (but simply apply the rule of thumb (31)), and (ii) the algorithm is subject to the curse of dimensionality (see also Section 5.3). However, since directional data in real-world applications mostly lie on a (hyper)sphere with dimension $q \leq 3$, Algorithm 1 is effective in practice, as we will demonstrate in Section 6.2.

6.2 Real-World Applications

We illustrate the practical relevance of the directional mean shift algorithm (Algorithm 1) with two applications in astronomy and seismology.

6.2.1 Craters on Mars

The distribution and cluster configuration of craters on Mars shed light on the planetary subsurface structure (water or ice), relative surfaces ages, resurfacing history, and past geologic processes (Cabrol and Grin, 2010; Barlow, 2015). García-Portugués et al. (2020) conducted three different statistical tests (Cramer-von Mises, Rothman, and Anderson-Darling-like tests) on Martian crater data to statistically validate the non-uniformity of the crater distribution on Mars. We apply the directional KDE (2) together with the directional mean shift algorithm to further estimate the density of craters and determine crater clusters on the surface of Mars. Martian crater data are publicly available on the Gazetteer of Planetary Nomenclature database (`https://planetarynames.wr.usgs.gov/AdvancedSearch`) of the International Astronomical Union (IUA). The positions of craters are recorded in areocentric coordinates (the planetocentric coordinates on Mars) so that the areocentric longitudes range from 0° to 360° and areocentric latitudes range from -90° to

90°. As craters with areocentric longitudes greater than 180° are on the western hemisphere of Mars, we transform their longitudes back to the interval $(-180°, 0°)$. (Note that 360° in longitude corresponds to 0° west/east after transformation.) In addition, we remove those small craters whose diameters are less than 5 kilometers from the crater data, as their presence may provide spurious information. After trimming, the data set contains 1653 craters. The bandwidth parameter is selected using (31), which becomes $h_{\mathrm{ROT}} \approx 0.338$ for the trimmed data set. The estimated distribution of craters on Mars and clustering results are presented in Figure 8.

As illustrated in Figure 8, the directional mean shift algorithm is capable of recovering the local modes of the estimated Martian crater density. Because we do not properly tune the bandwidth parameter, there is a spurious local mode around $(180°W, 30°S)$. Nevertheless, the mode clustering based on Algorithm 1 succeeds in capturing two major crater clusters (or basins of attraction) on Mars, in which one cluster is densely cratered while the other is lightly catered. This finding aligns with prior research on the Martian crater distribution, stating that Mars can be divided into two general classes of terrain (Soderblom et al., 1974). In Appendix A.2, we perform mode clustering with various smoothing bandwidths to illustrate multi-scale structures in the data.

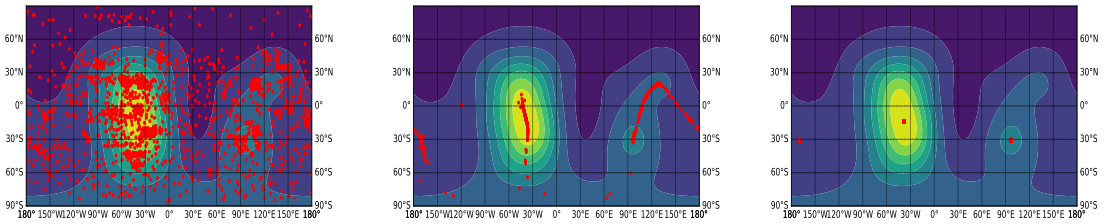
6.2.2 Earthquakes on Earth

Earthquakes on Earth tend to occur more frequently in some regions than others. We again leverage the directional KDE (2) as well as the directional mean shift algorithm to analyze earthquakes with magnitudes of 2.5+ occurring between 2020-08-21 00:00:00 UTC and 2020-09-21 23:59:59 UTC around the world. The earthquake data can be obtained from the Earthquake Catalog (`https://earthquake.usgs.gov/earthquakes/search/`) of the United States Geological Survey. The data set contains 1666 earthquakes worldwide for this one-month period. We use the default bandwidth estimator (31), which yields $h_{\mathrm{ROT}} \approx 0.245$ on the earthquake data set, and set the tolerance level to $\epsilon = 10^{-7}$ throughout the analysis.

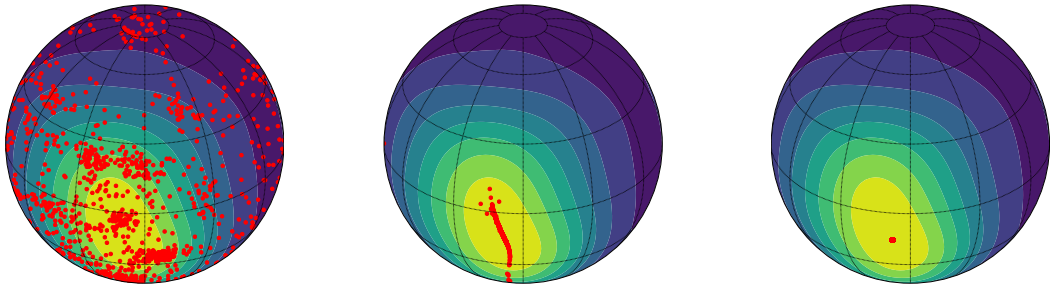
Figure 9 displays the results. There are seven local modes recovered by the directional mean shift algorithm, and they are located near (from left to right and top to bottom in Panel (g) of Figure 9) the Gulf of Alaska, the west side of the Rocky Mountain in Nevada, the Caribbean Sea, the west side of the Andes mountains in Chile, the Middle East, Indonesia, and Fiji. These regions are well-known active seismic areas along subduction zones, and our clustering of earthquake data elegantly partitions earthquakes into these regions without any prior knowledge from seismology.

7. Conclusion

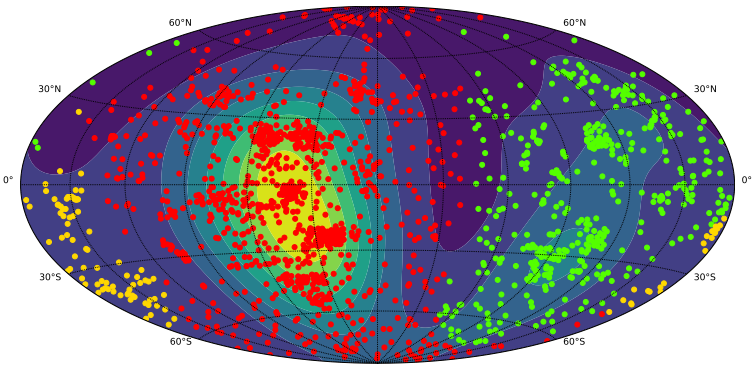
In this paper, we generalize the standard mean shift algorithm to directional data based on a total gradient (or differential) of the directional KDE and formulate it as a fixed-point iteration. We derive the explicit forms of the (Riemannian) gradient and Hessian estimators from a general directional KDE and establish pointwise and uniform rates of convergence for the two derivative estimators. With these powerful uniform consistency results, we demonstrate that the collection of estimated local modes obtained by the directional mean shift algorithm is a statistically consistent estimator of the set of true local modes under



(a) Step 0 (b) Step 17 (c) Step 178 (converged)

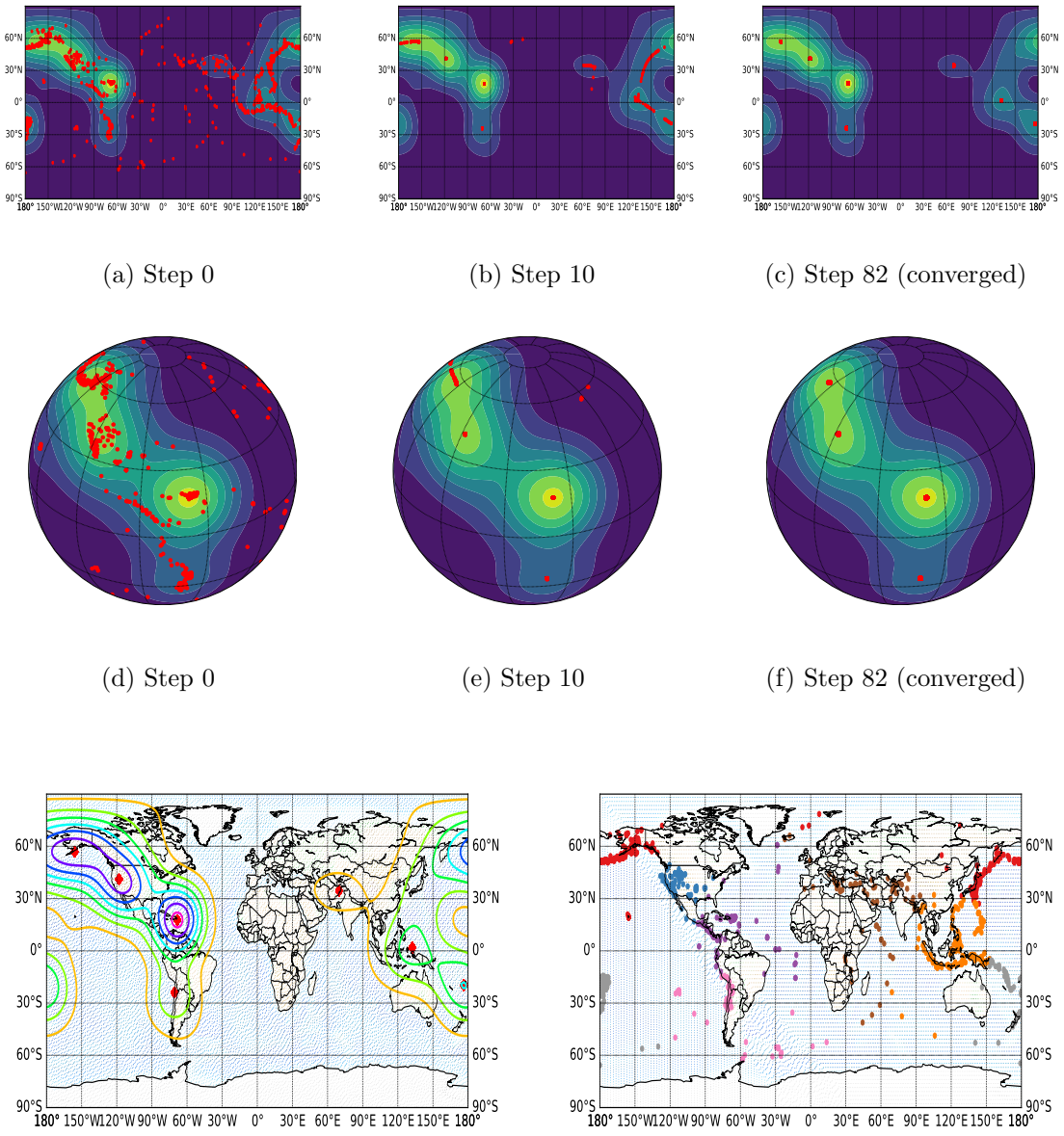


(d) Step 0 (e) Step 17 (f) Step 178 (converged)



(g) Mode clustering (Hammer projection view)

Figure 8: Directional mean shift algorithm performed on Martian crater data. The analysis is displayed in a similar way to Figure 6. **Panel (a)-(c):** Outcomes under different iterations of the algorithm displayed in a cylindrical equidistant view. **Panel (d)-(f):** Corresponding locations of points in panels (a)-(c) in an orthographic view. **Panel (g):** Clustering result under the Hammer projection.



(a) Step 0 (b) Step 10 (c) Step 82 (converged)

(d) Step 0 (e) Step 10 (f) Step 82 (converged)

(g) Local modes and contour lines on the world map (h) Mode clustering on the world map

Figure 9: Directional mean shift algorithm performed on earthquake data for a one-month period. The first two rows display the analysis similar to Figure 6. **Panel (a)-(c):** Outcomes under different iterations of the algorithm displayed in a cylindrical equidistant view. **Panel (d)-(f):** Corresponding locations of points in panels (a)-(c) in an orthographic view. **Panel (g):** Contour plots of estimated density. **Panel (h):** Clustering result using the directional mean shift algorithm.

some mild regularity conditions. Additionally, the ascending property and convergence of the proposed algorithm are proved. Finally, given a proper bandwidth parameter (or step size for other general gradient ascent algorithms on Ω_q), we argue that the directional mean shift algorithm (or other general gradient ascent algorithms on Ω_q) converge(s) linearly to the (estimated) local modes within their small neighborhoods regardless of whether a population-level or sample-based version of gradient ascent is applied.

Possible future extensions of our work are as follows.

- **Bandwidth Selection**. Current studies on bandwidth selectors for directional kernel smoothing settings primarily optimize the directional KDE itself. Research on bandwidth selection for derivatives of the directional KDE, especially gradient and Hessian estimators, has lagged behind. A well-designed bandwidth selector for the first-order derivatives of the directional KDE can further improve the algorithmic convergence rate of our algorithm in real-world applications. There are at least two common approaches to perform such bandwidth selection. The first is to calculate the explicit forms of dominating constants in the bias and stochastic variation terms when we derive pointwise convergence rates of the (Riemannian) gradient and Hessian estimators in Theorem 2. Then, under some assumptions on the underlying directional distribution, such as the von Mises Fisher distribution, a directional analogue to the rule of thumb of Silverman (1986) for gradient and Hessian estimators can be explicated, although the calculations may be heavy. Another approach is to rely on data-adaptive methods, such as cross-validation (Hall et al., 1987) and bootstrap (Marzio et al., 2011; Saavedra-Nieves and Crujeiras, 2020), which should be suitable for estimating the derivatives of the directional KDE. In addition, a bandwidth selector that is locally adaptive to the distribution of directional data is of great significance when the dimension is high.

- **Accelerated Directional Mean Shift**. Another future direction is to accelerate the current directional mean shift algorithm when the sample size is large, as the number of iterations for convergence is over 150 in one of our real-world applications. There are several feasible approaches mentioned in Section 1 for the Euclidean mean shift algorithm. One of the most notable methods is the blurring procedure (Carreira-Perpiñán, 2006, 2008), in which the (Gaussian) mean shift algorithm is performed with a crucial modification that successively updates the data set for density estimation after each mean shift iteration. It has been demonstrated that the blurring procedure improves the convergence rate of the (Gaussian) mean shift algorithm to be cubic or even higher order with Gaussian clusters and an appropriate step size. We present preliminary results of introducing blurring procedures into the directional mean shift algorithm with the von-Mises kernel in Appendix A.3, where the blurring procedures are able to substantially reduce the total number of iterations. However, in addition to those valid estimated local modes identified by the original directional mean shift algorithm, the blurring version also recovers some spurious local mode estimates (see Table 1 in Appendix A.3 for additional details). Because the current stopping criterion applied in the blurring directional mean shift algorithm is adopted from the criterion for Gaussian blurring mean shift (Carreira-Perpiñán, 2006), we plan to develop an

improved stopping criterion for the blurring directional mean shift algorithm and investigate its rate of convergence.

- **Connections to the EM Algorithm**. As pointed out by Carreira-Perpiñán (2007), the Gaussian mean shift algorithm for Euclidean data is an EM algorithm, while the mean shift algorithm with a non-Gaussian kernel is a generalized EM algorithm. It is unclear whether the directional mean shift algorithm with the von Mises kernel is also an EM algorithm on a mixture of von Mises-Fisher distributions on Ω_q (Banerjee et al., 2005) or even a generalized EM algorithm when other kernels are used in Algorithm 1. Bridging this connection can help establish the linear rate of convergence for the algorithm from a different angle.

Acknowledgments

We thank the anonymous reviewers and AE for their constructive comments that improved the quality of this paper. We also thank members in the Geometric Data Analysis Reading Group at the University of Washington for their helpful comments. YC is supported by NSF DMS - 1810960 and DMS - 1952781, NIH U01 - AG0169761.

Appendix A. Additional Experimental Results

A.1 Spherical Case with One Mode (Simulation Study)

We simulate 1000 data points from the density $f_2(\boldsymbol{x}) = f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \nu)$ with $\boldsymbol{\mu} = (1, 0, 0)$ and $\nu = 5$. The bandwidth parameter is selected using (31) and the tolerance level for terminating the algorithm is set to $\epsilon = 10^{-7}$. As presented in Figure 10, all the simulated data points converge to the mode of the estimated directional density except for a small portion of outliers. In addition, the misclassification rate in this example is 0.011, because there are some spurious local modes in the low density region. The total number of iterative steps in this case is greater than the case with three local modes in Figure 6.

A.2 Additional Mode Clustering Results on the Martian Crater Data

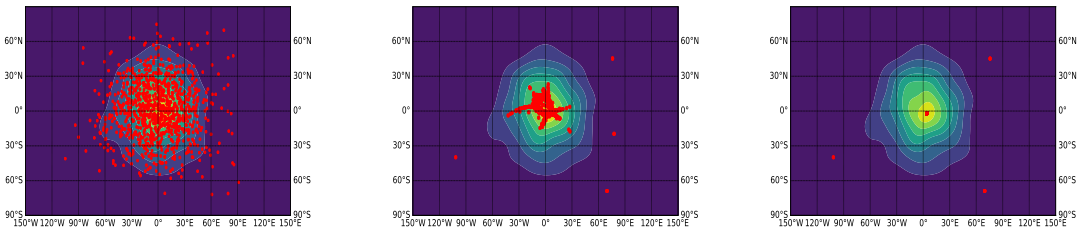
We varies the bandwidth parameter h from 0.1 to 0.6 with a step size 0.05 when conducting mode clustering on the trimmed Martian crater data set. The tolerance level for stopping Algorithm 1 is set to $\epsilon = 10^{-7}$. The number of crater clusters on Mars yielded from Algorithm 1 ranges from 37 when $h = 0.1$ to 1 when $h = 0.6$ in Figure 11. Those small crater clusters yielded by the directional mean shift algorithm with a small bandwidth parameter are of practical significance, since it may give astronomers more insight into the planetary subsurface structure and geologic processes on Mars.

A.3 Preliminary Experiments on Blurring Directional Mean Shift Algorithm with the von-Mises Kernel

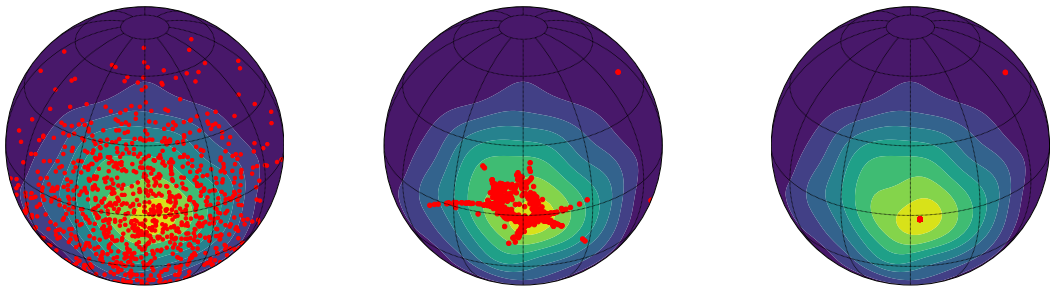
We randomly generate 1000 data points from von Mises-Fisher distributions with one ([1,0,0]), two ([0,1,0], [0,0,1]), and three ([0,1,0], [1,0,0], [0,-1,0]) true local modes via rejection sampling, respectively. Both the original directional mean shift algorithm with the von Mises-Fisher kernel and its blurring version are implemented on these simulated data sets. The stopping criterion for the blurring directional mean shift algorithm is adopted from the Gaussian Blurring mean shift algorithm with Euclidean data (Carreira-Perpiñán, 2006), that is,

$$\left(\left|H\left(\boldsymbol{e}^{(s+1)}\right) - H\left(\boldsymbol{e}^{(s)}\right)\right| \leq 10^{-8}\right) \quad \text{OR} \quad \left(\frac{1}{n}\sum_{i=1}^{n} e_i^{(s+1)} \leq \epsilon\right),$$

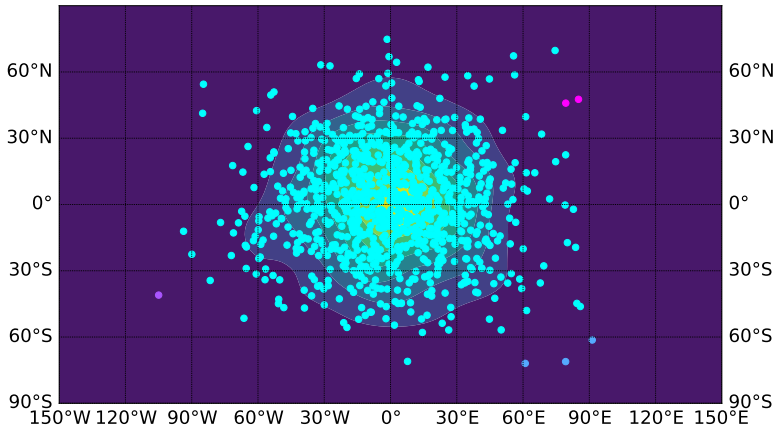
where $\boldsymbol{e}^{(s)} = (e_1^{(s)}, ..., e_n^{(s)})$, $e_i^{(s)} = \left|\left|\widehat{\boldsymbol{y}}_i^{(s)} - \widehat{\boldsymbol{y}}_i^{(s+1)}\right|\right|_2$, $\left\{\widehat{\boldsymbol{y}}_i^{(0)}\right\}_{i=1}^{n} = \{\boldsymbol{X}_i\}_{i=1}^{n}$ is the original data set, $H(\boldsymbol{e}) = -\sum_{i=1}^{B} f_i \log f_i$ is the entropy, f_i is the relative frequency of bin i (so $\sum_{i=1}^{B} f_i = 1$), and the bins span the interval $[0, \max(\boldsymbol{e})]$. The number of bins B is chosen as $B = 0.9n$, where n is the number of data points in the original data set. Among all the experiments, the bandwidth parameter is selected using the rule of thumb (31). The tolerance level is set to $\epsilon = 10^{-7}$. The repeated experimental results are recorded in Table 1, where the column "Avg. Err. of Est. Modes" presents the average distances between all the estimated local modes (identified by the original directional mean shift algorithm) and the nearest local mode estimates yielded by the blurring directional mean shift algorithm. As shown by Table 1, the blurring procedure is able to substantially reduce the total number of iterations for the directional mean shift algorithm with the von-Mises kernel. However,



(a) Step 0 (b) Step 19 (c) Step 79 (converged)

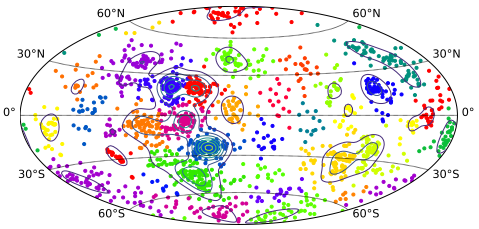


(d) Step 0 (e) Step 19 (f) Step 79 (converged)

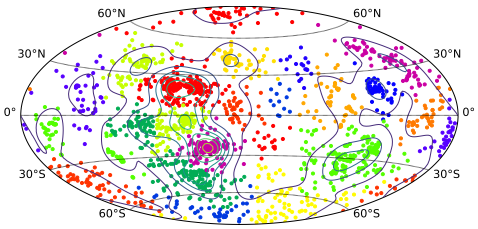


(g) Mode clustering (Cylindrical equidistant view)

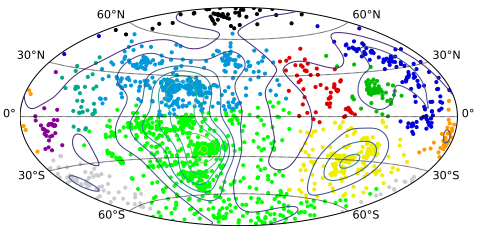
Figure 10: Directional mean shift algorithm performed on simulated data with one mode on Ω_2. The analysis is displayed similar to Figure 6. **Panel (a)-(c):** Outcomes under different iterations of the algorithm displayed in a cylindrical equidistant view. **Panel (d)-(f):** Corresponding locations of points in panels (a-c) in an orthographic view. **Panel (g):** Clustering result in a cylindrical equidistant view.



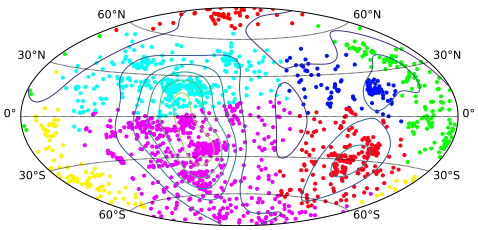
(a) 37 clusters (Step 137 (converged), $h = 0.1$)



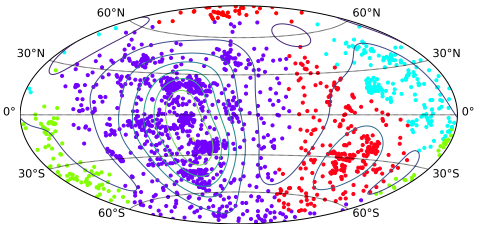
(b) 24 clusters (Step 62 (converged), $h = 0.15$)



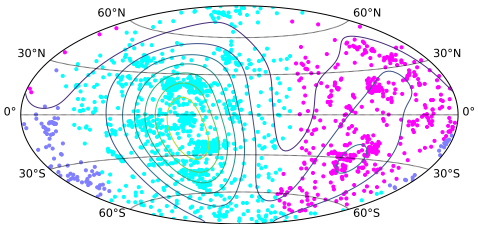
(c) 11 clusters (Step 135 (converged), $h = 0.2$)



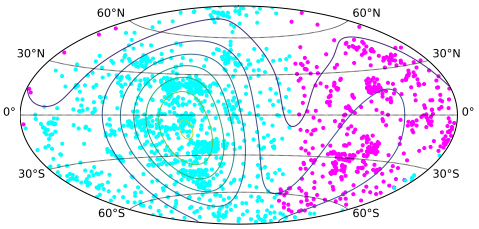
(d) 7 clusters (Step 234 (converged), $h = 0.25$)



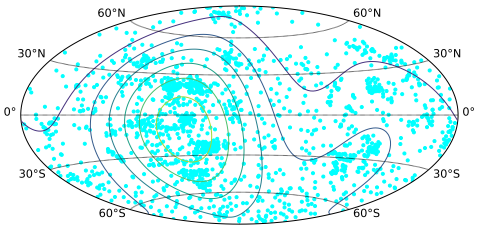
(e) 5 clusters (Step 237 (converged), $h = 0.3$)



(f) 3 clusters (Step 118 (converged), $h = 0.35$)



(g) 2 clusters (Step 155, 68, 57, and 61 (both converged) when $h = 0.4, 0.45, 0.5, 0.55$). The plot here is for $h = 0.4$.



(h) 1 cluster (Step 69 (converged), $h = 0.6$)

Figure 11: Directional mean shift algorithm with various bandwidth parameters performed on Martian crater data. The figures are visualized in their Hammer projections.

Method (Scenario)	# Est. Modes	# Steps	Avg. Err. of Est. Modes
DMS (One mode)	4.25 (1.670)	86.30 (48.774)	–
BDMS (One mode)	11.95 (2.156)	17.10 (2.700)	0.074 (0.0492)
DMS (Two modes)	2.40 (0.490)	30.55 (5.757)	–
BDMS (Two modes)	3.60 (1.114)	9.90 (1.868)	0.045 (0.0240)
DMS (Three modes)	3.00 (0.000)	28.65 (5.790)	–
BDMS (Three modes)	3.10 (0.300)	7.75 (0.698)	0.034 (0.0090)

Table 1: Comparisons between the Directional Mean Shift (DMS) and Blurring Directional Mean Shift (BDMS) algorithms. The means and standard errors (within round brackets) are calculated with 20 repeated experiments.

besides those valid estimated local modes identified by the original directional mean shift algorithm, the blurring version also recovers some spurious local mode estimates. The number of spurious local mode estimates from the blurring directional mean shift algorithm tends to decrease as the number of true local modes increases, given that the true local modes are well-separated. It illuminates a promising avenue to further accelerate the directional mean shift algorithm if a more delicate stopping criterion is designed.

Appendix B. Review of Geometry of Riemannian Manifolds

• **(Riemannian) Manifold**. A m-dimensional *manifold* $M \subset \mathbb{R}^D$ with $D > m$ is a second countable Hausdorff space where each point has a neighborhood that is homeomorphic to the m-dimensional Euclidean space. For each point $p \in M$, it is possible to define a *coordinate chart* (U, φ) centered at p as a homeomorphism $\varphi : U \to \varphi(U) \subset \mathbb{R}^m$, where U is an open subset of M containing p. Somewhat informally, if two coordinate charts (U, φ) and (V, ψ) are smoothly compatible, that is, either $U \cap V = \emptyset$ or the transition map $\psi \circ \varphi : \varphi(U \cap V) \to \psi(U \cap V)$ is a diffeomorphism, then M is a smooth manifold. See Chapter 1 in Lee (2012) for more formal definitions and discussions on smooth manifolds. A Riemannian manifold $(M, \mathfrak{g})$ is a real smooth manifold equipped with an inner product $\mathfrak{g}_p$ on the *tangent space* $T_p(M)$ of every point $p \in M$, such that if u, v are two vector fields on M then $p \mapsto \langle u, v \rangle_p := \mathfrak{g}_p(u, v)$ is a smooth function.

• **Curvature**. The curvature of a Riemannian manifold is characterized by its Riemannian metric tensor at each point. *Sectional curvature* is the Gaussian curvature of a two dimensional submanifold formed as the image of a two-dimensional subspace of a tangent space after exponential mapping. See Section 3-2 in Do Carmo (2016) for detailed discussions on the Gaussian curvature. It is known that a two-dimensional submanifold with positive, zero, or negative sectional curvature is locally isometric to a two-dimensional sphere, a Euclidean plane, or a hyperbolic plane with the same Gaussian curvature (Zhang and Sra, 2016).

• **Differential**. Given a smooth m-dimensional manifold M, the *differential* (or *total gradient*) of a smooth function $f : U \subset M \to \mathbb{R}$ at $p \in U$ is defined as a linear map

$$df_p : T_p(M) \to T_{f(p)}(\mathbb{R}) \simeq \mathbb{R},$$

where U is an open subset of M, $T_p(M)$ is the tangent space of M at p, and $V_1 \simeq V_2$ means that these two vector spaces are isomorphic. Other commonly used notations for the differential are: $df_p(v) = v(f)(p) = v_p(f) = (v \cdot f)(p)$ for $v \in T_p(M)$. See Section 2-4 in Do Carmo (2016) and Section 3.1 in Banyaga and Hurtubise (2004) for more details.

With an inner product structure on tangent spaces and the definition of differentials, one can define the gradient of a smooth function f on M.

Definition 13 (Riemannian Gradient) *The (Riemannian)* gradient *of a smooth function $f : M \to \mathbb{R}$ is a differentiable map $\texttt{grad}\, f : M \to \mathcal{T}M$ which assigns to each point $p \in M$ a vector $\texttt{grad}\, f(p) \in T_p(M) \subset \mathbb{R}^D$ such that*

$$\langle \texttt{grad}\, f(p), v\rangle_p = df_p(v) \quad \text{for all } v \in T_p(M). \tag{32}$$

Here $\mathcal{T}M$ is the tangent bundle, that is, the disjoint union of the tangent spaces at all points of M.

In terms of the following definition, the Hessian matrices on a manifold are only well-defined at critical points, that is, those points whose differentials vanish, though an extension of the definition to non-critical points is possible.

Definition 14 (Riemannian Hessian) *The* Hessian $\mathcal{H}_p f$ *of a smooth function $f : M \to \mathbb{R}$ at a critical point p is a symmetric bilinear map*

$$\mathcal{H}_p f : T_p(M) \times T_p(M) \to \mathbb{R}$$

defined as follows. For any tangent vectors $v, w \in T_p(M)$, we choose extensions $\widetilde{v}$ and $\widetilde{w}$ to vector fields on an open neighborhood of p and set

$$\mathcal{H}_p f(v, w) = \left(\widetilde{v} \cdot \left(\widetilde{w} \cdot f\right)\right)(p) = v_p\left(\widetilde{w} \cdot f\right).$$

The expression above is independent *of the extensions $\widetilde{v}$ of v and $\widetilde{w}$ of w, since*

$$\widetilde{v} \cdot \left(\widetilde{w} \cdot f\right)(p) - \widetilde{w} \cdot \left(\widetilde{v} \cdot f\right)(p) = [\widetilde{v}, \widetilde{w}]_p(f) = 0$$

at a critical point p, where $[\widetilde{v}, \widetilde{w}]_p$ is the commutator (or Lie bracket) of $\widetilde{v}$ and $\widetilde{w}$ at the point p. Thus, $\mathcal{H}_p f$ is a well-defined symmetric bilinear form on $T_p(M)$ at the critical point p.

Remark 15 *Note that in general, $\widetilde{v} \cdot \left(\widetilde{w} \cdot f\right)(p)$ and $\widetilde{w} \cdot \left(\widetilde{v} \cdot f\right)(p)$ might be of different values when p is not a critical point. This is essentially the definition of the vector $[\widetilde{v}, \widetilde{w}]_p = \widetilde{v} \cdot \left(\widetilde{w} \cdot f\right)(p) - \widetilde{w} \cdot \left(\widetilde{v} \cdot f\right)(p)$.*

Given a coordinate chart (U, φ) around $p \in M$, $\left\{\frac{\partial}{\partial x_1}\big|_p, ..., \frac{\partial}{\partial x_m}\big|_p\right\}$ forms a basis for $T_p(M)$, and the matrix of $\mathcal{H}_p f$ with respect to this basis can be expressed by the $m \times m$ matrix of second partial derivatives:

$$Q_p f := \left(\frac{\partial^2 (f \circ \phi^{-1})}{\partial x_i \partial x_j}\phi(p)\right).$$

It is possible to extend the definition of Hessian matrices of a smooth function $f : M \to \mathbb{R}$ to non-critical points based on the current definition (Milnor, 1963). Given a local coordinate chart (U, φ) near a non-critical point q and $v = \sum_{i=1}^{m} a_i \frac{\partial}{\partial x_i}\big|_q$, $w = \sum_{j=1}^{m} b_j \frac{\partial}{\partial x_j}\big|_q$, we take $\widetilde{w} = \sum_{j=1}^{m} b_j \frac{\partial}{\partial x_j}\big|_q$, where b_j now denotes a constant function. Then

$$\mathcal{H}_q f(v, w) = v(\widetilde{w}(f))(q) = v\left(\sum_{j=1}^{m} b_j \frac{\partial f}{\partial x_j}\Big|_q\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i b_j \frac{\partial^2 f}{\partial x_i \partial x_j}(q);$$

so the matrix $\left(\frac{\partial^2 f}{\partial x_i \partial x_j}(q)\right)_{i,j=1}^{m}$ represents the bilinear function $\mathcal{H}_q f$ with respect to the basis $\frac{\partial}{\partial x_1}\big|_q, ..., \frac{\partial}{\partial x_m}\big|_q$. Another feasible avenue to define the Hessian on a Riemannian manifold starts from the notion of Riemannian gradient (Definition 13) and covariant derivative (or affine connection). See Absil et al. (2013) for more details.

Definition 16 (Non-degenerate Critical Points and Morse Functions (Definition 3.1 in Banyaga *A critical point $p \in M$ of a differentiable function $f : M \to \mathbb{R}$ is* non-degenerate *if the Hessian $\mathcal{H}_p f$ is non-degenerate. In other words, the determinant of $Q_p f$ is non-zero. Otherwise, p is a* degenerate *critical point. A differentiable function on M is a* Morse function *if all its critical points are non-degenerate.*

A standard result for a Morse function on a finite dimensional compact smooth manifold M, including Ω_q, is that it has a finite number of critical points (Corollary 3.3 in Banyaga and Hurtubise 2004). Another remarkable fact in Morse theory is that integral curves on M (equivalently, gradient ascent paths with infinitely small step sizes) never intersect except at critical points, so they partition the space (Morse, 1925, 1930; Banyaga and Hurtubise, 2004). It thus serves as the backbone of mode clustering (Chen et al., 2016). We have presented some mode clustering results on Ω_q using both synthetic and real-world data in Section 6.

B.1 Function Classes on Riemannian Manifolds

The key definitions in this subsection are modified from Section 2 in Zhang and Sra (2016).

Definition 17 (Geodesic Concavity) *A function $f : M \to \mathbb{R}$ is said to be* geodesically concave *(or g-concave) if for any $p, q \in M$, a geodesic γ such that $\gamma(0) = p$ and $\gamma(1) = q$, and $t \in [0, 1]$, it holds that*

$$f(\gamma(t)) \geq (1 - t)f(p) + tf(q).$$

Equivalently, it can be shown that there exists a tangent vector $g_p \in T_p(M)$ such that

$$f(q) \leq f(p) + \langle g_p, \mathtt{Exp}_p^{-1}(q)\rangle_p, \tag{33}$$

where g_p is called a subgradient *of f at p, or the* gradient *if f is differentiable, and $\langle \cdot, \cdot \rangle_p$ denotes the inner product in the tangent space of p induced by the Riemannian metric.*

See, for instance, Section 1.2 in Bubeck (2015) for the definition of subgradients of convex functions.

Definition 18 (Geodesically Strong Concavity) *A function $f : M \to \mathbb{R}$ is said to be geodesically μ-strongly concave if for any $p, q \in M$,*

$$f(q) \leq f(p) + \langle g_p, \mathtt{Exp}_p^{-1}(q)\rangle_p - \frac{\mu}{2} \cdot d^2(p, q), \tag{34}$$

where $d(p, q) = \sqrt{\langle \mathtt{Exp}_q^{-1}(q), \mathtt{Exp}_p^{-1}(q)\rangle_p} = \big|\big|\mathtt{Exp}_p^{-1}(q)\big|\big|$.

Definition 19 (Lipschitzness) *A function $f : M \to \mathbb{R}$ is said to be geodesically L_f-Lipschitz if for any $p, q \in M$,*

$$|f(p) - f(q)| \leq L_f \cdot d(p, q). \tag{35}$$

Definition 20 (β-Smoothness) *A differentiable function $f : M \to \mathbb{R}$ is said to be geodesically β-smooth if its gradient is β-Lipschitz. That is, for any $p, q \in M$,*

$$||g_p - \Gamma_q^p(g_q)|| \leq \beta \cdot d(x, y), \tag{36}$$

where Γ_q^p is the parallel transport from q to p and $\beta > 0$ is a constant.

Appendix C. An alternative derivation of Algorithm 1

From the expression of the Riemannian/tangent gradient estimator (25), we obtain that

$$\begin{aligned}
\mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) &\equiv \mathtt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) \\
&= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\boldsymbol{x}^T \boldsymbol{X}_i \cdot \boldsymbol{x} - \boldsymbol{X}_i\right) \cdot L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) \\
&= \left[-\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{x}^T \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right)\right] \cdot \left[\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^n \boldsymbol{x}^T \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)} - \boldsymbol{x}\right] \\
&= \left[\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x})\right] \cdot \left[\frac{\nabla \widehat{f}_h(\boldsymbol{x})}{\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x})} - \boldsymbol{x}\right],
\end{aligned}$$

where we need to assume that $\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x}) \neq 0$. (This is true in small neighborhoods of estimated local modes under condition (C2), which in turn holds with high probability as the sample size increases and bandwidth parameter decreases accordingly. This is guaranteed by Lemma 10.) By equating the alternative directional mean shift vector $\Xi_h'(\boldsymbol{x}) = \frac{\nabla \widehat{f}_h(\boldsymbol{x})}{\boldsymbol{x}^T \nabla \widehat{f}_h(\boldsymbol{x})} - \boldsymbol{x}$ to 0, we obtain that

$$\widehat{\boldsymbol{y}}_{s+1}' = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)} \quad \text{and} \quad \widehat{\boldsymbol{y}}_{s+1} = \frac{\widehat{\boldsymbol{y}}_{s+1}'}{\big|\big|\widehat{\boldsymbol{y}}_{s+1}'\big|\big|_2} = \mathrm{sgn}\left(\widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right) \cdot \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}, \tag{37}$$

where $\mathrm{sgn}(x) = \mathbb{1}_{\{x \geq 0\}} - \mathbb{1}_{\{x \leq 0\}}$. Now, we discuss two mutually exclusive cases.

- (Case 1) If $\widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s) > 0$, then the directional mean shift vector $\Xi_h'(\widehat{\boldsymbol{y}}_s) = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)} - \widehat{\boldsymbol{y}}_s$ is parallel to the Riemannian gradient at $\widehat{\boldsymbol{y}}_s$ after being projected to the tangent space and points toward the direction of increasing the estimated density. Then, the preceding fixed-point iteration (37) is correct and can be simplified as

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s \boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^n L'\left(\frac{1-\widehat{\boldsymbol{y}}_s \boldsymbol{X}_i}{h^2}\right)}.$$

- (Case 2) If $\widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s) < 0$, then the mean shift vector $\Xi_h'(\widehat{\boldsymbol{y}}_s)$ is still parallel to the Riemannian gradient at $\widehat{\boldsymbol{y}}_s$ after being projected to the tangent space but points toward the direction of decreasing the estimated density. Thus, the preceding fixed-point equation (37) goes as

$$\widehat{\boldsymbol{y}}_{s+1} = -\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2}$$

  but is *not correct* in this case. We need to flip the sign of the fixed-point function and obtain that

$$\widehat{\boldsymbol{y}}_{s+1} = \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2} = -\frac{\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1-\widehat{\boldsymbol{y}}_s \boldsymbol{X}_i}{h^2}\right)}{\sum_{i=1}^n L'\left(\frac{1-\widehat{\boldsymbol{y}}_s \boldsymbol{X}_i}{h^2}\right)}.$$

In both cases, the final fixed-point iteration equations coincide with our previous result in Equation (20) or (23).

Appendix D. Proofs of Lemmas and Theorems

This section includes the proofs of our lemmas and theorems. Other auxiliary results are also presented along the way.

D.1 Proof of Lemma 1

Lemma 1 *Assume that kernel L is twice continuously differentiable. Then,*

$$\mathcal{H}\widetilde{f}_h(\boldsymbol{x}) = \mathcal{H}\widehat{f}_h(\boldsymbol{x})$$

for any point $\boldsymbol{x} \in \Omega_q$.

Proof Some straightforward matrix calculus shows that

$$\begin{aligned}\nabla\nabla \widetilde{f}_h(\boldsymbol{x}) = &\frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^n I_{q+1} \cdot L'\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right) \\ &+ \frac{c_{h,q}(L)}{nh^4} \sum_{i=1}^n (\boldsymbol{x}-\boldsymbol{X}_i)(\boldsymbol{x}-\boldsymbol{X}_i)^T \cdot L''\left(\frac{1-\boldsymbol{x}^T \boldsymbol{X}_i}{h^2}\right)\end{aligned}$$

$$:= \widehat{\mathcal{A}}_{\boldsymbol{x}} f$$

and

$$\nabla\nabla \widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^4} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T L'' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right).$$

According to the generalized form of the Hessian matrix on Ω_q in (13), we derive the Hessian estimator of the directional density f as

$$\begin{aligned}
&\left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\nabla\nabla \widetilde{f}_h(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla \widetilde{f}_h(\boldsymbol{x})\right] \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \\
&= \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \\
&\quad + \frac{c_{h,q}(L)}{nh^4} \sum_{i=1}^{n} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \boldsymbol{X}_i \boldsymbol{X}_i^T \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L'' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \\
&\quad - \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (1 - \boldsymbol{x}^T \boldsymbol{X}_i) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\frac{c_{h,q}(L)}{nh^4} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T L'' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \right. \\
&\quad \left. + \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{x}^T \boldsymbol{X}_i I_{q+1} \cdot L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \right] \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\nabla\nabla \widehat{f}_h(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla \widehat{f}_h(\boldsymbol{x})\right] \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right),
\end{aligned}$$

where we recall that $\nabla \widetilde{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} (\boldsymbol{x} - \boldsymbol{X}_i) \cdot L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right)$ from (21) in the first equality. Thus, we conclude that the directional Hessian estimator at a point $\boldsymbol{x} \in \Omega_q$ is defined to be

$$\begin{aligned}
\mathcal{H} \widehat{f}_h(\boldsymbol{x}) &= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\frac{c_{h,q}(L)}{nh^4} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T L'' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \right. \\
&\qquad \left. + \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \boldsymbol{x}^T \boldsymbol{X}_i I_{q+1} \cdot L' \left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_i}{h^2} \right) \right] \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \\
&= \mathcal{H} \widetilde{f}_h(\boldsymbol{x}).
\end{aligned} \tag{38}$$

The result follows. ■

D.2 Proof of Theorem 2

Before we dive into the (pointwise and uniform) consistency of the Riemannian gradient and Hessian estimators, we reiterate some common notation and terminology in directional

data. For a variable $\boldsymbol{x} \in \Omega_q$ and a fixed point $\boldsymbol{y} \in \Omega_q$, we denote $t = \boldsymbol{x}^T\boldsymbol{y}$ the inner product between $\boldsymbol{x}$ and $\boldsymbol{y}$ and write

$$\boldsymbol{x} = t\boldsymbol{y} + (1-t^2)^{\frac{1}{2}}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \in \Omega_q$ is a unit vector orthogonal to $\boldsymbol{y}$. Further, an area element on Ω_q can be written as

$$\omega_q(d\boldsymbol{x}) = (1-t^2)^{\frac{q}{2}-1}dt\,\omega_{q-1}(d\boldsymbol{\xi}).$$

We will make extensive use of Lemmas 1, 2 and 3 in García-Portugués et al. (2013) as well as their small extensions. Thus, we synthesize them in the following lemma.

Lemma 21 (A Change of Variables and Orthogonality in Ω_q) *The following results are extended from Lemmas 2 and 3 in García-Portugués et al. (2013):*

(a) Under condition (D2) or the stronger condition (D2'), we have that

$$\lim_{h\to 0}\lambda_{h,q}(L) = \lambda_q(L) = 2^{\frac{q}{2}-1}\bar{\omega}_{q-1}\int_0^\infty L(r)r^{\frac{q}{2}-1}dr,$$

where $\lambda_{h,q}(L) = \bar{\omega}_{q-1}\int_0^{2h^{-2}} L(r)r^{\frac{q}{2}-1}(2-rh^2)^{\frac{q}{2}-1}dr$ and $\bar{\omega}_q \equiv \omega_q(\Omega_q)$ is the surface area of Ω_q for $q \geq 1$. In other words, $\lambda_{h,q}(L) = \lambda_q(L) + o(1)$ as $h \to 0$.

(b) Let f be a function defined in Ω_q, and let $\boldsymbol{y} \in \Omega_q$ be a fixed point. The integral $\int_{\Omega_q} f(\boldsymbol{x})\omega_q(d\boldsymbol{x})$ can be expressed in one of the following equivalent integrals:

$$\begin{aligned}\int_{\Omega_q} f(\boldsymbol{x})\,\omega_q(d\boldsymbol{x}) &= \int_{-1}^1\int_{\Omega_{q-1}} f\left(t,(1-t^2)^{\frac{1}{2}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt \\ &= \int_{-1}^1\int_{\Omega_{q-1}} f\left(t\boldsymbol{y}+(1-t^2)^{\frac{1}{2}}\boldsymbol{B}_{\boldsymbol{y}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt,\end{aligned} \tag{39}$$

where $\boldsymbol{B}_{\boldsymbol{y}} = (\boldsymbol{b}_1,...,\boldsymbol{b}_q)_{(q+1)\times q}$ is the semi-orthonormal matrix ($\boldsymbol{B}_{\boldsymbol{y}}^T\boldsymbol{B}_{\boldsymbol{y}} = I_q$ and $\boldsymbol{B}_{\boldsymbol{y}}\boldsymbol{B}_{\boldsymbol{y}}^T = I_{q+1}$) resulting from the completion of $\boldsymbol{y}$ to the orthonormal basis $\{\boldsymbol{y},\boldsymbol{b}_1,...,\boldsymbol{b}_q\}$.

(c) For any variable $\boldsymbol{x} = (x_1,...,x_{q+1})^T \in \Omega_q$, it holds that

$$\int_{\Omega_q} x_i\omega_q(d\boldsymbol{x}) = 0, \quad \int_{\Omega_q} x_ix_j\,\omega_q(d\boldsymbol{x}) = \begin{cases}0, & i\neq j,\\ \frac{\bar{\omega}_q}{q+1}, & i=j,\end{cases} \quad \int_{\Omega_q} x_ix_jx_k\,\omega_q(d\boldsymbol{x}) = 0,$$

$$\int_{\Omega_q} x_ix_jx_kx_m\,\omega_q(d\boldsymbol{x}) = \begin{cases}\frac{3\bar{\omega}_q}{(q+1)(q+3)}, & i=j=k=m,\\ \frac{\bar{\omega}_q}{(q+1)(q+3)}, & i=k,j=m,i\neq j,\\ 0 & \text{otherwise},\end{cases} \quad \int_{\Omega_q} x_ix_jx_kx_mx_\ell\,\omega_q(d\boldsymbol{x}) = 0$$

for all $i,j,k,m,\ell = 1,...,q+1$, where $\bar{\omega}_q$ is the surface area of Ω_q for $q \geq 1$. In particular, using the notation in (b), we have that

$$\int_{\Omega_{q-1}} \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\,\omega_{q-1}(d\boldsymbol{\xi}) = 0.$$

Proof As we will use the argument of (a) in our proof of Theorem 2, we reproduce the proof of Lemma 1 in García-Portugués et al. (2013) here.
(a) Consider the functions

$$\varpi_h(r) = L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\mathbb{1}_{[0,2h^{-2})}(r),$$
$$\varpi(r) = \lim_{h\to 0}\varpi_h(r) = L(r)r^{\frac{q}{2}-1}2^{\frac{q}{2}-1}\mathbb{1}_{[0,\infty)}(r).$$

Then, proving $\lim_{h\to 0}\lambda_{h,q}(L) = \lambda_q(L)$ is equivalent to proving $\lim_{h\to 0}\int_0^\infty \varpi_h(r)\,dr = \int_0^\infty \varpi(r)\,dr$.

Consider first the case $q \geq 2$. As $\frac{q}{2}-1 \geq 0$, then $(2-h^2r)^{\frac{q}{2}-1} \leq 2^{\frac{q}{2}-1}$, $\forall h \geq 0, \forall r \in [0, 2h^{-2})$. Then,

$$|\varpi_h(r)| \leq L(r)r^{\frac{q}{2}-1}2^{\frac{q}{2}-1}\mathbb{1}_{[0,2h^{-2})}(r) \leq \varpi(r), \quad \forall r \in [0,\infty), \forall h > 0.$$

Since $\int_0^\infty \varpi(r)dr < \infty$ by condition (D2) on kernel L, by the Dominated Convergence Theorem, it follows that $\lim_{h\to 0}\int_0^\infty \varpi_h(r)dr = \int_0^\infty \varpi(r)dr$.

For the case $q = 1$, $\varpi_h(r) = L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[0,2h^{-2})}(r)$. Consider now the following decomposition:

$$\int_0^\infty \varpi_h(r)dr = \int_0^\infty L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[0,h^{-2})}(r)dr + \int_0^\infty L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[h^{-2},2h^{-2})}(r)dr.$$

The limit of the first integral can be derived analogously with the Dominated Convergence Theorem. As $(2-h^2r)^{-\frac{1}{2}}$ is monotonically increasing with respect to $r \in [0, h^{-2})$, we know that $(2-h^2r)^{-\frac{1}{2}} \leq 1$, $\forall r \in [0, h^{-2})$, $\forall h > 0$. Therefore,

$$\left|L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[0,h^{-2})}(r)\right| \leq L(r)r^{-\frac{1}{2}}\mathbb{1}_{[0,h^{-2})}(r) \leq \varpi(r), \quad \forall r \in [0,\infty), \forall h > 0.$$

Then, as $\lim_{h\to 0} L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[0,h^{-2})}(r) = \varpi(r)$ and $\int_0^\infty \varpi(r)dr < \infty$ by condition (D2), the Dominated Convergence Theorem guarantees that

$$\lim_{h\to 0}\int_0^\infty L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}\mathbb{1}_{[0,h^{-2})}(r)dr = \int_0^\infty \varpi(r)dr.$$

For the second integral, as a consequence of condition (D2), L must be decrease faster than any power function in order for $0 < \int_0^\infty L^k(r)r^{\frac{q}{2}-1}dr < \infty$ for all $q \geq 1$ and $k = 1, 2$. In particular, for some fixed $h_0 > 0$, $L(r) \leq r^{-1}$, $\forall r \in [h^{-2}, 2h^{-2})$, $\forall h \in (0, h_0)$. Using this, it results in:

$$\lim_{h\to 0}\int_{h^{-2}}^{2h^{-2}} L(r)r^{-\frac{1}{2}}(2-h^2r)^{-\frac{1}{2}}dr \leq \lim_{h\to 0}\int_{h^{-2}}^{2h^{-2}} r^{-\frac{3}{2}}(2-h^2r)^{-\frac{1}{2}}dr = \lim_{h\to 0} h = 0.$$

This completes the proof.

The proofs of (b) and the first two integral results in (c) can be found in García-Portugués et al. (2013) and thus omitted. We adopt some of the argument of Lemma 3 in García-Portugués et al. (2013) to prove the last three integrals in (c).
Recall that the n-dimensional spherical coordinates of $\boldsymbol{x} = (x_1, ..., x_n)^T$ with norm $r := ||\boldsymbol{x}||_2$ are given by

$$\begin{cases} x_1 = r\cos\phi_1, \\ x_j = r\cos\phi_j \prod_{k=1}^{j-1} \sin\phi_k, \quad j = 2, ..., n-2, \\ x_{n-1} = r\sin\theta \prod_{k=1}^{n-2} \sin\phi_k, \\ x_n = r\cos\theta \prod_{k=1}^{n-2} \sin\phi_k, \end{cases} \qquad J = r^{n-1} \prod_{k=1}^{n-2} \sin^k \phi_{n-1-k}, \tag{40}$$

where $0 \leq \phi_j \leq \pi$, $j = 1, ..., n-2$, $0 \leq \theta \leq 2\pi$, and $0 \leq r < \infty$. J denotes the Jacobian of the transformation. Without loss of generality, we assume, by the q-spherical coordinates (40), that $x_i = \cos\phi_1$, $x_j = \cos\phi_2 \sin\phi_1$, and $x_k = \cos\phi_3 \sin\phi_2 \sin\phi_1$. Then,

$$\begin{aligned} \int_{\Omega_q} x_i^3 \, \omega_q(d\boldsymbol{x}) &= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-1)}{\cdots} \times \int_0^{\pi} \cos^3\phi_1 \prod_{k=1}^{q-2} \sin^k \phi_{q-k} \sin^{q-1}\phi_1 \prod_{j=q-1}^{1} d\phi_j d\theta \\ &= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-2)}{\cdots} \times \int_0^{\pi} \prod_{k=1}^{q-2} \sin^k \phi_{q-k} \prod_{j=q-1}^{2} d\phi_j d\theta \times \int_0^{\pi} \cos^3\phi_1 \sin^{q-1}\phi_1 d\phi_1 \\ &= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-2)}{\cdots} \times \int_0^{\pi} \prod_{k=1}^{q-2} \sin^k \phi_{q-k} \prod_{j=q-1}^{2} d\phi_j d\theta \times \int_0^{\pi} (1-\sin^2\phi_1) \sin^{q-1}\phi_1 d(\sin\phi_1) \\ &= \bar{\omega}_{q-1} \times 0 = 0, \end{aligned}$$

$$\begin{aligned} &\int_{\Omega_q} x_i^2 x_j \omega_q(d\boldsymbol{x}) \\ &= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-1)}{\cdots} \times \int_0^{\pi} \cos^2\phi_1 \cos\phi_2 \sin\phi_1 \prod_{k=1}^{q-3} \sin^k \phi_{q-k} \sin^{q-2}\phi_2 \sin^{q-1}\phi_1 \prod_{j=q-1}^{1} d\phi_j d\theta \\ &= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-3)}{\cdots} \times \int_0^{\pi} \prod_{k=1}^{q-3} \sin^k \phi_{q-k} \prod_{j=q-1}^{3} d\phi_j d\theta \\ &\quad \times \int_0^{\pi} \cos^2\phi_1 \sin^q\phi_1 d\phi_1 \int_0^{\pi} \cos\phi_2 \sin^{q-2}\phi_2 d\phi_2 \\ &= \bar{\omega}_{q-2} \times \int_0^{\pi} \cos^2\phi_1 \sin^q\phi_1 d\phi_1 \times 0 = 0, \end{aligned}$$

and

$$\int_{\Omega_q} x_i x_j x_k \omega_q(d\boldsymbol{x}) = \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-1)}{\cdots} \times \int_0^{\pi} \cos\phi_1 \cos\phi_2 \sin\phi_1 \cos\phi_3 \sin\phi_2 \sin\phi_1$$

$$
\begin{aligned}
&\times \prod_{k=1}^{q-4} \sin^k \phi_{q-k} \sin^{q-3} \phi_3 \sin^{q-2} \phi_2 \sin^{q-1} \phi_1 \prod_{j=q-1}^{1} d\phi_j d\theta \\
&= \int_0^{2\pi} \int_0^{\pi} \times \overset{(q-4)}{\cdots} \times \int_0^{\pi} \prod_{k=1}^{q-4} \sin^k \phi_{q-k} \prod_{j=q-1}^{4} d\phi_j d\theta \\
&\quad \times \int_0^{\pi} \cos\phi_1 \sin^q \phi_1 d\phi_1 \int_0^{\pi} \cos\phi_2 \sin^{q-1} \phi_2 d\phi_2 \int_0^{\pi} \cos\phi_3 \sin^{q-2} \phi_3 d\phi_3 \\
&= \bar{\omega}_{q-3} \times 0 \times 0 \times 0 = 0.
\end{aligned}
$$

The preceding argument teaches us that

$$
\int_{\Omega_q} x_i x_j x_k x_m \, \omega_q(d\boldsymbol{x}) = \int_{\Omega_q} x_i x_j x_k x_m x_\ell \, \omega_q(d\boldsymbol{x}) = 0
$$

as long as one of the unique factors in the integrand has an odd multiplicity. (Indeed, any integration of a monomial with an odd degree on Ω_q will yield 0.) Thus, the only nonzero integrals in $\int_{\Omega_q} x_i x_j x_k x_m \, \omega_q(d\boldsymbol{x})$ and $\int_{\Omega_q} x_i x_j x_k x_m x_\ell \, \omega_q(d\boldsymbol{x})$ are

$$
\int_{\Omega_q} x_i^4 \, \omega_q(d\boldsymbol{x}) \quad \text{and} \quad \int_{\Omega_q} x_i^2 x_j^2 \, \omega_q(d\boldsymbol{x})
$$

with $i \neq j$. To compute the first integral, we define a vector field as

$$
\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), ..., F_{q+1}(\boldsymbol{x})) = (x_1^3, ..., x_{q+1}^3)
$$

with $\boldsymbol{x} = (x_1, ..., x_{q+1}) \in \Omega_q$. By the divergence theorem (Theorem 10.51 in Rudin 1976),

$$
\begin{aligned}
\int_{\Omega_q} x_i^4 \, \omega_q(d\boldsymbol{x}) &= \frac{1}{q+1} \int_{\Omega_q} \left(\sum_{i=1}^{q+1} x_i^4 \right) \omega_q(d\boldsymbol{x}) \\
&= \frac{1}{q+1} \int_{\Omega_q} \langle \boldsymbol{F}, \boldsymbol{x} \rangle \, \omega_q(d\boldsymbol{x}) \\
&= \frac{1}{q+1} \int_{V_q} \mathtt{div}\, \boldsymbol{F} \, dV \\
&= \frac{3}{q+1} \int_0^1 r^2 \cdot r^q \bar{\omega}_q dr = \frac{3\bar{\omega}_q}{(q+1)(q+3)},
\end{aligned}
$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^{q+1}$, $\mathtt{div}\, \boldsymbol{F} = \sum_{i=1}^{q+1} \frac{\partial F_i}{\partial x_i}$, and $\int_{V_q} \cdots dV$ is integrating the solid q-dimensional sphere V_q in $\mathbb{R}^{q+1}$. The second integral can be evaluated based on the preceding results as

$$
\begin{aligned}
\int_{\Omega_q} x_i^2 x_j^2 \, \omega_q(d\boldsymbol{x}) &= \frac{1}{q} \int_{\Omega_q} x_i^2 \left(\sum_{j \neq i} x_j^2 \right) \omega_q(d\boldsymbol{x}) \\
&= \frac{1}{q} \int_{\Omega_q} (x_i^2 - x_i^4) \, \omega_q(d\boldsymbol{x})
\end{aligned}
$$

$$= \frac{1}{q}\left[\frac{\bar{\omega}_q}{q+1} - \frac{3\bar{\omega}_q}{(q+1)(q+3)}\right] = \frac{\bar{\omega}_q}{(q+1)(q+3)}.$$

As a specific application of our above results, we know that $\int_{\Omega_{q-1}} \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi}) = 0$. ■

Remark 22 *García-Portugués et al. (2013) also provided a key remark about how to generalize the arguments in (a) of Lemma 21. Under condition (D2'), one can apply the same techniques in (a) to prove the result with the functions*

$$\begin{cases} \varpi_{h,i,j,k}(r) = L^k(r) r^{\frac{q}{2}+i}(2-h^2 r)^{\frac{q}{2}-j}\mathbb{1}_{[0,2h^{-2})}(r), \\ \varpi_{i,j,k}(r) = \lim\limits_{h\to 0}\varpi_{h,i,j,k}(r) = L^k(r) r^{\frac{q}{2}+i} 2^{\frac{q}{2}-j}\mathbb{1}_{[0,\infty)}(r); \end{cases}$$

$$\begin{cases} \varpi'_{h,i,j,k}(r) = [L'(r)]^k r^{\frac{q}{2}+i}(2-h^2 r)^{\frac{q}{2}-j}\mathbb{1}_{[0,2h^{-2})}(r), \\ \varpi'_{i,j,k}(r) = \lim\limits_{h\to 0}\varpi'_{h,i,j,k}(r) = [L'(r)]^k r^{\frac{q}{2}+i} 2^{\frac{q}{2}-j}\mathbb{1}_{[0,\infty)}(r); \end{cases}$$

$$\begin{cases} \varpi''_{h,i,j,k}(r) = [L''(r)]^k r^{\frac{q}{2}+i}(2-h^2 r)^{\frac{q}{2}-j}\mathbb{1}_{[0,2h^{-2})}(r), \\ \varpi''_{i,j,k}(r) = \lim\limits_{h\to 0}\varpi''_{h,i,j,k}(r) = [L''(r)]^k r^{\frac{q}{2}+i} 2^{\frac{q}{2}-j}\mathbb{1}_{[0,\infty)}(r) \end{cases}$$

with $i \geq -1$, $j \leq 1$, and $k = 1, 2$. For the case where $\frac{q}{2} - j \geq 0$, use the Dominated Convergence Theorem. For the other cases, subdivide the integral over $[0, 2h^{-2})$ into the intervals $[0, h^{-2})$ and $[h^{-2}, 2h^{-2})$. Then apply the Dominated Convergence Theorem in the former and use a suitable power function to make the latter tend to 0 in the same way as described in the proof of (a) in Lemma 21.

Theorem 2 *Assume conditions (D1) and (D2'). For any fixed $\boldsymbol{x} \in \Omega_q$, we have*

$$\texttt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$$

as $h \to 0$ and $nh^{q+2} \to \infty$.
Under the same condition, for any fixed $\boldsymbol{x} \in \Omega_q$, we have

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$$

as $h \to 0$ and $nh^{q+4} \to \infty$.

Proof Part A: Pointwise convergence rate of the Riemannian gradient estimator $\texttt{grad}\,\widehat{f}_h(\boldsymbol{x})$. Recall from Section 4.1 that the tangent/Riemannian gradient estimator of a directional KDE is uniquely defined under a given kernel function L. Thus, we can establish the pointwise convergence rate under any total gradient (or differential) estimator, that is, $\texttt{grad}\,\widehat{f}_h(\boldsymbol{x}) = \texttt{grad}\,\widetilde{f}_h(\boldsymbol{x}) \equiv \texttt{Tang}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right)$. Here, we stick on the differential form (21), $\nabla\widetilde{f}_h(\boldsymbol{x})$. (One may also prove Theorem 2 with $\nabla\widehat{f}_h(\boldsymbol{x})$. The proof of Lemma 10 provides a starting point for this direction.)

• **Result 1**: The expectation of the Riemannian gradient estimator, $\mathbb{E}\left[\texttt{grad}\,\widehat{f}_h(\boldsymbol{x})\right]$, has the following asymptotic behavior as $h \to 0$:

$$\begin{aligned}
\mathbb{E}\left[\texttt{grad}\,\widetilde{f}_h(\boldsymbol{x})\right] &= \mathbb{E}\left[\texttt{Tang}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right)\right] = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right] \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla f(\boldsymbol{x}) + \frac{h^2}{2}\left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla f(\boldsymbol{x}) \cdot \frac{\int_0^\infty L(r) r^{\frac{q}{2}} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&\quad + \frac{2h^2}{q}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i \cdot \frac{\int_0^\infty L'(r) r^{\frac{q}{2}+1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} + O(h^2) + o(h^2) \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\nabla f(\boldsymbol{x}) + O(h^2).
\end{aligned}$$

Derivation of Result 1. With the definition of $\boldsymbol{B}_{\boldsymbol{x}}$ from Lemma 21, the expected value of $\nabla\widetilde{f}_h(\boldsymbol{x})$ is

$$\begin{aligned}
\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right] &= \frac{c_{h,q}(L)}{h^2}\int_{\Omega_q}(\boldsymbol{x}-\boldsymbol{y})\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right) f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= \frac{c_{h,q}(L)}{h^2}\int_{-1}^{1}\int_{\Omega_{q-1}}\left(\boldsymbol{x} - t\boldsymbol{x} - \sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right) L'\left(\frac{1-t}{h^2}\right) \\
&\quad \times f\left(t\boldsymbol{x} + \sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dt \\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(rh^2\boldsymbol{x} - h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right) L'(r) \\
&\quad \times f(\boldsymbol{x} + \alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr
\end{aligned} \tag{41}$$

by (a) in Lemma 21 and a change of variable $r = \frac{1-t}{h^2}$, where $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$. By condition (D1), the Taylor's expansion of f at $\boldsymbol{x}$ is

$$\begin{aligned}
&f(\boldsymbol{x} + \alpha_{\boldsymbol{x},\boldsymbol{\xi}}) \\
&= f(\boldsymbol{x}) + \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T\nabla f(\boldsymbol{x}) + \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T\nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} + \frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x}) + o\left(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2^3\right) \\
&\equiv \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)} + o(h^3),
\end{aligned}$$

where $||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2^2 = r^2h^4 + h^2r(2-h^2r) = 2rh^2$ by the orthogonality of $\boldsymbol{x}$ and columns of $\boldsymbol{B}_{\boldsymbol{x}}$, and $(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i$ stands for the i^{th} entry of the vector $\alpha_{\boldsymbol{x},\boldsymbol{\xi}}$. Now we plug (I), (II), (III), (IV), and $o(h^3)$ back into (41) respectively to compute the dominating term of $\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]$.

Plug in (I)

$$\begin{aligned}
&= c_{h,q}(L)h^{q-2}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(rh^2\boldsymbol{x} - h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right) \\
&\qquad \times L'(r)\,r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\overset{\text{(i)}}{=} \bar{\omega}_{q-1}\cdot\boldsymbol{x}f(\boldsymbol{x})\int_0^{2h^{-2}} c_{h,q}(L)h^qL'(r)\cdot r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}-1}dr + 0
\end{aligned}$$

$$\overset{\text{(ii)}}{=} \bar{\omega}_{q-1} \cdot \boldsymbol{x} f(\boldsymbol{x}) \cdot c_{h,q}(L) h^q \Bigg\{ L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} \Big|_0^{2h^{-2}} - \int_0^{2h^{-2}} L(r) \left[\frac{q}{2} \cdot r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} - h^2 \left(\frac{q}{2} - 1 \right) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-2} \right] dr \Bigg\}$$

$$= -\frac{q}{2} \cdot \bar{\omega}_{q-1} \cdot \boldsymbol{x} f(\boldsymbol{x}) \cdot c_{h,q}(L) h^q \int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr + \left(\frac{q-2}{2} \right) \bar{\omega}_{q-1} \cdot \boldsymbol{x} f(\boldsymbol{x}) \cdot c_{h,q}(L) h^{q+2} \int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-2} dr$$

$$\overset{\text{(iii)}}{=} -\frac{q}{2} \cdot \boldsymbol{x} f(\boldsymbol{x}) + \left(\frac{q-2}{2} \right) \boldsymbol{x} f(\boldsymbol{x}) h^2 \cdot \frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-2} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr}$$

$$\overset{\text{(iv)}}{=} -\frac{q}{2} \cdot \boldsymbol{x} f(\boldsymbol{x}) + \left(\frac{q-2}{4} \right) \boldsymbol{x} f(\boldsymbol{x}) h^2 \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} + o(h^2),$$

as $h \to 0$, where we use (c) of Lemma 21 with $\boldsymbol{B}_{\boldsymbol{x}} \boldsymbol{\xi} = \sum_{i=1}^q \xi_i \boldsymbol{b}_i$ in (i), conduct integration by parts in (ii), plug in the expression (3) of $c_{h,q}(L)$ in (iii), and take $h \to 0$ with our argument in (a) of Lemma 21 and Remark 22 to obtain (iv). The $o(h^2)$-term in (iv) takes into account those small error terms as $h \to 0$. Likewise,

Plug in (II)

$$= c_{h,q}(L) h^{q-2} \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \left(r h^2 \boldsymbol{x} - h \sqrt{r(2 - h^2 r)} \boldsymbol{B}_{\boldsymbol{x}} \boldsymbol{\xi} \right) \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T f(\boldsymbol{x}) \times L'(r) \cdot r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$= -c_{h,q}(L) h^{q+2} \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) L'(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-1} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$+ c_{h,q}(L) h^{q+1} \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \boldsymbol{x}\boldsymbol{\xi}^T \boldsymbol{B}_{\boldsymbol{x}}^T \nabla f(\boldsymbol{x}) L'(r) r^{\frac{q+1}{2}} (2 - h^2 r)^{\frac{q-1}{2}} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$+ c_{h,q}(L) h^{q+1} \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \boldsymbol{B}_{\boldsymbol{x}} \boldsymbol{\xi} \cdot \boldsymbol{x}^T \nabla f(\boldsymbol{x}) L'(r) r^{\frac{q+1}{2}} (2 - h^2 r)^{\frac{q-1}{2}} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$- c_{h,q}(L) h^q \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \boldsymbol{B}_{\boldsymbol{x}} \boldsymbol{\xi} \cdot \boldsymbol{\xi}^T \boldsymbol{B}_{\boldsymbol{x}}^T \nabla f(\boldsymbol{x}) L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$\overset{\text{(i)}}{=} -c_{h,q}(L) h^{q+2} \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \bar{\omega}_{q-1} \int_0^{2h^{-2}} L'(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-1} dr + 0 + 0$$

$$- c_{h,q}(L) h^q \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \left(\sum_{i=1}^q \xi_i \boldsymbol{b}_i \right) \left(\sum_{i=1}^q \xi_i \boldsymbol{b}_i^T \nabla f(\boldsymbol{x}) \right) L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} \omega_{q-1}(d\boldsymbol{\xi}) dr$$

$$\overset{\text{(ii)}}{=} -c_{h,q}(L) h^{q+2} \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \bar{\omega}_{q-1} \Bigg\{ r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-1} L(r) \Big|_0^{2h^{-2}} - \int_0^{2h^{-2}} L(r) \left[\left(\frac{q+2}{2} \right) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} - h^2 \left(\frac{q-2}{2} \right) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-2} \right] dr \Bigg\}$$

$$
\begin{aligned}
&- \frac{\bar{\omega}_{q-1}}{q}\left(\sum_{i=1}^{q} \boldsymbol{b}_i \boldsymbol{b}_i^T\right) \nabla f(\boldsymbol{x}) \cdot c_{h,q}(L) h^q \int_0^{2h^{-2}} L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} dr \\
&\overset{\text{(iii)}}{=} c_{h,q}(L) h^{q+2} \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \bar{\omega}_{q-1}\left(\frac{q+2}{2}\right) \int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} dr \\
&\quad - c_{h,q}(L) h^{q+4} \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \bar{\omega}_{q-1}\left(\frac{q-2}{2}\right) \int_0^{2h^{-2}} r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-2} dr \\
&\quad - \frac{\bar{\omega}_{q-1}}{q} (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \nabla f(\boldsymbol{x}) \cdot c_{h,q}(L) h^q \Bigg[L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} \Big|_0^{2h^{-2}} \\
&\qquad\qquad - \frac{q}{2} \int_0^{2h^{-2}} L(r) \left(r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} - h^2 r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} \right) dr \Bigg] \\
&\overset{\text{(iv)}}{=} \left(\frac{q+2}{2}\right) h^2 \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} \\
&\quad - \left(\frac{q-2}{2}\right) h^4 \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-2} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} \\
&\quad + \frac{1}{2} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) \Bigg[\frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} \\
&\qquad\qquad - h^2 \cdot \frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} \Bigg] \\
&\overset{\text{(v)}}{=} \left(\frac{q+2}{2}\right) h^2 \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} - \left(\frac{q-2}{4}\right) h^4 \cdot \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}+1} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} \\
&\quad + \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) + O(h^2) - \frac{h^2}{2} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} + o(h^2) \\
&= (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \nabla f(\boldsymbol{x}) + h^2 \left(\frac{q+2}{2}\right) \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} \\
&\quad + \frac{h^2}{2} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^{\infty} L(r) r^{\frac{q}{2}} dr}{\int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr} + O(h^2) + o(h^2),
\end{aligned}
$$

where we use (c) of Lemma 21 in (i) and (ii), leverage the fact that $\sum_{i=1}^{q} \boldsymbol{b}_i \boldsymbol{b}_i^T = \boldsymbol{B}_{\boldsymbol{x}} \boldsymbol{B}_{\boldsymbol{x}}^T = I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T$ in (iii), plug in the expression (3) of $c_{h,q}(L)$ in (iv), and take $h \to 0$ with arguments in Lemma 21 and Remark 22 to obtain (v). The $o(h^2)$-term incorporates higher-order error terms, while the $O(h^2)$-term in (v) comes from the following arguments:

$$
\frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} - 2 = -h^2 \cdot \frac{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} dr}{\int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}-1} dr} = O(h^2). \quad (42)
$$

We now move on to the calculation of (III), which is more complicated.

$$\begin{aligned}\text{Plug in (III)} = c_{h,q}(L)h^q \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} \cdot \boldsymbol{x} L'(r) r^{\frac{q}{2}} (2-h^2 r)^{\frac{q}{2}-1} \omega_{q-1}(d\boldsymbol{\xi}) dr \\ - c_{h,q}(L)h^{q-1} \int_0^{2h^{-2}} \int_{\Omega_{q-1}} \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} \cdot \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} L'(r) r^{\frac{q-1}{2}} (2-h^2 r)^{\frac{q-1}{2}} \omega_{q-1}(d\boldsymbol{\xi}) dr.\end{aligned} \tag{43}$$

Notice that

$$\begin{aligned}&\int_{\Omega_{q-1}} \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} \cdot \boldsymbol{x}\, \omega_{q-1}(\boldsymbol{\xi}) \\ &= r^2 h^4 \int_{\Omega_{q-1}} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{x}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &\quad - 2rh^3\sqrt{r(2-h^2 r)} \int_{\Omega_{q-1}} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} \cdot \boldsymbol{x}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &\quad + h^2 r(2-h^2 r) \int_{\Omega_{q-1}} \boldsymbol{\xi}^T \boldsymbol{B}_{\boldsymbol{x}}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} \cdot \boldsymbol{x}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &= r^2 h^4 \bar{\omega}_{q-1} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{x} + h^2 r(2-h^2 r) \int_{\Omega_{q-1}} \left(\sum_{i,j=1}^q \boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j \xi_i \xi_j \right) \boldsymbol{x}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &= r^2 h^4 \bar{\omega}_{q-1} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{x} + h^2 r(2-h^2 r) \int_{\Omega_{q-1}} \left(\sum_{i=1}^q \boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i \xi_i^2 \right) \boldsymbol{x}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &= r^2 h^4 \bar{\omega}_{q-1} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{x} + h^2 r(2-h^2 r) \cdot \frac{\bar{\omega}_{q-1}}{q} \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \right] \cdot \boldsymbol{x},\end{aligned} \tag{44}$$

where we use (c) of Lemma 21 in the second, third, and fourth equations and the fact that

$$\begin{aligned}\sum_{i=1}^q \boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i = \operatorname{tr}\left[\nabla\nabla f(\boldsymbol{x}) \sum_{i=1}^q \boldsymbol{b}_i \boldsymbol{b}_i^T \right] &= \operatorname{tr}\left[\nabla\nabla f(\boldsymbol{x})(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T) \right] \\ &= \Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x}.\end{aligned}$$

Here, $\Delta f(\boldsymbol{x}) = \sum_{i=1}^{q+1} \frac{\partial^2}{\partial x_i^2} f(\boldsymbol{x})$ is the Laplace of function f. At the same time,

$$\begin{aligned}&\int_{\Omega_{q-1}} \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} \cdot \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &= r^2 h^4 \int_{\Omega_{q-1}} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{x} \cdot \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &\quad - 2rh^3\sqrt{r(2-h^2 r)} \int_{\Omega_{q-1}} \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} \cdot \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi}) \\ &\quad + h^2 r(2-h^2 r) \int_{\Omega_{q-1}} \boldsymbol{\xi}^T \boldsymbol{B}_{\boldsymbol{x}}^T \nabla\nabla f(\boldsymbol{x})\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} \cdot \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi})\end{aligned}$$

$$
\begin{aligned}
&= -2rh^3\sqrt{r(2-h^2r)}\int_{\Omega_{q-1}}\left(\sum_{i=1}^{q}\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\cdot\boldsymbol{b}_i\xi_i^2\right)\omega_{q-1}(d\boldsymbol{\xi})\\
&\quad + h^2r(2-h^2r)\int_{\Omega_{q-1}}\left(\sum_{i,j=1}^{q}\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\xi_i\xi_j\right)\left(\sum_{k=1}^{q}\boldsymbol{b}_k\xi_k\right)\omega_{q-1}(d\boldsymbol{\xi})\\
&= -2rh^3\sqrt{r(2-h^2r)}\cdot\frac{\bar{\omega}_{q-1}}{q}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i,
\end{aligned}
$$

where we apply (c) of Lemma 21 in the last two equations. That is,

$$
\int_{\Omega_{q-1}}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T\nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}}\cdot\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\,\omega_{q-1}(d\boldsymbol{\xi}) = -2rh^3\sqrt{r(2-h^2r)}\cdot\frac{\bar{\omega}_{q-1}}{q}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i. \tag{45}
$$

Plugging (44) and (45) back into (43), we proceed "Plug in (III)" as

$$
\begin{aligned}
\text{Plug in (III)} &= \frac{c_{h,q}(L)h^{q+4}}{2}\cdot\bar{\omega}_{q-1}\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\cdot\boldsymbol{x}\int_0^{2h^{-2}} r^{\frac{q}{2}+2}(2-h^2r)^{\frac{q}{2}-1}L'(r)dr\\
&\quad + \frac{c_{h,q}(L)h^{q+2}}{2q}\cdot\bar{\omega}_{q-1}\left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\right]\boldsymbol{x}\int_0^{2h^{-2}} L'(r)r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}dr\\
&\quad + \frac{c_{h,q}(L)h^{q+2}}{q}\cdot\bar{\omega}_{q-1}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i\int_0^{2h^{-2}} L'(r)r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}dr\\
&\overset{\text{(i)}}{=} \frac{h^4}{2}\cdot\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\cdot\boldsymbol{x}\cdot\frac{\int_0^{\infty}L'(r)r^{\frac{q}{2}+2}(2-h^2r)^{\frac{q}{2}-1}dr}{\int_0^{\infty}L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr}\\
&\quad + \frac{h^2}{2q}\left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\right]\boldsymbol{x}\cdot\frac{\int_0^{\infty}L'(r)r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}dr}{\int_0^{\infty}L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr}\\
&\quad + \frac{h^2}{q}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i\cdot\frac{\int_0^{\infty}L'(r)r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}dr}{\int_0^{\infty}L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr}\\
&\overset{\text{(ii)}}{=} \frac{h^2}{q}\left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\right]\boldsymbol{x}\cdot\frac{\int_0^{\infty}L'(r)r^{\frac{q}{2}+1}dr}{\int_0^{\infty}L(r)r^{\frac{q}{2}-1}dr}\\
&\quad + \frac{2h^2}{q}\sum_{i=1}^{q}\left(\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right)\boldsymbol{b}_i\cdot\frac{\int_0^{\infty}L'(r)r^{\frac{q}{2}+1}dr}{\int_0^{\infty}L(r)r^{\frac{q}{2}-1}dr} + o(h^2),
\end{aligned}
$$

where we plug in the expression (3) of $c_{h,q}(L)$ in (i) and take $h\to 0$ with arguments in Lemma 21 and Remark 22 in (ii).

We argue that after plugging (IV)+$o(h^3)$ back into (41), it yields a $o(h^2)$ term.

$$
\begin{aligned}
&\text{Plug in (IV)} + o(h^3)\\
&= c_{h,q}(L)h^q\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\frac{1}{6}\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\boldsymbol{x}L'(r)r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr
\end{aligned}
$$

$$
\begin{aligned}
&+ c_{h,q}(L)h^{q-1}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\frac{1}{6}\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\\
&\qquad\qquad\times L'(r)r^{\frac{q-1}{2}}(2-h^2r)^{\frac{q-1}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&+ c_{h,q}(L)\cdot o(h^{q+1})\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(rh^2\boldsymbol{x}-h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)L'(r)r^{\frac{q-1}{2}}(2-h^2r)^{\frac{q-1}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&= c_{h,q}(L)h^q\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\frac{1}{6}\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\boldsymbol{x}L'(r)r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&+ c_{h,q}(L)h^{q-1}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\frac{1}{6}\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\\
&\qquad\qquad\times L'(r)r^{\frac{q-1}{2}}(2-h^2r)^{\frac{q-1}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&+ c_{h,q}(L)\cdot o(h^{q+3})\cdot\bar{\omega}_{q-1}\boldsymbol{x}\int_0^{2h^{-2}}L'(r)r^{\frac{q+1}{2}}(2-h^2r)^{\frac{q-1}{2}}dr
\end{aligned}
$$

by (c) of Lemma 21 in the last equality. As $c_{h,q}(L)=h^q\lambda_{h,q}(L)=O(h^q)$ by (3) and (a) of Lemma 21, we know that the third integral is of the order $o(h^3)$. Since $\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}=\sum_{i=1}^q\xi_i\boldsymbol{b}_i$ and $\alpha_{\boldsymbol{x},\boldsymbol{\xi}}=-rh^2\boldsymbol{x}+h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$, we derive that

$$
\begin{aligned}
&\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\\
&= A_{f,1}\cdot h^3r^{\frac{3}{2}}(2-h^2r)^{\frac{3}{2}}\sum_{i,j,k}\boldsymbol{b}_i\boldsymbol{b}_j\boldsymbol{b}_k\xi_i\xi_j\xi_k + A_{f,2}\cdot h^4r^2(2-h^2r)\sum_{i,j}\boldsymbol{b}_i\boldsymbol{b}_j\xi_i\xi_j\\
&\quad + A_{f,3}\cdot h^5r^{\frac{5}{2}}(2-h^2r)^{\frac{1}{2}}\sum_i\boldsymbol{b}_i\xi_i + A_{f,4}h^6r^3
\end{aligned}
$$

and

$$
\begin{aligned}
&\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\\
&= \widetilde{A}_{f,1}\cdot h^3r^{\frac{3}{2}}(2-h^2r)^{\frac{3}{2}}\sum_{i,j,k,\ell}\boldsymbol{b}_i\boldsymbol{b}_j\boldsymbol{b}_k\boldsymbol{b}_\ell\xi_i\xi_j\xi_k\xi_\ell + \widetilde{A}_{f,2}\cdot h^4r^2(2-h^2r)\sum_{i,j,k}\boldsymbol{b}_i\boldsymbol{b}_j\boldsymbol{b}_k\xi_i\xi_j\xi_k\\
&\quad + \widetilde{A}_{f,3}\cdot h^5r^{\frac{5}{2}}(2-h^2r)^{\frac{1}{2}}\sum_{i,j}\boldsymbol{b}_i\boldsymbol{b}_j\xi_i\xi_j + \widetilde{A}_{f,4}h^6r^3\sum_i\boldsymbol{b}_i\xi_i,
\end{aligned}
$$

where $A_{f,i}, i=1,...,4$ and $\widetilde{A}_{f,i}, i=1,...,4$ are some "constants" that depends on the partial derivatives of $f(\boldsymbol{x})$. Thus, by (c) of Lemma 21 (that is, any integration of a monomial of $\boldsymbol{\xi}$ with an odd degree on Ω_{q-1} will yield 0), we know that

$$
\int_{\Omega_{q-1}}\left[\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i\cdot\frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\omega_{q-1}(d\boldsymbol{\xi})\asymp h^4r^2(2-h^2r)+o(h^4),
$$

$$\int_{\Omega_{q-1}} \left[\left(\sum_{i=1}^{q+1} (\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i} \right)^3 f(\boldsymbol{x}) \right] \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\, \omega_{q-1}(d\boldsymbol{\xi}) \asymp h^3 r^{\frac{3}{2}} (2-h^2 r)^{\frac{3}{2}} + h^5 r\sqrt{2-h^2 r} + o(h^3),$$

where "$\asymp$" means an asymptotic equivalence. With condition (D2') and our arguments of (a) in Lemma 21 and Remark 22, we obtain that "Plug in (IV)+$o(h^3)$" yields a $o(h^2)$ term. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\nabla \widetilde{f}_h(\boldsymbol{x})\right] = &-\frac{q}{2}\cdot \boldsymbol{x} f(\boldsymbol{x}) + \left(\frac{q-2}{4}\right) \boldsymbol{x} f(\boldsymbol{x}) h^2 \cdot \frac{\int_0^\infty L(r) r^{\frac{q}{2}} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&+ (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\nabla f(\boldsymbol{x}) + h^2 \left(\frac{q+2}{2}\right) \boldsymbol{x}\boldsymbol{x}^T \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^\infty L(r) r^{\frac{q}{2}} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&+ \frac{h^2}{2} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^\infty L(r) r^{\frac{q}{2}} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&+ \frac{h^2}{q} \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] \boldsymbol{x} \cdot \frac{\int_0^\infty L'(r) r^{\frac{q}{2}+1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&+ \frac{2h^2}{q} \sum_{i=1}^{q} \left(\boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_i\right) \boldsymbol{b}_i \cdot \frac{\int_0^\infty L'(r) r^{\frac{q}{2}+1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} + O(h^2) + o(h^2),
\end{aligned} \tag{46}$$

which in turn shows that

$$\begin{aligned}
\mathbb{E}\left[\mathtt{grad}\, \widetilde{f}_h(\boldsymbol{x})\right] &= \mathbb{E}\left[\mathtt{Tang}\left(\nabla \widetilde{f}_h(\boldsymbol{x})\right)\right] = (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)\mathbb{E}\left[\nabla \widetilde{f}_h(\boldsymbol{x})\right] \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) + \frac{h^2}{2} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) \cdot \frac{\int_0^\infty L(r) r^{\frac{q}{2}} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} \\
&\quad + \frac{2h^2}{q} \sum_{i=1}^{q} \left(\boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_i\right) \boldsymbol{b}_i \cdot \frac{\int_0^\infty L'(r) r^{\frac{q}{2}+1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} + O(h^2) + o(h^2) \\
&= \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x}) + O(h^2)
\end{aligned}$$

as $h \to 0$. **Result 1** thus follows and the Riemannian gradient estimator is unbiased.

• **Result 2** The covariance matrix of $\nabla \widetilde{f}_h(\boldsymbol{x})$ has the following asymptotic rate as $h \to 0$ and $nh^{q+2} \to \infty$:

$$\mathrm{Cov}\left[\nabla \widetilde{f}_h(\boldsymbol{x})\right] = \frac{1}{nh^{q+2}} \cdot R(f, L) + o\left(\frac{1}{nh^{q+2}}\right),$$

where $R(f,L) = \frac{f(\boldsymbol{x}) \int_0^\infty L'(r)^2 r^{\frac{q}{2}} dr}{2^{\frac{q}{2}-2} q \cdot \bar{\omega}_{q-1} \left(\int_0^\infty L(r) r^{\frac{q}{2}-1} dr\right)^2} \cdot (I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T)$.

Derivation of Result 2. By (46), the covariance matrix of $\nabla \widetilde{f}_h(\boldsymbol{x})$ can be calculated as

$$\begin{aligned}
&\mathrm{Cov}\left[\nabla \widetilde{f}_h(\boldsymbol{x})\right] \\
&= \frac{c_{h,q}(L)^2}{nh^4} \cdot \mathrm{Cov}\left[(\boldsymbol{x} - \boldsymbol{X}_1) L'\left(\frac{1 - \boldsymbol{x}^T \boldsymbol{X}_1}{h^2}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{c_{h,q}(L)^2}{nh^4} \cdot \mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{X}_1)(\boldsymbol{x}-\boldsymbol{X}_1)^T L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_1}{h^2}\right)^2\right] - \frac{1}{n}\cdot\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]\cdot\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]^T \\
&= \frac{c_{h,q}(L)^2}{nh^4}\int_{\Omega_q}(\boldsymbol{x}-\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{y})^T L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)^2 f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) + O\left(\frac{1}{n}\right) \\
&= \frac{c_{h,q}(L)^2}{n}h^{q-4}\int_0^{2h^{-2}}\int_{\Omega_{q-1}} \alpha_{\boldsymbol{x},\boldsymbol{\xi}}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T L'(r)^2 f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}}) r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + O\left(\frac{1}{n}\right),
\end{aligned}
$$

where $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$. By condition (D1), the first-order Taylor's expansion of f at $\boldsymbol{x}\in\Omega_q$ is

$$
f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}}) = f(\boldsymbol{x}) + O(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2) = f(\boldsymbol{x}) + O(h).
$$

Thus,

$$
\begin{aligned}
&\mathrm{Cov}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right] \\
&= \frac{c_{h,q}(L)^2}{n}h^{q-4}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\Big[r^2h^4\boldsymbol{x}\boldsymbol{x}^T - rh^3\sqrt{r(2-h^2r)}\boldsymbol{x}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T \\
&\quad - rh^3\sqrt{r(2-h^2r)}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\boldsymbol{x}^T + h^2r(2-h^2r)\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T\Big]L'(r)^2 r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + o\left(\frac{1}{nh^{q+2}}\right) \\
&\overset{\text{(i)}}{=} \frac{c_{h,q}(L)^2}{n}h^q f(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\bar{\omega}_{q-1}\int_0^{2h^{-2}} L'(r)^2 r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad + \frac{c_{h,q}(L)^2}{n}h^{q-2}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \xi_i\boldsymbol{b}_i\right)\left(\sum_{i=1}^q \xi_i\boldsymbol{b}_i^T\right)L'(r)^2 r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + o\left(\frac{1}{nh^{q+2}}\right) \\
&\overset{\text{(ii)}}{=} \frac{c_{h,q}(L)^2}{n}h^q f(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\bar{\omega}_{q-1}\int_0^{2h^{-2}} L'(r)^2 r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad + \frac{c_{h,q}(L)^2\bar{\omega}_{q-1}}{nq}\cdot h^{q-2}f(\boldsymbol{x})\int_0^{2h^{-2}}\left(\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\right)L'(r)^2 r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}dr + o\left(\frac{1}{nh^{q+2}}\right) \\
&\overset{\text{(iii)}}{=} \frac{c_{h,q}(L)}{n}f(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\cdot\frac{\int_0^{2h^{-2}} L'(r)^2 r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}-1}dr}{\int_0^{2h^{-2}} L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr} \\
&\quad + \frac{c_{h,q}(L)}{nq}\cdot h^{q-2}f(\boldsymbol{x})\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\frac{\int_0^{2h^{-2}} L'(r)^2 r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}dr}{\int_0^{2h^{-2}} L(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr} + o\left(\frac{1}{nh^{q+2}}\right) \\
&\overset{\text{(iv)}}{=} \frac{1}{nh^{q+2}}\cdot R(f,L) + o\left(\frac{1}{nh^{q+2}}\right),
\end{aligned}
$$

where $R(f,L)=\frac{f(\boldsymbol{x})\int_0^\infty L'(r)^2 r^{\frac{q}{2}}dr}{2^{\frac{q}{2}-2}q\cdot\bar{\omega}_{q-1}\left(\int_0^\infty L(r)r^{\frac{q}{2}-1}dr\right)^2}\cdot(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T)$ is a matrix whose columns lie in the tangent space of Ω_q at $\boldsymbol{x}$. During the derivation, we use (c) of Lemma 21 in (i) and (ii), plug in the expression (3) of $c_{h,q}(L)$ in (iii), and take $h\to 0$ with arguments in (a) of Lemma 21 and Remark 22 in (iv). **Result 2** thus follows.

By the central limit theorem,

$$\begin{aligned}\nabla\widetilde{f}_h(\boldsymbol{x})-\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]&=\mathrm{Cov}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]^{\frac{1}{2}}\cdot\mathrm{Cov}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]^{-\frac{1}{2}}\left\{\nabla\widetilde{f}_h(\boldsymbol{x})-\mathbb{E}\left[\nabla\widetilde{f}_h(\boldsymbol{x})\right]\right\}\\&=\left[\frac{1}{nh^{q+2}}\cdot R(f,L)+o\left(\frac{1}{nh^{q+2}}\right)\right]^{\frac{1}{2}}\cdot\boldsymbol{Z}_n(\boldsymbol{x})\\&=O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right),\end{aligned}$$

where $\boldsymbol{Z}_n(\boldsymbol{x})\xrightarrow{d}N_{q+1}(\boldsymbol{0},I_{q+1})$.
The asymptotic rate for $\mathtt{grad}\,\widetilde{f}_h(\boldsymbol{x})-\mathbb{E}\left[\mathtt{grad}\,\widetilde{f}_h(\boldsymbol{x})\right]=\mathtt{Tang}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right)-\mathbb{E}\left[\mathtt{Tang}\left(\nabla\widetilde{f}_h(\boldsymbol{x})\right)\right]$ remains unchanged, since the dominating constant $R(f,L)$ are within the tangent space of Ω_q at $\boldsymbol{x}$.
In a nutshell, we conclude with bias (**Result 1**) and variance (**Result 2**) estimation that

$$\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})-\mathtt{grad}\,f(\boldsymbol{x})=\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)-\mathtt{Tang}\left(\nabla f(\boldsymbol{x})\right)=O(h^2)+O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right)$$

for any fixed $\boldsymbol{x}\in\Omega_q$, as $h\to 0$ and $nh^{q+2}\to\infty$.

Part B: Pointwise convergence rate of the Riemannian Hessian estimator $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$.
As shown in Lemma 1, $\mathcal{H}\widehat{f}_h(\boldsymbol{x})=\mathcal{H}\widetilde{f}_h(\boldsymbol{x})$ for any $\boldsymbol{x}\in\Omega_q$ and we can establish the pointwise convergence rate using either Riemannian Hessian estimator. Here, we stick to the Riemannian Hessian estimator $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ in (38).

• **Result 3**. The expectation of the Riemannian Hessian estimator, $\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]$, has the following asymptotic behavior as $h\to 0$:

$$\begin{aligned}\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]&=\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\mathbb{E}\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right]\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\\&=\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\left[\nabla\nabla f(\boldsymbol{x})-\boldsymbol{x}^T\nabla f(\boldsymbol{x})\right]\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)+O(h^2),\end{aligned}$$

where $\mathcal{A}\widehat{f}_h(\boldsymbol{x})=\frac{c_{h,q}(L)}{nh^4}\sum_{i=1}^n\boldsymbol{X}_i\boldsymbol{X}_i^TL''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)+\frac{c_{h,q}(L)}{nh^2}\sum_{i=1}^n\boldsymbol{x}^T\boldsymbol{X}_iI_{q+1}\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$.
Derivation of Result 3. We first compute the expectation of $\mathcal{A}\widehat{f}_h(\boldsymbol{x})$ and apply the left and right multiplications of $\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)$ to simplify our calculation. Notice that

$$\begin{aligned}&\mathbb{E}\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right]\\&=\frac{c_{h,q}(L)}{h^4}\int_{\Omega_q}\boldsymbol{y}\boldsymbol{y}^TL''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y})\end{aligned}$$

$$
\begin{aligned}
&+\frac{c_{h,q}(L)}{h^2}\int_{\Omega_q}\boldsymbol{x}^T\boldsymbol{y}\cdot I_{q+1}\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y})\\
&=\frac{c_{h,q}(L)}{h^4}\int_{-1}^{1}\int_{\Omega_{q-1}}\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)^T\\
&\qquad\qquad\times L''\left(\frac{1-t}{h^2}\right)\cdot f\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt\\
&+\frac{c_{h,q}(L)}{h^2}\int_{-1}^{1}\int_{\Omega_{q-1}}tI_{q+1}\cdot L'\left(\frac{1-t}{h^2}\right)\cdot f\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt\\
&=c_{h,q}(L)h^{q-4}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T L''(r)\\
&\qquad\qquad\times f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&+c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(1-h^2r)I_{q+1}\cdot L'(r)f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr
\end{aligned}
$$

by $r=\frac{1-t}{h^2}$ and $\alpha_{\boldsymbol{x},\boldsymbol{\xi}}=-rh^2\boldsymbol{x}+h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$. Since

$$
\begin{aligned}
(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T=&(1-rh^2)^2\boldsymbol{x}\boldsymbol{x}^T+h(1-rh^2)\sqrt{r(2-h^2r)}\left[\boldsymbol{x}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T+(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\boldsymbol{x}^T\right]\\
&+h^2r(2-h^2r)\cdot(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T,
\end{aligned}
$$

the preceding calculation proceeds as

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right]\\
&=c_{h,q}(L)h^{q-4}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\boldsymbol{x}\boldsymbol{x}^T L''(r)\cdot f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q}{2}-1}(1-rh^2)^2(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad+c_{h,q}(L)h^{q-3}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left[\boldsymbol{x}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T+(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\boldsymbol{x}^T\right]L''(r)\\
&\qquad\qquad\times f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q-1}{2}}(1-rh^2)(2-h^2r)^{\frac{q-1}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad+c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T L''(r)\cdot f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad+c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(1-h^2r)I_{q+1}\cdot L'(r)f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\,r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\equiv\text{(I)}+\text{(II)}+\text{(III)}+\text{(IV)}.
\end{aligned}
\tag{47}
$$

The above terms (I) and (II), after we apply the congruence operation

$$
\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\mathbb{E}\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right]\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right),
$$

yield zero. Hence, we will not continue to compute them. By condition (D1), the Taylor's expansion of f at $\boldsymbol{x}$ is

$$
f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})
$$

$$= f(\boldsymbol{x}) + \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla f(\boldsymbol{x}) + \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} + \frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x}) + O\left(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2^4\right)$$

$$= f(\boldsymbol{x}) + \alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla f(\boldsymbol{x}) + \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T \nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}} + \frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x}) + O(h^4),$$

where $||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2^2 = r^2h^4 + h^2r(2-h^2r) = 2rh^2$ by the orthogonality of $\boldsymbol{x}$ and columns of $\boldsymbol{B}_{\boldsymbol{x}}$, and $(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i$ stands for the i^{th} entry of the vector $\alpha_{\boldsymbol{x},\boldsymbol{\xi}}$. Note that plugging $O(h^4)$ into (III) or (IV) in (47) both leads to a $O(h^2)$ after integration, since with condition (D2') and our arguments in (a) of Lemma 21 and Remark 22,

$$O(h^4) \cdot c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}} \phi(r,\boldsymbol{\xi})L'(r)\,\omega_{q-1}(d\boldsymbol{\xi})dr \asymp O(h^2),$$

where $\phi(r,\boldsymbol{\xi})$ is a square integrable function of $(r,\boldsymbol{\xi})$ and "$\asymp$" stands for the asymptotic equivalence. It shows that carrying out the Taylor's expansion of f at $\boldsymbol{x}$ to the third order is sufficient in our context.

More importantly, plugging the term $\frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})$ into (II) and (III) also gives rise to a $O(h^2)$ term. Because $\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi} = \sum_{i=1}^q \xi_i \boldsymbol{b}_i$ and $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$,

$$\begin{aligned}
&\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T\left[\frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\omega_{q-1}(d\boldsymbol{\xi}) \\
&= \int_{\Omega_{q-1}}\left[P_r(\boldsymbol{\xi},2)h^6 + P_r(\boldsymbol{\xi},3)h^5 + P_r(\boldsymbol{\xi},4)h^4 + P_r(\boldsymbol{\xi},5)h^3\right]\omega_{q-1}(d\boldsymbol{\xi}) \qquad (48)\\
&= O(h^4) + \underbrace{\int_{\Omega_{q-1}} h^3 P_r(\boldsymbol{\xi},5)\,\omega_{q-1}(d\boldsymbol{\xi})}_{=0},
\end{aligned}$$

$$\begin{aligned}
&\int_{\Omega_{q-1}}\left[\frac{1}{6}\left(\sum_{i=1}^{q+1}(\alpha_{\boldsymbol{x},\boldsymbol{\xi}})_i \cdot \frac{\partial}{\partial x_i}\right)^3 f(\boldsymbol{x})\right]\omega_{q-1}(d\boldsymbol{\xi}) \\
&= \int_{\Omega_{q-1}}\left[P_r(\boldsymbol{\xi},0)h^6 + P_r(\boldsymbol{\xi},1)h^5 + P_r(\boldsymbol{\xi},2)h^4 + P_r(\boldsymbol{\xi},3)h^3\right]\omega_{q-1}(d\boldsymbol{\xi}) \qquad (49)\\
&= O(h^4) + \underbrace{\int_{\Omega_{q-1}} h^3 P_r(\boldsymbol{\xi},3)\,\omega_{q-1}(d\boldsymbol{\xi})}_{=0},
\end{aligned}$$

where $P_r(\boldsymbol{\xi},n)$ is a polynomial of elements of $\boldsymbol{\xi} = (\xi_1,...,\xi_q)$ with only degree n terms, whose coefficients may involve the variable r. The integral $\int_{\Omega_{q-1}} h^3 P_r(\boldsymbol{\xi},5)\,\omega_{q-1}(d\boldsymbol{\xi}) = \int_{\Omega_{q-1}} h^3 P_r(\boldsymbol{\xi},3)\,\omega_{q-1}(d\boldsymbol{\xi}) = 0$ is due to (c) of Lemma 21 and the fact that the integrand is a linear combination of degree 5 or 3 monomials of elements of $\boldsymbol{\xi}$. With condition (D2') and our arguments in (a) of Lemma 21 and Remark 22, the final $O(h^4)$ terms in (48) and (49)

both yield $O(h^2)$ terms after being plugged into (III) and (IV).
We now plug the Taylor's expansion of $f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})$ back into (III) and obtain that

$$
\begin{aligned}
&\text{Plug in (III) in (47)}\\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)f(\boldsymbol{x})r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad + c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\nabla f(\boldsymbol{x})^T\alpha_{\boldsymbol{x},\boldsymbol{\xi}}\\
&\qquad\qquad \times r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad + c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\\
&\qquad\qquad \times \frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T\nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}}\cdot r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr + O(h^2)\\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\frac{\bar{\omega}_{q-1}}{q}\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)L''(r)f(\boldsymbol{x})r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,dr\\
&\quad - c_{h,q}(L)h^{q}\int_0^{2h^{-2}}\frac{\bar{\omega}_{q-1}}{q}\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)L''(r)\nabla f(\boldsymbol{x})^T\boldsymbol{x}r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}\,dr\\
&\quad + c_{h,q}(L)h^{q+2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\\
&\qquad\qquad \times \frac{1}{2}\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\cdot r^{\frac{q}{2}+2}(2-h^2r)^{\frac{q}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad - c_{h,q}(L)h^{q+1}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\\
&\qquad\qquad \times \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\cdot r^{\frac{q+3}{2}}(2-h^2r)^{\frac{q+1}{2}}\,\omega_{q-1}(d\boldsymbol{\xi})dr\\
&\quad + c_{h,q}(L)h^{q}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\\
&\qquad\qquad \times \frac{1}{2}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T\nabla\nabla f(\boldsymbol{x})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\cdot r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}+1}\,\omega_{q-1}(d\boldsymbol{\xi})dr + O(h^2)\\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\frac{\bar{\omega}_{q-1}}{q}\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)L''(r)f(\boldsymbol{x})\cdot r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}}\,dr\\
&\quad - c_{h,q}(L)h^{q}\int_0^{2h^{-2}}\frac{\bar{\omega}_{q-1}}{q}\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)L''(r)\nabla f(\boldsymbol{x})^T\boldsymbol{x}\cdot r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}\,dr \quad - \quad 0\\
&\quad + c_{h,q}(L)h^{q}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T L''(r)\\
&\qquad\qquad \times \frac{1}{2}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T\nabla\nabla f(\boldsymbol{x})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\cdot r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}+1}\,\omega_{q-1}(d\boldsymbol{\xi})dr + O(h^2),
\end{aligned}
\tag{50}
$$

where we apply (c) of Lemma 21 and the fact that $\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T = I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T$. We also absorb the third integral in the second equality into $O(h^2)$ to obtain the third equality, given condition (D2') and our arguments in (a) of Lemma 21. The "0" term in the third equality is due to the fact that

$$\int_{\Omega_{q-1}} \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})\,\omega_{q-1}(d\boldsymbol{\xi}) = \int_{\Omega_{q-1}} P_r(\boldsymbol{\xi},3)\,\omega_{q-1}(d\boldsymbol{\xi}) = 0$$

by (c) of Lemma 21, where the notation $P_r(\boldsymbol{\xi},3)$ is defined in (49). Now, we consider the inner integral inside the last integration.

$$\begin{aligned}
&\int_{\Omega_{q-1}} \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T \cdot \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T \nabla\nabla f(\boldsymbol{x}) \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\omega_{q-1}(d\boldsymbol{\xi})\\
&= \int_{\Omega_{q-1}} \left(\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\xi_i^2 + \sum_{i\neq j}\boldsymbol{b}_i\boldsymbol{b}_j^T\xi_i\xi_j\right)\left(\sum_{i=1}^q \boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\xi_i^2 + \sum_{i\neq j}\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\xi_i\xi_j\right)\omega_{q-1}(d\boldsymbol{\xi})\\
&\overset{(*)}{=} \int_{\Omega_{q-1}} \left[\left(\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\xi_i^2\right)\left(\sum_{i=1}^q \boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\xi_i^2\right)\right.\\
&\quad \left. + \left(\sum_{i\neq j}\boldsymbol{b}_i\boldsymbol{b}_j^T\xi_i\xi_j\right)\left(\sum_{i\neq j}\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\xi_i\xi_j\right)\right]\omega_{q-1}(d\boldsymbol{\xi})\\
&= \int_{\Omega_{q-1}} \left[\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\cdot\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\xi_i^4 + \sum_{i\neq j}\boldsymbol{b}_i\boldsymbol{b}_i^T\cdot\boldsymbol{b}_j^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\xi_i^2\xi_j^2\right.\\
&\quad \left. + 2\sum_{i\neq j}\boldsymbol{b}_i\boldsymbol{b}_j^T\cdot\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\xi_i^2\xi_j^2\right]\omega_{q-1}(d\boldsymbol{\xi}),
\end{aligned}$$

where the cross product terms vanish after integration by Lemma 21 and the factor 2 in front of the last summation emerges because any fixed (i,j) term $(i\neq j)$ in the first factor of the second product in equality $(*)$ can be matched up with both (i,j) and (j,i) terms in the second factor to yield a summand. Using Lemma 21 and the facts that $\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T = I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T$ and

$$\sum_{i=1}^q \boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i = \operatorname{tr}\left[\nabla\nabla f(\boldsymbol{x})\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\right] = \Delta f(\boldsymbol{x}) - \boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x},$$

the preceding display continues as

$$\begin{aligned}
&\int_{\Omega_{q-1}} \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T \cdot \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)^T \nabla\nabla f(\boldsymbol{x}) \left(\sum_{i=1}^q \boldsymbol{b}_i\xi_i\right)\omega_{q-1}(d\boldsymbol{\xi})\\
&= \frac{3\bar{\omega}_{q-1}}{q(q+2)}\left(\sum_{i=1}^q \boldsymbol{b}_i\boldsymbol{b}_i^T\cdot\boldsymbol{b}_i^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_i\right) + \frac{\bar{\omega}_{q-1}}{q(q+2)}\left(\sum_{i\neq j}\boldsymbol{b}_i\boldsymbol{b}_i^T\cdot\boldsymbol{b}_j^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{b}_j\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{2\bar{\omega}_{q-1}}{q(q+2)} \left(\sum_{i \neq j} \boldsymbol{b}_i \boldsymbol{b}_j^T \cdot \boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_j \right) \\
&= \frac{\bar{\omega}_{q-1}}{q(q+2)} \left[\sum_{i=1}^{q} \boldsymbol{b}_i \boldsymbol{b}_i^T \left(\sum_{j=1}^{q} \boldsymbol{b}_j^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_j \right) \right] \\
&\quad + \frac{2\bar{\omega}_{q-1}}{q(q+2)} \left[\sum_{i=1}^{q} \boldsymbol{b}_i \left(\boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_i \right) \boldsymbol{b}_i^T + \sum_{i \neq j} \boldsymbol{b}_i \left(\boldsymbol{b}_i^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{b}_j \right) \boldsymbol{b}_j^T \right] \\
&= \frac{\bar{\omega}_{q-1}}{q(q+2)} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x} \right] \\
&\quad + \frac{2\bar{\omega}_{q-1}}{q(q+2)} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right).
\end{aligned}
$$

Plugging this result back into (50) and conduct some integration by parts, we obtain that

Plug in (III) in (47)

$$
\begin{aligned}
&= c_{h,q}(L) h^{q-2} \int_0^{2h^{-2}} \frac{\bar{\omega}_{q-1}}{q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) f(\boldsymbol{x}) L''(r) \cdot r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} \, dr \\
&\quad - c_{h,q}(L) h^{q} \int_0^{2h^{-2}} \frac{\bar{\omega}_{q-1}}{q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L''(r) \cdot r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}} \, dr \\
&\quad + c_{h,q}(L) h^{q} \cdot \frac{\bar{\omega}_{q-1}}{2q(q+2)} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x} \right] L''(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}+1} dr \\
&\quad + c_{h,q}(L) h^{q} \cdot \frac{\bar{\omega}_{q-1}}{q(q+2)} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) L''(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}+1} dr \\
&\quad + O(h^2) \\
&= c_{h,q}(L) h^{q-2} \cdot \frac{\bar{\omega}_{q-1}}{q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) L'(r) f(\boldsymbol{x}) \cdot r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} \Big|_0^{2h^{-2}} \\
&\quad - c_{h,q}(L) h^{q-2} \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) f(\boldsymbol{x}) L'(r) \left[r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} - h^2 r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} \right] dr \\
&\quad - c_{h,q}(L) h^{q} \cdot \frac{\bar{\omega}_{q-1}}{q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L'(r) \cdot r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}} \Big|_0^{2h^{-2}} \\
&\quad + c_{h,q}(L) h^{q} \int_0^{2h^{-2}} \frac{\bar{\omega}_{q-1}}{q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L'(r) \\
&\qquad\qquad \times \left[\left(\frac{q+2}{2} \right) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} - \frac{q h^2}{2} r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-1} \right] dr \\
&\quad + c_{h,q}(L) h^{q} \cdot \frac{\bar{\omega}_{q-1}}{2q(q+2)} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x} \right] L'(r) r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}+1} \Big|_0^{2h^{-2}} \\
&\quad - c_{h,q}(L) h^{q} \cdot \frac{\bar{\omega}_{q-1}}{4q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T \right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x} \right] L'(r) \\
&\qquad\qquad \times \left[r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}+1} - h^2 r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}} \right] dr
\end{aligned}
$$

$$
\begin{aligned}
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{q(q+2)} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L'(r) r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}+1}\Big|_0^{2h^{-2}} \\
&- c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L'(r) \\
&\qquad\qquad \times \left[r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}+1} - h^2 r^{\frac{q}{2}+1}(2-h^2r)^{\frac{q}{2}}\right] dr \\
&+ O(h^2) \\
= &-c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) f(\boldsymbol{x}) L'(r) r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}(1-h^2r) dr \\
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} \cdot L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}} dr \\
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} \cdot L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}} dr \\
&- c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}+1} dr \\
&- c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}+1} dr \\
&+ O(h^2),
\end{aligned}
$$

where we use the fact that

$$
c_{h,q}(L)h^{q+2} \int_0^{2h^{-2}} L'(r) r^j (2-h^2r)^k dr = O(h^2) \tag{51}
$$

for any $k, j > 0$ via Remark 22. With extra integration by parts on the third and fifth term in the preceding display, we obtain that

$$
\begin{aligned}
&\text{Plug in (III) in (47)} \\
&= -c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) f(\boldsymbol{x}) L'(r) r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}(1-h^2r) dr \\
&\quad + c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}} dr \\
&\quad - c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L(r) r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}} dr \\
&\quad - c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] L'(r) r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}+1} dr \\
&\quad + c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L(r) r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}+1} dr \\
&\quad + O(h^2).
\end{aligned} \tag{52}
$$

We now plug the Taylor's expansion of $f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})$ at $\boldsymbol{x}$ back into (IV) in (47) and deduce that

Plug in (IV) in (47)

$$
\begin{aligned}
&= c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1}\int_0^{2h^{-2}} I_{q+1}f(\boldsymbol{x})L'(r)\cdot r^{\frac{q}{2}-1}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad + c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}} I_{q+1}\nabla f(\boldsymbol{x})^T\alpha_{\boldsymbol{x},\boldsymbol{\xi}}L'(r)\cdot r^{\frac{q}{2}-1}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}} I_{q+1}\cdot\frac{1}{2}\alpha_{\boldsymbol{x},\boldsymbol{\xi}}^T\nabla\nabla f(\boldsymbol{x})\alpha_{\boldsymbol{x},\boldsymbol{\xi}}r^{\frac{q}{2}-1}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + O(h^2) \\
&= c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1}\int_0^{2h^{-2}} I_{q+1}f(\boldsymbol{x})L'(r)\cdot r^{\frac{q}{2}-1}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad - c_{h,q}(L)h^{q}\bar{\omega}_{q-1}\int_0^{2h^{-2}} I_{q+1}\nabla f(\boldsymbol{x})^T\boldsymbol{x}L'(r)r^{\frac{q}{2}}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad + c_{h,q}(L)h^{q-1}\int_0^{2h^{-2}} \underbrace{\int_{\Omega_{q-1}} I_{q+1}\cdot\nabla f(\boldsymbol{x})^T(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})r^{\frac{q-1}{2}}(2-h^2r)^{\frac{q-1}{2}}(1-h^2r)\,\omega_{q-1}(d\boldsymbol{\xi})}_{=0}\,dr \\
&\quad + c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}} I_{q+1}\cdot\frac{1}{2}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T\nabla\nabla f(\boldsymbol{x})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})L'(r)r^{\frac{q}{2}}(1-h^2r)(2-h^2r)^{\frac{q}{2}} \\
&\quad + O(h^2) \\
&= c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1}\int_0^{2h^{-2}} I_{q+1}f(\boldsymbol{x})L'(r)\cdot r^{\frac{q}{2}-1}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad - c_{h,q}(L)h^{q}\bar{\omega}_{q-1}\int_0^{2h^{-2}} I_{q+1}\nabla f(\boldsymbol{x})^T\boldsymbol{x}L'(r)r^{\frac{q}{2}}(1-h^2r)(2-h^2r)^{\frac{q}{2}-1}dr \\
&\quad + c_{h,q}(L)h^{q}\cdot\frac{\bar{\omega}_{q-1}}{2q}\int_0^{2h^{-2}} I_{q+1}\left[\Delta f(\boldsymbol{x})-\boldsymbol{x}^T\nabla\nabla f(\boldsymbol{x})\boldsymbol{x}\right]L'(r)r^{\frac{q}{2}}(1-h^2r)(2-h^2r)^{\frac{q}{2}}dr \\
&\quad + O(h^2),
\end{aligned}
\tag{53}
$$

where we expand $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$, absorb $O(h^2)$ terms via (51), make use of (c) in Lemma 21, and leverage our argument in (44). Combining (47), (52), and (53), we conclude that

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right] \\
&= \left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\mathbb{E}\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right]\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right) \\
&= \left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\cdot(\mathrm{III})\cdot\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right) + \left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\cdot(\mathrm{IV})\cdot\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right) \\
&= -c_{h,q}(L)h^{q-2}\bar{\omega}_{q-1}\int_0^{2h^{-2}}\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)f(\boldsymbol{x})L'(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}(1-h^2r)dr
\end{aligned}
$$

$$
\begin{aligned}
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} dr \\
&- c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} dr \\
&- c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}+1} dr \\
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}+1} dr \\
&+ c_{h,q}(L)h^{q-2} \cdot \bar{\omega}_{q-1} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) f(\boldsymbol{x}) L'(r) \cdot r^{\frac{q}{2}-1} (1 - h^2 r)(2 - h^2 r)^{\frac{q}{2}-1} dr \\
&- c_{h,q}(L)h^q \cdot \bar{\omega}_{q-1} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} L'(r) r^{\frac{q}{2}} (1 - h^2 r)(2 - h^2 r)^{\frac{q}{2}-1} dr \\
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2q} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] \\
&\qquad\qquad \times L'(r) r^{\frac{q}{2}} (1 - h^2 r)(2 - h^2 r)^{\frac{q}{2}} dr \\
&+ O(h^2) \\
\overset{(**)}{=}\ & 0 + c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} \\
&\qquad\qquad \times L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}-1} (2 - h^2 r - 2 + 2rh^2) dr \\
&- \lambda_{h,q}(L)^{-1} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} \cdot \frac{\bar{\omega}_{q-1}}{2} \int_0^{2h^{-2}} L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}} dr \\
&+ c_{h,q}(L)h^q \cdot \frac{\bar{\omega}_{q-1}}{4q} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \left[\Delta f(\boldsymbol{x}) - \boldsymbol{x}^T \nabla\nabla f(\boldsymbol{x}) \boldsymbol{x}\right] \\
&\qquad\qquad \times \int_0^{2h^{-2}} L'(r) r^{\frac{q}{2}} (2 - h^2 r)^{\frac{q}{2}} (2 - 2h^2 r - 2 + h^2 r) dr \\
&+ \lambda_{h,q}(L)^{-1} \cdot \frac{\bar{\omega}_{q-1}}{4} \int_0^{2h^{-2}} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \\
&\qquad\qquad \times L(r) r^{\frac{q}{2}-1} (2 - h^2 r)^{\frac{q}{2}+1} dr \\
&+ O(h^2) \\
\overset{(***)}{=}\ & O(h^2) - \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} \cdot \frac{\bar{\omega}_{q-1}}{2} \lambda_q(L)^{-1} \int_0^{\infty} L(r) r^{\frac{q}{2}-1} 2^{\frac{q}{2}} dr + O(h^2) \\
&+ \lambda_q(L)^{-1} \cdot \frac{\bar{\omega}_{q-1}}{4} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \int_0^{\infty} L(r) r^{\frac{q}{2}-1} 2^{\frac{q}{2}+1} dr + O(h^2) \\
=\ & - \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla f(\boldsymbol{x})^T \boldsymbol{x} + \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) + O(h^2),
\end{aligned}
$$

where the first term matches up with the sixth term, the second term with the seventh term, the fourth term with the eighth term, and (3) is applied to the rest terms when $h \to 0$ in $(**)$. In addition, we leverage the asymptotic rates (51) and (42) as well as recall that $\lambda_q(L) = 2^{\frac{q}{2}-1} \bar{\omega}_{q-1} \int_0^{\infty} L(r) r^{\frac{q}{2}-1} dr$ from (a) of Lemma 21 in $(***)$. **Result 3** thus follows. It

implies that the bias $\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right] - \mathcal{H}f(\boldsymbol{x})$ is of the rate $O(h^2)$ and the Riemannian Hessian estimator is asymptotically unbiased.

Now, we proceed to bound

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right] = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right)\left[\mathcal{A}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]\left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right).$$

• **Result 4**. The covariance matrix $\mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right]$ has the following asymptotic rate as $h \to 0$ and $nh^{q+4} \to \infty$:

$$\mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right] = O\left(\frac{1}{nh^{q+4}}\right),$$

where we define the matrix $\mathtt{vec}$ operator, which converts a matrix into a vector by stacking the columns. That is, given a matrix $A \in \mathbb{R}^{m\times n}$, $\mathtt{vec}(A)$ is a vector of length mn.

Derivation of Result 4. We first calculate the covariance matrix of $\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)$ as

$$\begin{aligned}
&\mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right] \\
&= \frac{c_{h,q}(L)^2}{nh^8}\cdot\mathbb{E}\left[\mathtt{vec}(\boldsymbol{X}_1\boldsymbol{X}_1^T)\cdot\mathtt{vec}(\boldsymbol{X}_1\boldsymbol{X}_1^T)^T L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)^2\right] \\
&\quad + \frac{2c_{h,q}(L)^2}{nh^6}\cdot\mathbb{E}\left[\mathtt{vec}(\boldsymbol{X}_i\boldsymbol{X}_i^T)\cdot\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{X}_1 I_{q+1})^T L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)\right] \\
&\quad + \frac{c_{h,q}(L)^2}{nh^4}\cdot\mathbb{E}\left[\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{X}_1 I_{q+1})\cdot\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{X}_1 I_{q+1})^T L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)^2\right] \\
&\quad - \frac{1}{n}\cdot\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]^T \\
&= \frac{c_{h,q}(L)^2}{nh^8}\int_{\Omega_q}\mathtt{vec}(\boldsymbol{y}\boldsymbol{y}^T)\cdot\mathtt{vec}(\boldsymbol{y}\boldsymbol{y}^T)^T L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)^2 f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&\quad + \frac{2c_{h,q}(L)^2}{nh^6}\int_{\Omega_q}\mathtt{vec}(\boldsymbol{y}\boldsymbol{y}^T)\cdot\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{y} I_{q+1})^T L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&\quad + \frac{c_{h,q}(L)^2}{nh^4}\int_{\Omega_q}\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{y} I_{q+1})\cdot\mathtt{vec}(\boldsymbol{x}^T\boldsymbol{y} I_{q+1})^T L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)^2 f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&\quad - \frac{1}{n}\cdot\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]^T \\
&= \frac{c_{h,q}(L)^2 h^q}{nh^8}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\mathtt{vec}\left[(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\right]\mathtt{vec}\left[(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\right]^T \\
&\qquad\qquad \times L''(r)^2 f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2 r)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + \frac{2c_{h,q}(L)^2 h^q}{nh^6}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\mathtt{vec}\left[(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\right]\mathtt{vec}\left[I_{q+1}(1-rh^2)\right]^T \\
&\qquad\qquad \times L''(r)L'(r) f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2 r)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dr
\end{aligned}$$

$$+\frac{c_{h,q}(L)^2h^q}{nh^4}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}\mathtt{vec}\left[I_{q+1}(1-rh^2)\right]\mathtt{vec}\left[I_{q+1}(1-rh^2)\right]^T \times L'(r)^2 f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dr$$
$$-\frac{1}{n}\cdot\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]\mathbb{E}\left[\mathtt{vec}\left(\mathcal{A}\widehat{f}_h(\boldsymbol{x})\right)\right]^T,$$

where $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$. Note that by condition (D1), the first-order Taylor's expansion of f at $\boldsymbol{x}\in\Omega_q$ is

$$f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}}) = f(\boldsymbol{x}) + O(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2) = f(\boldsymbol{x}) + O(h).$$

In addition,

$$(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T = (1-rh^2)^2\boldsymbol{x}\boldsymbol{x}^T + h(1-r^2h)\sqrt{r(2-h^2r)}\left[\boldsymbol{x}(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T + (\boldsymbol{B}_{\boldsymbol{x}})\boldsymbol{x}^T\right] + h^2r(2-h^2r)(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})(\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi})^T.$$

Moreover, when the congruence operation $\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\mathcal{A}\widehat{f}_h(\boldsymbol{x})\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)$ is introduced, it will be applied inside the vec operation. Thus, after applying the congruence operation,

$$\mathtt{vec}\left[\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\right] \times \mathtt{vec}\left[\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\right]^T = O(h^4)$$

and

$$\mathtt{vec}\left[\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^T\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)\right] \times \mathtt{vec}\left[\left(I_{q+1}-\boldsymbol{x}\boldsymbol{x}^T\right)(1-rh^2)\right]^T = O(h^2).$$

Together with condition (D2'), (3), (51), and the bias bound $\mathbb{E}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right] = \mathcal{H}f(\boldsymbol{x}) + O(h^2)$, we conclude that

$$\mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right] = O\left(\frac{1}{nh^{q+4}}\right).$$

Result 4 is thus proved. Finally, by the central limit theorem,

$$\begin{aligned}
&\mathtt{vec}\left\{\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]\right\}\\
&= \mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right]^{\frac{1}{2}}\mathrm{Cov}\left[\mathtt{vec}\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right)\right]^{-\frac{1}{2}}\mathtt{vec}\left\{\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]\right\}\\
&= O\left(\sqrt{\frac{1}{nh^{q+4}}}\right)\cdot\widetilde{\boldsymbol{Z}}_n(\boldsymbol{x})\\
&= O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right),
\end{aligned}$$

where $\widetilde{\boldsymbol{Z}}_n(\boldsymbol{x}) \xrightarrow{d} N_{(q+1)^2}\left(\boldsymbol{0}, I_{(q+1)^2}\right)$. In total, we conclude with our bias and stochastic variation bounds that

$$\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x}) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right)$$

for any fixed $\boldsymbol{x} \in \Omega_q$ as $h \to 0$ and $nh^{q+4} \to \infty$, where

$$\mathcal{H}f(\boldsymbol{x}) = \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) \nabla\nabla f(\boldsymbol{x}) \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right) - \nabla f(\boldsymbol{x})^T \boldsymbol{x} \left(I_{q+1} - \boldsymbol{x}\boldsymbol{x}^T\right).$$

■

D.3 Proof of Theorem 4

Theorem 4 *Assume (D1), (D2'), and (K1). The uniform convergence rate of $\widehat{f}_h$ is given by*

$$\sup_{\boldsymbol{x}\in\Omega_q} |\widehat{f}_h(\boldsymbol{x}) - f(\boldsymbol{x})| = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^q}}\right)$$

as $h \to 0$ and $\frac{nh^q}{|\log h|} \to \infty$.
Furthermore, the uniform convergence rate of $\texttt{grad}\, \widehat{f}_h(\boldsymbol{x})$ on Ω_q is

$$\sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right),$$

as $h \to 0$ and $\frac{nh^{q+2}}{|\log h|} \to \infty$. Finally, the uniform convergence rate of $\mathcal{H}\widehat{f}_h(\boldsymbol{x})$ on Ω_q is

$$\sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x})\right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+4}}}\right),$$

as $h \to 0$ and $\frac{nh^{q+4}}{|\log h|} \to \infty$, where $||\cdot||_{\max}$ is the elementwise maximum norm for a vector in $\mathbb{R}^{q+1}$ or a matrix in $\mathbb{R}^{(q+1)\times(q+1)}$.

Proof Note that with the directional KDE form (15), we have that

$$D^{\tau_j}\widehat{f}_h(\boldsymbol{x}) = \frac{c_{h,q}(L)}{nh} \sum_{i=1}^{n} \left(\frac{x_{\tau_j} - X_{\tau_j}}{h}\right) L'\left(\frac{1 - \boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right),$$

$$D^{[\tau_j,\tau_k]}\widehat{f}_h(\boldsymbol{x}) = \begin{cases} \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\frac{x_{\tau_j} - X_{\tau_j}}{h}\right)\left(\frac{x_{\tau_k} - X_{\tau_k}}{h}\right) L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) & j \neq k, \\ \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} \left(\frac{x_{\tau_j} - X_{\tau_j}}{h}\right)^2 L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) + \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right) & j = k, \end{cases}$$

and

$$\left|\left|D^{[\tau]}\widehat{f}_h - D^{[\tau]}f\right|\right|_\infty \leq \left|\left|\mathbb{E}\left[D^{[\tau]}\widehat{f}_h\right] - D^{[\tau]}f\right|\right|_\infty + \left|\left|D^{[\tau]}\widehat{f}_h - \mathbb{E}\left[D^{[\tau]}\widehat{f}_h\right]\right|\right|_\infty$$

for $[[\tau]] = 0, 1, 2$. The first term in the preceding display is of order $O(h^2)$ inside the tangent space by Theorem 2 and the differentiability of f under condition (D1). The proof of $\left|\left|D^{[\tau]}\widehat{f}_h - \mathbb{E}\left[D^{[\tau]}\widehat{f}_h\right]\right|\right|_\infty = O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2[[\tau]]}}}\right)$ follows directly from the argument of Theorem 2.3 in Giné and Guillou (2002) and the following calculations:

$$\begin{aligned}
\mathbb{E}\left[L^2\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)\right] &= \int_{\Omega_q} L^2\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)\cdot f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= \int_{-1}^{1}\int_{\Omega_{q-1}} L^2\left(\frac{1-t}{h^2}\right)\cdot f\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt \\
&= h^q\int_0^{2h^{-2}}\int_{\Omega_{q-1}} f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot L^2(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt \\
&\le h^q||f||_\infty\omega_{q-1}2^{\frac{q}{2}-1}\int_0^\infty L^2(r)r^{\frac{q}{2}-1}dr,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{x_i-X_i}{h}\right)^2\left|L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)\right|^2\right] \\
&\le \mathbb{E}\left[\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\cdot\left|L'\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)\right|^2\right] \\
&= 2\int_{\Omega_q}\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)\left|L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)\right|^2 f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= 2\int_{-1}^{1}\int_{\Omega_{q-1}}\left(\frac{1-t}{h^2}\right)\left|L'\left(\frac{1-t}{h^2}\right)\right|^2 f\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dt \\
&= 2h^q\int_0^{2h^{-2}}\int_{\Omega_{q-1}} f(x+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot|L'(r)|^2\cdot r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\le 2h^q||f||_\infty\omega_{q-1}2^{\frac{q}{2}-1}\int_0^\infty|L'(r)|^2r^{\frac{q}{2}}dr,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\max\left\{\left(\frac{x_i-X_i}{h}\right)^4\left|L''\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)\right|^2,\left(\frac{x_i-X_i}{h}\right)^2\left(\frac{x_j-X_j}{h}\right)^2\left|L''\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)\right|^2\right\}\right] \\
&\le \mathbb{E}\left[\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^4 L''\left(\frac{1}{2}\left|\left|\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right|\right|_2^2\right)^2\right] \\
&= 4\int_{\Omega_q}\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)^2\left|L''\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)\right|^2 f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= 4\int_{-1}^{1}\int_{\Omega_{q-1}}\left(\frac{1-t}{h^2}\right)^2\left|L''\left(\frac{1-t}{h^2}\right)\right|^2 f\left(t\boldsymbol{x}+\sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\,\omega_{q-1}(d\boldsymbol{\xi})dt
\end{aligned}$$

$$= 4h^q \int_0^{2h^{-2}} \int_{\Omega_{q-1}} f(\boldsymbol{x} + \alpha_{\boldsymbol{x},\boldsymbol{\xi}}) \cdot L''(r)^2 r^{\frac{q}{2}+1} (2 - h^2 r)^{\frac{q}{2}-1} \, \omega_{q-1}(d\boldsymbol{\xi}) dr$$
$$\leq 4h^q ||f||_\infty \omega_{q-1} 2^{\frac{q}{2}-1} \int_0^\infty r^{\frac{q}{2}+1} L''(r)^2 dr$$

for $i = 1, ..., q+1$, where we apply (a) in Lemma 21, the change of variable $r = \frac{1-t}{h^2}$, and $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$ in the preceding three displays. ■

D.4 Proof of Theorem 6

Theorem 6 *Assume (D1), (D2'), (K1), and (M1-2). For any $\delta \in (0,1)$, when h is sufficiently small and n is sufficiently large,*

(a) there must be at least one estimated local mode $\widehat{\boldsymbol{m}}_k$ within $S_k = \boldsymbol{m}_k \oplus \rho_*$ for every $\boldsymbol{m}_k \in \mathcal{M}$, and

(b) the collection of estimated modes satisfies $\widehat{\mathcal{M}}_n \subset \mathcal{M} \oplus \rho_*$ and there is a unique estimated local mode $\widehat{\boldsymbol{m}}_k$ within $S_k = \boldsymbol{m}_k \oplus \rho_*$

with probability at least $1 - \delta$. In total, when h is sufficiently small and n is sufficiently large, there exist some constants $A_3, B_3 > 0$ such that

$$\mathbb{P}\left(\widehat{K}_n \neq K\right) \leq B_3 e^{-A_3 n h^{q+4}}.$$

(c) The Hausdorff distance between the collection of local modes and its estimator satisfies

$$\texttt{Haus}\left(\mathcal{M}, \widehat{\mathcal{M}}_n\right) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right),$$

as $h \to 0$ and $nh^{q+2} \to \infty$.

Proof The proof is partially adopted from the proof of Theorem 1 in Chen et al. (2016).

Statement (a). Without loss of generality, we consider the local mode $\boldsymbol{m}_k$ and the set $S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \boldsymbol{m}_k|| \leq \rho_*\}$. With condition (D1), we can apply the Taylor's expansion on the exponential map $\texttt{Exp}_{\boldsymbol{m}_k} : D_\epsilon \subset T_{\boldsymbol{m}_k}(\Omega_q) \to \Omega_q$ with $\texttt{Exp}_{\boldsymbol{m}_k}(0) = \boldsymbol{m}_k$, where D_ϵ is a disk of radius ϵ in $T_{\boldsymbol{m}_k}(\Omega_q)$ with center in the origin and $\epsilon > \arccos\left(1 - \frac{\rho_*^2}{2}\right)$. Here, $\arccos\left(1 - \frac{\rho_*^2}{2}\right)$ is the geodesic distance from the center $\boldsymbol{m}_k$ to ∂S_k on Ω_q, where $\partial S_k = \{\boldsymbol{x} \in \Omega_q : ||\boldsymbol{x} - \boldsymbol{m}_k||_2 = \rho_*\}$ is the boundary of S_k. With (M1) and the fact that the third order partial derivatives of f are upper bounded by C_3,

$$\begin{aligned}
\sup_{\boldsymbol{x} \in \partial S_k} f(\boldsymbol{x}) \leq \sup_{\boldsymbol{x} \in \partial S_k} \Bigg[& f(\boldsymbol{m}_k) + [\texttt{grad}\, f(\boldsymbol{m}_k)]^T \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{x}) \\
& + \frac{1}{2} \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{x})^T \left(\mathcal{H}_{\boldsymbol{m}_k} f\right) \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{x}) + \frac{C_3}{6} ||\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{x})||_2^3 \Bigg] \qquad (54) \\
\leq f(\boldsymbol{m}_k) - \frac{\lambda_*}{2} \left(\frac{3\lambda_*}{2C_3}\right)^2 + \frac{C_3}{6} \left(\frac{3\lambda_*}{2C_3}\right)^3 & = f(\boldsymbol{m}_k) - \frac{9\lambda_*^3}{8C_3^2},
\end{aligned}$$

where recall that $\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{x}) \in T_{\boldsymbol{m}_k}(\Omega_q)$ is in the direction from $\boldsymbol{m}_k$ to $\boldsymbol{x}$ with the length equal to the great-circle (or geodesic) distance on Ω_q. (We indeed apply the Cauchy-Schwarz inequality implicitly to obtain the first inequality in (54).) Then, by the uniform consistency of $\widehat{f}_h$ (Theorem 4), when h is sufficiently small and $\frac{nh^q}{|\log h|}$ is large enough,

$$\left|\left|\widehat{f}_h - f\right|\right|_\infty < \frac{9\lambda_*}{16C_3^2} \tag{55}$$

with probability at least $1-\delta$ for any $0 < \delta < 1$. We thus conclude that there must be at least one estimated local mode $\widehat{\boldsymbol{m}}_k$ within S_k. (If, on the contrary, there exists no $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$ within S_k, then the maximum of $\widehat{f}_h$ on S_k is attained at the boundary ∂S_k, that is, $\max_{\boldsymbol{x}\in S_k} \widehat{f}_h(\boldsymbol{x}) = \max_{\boldsymbol{x}\in \partial S_k} \widehat{f}_h(\boldsymbol{x})$. However, $\max_{\boldsymbol{x}\in \partial S_k} \widehat{f}_h(\boldsymbol{x}) < \max_{\boldsymbol{x}\in \partial S_k} f(\boldsymbol{x}) + \frac{9\lambda_*}{16C_3^2} \leq f(\boldsymbol{m}_k) - \frac{9\lambda_*}{16C_3^2} < \widehat{f}_h(\widehat{\boldsymbol{m}}_k)$ by (54), contradiction.) Note that this argument can be generalized to each $k = 1, ..., K$.

Statement (b). With (M2), we know that whenever

$$\begin{aligned} \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} &= \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\texttt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) - \texttt{Tang}\left(\nabla f(\boldsymbol{x})\right)\right|\right|_{\max} \leq \Theta_1, \\ \sup_{\boldsymbol{x}\in\Omega_q} \left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathcal{H}f(\boldsymbol{x})\right|\right|_{\max} &\leq \Theta_2 \end{aligned} \tag{56}$$

for some $\Theta_2 > 0$, the followings hold simultaneously:

(i) $||\texttt{grad}\, f(\widehat{\boldsymbol{m}}_k)||_{\max} = \left|\left|\texttt{grad}\, f(\widehat{\boldsymbol{m}}_k) - \underbrace{\texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{m}}_k)}_{=0}\right|\right|_{\max} \leq \Theta_1,$

(ii) $\sup_{\boldsymbol{x}\in S_k} \lambda_1\left(\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right) < 0$ and $\lambda_1\left(\mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right) \leq -\frac{\lambda_*}{2} - (q+1)\Theta_2$ by choosing $\Theta_2 > 0$ properly,

(iii) and

$$\begin{aligned} \lambda_1\left(\mathcal{H}f(\widehat{\boldsymbol{m}}_k)\right) &\leq \lambda_1\left(\mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right) + \lambda_{q-1}\left(\mathcal{H}f(\widehat{\boldsymbol{m}}_k) - \mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right) \\ &\leq -\frac{\lambda_*}{2} - (q+1)\Theta_2 + (q+1)\Theta_2 = -\frac{\lambda_*}{2} \end{aligned}$$

by Weyl's theorem (Theorem 4.3.1 in Horn and Johnson 2012) and the fact that

$$\begin{aligned} \left|\lambda_{q-1}(\mathcal{H}f(\widehat{\boldsymbol{m}}_k) - \mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right| &\leq \sup_{||v||_2=1} \left|\left|\left[\mathcal{H}f(\widehat{\boldsymbol{m}}_k) - \mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right] v\right|\right|_2 \\ &\leq \sqrt{(q+1)\times(q+1)}\left|\left|\mathcal{H}f(\widehat{\boldsymbol{m}}_k) - \mathcal{H}\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_{\max} \\ &\leq (q+1)\Theta_2. \end{aligned}$$

See Section 3.3 in Genovese et al. (2014) for detailed relations between different types of matrix norms.

Notice that (ii) is true because $\lambda_1(\boldsymbol{m}_k) \leq \lambda_*$ by (M1) and the difference between $\mathcal{H}\widehat{f}_h$ and $\mathcal{H}f$ will be minute given a small Θ_2. By (i) and (iii), we conclude that $\widehat{\mathcal{M}}_n \subset \mathcal{M} \oplus \rho_*$. By (ii) and Lemma 3.2 in Banyaga and Hurtubise (2004), there is only one estimated local mode $\widehat{\boldsymbol{m}}_k$ within S_k. They both hold with probability at least $1-\delta$ for any $\delta \in (0,1)$.

In total, a sufficient condition for the number of true local modes and estimated local modes being the same is a combination of the inequalities in (55) and (56). That is,

$$\begin{aligned} \left|\left|\widehat{f}_h - f\right|\right|_\infty &< \frac{9\lambda_*}{16C_3^2} \\ \left|\left|\mathtt{Tang}\left(\nabla\widehat{f}_h\right) - \mathtt{Tang}\left(\nabla f\right)\right|\right|_{\max,\infty} &\leq \Theta_1 \\ \left|\left|\widehat{\mathcal{H}}f - \mathcal{H}f\right|\right|_{\max,\infty} &\leq \Theta_2. \end{aligned} \tag{57}$$

By bias bounds in Theorem 2 or Theorem 4, as h is sufficiently small, we have

$$\begin{gathered} \left|\left|\mathbb{E}\left[\widehat{f}_h\right] - f\right|\right|_\infty < \frac{9\lambda_*}{32C_3^2}, \quad \sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathbb{E}\left[\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})\right] - \mathtt{grad}\,f(\boldsymbol{x})\right|\right|_{\max} \leq \frac{\Theta_1}{2}, \\ \text{and } \sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right] - \mathcal{H}f(\boldsymbol{x})\right|\right|_{\max} \leq \frac{\Theta_2}{2}. \end{gathered}$$

Therefore, (57) holds whenever

$$\begin{aligned} \left|\left|\widehat{f}_h - \mathbb{E}\left[\widehat{f}_h\right]\right|\right|_\infty &< \frac{9\lambda_*}{32C_3^2} \\ \sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})\right]\right|\right|_{\max} &\leq \frac{\Theta_1}{2} \\ \sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]\right|\right|_{\max} &\leq \frac{\Theta_2}{2} \end{aligned} \tag{58}$$

and h is sufficiently small. Now applying Talagrand's inequality Talagrand (1996); Giné and Guillou (2002), there exist constants $A_0, A_1, A_2 > 0$ and $B_0, B_1, B_2 > 0$ such that when n is large enough,

$$\begin{aligned} \mathbb{P}\left(\left|\left|\widehat{f}_h - \mathbb{E}\left[\widehat{f}_h\right]\right|\right|_\infty \geq \epsilon\right) &\leq B_0 e^{-A_0\epsilon^2 nh^q} \\ \mathbb{P}\left(\sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathtt{grad}\,\widehat{f}_h(\boldsymbol{x})\right]\right|\right|_{\max} \geq \epsilon\right) &\leq B_1 e^{-A_1\epsilon^2 nh^{q+2}} \\ \mathbb{P}\left(\sup_{\boldsymbol{x}\in\Omega_q}\left|\left|\mathcal{H}\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\mathcal{H}\widehat{f}_h(\boldsymbol{x})\right]\right|\right|_{\max} \geq \epsilon\right) &\leq B_2 e^{-A_2\epsilon^2 nh^{q+4}}. \end{aligned} \tag{59}$$

Combining (58) and (59), we conclude that there exist some constants $A_3, B_3 > 0$ such that

$$\mathbb{P}\Big((57) \text{ holds}\Big) > 1 - B_3 e^{-A_3 nh^{q+4}}$$

when h is sufficiently small. Since the condition (57) implies $\widehat{K}_n = K$, we conclude that

$$\mathbb{P}\left(\widehat{K}_n \neq K\right) \leq B_3 e^{-A_3 n h^{q+4}}$$

for some constants $A_3, B_3 > 0$ as h is sufficiently small. This proves the so-called modal consistency.

Statement (c). To establish the convergence rate of the Hausdorff distance between $\widehat{\mathcal{M}}_n$ and $\mathcal{M}$, we assume that (57) holds so that $K = \widehat{K}_n$ and each local mode is approximating by an unique estimated local mode. Notice that $||\boldsymbol{m}_k - \widehat{\boldsymbol{m}}_k||_2$ is upper bounded by the great-circle distance between these two points. Then,

$$\begin{aligned}
&\texttt{grad}\, f(\widehat{\boldsymbol{m}}_k) \\
&= \texttt{Tang}\left(\nabla f(\widehat{\boldsymbol{m}}_k)\right) - \underbrace{\texttt{Tang}\left(\nabla f(\boldsymbol{m}_k)\right)}_{=0} \\
&= \nabla \texttt{Tang}\left(\nabla f(\boldsymbol{m}_k)\right) \cdot \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k) + o\left(\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k)\right|\right|_2\right) \\
&= \left[(I_{q+1} - \boldsymbol{m}_k \boldsymbol{m}_k^T)\nabla\nabla f(\boldsymbol{m}_k) - \boldsymbol{m}_k^T \nabla f(\boldsymbol{m}_k) I_{q+1} - \boldsymbol{m}_k \nabla f(\boldsymbol{m}_k)^T\right] \cdot \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k) \\
&\quad + o\left(\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k)\right|\right|_2\right) \\
&= \left[\mathcal{H} f(\boldsymbol{m}_k)\right] \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k) + o\left(\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k)\right|\right|_2\right),
\end{aligned} \tag{60}$$

because $\nabla f(\boldsymbol{m}_k) = ||\nabla f(\boldsymbol{m}_k)||_2 \cdot \boldsymbol{m}_k$ when $\boldsymbol{m}_k$ is a local mode and $\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k) \in T_{\boldsymbol{m}_k}$ is orthogonal to $\boldsymbol{m}_k$. Under (M1), the matrices $\mathcal{H} f \boldsymbol{m}_k$ are nonsingular for all $\boldsymbol{m}_k \in \mathcal{M}$ inside the tangent space $T_{\boldsymbol{m}_k}$, respectively. As the chord distance between two points on Ω_q is bounded by their great-circle distance, we multiply $[\mathcal{H} f(\boldsymbol{m}_k)]^{-1}$ on both sides of (60) and obtain that

$$||\widehat{\boldsymbol{m}}_k - \boldsymbol{m}_k||_2 \leq \left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k)\right|\right|_2 = [\mathcal{H} f(\boldsymbol{m}_k)]^{-1} \texttt{grad}\, f(\widehat{\boldsymbol{m}}_k) + o\left(\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\widehat{\boldsymbol{m}}_k)\right|\right|_2\right),$$

where the matrix inverse, strictly speaking, is taken with respect to the local coordinate system near $\boldsymbol{m}_k$. Note that $\left|\left|[\mathcal{H} f(\boldsymbol{m}_k)]^{-1}\right|\right|_{\max}$ is bounded within $T_{\boldsymbol{m}_k}$ for all $\boldsymbol{m}_k \in \mathcal{M}$ under the assumption (57). Moreover, by Theorem 2,

$$\begin{aligned}
\texttt{grad}\, f(\widehat{\boldsymbol{m}}_k) &= \texttt{grad}\, f(\widehat{\boldsymbol{m}}_k) - \underbrace{\texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{m}}_k)}_{=0} \\
&= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{q+2}}}\right).
\end{aligned}$$

Now applying this rate of convergence to each local mode and using the fact that

$$\texttt{Haus}\left(\widehat{\mathcal{M}}_n, \mathcal{M}\right) = \max_{k=1,\ldots,K} ||\widehat{\boldsymbol{m}}_k - \boldsymbol{m}_k||_2,$$

we obtain the final conclusion. ■

D.5 Proofs of Theorem 8, Lemma 10, and Theorem 11

Theorem 8 (Ascending Property) *If kernel $L : [0,\infty) \to [0,\infty)$ is monotonically decreasing, differentiable, and convex with $L(0) < \infty$, then the sequence $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0}^{\infty}$ is monotonically increasing and thus converges.*

Proof Obviously, the sequence $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0,1,\ldots}$ is bounded if the kernel function L is monotonically decreasing with $L(0) < \infty$. Hence, it suffices to show that it is monotonically increasing. The convexity and differentiability of kernel L imply that

$$L(x_2) - L(x_1) \geq L'(x_1) \cdot (x_2 - x_1) \tag{61}$$

for all $x_1, x_2 \in [0,\infty), x_1 \neq x_2$. Using the fact that (rearrangement from Algorithm 1)

$$\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1 - \widehat{\boldsymbol{y}}_s^T \boldsymbol{X}_i}{h^2}\right) = -\widehat{\boldsymbol{y}}_{s+1} \left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2$$

we have that

$$\begin{aligned}
\widehat{f}_h(\boldsymbol{y}_{s+1}) - \widehat{f}_h(\boldsymbol{y}_s) &= \frac{c_{h,q}(L)}{n} \sum_{i=1}^{n} \left[L\left(\frac{1 - \boldsymbol{y}_{s+1}^T \boldsymbol{X}_i}{h^2}\right) - L\left(\frac{1 - \boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right)\right] \\
&\geq \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^{n} L'\left(\frac{1 - \boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right) \cdot (\boldsymbol{y}_s - \boldsymbol{y}_{s+1})^T \boldsymbol{X}_i \\
&= \frac{c_{h,q}(L)}{nh^2} \cdot (\boldsymbol{y}_{s+1} - \boldsymbol{y}_s)^T \boldsymbol{y}_{s+1} \cdot \left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2 \\
&= \frac{c_{h,q}(L)}{2nh^2} \left|\left|\boldsymbol{y}_{s+1} - \boldsymbol{y}_s\right|\right|_2^2 \cdot \left|\left|\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{y}_s^T \boldsymbol{X}_i}{h^2}\right)\right|\right|_2 \\
&\geq 0,
\end{aligned} \tag{62}$$

where we use the fact that $2(\boldsymbol{y}_{s+1} - \boldsymbol{y}_s)^T \boldsymbol{y}_{s+1} = 2 - 2\boldsymbol{y}_s^T \boldsymbol{y}_{s+1} = ||\boldsymbol{y}_{s+1} - \boldsymbol{y}_s||_2^2$ between the third and fourth lines, given that $||\boldsymbol{y}_s||_2 = ||\boldsymbol{y}_{s+1}||_2 = 1$. ■

Lemma 10 *Assume conditions (D1) and (D2'). For any fixed $\boldsymbol{x} \in \Omega_q$, we have*

$$h^2 \cdot \mathtt{Rad}\left(\nabla \widehat{f}_h(\boldsymbol{x})\right) \asymp h^2 \cdot \nabla \widehat{f}_h(\boldsymbol{x}) = \boldsymbol{x} f(\boldsymbol{x}) C_{L,q} + o(1) + O_P\left(\sqrt{\frac{1}{nh^q}}\right)$$

as $nh^q \to \infty$ and $h \to 0$, where $C_{L,q} = -\frac{\int_0^\infty L'(r) r^{\frac{q}{2}-1} dr}{\int_0^\infty L(r) r^{\frac{q}{2}-1} dr} > 0$ is a constant depending only on kernel L and dimension q and "$\asymp$" stands for an asymptotic equivalence.

Proof The proof follows the same logic as the one for Theorem 2. Note that

$$\nabla \widehat{f}_h(\boldsymbol{x}) = \mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right] + \nabla \widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right]. \tag{63}$$

Recall that $\nabla \widehat{f}_h(\boldsymbol{x}) = -\frac{c_{h,q}(L)}{nh^2}\sum_{i=1}^{n} \boldsymbol{X}_i L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_i}{h^2}\right)$. The expectation of $\nabla \widehat{f}_h(\boldsymbol{x})$ is

$$\begin{aligned}
\mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right] &= \frac{c_{h,q}(L)}{h^2}\int_{\Omega_q}(-\boldsymbol{y})L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right) f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y}) \\
&= \frac{c_{h,q}(L)}{h^2}\int_{-1}^{1}\int_{\Omega_{q-1}}\left(-t\boldsymbol{x} - \sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right) L'\left(\frac{1-t}{h^2}\right) \\
&\quad \times f\left(-t\boldsymbol{x} - \sqrt{1-t^2}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\right)(1-t^2)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dt \\
&= c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(-\boldsymbol{x}-\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot L'(r) \\
&\quad \times f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr
\end{aligned} \tag{64}$$

by (a) in Lemma 21 and a change of variable $r = \frac{1-t}{h^2}$, where $\alpha_{\boldsymbol{x},\boldsymbol{\xi}} = -rh^2\boldsymbol{x} + h\sqrt{r(2-h^2r)}\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}$. By condition (D1), the first-order Taylor's expansion of f at $\boldsymbol{x} \in \Omega_q$ is

$$f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}}) = f(\boldsymbol{x}) + O\left(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2\right),$$

where $||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2^2 = 2rh^2$ by the orthogonality of $\boldsymbol{x}$ and columns of $\boldsymbol{B}_{\boldsymbol{x}}$. Now we plug it back into (64) respectively to compute the dominating term of $\mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right]$.

$$\begin{aligned}
&\mathbb{E}\left[\nabla \widehat{f}_h(\boldsymbol{x})\right] \\
&= -c_{h,q}(L)h^{q-2}\boldsymbol{x}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}} L'(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad - c_{h,q}(L)h^{q-2}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}} \alpha_{\boldsymbol{x},\boldsymbol{\xi}}L'(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\quad + O(h)\cdot c_{h,q}(L)h^{q-2}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(-\boldsymbol{x}-\alpha_{\boldsymbol{x},\boldsymbol{\xi}})L'(r)r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr \\
&\overset{\text{(i)}}{=} -c_{h,q}(L)h^{q-2}\boldsymbol{x}f(\boldsymbol{x})\cdot\omega_{q-1}\int_0^{2h^{-2}} L'(r)\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\,dr \\
&\quad + c_{h,q}(L)h^{q}\boldsymbol{x}f(\boldsymbol{x})\cdot\omega_{q-1}\int_0^{2h^{-2}} L'(r)\cdot r^{\frac{q}{2}}(2-h^2r)^{\frac{q}{2}-1}\,dr \\
&\quad - c_{h,q}(L)h^{q-1}f(\boldsymbol{x})\int_0^{2h^{-2}}\int_{\Omega_{q-1}} \boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}\cdot L'(r)r^{\frac{q-1}{2}}(2-h^2r)^{\frac{q-1}{2}}\omega_{q-1}(d\boldsymbol{\xi})dr + O(h^{-1}), \\
&\overset{\text{(ii)}}{=} -c_{h,q}(L)h^{q-2}\boldsymbol{x}f(\boldsymbol{x})\cdot\omega_{q-1}\int_0^{2h^{-2}} L'(r)\cdot r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\,dr + O(1) + 0 + O(h^{-1})
\end{aligned}$$

$$\overset{\text{(iii)}}{=} -\frac{\boldsymbol{x}f(\boldsymbol{x})}{h^2}\cdot\frac{\int_0^\infty L'(r)r^{\frac{q}{2}-1}dr}{\int_0^\infty L(r)r^{\frac{q}{2}-1}dr}+o(h^{-2})$$
$$\equiv C_{L,q}\cdot\boldsymbol{x}f(\boldsymbol{x})h^{-2}+o(h^{-2}),$$

where we use (b) of Lemma 21 and the fact that $\boldsymbol{B}_{\boldsymbol{x}}\boldsymbol{\xi}=\sum_{i=1}^q\xi_i\boldsymbol{b}_i$ in (ii), and apply condition (D2') and (3) to argue that

$$c_{h,q}(L)h^q\int_0^{2h^{-2}}\int_{\Omega_{q-1}}L'(r)\cdot\phi(r,\boldsymbol{\xi})\;\omega_{q-1}(d\boldsymbol{\xi})dr\asymp O(1) \tag{65}$$

in both (i) and (ii). We also use the asymptotic relation (3) in (iii) and denote $C_{L,q}=-\frac{\int_0^\infty L'(r)r^{\frac{q}{2}-1}dr}{\int_0^\infty L(r)r^{\frac{q}{2}-1}dr}$ in the last equality. Thus,

$$\mathbb{E}\left[\mathtt{Rad}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)\right]=\boldsymbol{x}\boldsymbol{x}^T\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]=C_{L,q}\cdot\boldsymbol{x}f(\boldsymbol{x})h^{-2}+o(h^{-2}).$$

Based on the asymptotic rate of $\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]$, we calculate the covariance matrix of $\nabla\widehat{f}_h(\boldsymbol{x})$ as

$$\begin{aligned}
\mathrm{Cov}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]&=\frac{c_{h,q}(L)^2}{nh^4}\cdot\mathrm{Cov}\left[\boldsymbol{X}_1\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_1}{h^2}\right)\right]\\
&=\frac{c_{h,q}(L)^2}{nh^4}\cdot\mathbb{E}\left[\boldsymbol{X}_1\boldsymbol{X}_1^T\cdot L'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{X}_1}{h^2}\right)^2\right]-\frac{1}{n}\cdot\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]\mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]^T\\
&=\frac{c_{h,q}(L)^2}{nh^4}\int_{\Omega_q}\boldsymbol{y}\boldsymbol{y}^TL'\left(\frac{1-\boldsymbol{x}^T\boldsymbol{y}}{h^2}\right)^2f(\boldsymbol{y})\,\omega_q(d\boldsymbol{y})+O\left(\frac{1}{nh^4}\right)\\
&=\frac{c_{h,q}(L)^2}{n}h^{q-4}\int_0^{2h^{-2}}\int_{\Omega_{q-1}}(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})^TL'(r)^2\\
&\quad\times f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}\omega_{q-1}(d\boldsymbol{\xi})dr+O\left(\frac{1}{nh^4}\right).
\end{aligned}$$

With condition (D1), we carry out the first-order Taylor's expansion of f at $\boldsymbol{x}\in\Omega_q$ as

$$f(\boldsymbol{x}+\alpha_{\boldsymbol{x},\boldsymbol{\xi}})=f(\boldsymbol{x})+O\left(||\alpha_{\boldsymbol{x},\boldsymbol{\xi}}||_2\right)=f(\boldsymbol{x})+O(h).$$

Therefore,

$$\begin{aligned}
\mathrm{Cov}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]&=\frac{c_{h,q}(L)^2}{n}h^{q-4}\boldsymbol{x}\boldsymbol{x}^Tf(\boldsymbol{x})\omega_{q-1}\int_0^{2h^{-2}}L'(r)^2r^{\frac{q}{2}-1}(2-h^2r)^{\frac{q}{2}-1}dr+o\left(\frac{1}{nh^{q+4}}\right)\\
&=\frac{\boldsymbol{x}\boldsymbol{x}^Tf(\boldsymbol{x})}{nh^{q+4}}\cdot\frac{\int_0^\infty L'(r)^2r^{\frac{q}{2}-1}dr}{\omega_{q-1}2^{\frac{q}{2}-1}\left(\int_0^\infty L(r)r^{\frac{q}{2}-1}dr\right)^2}+o\left(\frac{1}{nh^{q+4}}\right),
\end{aligned}$$

where we use (b) of Lemma 21, asymptotic rate (3), and (65) to absorb some higher order terms into $o\left(\frac{1}{nh^{q+4}}\right)$. The dominating term of $\mathrm{Cov}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]$ is in the radial direction, so by the central limit theorem,

$$\begin{aligned}
&\mathtt{Rad}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right) - \mathbb{E}\left[\mathtt{Rad}\left(\nabla\widehat{f}_h(\boldsymbol{x})\right)\right] \\
&\asymp \nabla\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right] \\
&= \mathrm{Cov}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]^{\frac{1}{2}} \cdot \mathrm{Cov}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]^{-\frac{1}{2}} \left\{\nabla\widehat{f}_h(\boldsymbol{x}) - \mathbb{E}\left[\nabla\widehat{f}_h(\boldsymbol{x})\right]\right\} \\
&= \left[\frac{\boldsymbol{x}\boldsymbol{x}^T f(\boldsymbol{x})}{nh^{q+4}} \cdot \frac{\int_0^\infty L'(r)^2 r^{\frac{q}{2}-1} dr}{\omega_{q-1} 2^{\frac{q}{2}-1}\left(\int_0^\infty L(r) r^{\frac{q}{2}-1} dr\right)^2} + o\left(\frac{1}{nh^{q+4}}\right)\right]^{\frac{1}{2}} \cdot \widehat{\boldsymbol{Z}}_n(\boldsymbol{x}) \\
&= O_P\left(\sqrt{\frac{1}{nh^{q+4}}}\right),
\end{aligned}$$

where $\widehat{\boldsymbol{Z}}_n(\boldsymbol{x}) \xrightarrow{d} N_{q+1}(\boldsymbol{0}, I_{q+1})$. In total, we conclude with (63) that

$$\mathtt{Rad}\left(h^2\nabla\widehat{f}_h(\boldsymbol{x})\right) \asymp h^2\nabla\widehat{f}_h(\boldsymbol{x}) = \boldsymbol{x} f(\boldsymbol{x}) C_{L,q} + o(1) + O_P\left(\sqrt{\frac{1}{nh^q}}\right)$$

for any fixed $\boldsymbol{x} \in \Omega_q$, as $h \to 0$ and $nh^q \to \infty$. ■

Before we prove Theorem 11, we first note the following useful result.

Proposition 23 *Assume (C1) and the conditions on the kernel L in Theorem 8. Then for any mode $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$ satisfying (C2), we have that $\widehat{\boldsymbol{m}}_k^T \nabla\widehat{f}_h(\widehat{\boldsymbol{m}}_k) > 0$.*

Proof Suppose, on the contrary, that $\widehat{\boldsymbol{m}}_k^T \nabla\widehat{f}_h(\widehat{\boldsymbol{m}}_k) < 0$. By the definition of a local mode $\widehat{\boldsymbol{m}}_k$ of $\widehat{f}_h$ on Ω_q, we know that $\left|\left|\mathtt{grad}\,\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2 = \left|\left|\mathtt{Tang}\left(\nabla\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right)\right|\right|_2 = 0$. Then

$$\frac{\nabla\widehat{f}_h(\widehat{\boldsymbol{m}}_k)}{\left|\left|\nabla\widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2} = -\widehat{\boldsymbol{m}}_k$$

and by (C1), there exist a $\widehat{r}_k \in (0, 2]$ such that $\mathtt{Tang}\left(\nabla\widehat{f}_h(\boldsymbol{y})\right) \neq 0$ and $\widehat{f}_h(\boldsymbol{y}) \leq \widehat{f}_h(\widehat{\boldsymbol{m}}_k)$ for any $\boldsymbol{y} \in \left\{\boldsymbol{z} \in \Omega_q : \boldsymbol{z}^T\widehat{\boldsymbol{m}}_k \geq 1 - \frac{\widehat{r}_k^2}{2}\right\} \setminus \{\widehat{\boldsymbol{m}}_k\} = \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \leq \widehat{r}_k\} \setminus \{\widehat{\boldsymbol{m}}_k\}$. That is, $\widehat{\boldsymbol{m}}_k$ is the unique mode inside its neighborhood $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \leq \widehat{r}_k\}$. Since the sum of convex functions is convex, $\widehat{f}_h$ is indeed convex and we deduce that when $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \leq \widehat{r}_k\} \setminus \{\widehat{\boldsymbol{m}}_k\}$,

$$\widehat{f}_h(\widehat{\boldsymbol{m}}_k) - \widehat{f}_h(\boldsymbol{y}) \leq \frac{c_{h,q}(L)}{nh^2} \sum_{i=1}^n L'\left(\frac{1 - \widehat{\boldsymbol{m}}_k^T \boldsymbol{X}_i}{h^2}\right) \boldsymbol{X}_i^T(\boldsymbol{y} - \widehat{\boldsymbol{m}}_k)$$

$$
\begin{aligned}
&= \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2 \cdot (-\widehat{\boldsymbol{m}}_k)^T(\widehat{\boldsymbol{m}}_k - \boldsymbol{y}) \\
&= \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2 \cdot (\widehat{\boldsymbol{m}}_k^T\boldsymbol{y} - 1) \\
&< 0
\end{aligned}
$$

contradicting to the fact that $\widehat{\boldsymbol{m}}_k$ is the unique local mode in $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$. The result follows. ■

Theorem 11 *Assume (C1) and (C2) and the conditions on kernel L in Theorem 8. We further assume that L is continuously differentiable. Then, for each local mode $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$, there exists a $\widehat{r}_k > 0$ such that the sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$ converges to $\widehat{\boldsymbol{m}}_k$ whenever the initial point $\widehat{\boldsymbol{y}}_0 \in \Omega_q$ satisfies $||\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k$. Moreover, under conditions (D1) and (D2'), there exists a fixed constant $r^* > 0$ such that $\mathbb{P}(\widehat{r}_k \ge r^*) \to 1$ as $h \to 0$ and $nh^q \to \infty$.*

Proof By the definition of a local mode $\widehat{\boldsymbol{m}}_k$ of $\widehat{f}_h$ on Ω_q,

$$
\left|\left|\mathtt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2 = \left|\left|\mathtt{Tang}\left(\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right)\right|\right|_2 = 0.
$$

Hence, with condition (C2) imposed on $\widehat{\boldsymbol{m}}_k$ and Proposition 23,

$$
\frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2} = \widehat{\boldsymbol{m}}_k \quad \text{and} \quad \widehat{\boldsymbol{m}}_k^T \cdot \frac{\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)}{\left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k)\right|\right|_2} = 1.
$$

It indicates that our one-step fixed-point iteration of Algorithm 1 on the local mode $\widehat{\boldsymbol{m}}_k$ will yield $\widehat{\boldsymbol{m}}_k$ itself. (This is the so-called consistency of fixed-point iterations.) Moreover, there exists a $\widehat{r}_k > 0$ such that $\widehat{\boldsymbol{m}}_k$ is the only point in $\{\boldsymbol{y} \in \Omega_q : ||\boldsymbol{y} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$ satisfying $\left|\left|\mathtt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{y})\right)\right|\right|_2 = 0$. See Figure 4 for a graphical illustration.
In addition, given that L is continuously differentiable, we may shrink $\widehat{r}_k > 0$ if necessary so that

$$
\widehat{\boldsymbol{m}}_k^T \cdot \frac{\nabla \widehat{f}_h(\boldsymbol{y})}{\left|\left|\nabla \widehat{f}_h(\boldsymbol{y})\right|\right|_2} \ge 1 - \frac{\widehat{r}_k^2}{2} \quad \text{and} \quad \left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1 - \boldsymbol{y}^T\boldsymbol{X}_i}{h^2}\right)\right|\right|_2 = \frac{nh^2}{c_{h,q}(L)}\left|\left|\nabla \widehat{f}_h(\boldsymbol{y})\right|\right|_2 \ge \widehat{C}_k \tag{66}
$$

for all $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$ and some constant $\widehat{C}_k > 0$. The first inequality in (66) ensures that the sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$ yielded by our fixed-point iteration will not jump outside of the set $\{\boldsymbol{y} \in \Omega_q : ||\boldsymbol{y} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$ as long as the initial point $\widehat{\boldsymbol{y}}_0$ is in the set. It also guarantees the correctness of the second inequality in (66) for the iterative sequence $\{\widehat{\boldsymbol{y}}_s\}_{s=0}^{\infty}$. By (62) in the proof of Theorem 8, we know that

$$
\begin{aligned}
\widehat{f}_h(\widehat{\boldsymbol{y}}_{s+1}) - \widehat{f}_h(\widehat{\boldsymbol{y}}_s) &\ge \frac{c_{h,q}(L)}{2nh^2}||\widehat{\boldsymbol{y}}_{s+1} - \widehat{\boldsymbol{y}}_s||_2^2 \cdot \left|\left|\sum_{i=1}^n \boldsymbol{X}_i L'\left(\frac{1 - \widehat{\boldsymbol{y}}_s^T\boldsymbol{X}_i}{h^2}\right)\right|\right|_2 \\
&\ge \frac{c_{h,q}(L) \cdot \widehat{C}_k}{2nh^2} \cdot ||\widehat{\boldsymbol{y}}_{s+1} - \widehat{\boldsymbol{y}}_s||_2^2,
\end{aligned}
$$

where we used (66) in the last strict inequality. Since $\left\{\widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right\}_{s=0}^{\infty}$ converges by Theorem 8 as $s \to \infty$, we conclude that

$$\lim_{s\to\infty} ||\widehat{\boldsymbol{y}}_{s+1} - \widehat{\boldsymbol{y}}_s||_2^2 = 0 \quad \text{or equivalently,} \quad \lim_{s\to\infty} \widehat{\boldsymbol{y}}_{s+1}^T \widehat{\boldsymbol{y}}_s = 1. \tag{67}$$

Now with the expression (25),

$$\begin{aligned}
\left|\left|\texttt{Tang}\left(\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)\right|\right|_2^2 &= \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s) - \widehat{\boldsymbol{y}}_s^T \nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s) \cdot \widehat{\boldsymbol{y}}_s\right|\right|_2^2 \\
&= \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2^2 \cdot \left|\left|\widehat{\boldsymbol{y}}_{s+1} - \widehat{\boldsymbol{y}}_{s+1}^T \widehat{\boldsymbol{y}}_s \cdot \widehat{\boldsymbol{y}}_s\right|\right|_2^2 \\
&= \left|\left|\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right|\right|_2^2 \cdot \left[1 - \left(\widehat{\boldsymbol{y}}_{s+1}^T \widehat{\boldsymbol{y}}_s\right)^2\right],
\end{aligned}$$

where we plug in (23) in the second equality. As the function $\boldsymbol{u} \mapsto \left|\left|\nabla \widehat{f}_h(\boldsymbol{u})\right|\right|_2^2$ is continuous on a compact set Ω_q, it is upper bounded on Ω_q. As $s \to \infty$, $\left|\left|\texttt{Tang}\left(\nabla \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)\right|\right|_2 \to 0$ by (67). Given that $\widehat{\boldsymbol{m}}_k$ is the unique point in $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$ satisfying this $\left|\left|\texttt{Tang}\left(\nabla \widehat{f}_h(\boldsymbol{y})\right)\right|\right|_2 = 0$, we conclude that $\widehat{\boldsymbol{y}}_s \to \widehat{\boldsymbol{m}}_k$ as $s \to \infty$ and $\widehat{\boldsymbol{y}}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \widehat{\boldsymbol{m}}_k||_2 \le \widehat{r}_k\}$.
Now with Lemma 10, we know that $\widehat{\boldsymbol{m}}_k^T \nabla \widehat{f}_h(\widehat{\boldsymbol{m}}_k) > 0$ for any $\widehat{\boldsymbol{m}}_k \in \widehat{\mathcal{M}}_n$ with probability tending to 1 as $h \to 0$ and $nh^q \to \infty$. Therefore, as h is small enough and n is sufficiently large, there exists a fixed constant $r^* > 0$ such that $r^* \le \min_k \widehat{r}_k$ with high probability. The results follow. ■

D.6 Proof of Theorem 12

Before proving Theorem 12, we introduce the following two useful results. As pointed out in Zhang and Sra (2016), a main hurdle in analyzing non-asymptotic convergence of first-order methods on smooth manifolds is that the Euclidean law of cosines does not hold. Fortunately, there is a trigonometric distance bound stated below for Alexandrov space (Burago et al., 1992) with curvature bounded below.

Lemma 24 (Lemma 5 in Zhang and Sra 2016; see also Bonnabel 2013) *If a, b, c are the sides (that is, side lengths) of a geodesic triangle in an Alexandrov space with sectional curvature (see Appendix B) lower bounded by κ, and A is the angle between sides b and c, then*

$$a^2 \le \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)} b^2 + c^2 - 2bc\cos(A). \tag{68}$$

The sketching proof of Lemma 24 can be founded in Lemma 5 of Zhang and Sra (2016). Note that $\kappa = 1$ on Ω_q. We inherit the notation in Zhang and Sra (2016) and denote $\frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)}$ by $\zeta(\kappa, c)$ for the curvature dependent quantity from inequality (68). One can show by differentiating $\zeta(\kappa, c)$ with respect to c that $\zeta(\kappa, c)$ is strictly increasing and greater

than 1 for any $c > 0$ and fixed $\kappa \neq 0$. With Lemma 24 in hand, we are able to state a straightforward corollary indicating an important relation between two consecutive updates of a gradient ascent algorithm on Ω_q.

Corollary 25 *For any point $\boldsymbol{x}, \boldsymbol{y}_s$ in a convex set on Ω_q, the update in (30) satisfies*

$$2\eta\langle \texttt{grad}\, f(\boldsymbol{y}_s), \texttt{Exp}_{\boldsymbol{y}_s}^{-1}(\boldsymbol{x})\rangle \leq d^2(\boldsymbol{y}_s, \boldsymbol{x}) - d^2(\boldsymbol{y}_{s+1}, \boldsymbol{x}) + \zeta(1, d(\boldsymbol{y}_s, \boldsymbol{x})) \cdot \eta^2 ||\texttt{grad}\, f(\boldsymbol{y}_s)||_2^2,$$

recalling that $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\langle \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y}), \texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\rangle} = \left|\left|\texttt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2$ on Ω_q.

The proof is similar to Corollary 8 in Zhang and Sra (2016) and thus omitted.

Theorem 12 *Assume (D1) and (M1).*

(a) ***Linear convergence of gradient ascent with*** *f: Given a convergence radius r_0 with $0 < r_0 \leq \sqrt{2 - 2\cos\left[\frac{3\lambda_*}{2(q+1)^{\frac{3}{2}}C_3}\right]}$, the iterative sequence $\{\boldsymbol{y}_s\}_{s=0}^{\infty}$ defined by the population-level gradient ascent algorithm* (30) *satisfies*

$$d(\boldsymbol{y}_s, \boldsymbol{m}_k) \leq \Upsilon^s \cdot d(\boldsymbol{y}_0, \boldsymbol{m}_k) \quad \text{with} \quad \Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}},$$

whenever $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)C_3\zeta(1,r_0)}\right\}$ and the initial point $\boldsymbol{y}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ for some $\boldsymbol{m}_k \in \mathcal{M}$. We recall from Section 4.5 that C_3 is an upper bound for the derivatives of the directional density f up to the third order, $\lambda_* > 0$ is defined in (M1), and $\mathcal{M}$ is the set of local modes of the directional density f.

We further assume (D2') and (K1) in the sequel.

(b) ***Linear convergence of gradient ascent with*** *$\widehat{f}_h$: Let the sample-based gradient ascent update on Ω_q be $\widehat{\boldsymbol{y}}_{s+1} = \texttt{Exp}_{\boldsymbol{y}_s}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s)\right)$. With the same choice of the convergence radius $r_0 > 0$ and $\Upsilon = \sqrt{1 - \frac{\eta\lambda_*}{2}}$ as in (a), if $h \to 0$ and $\frac{nh^{q+2}}{|\log h|} \to \infty$, then for any $\delta \in (0, 1)$,*

$$d\left(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k\right) \leq \Upsilon^s \cdot d\left(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k\right) + O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$$

with probability at least $1 - \delta$, whenever $\eta \leq \min\left\{\frac{2}{\lambda_*}, \frac{1}{(q+1)C_3 \cdot \zeta(1,r_0)}\right\}$ and the initial point $\widehat{\boldsymbol{y}}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ for some $\boldsymbol{m}_k \in \mathcal{M}$.

Proof (a) **Linear convergence of gradient ascent with** f: The proof of the linear convergence of the population-level gradient ascent algorithm (30) is similar to some standard results in optimization theory, except that we are under the manifold context now. Recall from (30) that the iterative formula reads $\boldsymbol{y}_{s+1} = \texttt{Exp}_{\boldsymbol{y}_s}(\eta \cdot \texttt{grad}\, f(\boldsymbol{y}_s))$ for $s = 0, 1, ...$ We begin by deriving the following three facts.

• *Fact 1*: Given (M1), f is geodesically strongly concave around some small neighborhoods of $\mathcal{M}$. In particular, when $0 < r_0 \leq \sqrt{2 - 2\cos\left[\frac{3\lambda_*}{2(q+1)^{\frac{3}{2}}C_3}\right]}$,

$$f(\boldsymbol{y}) - f(\boldsymbol{m}_k) - \langle \texttt{grad}\, f(\boldsymbol{m}_k), \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle \leq -\frac{\lambda_*}{4}\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 \tag{69}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and any $\boldsymbol{m}_k \in \mathcal{M}$.
• *Fact 2*. Given (D1) and (M1), we know that $||\texttt{grad}\, f(\boldsymbol{x})||_2 \equiv ||\texttt{Tang}\,(\nabla f(\boldsymbol{x}))||_2 > 0$ and

$$f(\boldsymbol{m}_k) - f\left(\texttt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)C_3}\texttt{grad}\, f(\boldsymbol{x})\right)\right) \geq 0$$

for all $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\} \setminus \{\boldsymbol{m}_k\}$ and any $\boldsymbol{m}_k \in \mathcal{M}$.
• *Fact 3*. Given (D1), the directional density f is $(q+1)C_3$-smooth.

As for *Fact 1*, it follows from the differentiability of f guaranteed by (D1) and the eigenvalue condition (M1). By Taylor's expansion on manifolds (Pennec, 2006) and (M1),

$$\begin{aligned}
&f(\boldsymbol{y}) - f(\boldsymbol{m}_k) \\
&= \langle \texttt{grad}\, f(\boldsymbol{m}_k), \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle + \frac{1}{2}\cdot \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})^T\left[\mathcal{H}f(\boldsymbol{m}_k)\right]\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y}) + o\left(\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2\right) \\
&\leq \langle \texttt{grad}\, f(\boldsymbol{m}_k), \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle - \frac{\lambda_*}{2}\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 + \frac{(q+1)^{\frac{3}{2}}C_3}{6}\cdot\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^3
\end{aligned} \tag{70}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and $\boldsymbol{m}_k \in \mathcal{M}$. Since $||\boldsymbol{y} - \boldsymbol{m}_k||_2 \leq r_0$, the geodesic distance between $\boldsymbol{y}$ and $\boldsymbol{m}_k$ satisfies $d_g(\boldsymbol{y}, \boldsymbol{m}_k) = \left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2 = \arccos(\boldsymbol{y}^T\boldsymbol{m}_k) \leq \frac{3\lambda_*}{2(q+1)^{\frac{3}{2}}C_3}$. Plugging this result back into (70) yields that

$$f(\boldsymbol{y}) - f(\boldsymbol{m}_k) \leq \langle \texttt{grad}\, f(\boldsymbol{m}_k), \texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle - \frac{\lambda_*}{4}\left|\left|\texttt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2.$$

For our purpose, it suffices to only prove (69) as above. One can shrink the upper bound of the convergence radius $r_0 > 0$ so that the geodesically strong concavity is valid for any pair of points within $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$. Indeed, the local strong concavity of f is a natural consequence of Morse Lemma (Lemma 3.11 in Banyaga and Hurtubise (2004)) given (M1).

Fact 2 is an obvious result under the eigenvalue condition (M1) and differentiable condition (D1). This is because $\boldsymbol{m}_k$ is an unique local mode of f within the neighborhood $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and the geodesic distance between $\boldsymbol{x}$ and one-step gradient ascent iteration from $\boldsymbol{x}$ with the step size $\frac{1}{(q+1)C_3}$ satisfies

$$\begin{aligned}
d\left(\texttt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)C_3}\texttt{grad}\, f(\boldsymbol{x})\right), \boldsymbol{x}\right) &= \frac{1}{(q+1)C_3}\left|\left|\texttt{grad}\, f(\boldsymbol{x})\right|\right|_2 \\
&= \frac{1}{(q+1)C_3}\left|\left|\texttt{grad}\, f(\boldsymbol{x}) - \underbrace{\Gamma_{\boldsymbol{m}_k}^{\boldsymbol{x}}\left(\texttt{grad}\, f(\boldsymbol{m}_k)\right)}_{=0}\right|\right|_2
\end{aligned}$$

$$\leq \frac{1}{(q+1)C_3} ||\mathcal{H}f(\boldsymbol{x})||_2 \cdot \left|\left|\mathtt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{m}_k)\right|\right|_2$$
$$\leq \left|\left|\mathtt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{m}_k)\right|\right|_2 = d_g(\boldsymbol{x}, \boldsymbol{m}_k),$$

where we use the fact that $||\mathcal{H}f(\boldsymbol{x})||_2 \leq (q+1)\,||\mathcal{H}f(\boldsymbol{x})||_{\max} \leq (q+1)C_3$ to deduce the last inequality. This shows that the one-step gradient ascent iteration $\mathtt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)C_3} \cdot \mathtt{grad}\, f(\boldsymbol{x})\right)$ on Ω_q will stay within the neighborhood $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ whenever $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\} \setminus \{\boldsymbol{m}_k\}$. Therefore, $f(\boldsymbol{m}_k) - f\left(\mathtt{Exp}_{\boldsymbol{x}}\left(\frac{1}{(q+1)C_3} \cdot \mathtt{grad}\, f(\boldsymbol{x})\right)\right) \geq 0$.

As for *Fact 3*, note that $||\mathcal{H}f(\boldsymbol{x})||_{\max} \leq C_3$ for all $\boldsymbol{x} \in \Omega_q$. Thus,

$$\begin{aligned}\left|\left|\mathtt{grad}\, f(\boldsymbol{x}) - \Gamma_{\boldsymbol{y}}^{\boldsymbol{x}}\left(\mathtt{grad}\, f(\boldsymbol{y})\right)\right|\right|_2 &= \left|\left|\left(\mathcal{H}f(\boldsymbol{x})\right)\mathtt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}^*)\right|\right|_2 \\ &\leq ||\mathcal{H}f(\boldsymbol{x})||_2 \cdot \left|\left|\mathtt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2 \\ &\leq (q+1)C_3 \left|\left|\mathtt{Exp}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right|\right|_2,\end{aligned}$$

where $\boldsymbol{x}^* \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{x}||_2 \leq ||\boldsymbol{y} - \boldsymbol{x}||_2\}$, and we use the fact that $||A||_2 \leq \sqrt{mn}||A||_{\max}$ for any $A \in \mathbb{R}^{m\times n}$. See Section 3.3 in Genovese et al. (2014) or Section 5.6 in Horn and Johnson (2012) for detailed relations between different types of matrix norms.

With *Fact 1* and *Fact 3*, we have that

$$\begin{aligned}&-\frac{(q+1)C_3}{2}\left|\left|\mathtt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{m}_k) - \langle \mathtt{grad}\, f(\boldsymbol{m}_k), \mathtt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle, \\ &f(\boldsymbol{y}) - f(\boldsymbol{m}_k) - \langle \mathtt{grad}\, f(\boldsymbol{m}_k), \mathtt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\rangle \leq -\frac{\lambda_*}{4}\left|\left|\mathtt{Exp}_{\boldsymbol{m}_k}^{-1}(\boldsymbol{y})\right|\right|_2^2 < 0\end{aligned} \tag{71}$$

for any $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and $\boldsymbol{m}_k \in \mathcal{M}$.
Hence, given a point $\boldsymbol{y} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and using *Fact 2*,

$$\begin{aligned}&f(\boldsymbol{y}) - f(\boldsymbol{m}_k) \\ &\leq f(\boldsymbol{y}) - f(\boldsymbol{m}_k) + f(\boldsymbol{m}_k) - f\left(\mathtt{Exp}_{\boldsymbol{y}}\left(\frac{1}{(q+1)C_3} \cdot \mathtt{grad}\, f(\boldsymbol{y})\right)\right) \\ &= -\left[f\left(\mathtt{Exp}_{\boldsymbol{y}}\left(\frac{1}{(q+1)C_3} \cdot \mathtt{grad}\, f(\boldsymbol{y})\right)\right) - f(\boldsymbol{y})\right] \\ &\leq -\left[\langle \mathtt{grad}\, f(\boldsymbol{y}), \frac{1}{(q+1)C_3}\mathtt{grad}\, f(\boldsymbol{y})\rangle - \frac{(q+1)C_3}{2}\left|\left|\mathtt{Exp}_{\boldsymbol{y}}^{-1}\left(\frac{1}{(q+1)C_3} \cdot \mathtt{grad}\, f(\boldsymbol{y})\right)\right|\right|_2^2\right] \\ &= -\frac{1}{2(q+1)C_3}\left|\left|\mathtt{grad}\, f(\boldsymbol{y})\right|\right|_2^2,\end{aligned}$$

where we apply the first inequality in (71) to obtain the fourth line. Thus, for any $\boldsymbol{x} \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$,

$$||\mathtt{grad}\, f(\boldsymbol{x})||_2^2 \leq 2(q+1)C_3\left[f(\boldsymbol{m}_k) - f(\boldsymbol{x})\right]. \tag{72}$$

With $\boldsymbol{y}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ and Corollary 25, we deduce that

$$d^2(\boldsymbol{y}_{s+1}, \boldsymbol{m}_k) \leq d^2(\boldsymbol{y}_s, \boldsymbol{m}_k) - 2\eta\langle \mathtt{grad}\, f(\boldsymbol{y}_s), \mathtt{Exp}_{\boldsymbol{y}_s}^{-1}(\boldsymbol{m}_k)\rangle + \zeta(1, d(\boldsymbol{y}_s, \boldsymbol{m}_k)) \cdot \eta^2\,||\mathtt{grad}\, f(\boldsymbol{y}_s)||_2^2$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} d^2(\boldsymbol{y}_s, \boldsymbol{m}_k) + 2\eta \left[f(\boldsymbol{y}_s) - f(\boldsymbol{m}_k) - \frac{\lambda_*}{4} d^2(\boldsymbol{y}_s, \boldsymbol{m}_k) \right] \\
&\quad + \zeta(1, r_0) \cdot \eta^2 \cdot 2(q+1) C_3 \left[f(\boldsymbol{m}_k) - f(\boldsymbol{y}_s) \right] \\
&= \left(1 - \frac{\eta \lambda_*}{2} \right) d^2(\boldsymbol{y}_s, \boldsymbol{m}_k) - 2\eta \left[1 - \zeta(1, r_0)(q+1) C_3 \eta \right] \cdot \underbrace{\left[f(\boldsymbol{m}_k) - f(\boldsymbol{y}_s) \right]}_{\geq 0} \\
&\leq \left(1 - \frac{\eta \lambda_*}{2} \right) d^2(\boldsymbol{y}_s, \boldsymbol{m}_k)
\end{aligned}
$$

whenever $\eta \leq \min\left\{ \frac{2}{\lambda_*}, \frac{1}{(q+1) C_3 \cdot \zeta(1, r_0)} \right\}$, where we use the second inequality of (71), the monotonicity of $\zeta(1, c)$ with respect to c, and (72) to obtain (i). By telescoping, we conclude that when $\eta \leq \min\left\{ \frac{2}{\lambda_*}, \frac{1}{(q+1) C_3 \cdot \zeta(1, r_0)} \right\}$ and $\boldsymbol{y}_0 \in \{ \boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0 \}$,

$$
d(\boldsymbol{y}_s, \boldsymbol{m}_k) = \left|\left| \mathtt{Exp}_{\boldsymbol{y}_s}^{-1}(\boldsymbol{m}_k) \right|\right|_2 \leq \left(1 - \frac{\eta \lambda_*}{2} \right)^{\frac{s}{2}} \cdot d(\boldsymbol{y}_0, \boldsymbol{m}_k) = \left(1 - \frac{\eta \lambda_*}{2} \right)^{\frac{s}{2}} \left|\left| \mathtt{Exp}_{\boldsymbol{y}_0}^{-1}(\boldsymbol{m}_k) \right|\right|_2 .
$$

The result follows.

(b) **Linear convergence of gradient ascent with $\widehat{f}_h$**: The proof here is partially adopted from the proof of Theorem 2 in Balakrishnan et al. (2017). By Theorem 4 and the continuity of exponential map, when h is sufficiently small and $\frac{nh^{q+2}}{|\log h|}$ is sufficiently large, we have that for any $\delta \in (0, 1)$,

$$
\begin{aligned}
d\left(\mathtt{Exp}_{\boldsymbol{x}}(\eta \cdot \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x})), \mathtt{Exp}_{\boldsymbol{x}}(\eta \cdot \mathtt{grad}\, f(\boldsymbol{x})) \right) &\leq \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x} \in \Omega_q} \left|\left| \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \mathtt{grad}\, f(\boldsymbol{x}) \right|\right|_{\max} \\
&\equiv \epsilon_{n,h} \\
&\leq (1 - \Upsilon) \cdot \arccos\left(1 - \frac{r_0^2}{2} \right)
\end{aligned} \tag{73}
$$

with probability at least $1 - \delta$, where $\bar{C}_4$ is some constant independent of $\boldsymbol{x} \in \Omega_q$, and $\epsilon_{n,h} = \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x} \in \Omega_q} \left|\left| \mathtt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \mathtt{grad}\, f(\boldsymbol{x}) \right|\right|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}} \right)$.
We now claim that $d(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k) \leq \arccos\left(1 - \frac{r_0^2}{2} \right)$ and

$$
d(\widehat{\boldsymbol{y}}_{s+1}, \boldsymbol{m}_k) \leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k) + \epsilon_{n,h} \tag{74}
$$

for any fixed $s = 0, 1, 2, ...$ with probability at least $1 - \delta$. We will prove this claim by induction on the iteration number. Recall that

$$
\widehat{\boldsymbol{y}}_{s+1} = \mathtt{Exp}_{\widehat{\boldsymbol{y}}_s}\left(\eta \cdot \mathtt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_s) \right).
$$

Then with $s = 1$, we have that

$$
d(\widehat{\boldsymbol{y}}_1, \boldsymbol{m}_k)
$$

$$
\begin{aligned}
&= d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}_0}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_0)\right), \boldsymbol{m}_k\right) \\
&\leq d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}_0}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}_0)\right), \boldsymbol{m}_k\right) + d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}_0}\left(\eta \cdot \texttt{grad}\, \widehat{f}_h(\widehat{\boldsymbol{y}}_0)\right), \texttt{Exp}_{\widehat{\boldsymbol{y}}_0}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}_0)\right)\right) \\
&\leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k) + \eta \bar{C}_4 \cdot \sup_{\boldsymbol{x} \in \Omega_q} \left|\left|\texttt{grad}\, \widehat{f}_h(\boldsymbol{x}) - \texttt{grad}\, f(\boldsymbol{x})\right|\right|_{\max} \\
&= \Upsilon \cdot d(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k) + \epsilon_{n,h},
\end{aligned}
$$

where the first inequality follows by the triangle inequality, the second one is from our result in (a), whereas the third equality is by (73). The triangle inequality is valid in this context because a geodesic measures the minimal distance between two points on Ω_q. In addition, the bound in (73) and our initialization $\widehat{\boldsymbol{y}}_0 \in \{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ ensure that $d(\widehat{\boldsymbol{y}}_1, \boldsymbol{m}_k) \leq \arccos\left(1 - \frac{r_0^2}{2}\right)$. In the induction from $s \to s+1$, suppose that $d(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k) \leq \arccos\left(1 - \frac{r_0^2}{2}\right)$ and the claim (74) holds at step s. With the fact proved in (a) that

$$
d\left(\texttt{Exp}_{\widehat{\boldsymbol{y}}_s}\left(\eta \cdot \texttt{grad}\, f(\widehat{\boldsymbol{y}}_s)\right), \boldsymbol{m}_k\right) \leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k),
$$

the same argument implies that the claim (74) holds for step $s+1$ and $d(\widehat{\boldsymbol{y}}_{s+1}, \boldsymbol{m}_k) \leq \arccos\left(1 - \frac{r_0^2}{2}\right)$. The claim (74) is thus proved.

As a result, $\widehat{\boldsymbol{y}}_s$ always lies within $\{\boldsymbol{z} \in \Omega_q : ||\boldsymbol{z} - \boldsymbol{m}_k||_2 \leq r_0\}$ for all $s = 0, 1, ...$ Now, with this claim and $\Upsilon = \sqrt{1 - \frac{\eta \lambda_*}{2}} < 1$, we iterate it to show that

$$
\begin{aligned}
d(\widehat{\boldsymbol{y}}_s, \boldsymbol{m}_k) &\leq \Upsilon \cdot d(\widehat{\boldsymbol{y}}_{s-1}, \boldsymbol{m}_k) + \epsilon_{n,h} \\
&\leq \Upsilon \cdot [\Upsilon \cdot d(\widehat{\boldsymbol{y}}_{s-2}, \boldsymbol{m}_k) + \epsilon_{n,h}] + \epsilon_{n,h} \\
&\leq \Upsilon^s \cdot d(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k) + \left\{\sum_{k=0}^{s-1} \Upsilon^k\right\} \cdot \epsilon_{n,h} \\
&\leq \Upsilon^s \cdot d(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k) + \frac{\epsilon_{n,h}}{1 - \Upsilon} \\
&\leq \Upsilon^s \cdot d(\widehat{\boldsymbol{y}}_0, \boldsymbol{m}_k) + O(h^2) + O_P\left(\frac{|\log h|}{nh^{q+2}}\right),
\end{aligned}
$$

where the fourth inequality follows by summing the geometric series, and the last one follows from our notation that $\epsilon_{n,h} = O(h^2) + O_P\left(\sqrt{\frac{|\log h|}{nh^{q+2}}}\right)$. It completes the proof. ■

References

Timothy Abbott, Filipe B. Abdalla, J. Aleksić, S. Allam, Adam Amara, D. Bacon, Eduardo Balbinot, M. Banerji, Keith Bechtol, Aurélien Benoit-Lévy, et al. The dark energy survey: more than dark energy–an overview. *Monthly Notices of the Royal Astronomical Society*, 460(2):1270–1299, 2016.

P-A Absil, Robert Mahony, and Jochen Trumpf. An extrinsic look at the riemannian hessian. In *Proceedings of the International Conference on Geometric Science of Information*, pages 361–368. Springer, 2013.

Youness Aliyari Ghassabeh. On the convergence of the mean shift algorithm in the one-dimensional space. *Pattern Recognition Letters*, 34(12):1423–1427, 2013.

Youness Aliyari Ghassabeh. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis*, 135:1–10, 2015.

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.

Mina Ashizawa, Hiroaki Sasaki, Tomoya Sakai, and Masashi Sugiyama. Least-squares log-density gradient clustering for riemannian manifolds. In *Proceedings of the Artificial Intelligence and Statistics*, pages 537–546. PMLR, 2017.

Z.D. Bai, C.Radhakrishna Rao, and L.C. Zhao. Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27(1):24–39, 1988.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1): 77–120, 2017.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46):1345–1382, 2005.

Augustin Banyaga and David Hurtubise. *Lectures on Morse Homology*. Springer Netherlands, 2004.

Nadine G. Barlow. Constraining geologic properties and processes through the use of impact craters. *Geomorphology*, 240:18–33, 2015.

Tyrus Berry and Timothy Sauer. Density estimation on manifolds with boundary. *Computational Statistics & Data Analysis*, 107:1–17, 2017.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, USA, 2004.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(4):231–357, 2015.

Yu Burago, Mikhail Gromov, and Gregory Perel'man. Ad alexandrov spaces with curvature bounded below. *Russian mathematical surveys*, 47(2):1–58, 1992.

Nathalie A. Cabrol and Edmond A. Grin, editors. *Lakes on Mars*. Elsevier, Amsterdam, 2010.

Miguel Á. Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 153–160, 2006.

Miguel Á Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.

Miguel Á. Carreira-Perpiñán. Generalised blurring mean-shift algorithms for nonparametric clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

Miguel Á Carreira-Perpiñán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

Rui Caseiro, João F. Henriques, Pedro Martins, and Jorge Batista. Semi-intrinsic mean shift on riemannian manifolds. In *Proceedings of the European Conference on Computer Vision*, pages 342–355. Springer Berlin Heidelberg, 2012.

Hasan Ertan Cetingul and René Vidal. Intrinsic mean shift for clustering on stiefel and grassmann manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1896–1902. IEEE, 2009.

José E Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

José E Chacón. The modal age of statistics. *International Statistical Review*, 88(1):122–141, 2020.

José E Chacón, Tarn Duong, and M. P. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807–840, 2011.

Shou-Jen Chang-Chien, Wen-Liang Hung, and Miin-Shen Yang. On mean shift-based clustering for circular data. *Soft Computing*, 16(6):1043–1060, 2012.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

Yen-Chi Chen, Christopher R. Genovese, and Larry Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

T.J. Craig, J.A. Jackson, K. Priestley, and D. McKenzie. Earthquake distribution patterns in africa: their relationship to variations in lithospheric and geological structure, and their rheological implications. *Geophysical Journal International*, 185(1):403–434, 2011.

Manfredo P. Do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Courier Dover Publications, 2016.

Costas Efthimiou and Christopher Frye. *Spherical Harmonics In p Dimensions*. World Scientific, 2014.

Uwe Einmahl and David M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics*, 33(3):1380–1403, 2005.

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

Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

Eduardo García-Portugués. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7:1655–1685, 2013.

Eduardo García-Portugués, Rosa M. Crujeiras, and Wenceslao González-Manteiga. Kernel density estimation for directional-linear data. *Journal of Multivariate Analysis*, 121:152–175, 2013.

Eduardo García-Portugués, Paula Navarro-Esteban, and Juan A Cuesta-Albertos. On a projection-based class of uniformity tests on the hypersphere. *arXiv preprint arXiv:2008.09897*, 2020.

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

Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002.

Peter Hall, G. S. Watson, and Javier Cabrara. Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762, 1987.

Harrie Hendriks. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *The Annals of Statistics*, 18(2):832–849, 1990.

Guillermo Henry and Daniela Rodriguez. Kernel density estimation on riemannian manifolds: Asymptotic results. *Journal of Mathematical Imaging and Vision*, 34(3):235–239, 2009.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, second edition, 2012.

Kejun Huang, Xiao Fu, and Nicholas Sidiropoulos. On convergence of epanechnikov mean shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ollivier Hyrien and Andrea Baran. Fast nonparametric density-based clustering of large data sets using a stochastic approximation mean-shift algorithm. *Journal of Computational and Graphical Statistics*, 25(3):899–916, 2016.

Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *Proceedings of the International Conference on Machine Learning*, pages 1694–1703. PMLR, 2017.

Mehran Kafai, Yiyi Miao, and Kazunori Okada. Directional mean shift and its application for topology classification of local 3d structures. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 170–177, 2010.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *Proceedings of the International Conference on Machine Learning*, pages 3398–3407. PMLR, 2019.

Jussi Klemelä. Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, 73(1):18–40, 2000.

Takumi Kobayashi and Nobuyuki Otsu. Von mises-fisher mean shift for clustering on a hypersphere. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 2130–2133, 2010.

John M. Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, second edition, 2012.

Christophe Ley and Thomas Verdebout. *Applied Directional Statistics: Modern Methods and Case Studies*. CRC Press, 2018.

Xiangru Li, Zhanyi Hu, and Fuchao Wu. A note on the convergence of the mean shift. *Pattern Recognition*, 40(6):1756–1762, 2007.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 2000.

Marco Di Marzio, Agnese Panzera, and Charles C. Taylor. Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, 141(6):2156–2173, 2011.

John Milnor. *Morse Theory. (AM-51), Volume 51*. Princeton University Press, 1963.

Marston Morse. Relations between the critical points of a real function of n independent variables. *Transactions of the American Mathematical Society*, 27(3):345–396, 1925.

Marston Morse. The foundations of a theory of the calculus of variations in the large in m-space (second paper). *Transactions of the American Mathematical Society*, 32(4): 599–631, 1930.

Deborah Nolan and David Pollard. U-processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.

Shigeyuki Oba, Kikuya Kato, and Shin Ishii. Multi-scale clustering for gene expression profiling data. In *Proceedings of the Fifth IEEE Symposium on Bioinformatics and Bioengineering*, pages 210–217, 2005.

María Oliveira, Rosa M Crujeiras, and Alberto Rodríguez-Casal. A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis*, 56(12):3898–3908, 2012.

Emanuel Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Bruno Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, 73(3):297–304, 2005.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Arthur Pewsey and Eduardo García-Portugués. Recent advances in directional statistics. *TEST*, pages 1–58, 2021.

Joseph P. Romano. On weak convergence and optimality of kernel density estimates of the mode. *Annals of Statistics*, 16(2):629–647, 06 1988.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Halsey Lawrence Royden and Patrick Fitzpatrick. *Real Analysis*. Pearson, fourth edition, 2010.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill New York, third edition, 1976.

Paula Saavedra-Nieves and Rosa María Crujeiras. Nonparametric estimation of directional highest density regions. *arXiv preprint arXiv:2009.08915*, 2020.

Hiroaki Sasaki, Takafumi Kanamori, Aapo Hyvärinen, Gang Niu, and Masashi Sugiyama. Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios. *Journal of Machine Learning Research*, 18(180):1–47, 2018.

David W. Scott. Multivariate density estimation and visualization. In *Handbook of computational statistics*, pages 549–569. Springer, 2012.

David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 2015.

Shou-Jen Chang-Chien, M. Yang, and Wen-Liang Hung. Mean shift-based clustering for directional data. In *Proceedings of the Third International Workshop on Advanced Computational Intelligence*, pages 367–372, Aug 2010.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

M. F. Skrutskie, R. M. Cutri, R. Stiening, M. D. Weinberg, S. Schneider, J. M. Carpenter, C. Beichman, R. Capps, T. Chester, J. Elias, J. Huchra, J. Liebert, C. Lonsdale, D. G. Monet, S. Price, P. Seitzer, T. Jarrett, J. D. Kirkpatrick, J. E. Gizis, E. Howard, T. Evans, J. Fowler, L. Fullmer, R. Hurt, R. Light, E. L. Kopan, K. A. Marsh, H. L. McCallon, R. Tam, S. Van Dyk, and S. Wheelock. The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2):1163–1183, 2006.

John P. Snyder, Philip M. Voxland, and Geological Survey (U.S.). *An Album of Map Projections*. Number 1453. U.S. Government Printing Office, 1989.

L.A. Soderblom, C.D. Condit, R.A. West, B.M. Herman, and T.J. Kreidler. Martian planetwide crater distributions: Implications for geologic history and surface processes. *Icarus*, 22(3):239–263, 1974.

Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of $i_s(x)$. *Computational Statistics*, 27(1):177–190, 2012.

Raghav Subbarao and Peter Meer. Nonlinear mean shift for clustering over analytic manifolds. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1168–1175. IEEE, 2006.

Raghav Subbarao and Peter Meer. Nonlinear mean shift over riemannian manifolds. *International Journal of Computer Vision*, 84(1):1–20, 2009.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, November 1996.

Charles C. Taylor. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52(7):3493–3500, 2008.

Michael Taylor and An Yin. Active structures of the himalayan-tibetan orogen and their relationships to earthquake distribution, contemporary strain field, and cenozoic volcanismactive structures on the tibetan plateau and surrounding regions. *Geosphere*, 5(3): 199–214, 2009.

Oncel Tuzel, Raghav Subbarao, and Peter Meer. Simultaneous multiple 3d motion estimation via mode finding on lie groups. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, volume 1, pages 18–25. IEEE, 2005.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Philippe Vieu. A note on density mode estimation. *Statistics & probability letters*, 26(4): 297–307, 1996.

Xiaogang Wang, Weiliang Qiu, and Jianhong Wu. Convergence and stability analysis of mean-shift algorithm on large data sets. *Statistics and Its Interface*, 9(2):159–170, 2016.

Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag, Berlin, Heidelberg, 2006.

Miin-Shen Yang, Shou-Jen Chang-Chien, and Hsun-Chih Kuo. On mean shift clustering for directional data on a hypersphere. In *Proceedings of the Artificial Intelligence and Soft Computing*, pages 809–818, Cham, 2014. Springer International Publishing.

Donald G. York, J. Adelman, John E. Anderson Jr, Scott F. Anderson, James Annis, Neta A. Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3): 1579–1587, 2000.

Xiao-Tong Yuan and Stan Z. Li. Stochastic gradient kernel density mode-seeking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1926–1931, 2009.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pages 1617–1638, Columbia University, New York, USA, 2016. PMLR.

Kai Zhang, Jamesk T. Kwok, and Ming Tang. Accelerated convergence using dynamic mean shift. In *Proceedings of the European Conference on Computer Vision*, pages 257–268. Springer Berlin Heidelberg, 2006.

Lincheng Zhao and Chengqing Wu. Central limit theorem for integrated square error of kernel estimators of spherical density. *Science in China Series A: Mathematics*, 44(4): 474–483, 2001.