Electronic Journal of Statistics
Vol. 17 (2023) 688–722
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2110

Analytical and statistical properties of local depth functions motivated by clustering applications*

Giacomo Francisci[1], Claudio Agostinelli[2], Alicia Nieto-Reyes[†3] and Anand N. Vidyashankar [1]

[1] *Department of Statistics, George Mason University*
e-mail: gfranci@gmu.edu; avidyash@gmu.edu

[2] *Department of Mathematics, University of Trento*
e-mail: claudio.agostinelli@unitn.it

[3] *Department of Mathematics, Statistics and Computer Science, Universidad de Cantabria*
e-mail: alicia.nieto@unican.es

Abstract: Local general depth (LGD) functions are used for describing the local geometric features and mode(s) in multivariate distributions. In this paper, we undertake a rigorous systematic study of LGD and establish several analytical and statistical properties. First, we show that, when the underlying probability distribution is absolutely continuous with density $f(\cdot)$, the scaled version of LGD (referred to as τ-approximation) converges, uniformly and in $L^d(\mathbb{R}^p)$ to $f(\cdot)$ when τ converges to zero. Second, we establish that, as the sample size diverges to infinity the centered and scaled sample LGD converge in distribution to a centered Gaussian process uniformly in the space of bounded functions on $\mathcal{H}_G$, a class of functions yielding LGD. Third, using the sample version of the τ-approximation ($S\tau A$) and the gradient system analysis, we develop a new clustering algorithm. The validity of this algorithm requires several results concerning the uniform finite difference approximation of the gradient system associated with $S\tau A$. For this reason, we establish *Bernstein*-type inequality for deviations between the centered and scaled sample LGD, which is also of independent interest. Finally, invoking the above results, we establish consistency of the clustering algorithm. Applications of the proposed methods to mode estimation and upper level set estimation are also provided. Finite sample performance of the methodology are evaluated using numerical experiments and data analysis.

Keywords and phrases: Local depth, extreme localization, Hoeffding's decomposition, sample local depth, uniform central limit theorem, clustering, modes, gradient system, Lyapunov's stability Theorem.

Received April 2022.

*A.N.-R.'s research is supported by Grant 21.VP67.64662 funded by "Proyectos Puente 2022" from the Spanish "Consejería de Universidades, Igualdad, Cultura y Deporte del Gobierno de Cantabria".

†Orcid 0000-0002-0268-3322

1. Introduction

Investigation of data depths is gaining momentum due to its applicability in a variety of machine learning problems such as non-parametric classification and clustering. This concept, formalized in Liu (1990) and Zuo and Serfling (2000a), serves to identify a center for multivariate distributions and a multidimensional center-outward order similar to that of a real line. The ordering enables a description of quantiles of multivariate distributions (see Zuo and Serfling (2000b)) and aids in using depth functions (DFs) for clustering. The current paper develops the intuitive notion that local depths possess properties that help in identifying peaks and valleys and hence clustering based on such identification can improve the quality and stability of the clustering algorithm.

The notion of local depth (Agostinelli and Romanazzi, 2011) provides a framework to describe the local multidimensional features of multivariate distributions. Section 2 of this paper provides a detailed study of local depth functions (LDFs) and their scaled versions, referred to as τ-approximation. Specifically, let $h_\tau^{(G)} : \mathbb{R}^p \times (\mathbb{R}^p)^{k_G} \to [0, \infty)$ be a bounded function satisfying the symmetry conditions

$$h_\tau^{(G)}(x + v; x_1 + v, \ldots, x_{k_G} + v) = h_\tau^{(G)}(x; x_1, \ldots, x_{k_G}), \quad v \in \mathbb{R}^p$$
$$\text{and} \quad h_\tau^{(G)}(-x; -x_1, \ldots, -x_{k_G}) = h_\tau^{(G)}(x; x_1, \ldots, x_{k_G}).$$

Then the local general depth (LGD) function is given by (see below for a precise definition)

$$LGD(x, \tau, P) = \int h_\tau^{(G)}(x; x_1, \ldots, x_{k_G}) dP(x_1) \ldots dP(x_{k_G})$$

and τ is referred to as the *localizing parameter*. This integral representation provides a unified treatment and analyses of several local depth functions available in the literature. We denote by $\mathcal{H}_G = \{h_\tau^{(G)}(x; \cdot) : x \in \mathbb{R}^p, \tau \in [0, \infty]\}$ the class of functions yielding LGD. Typically studied LDFs can be obtained by taking $h_\tau^{(G)}(\cdot; \cdot)$ to be indicators of appropriate Borel sets; that is,

$$h_\tau^{(G)}(x; \cdot) = \mathbf{I}(\cdot \in Z_\tau^G(x)),$$

where $Z_\tau^G(x)$ is referred to as the local set. The set associated with local lens depth (Liu and Modarres, 2011), denoted by LLD (G in LGD is replaced by L), is

$$Z_\tau^L(x) = \{(x_1, x_2) \in (\mathbb{R}^p)^2 : \max_{i=1,2} \|x - x_i\| \leq \|x_1 - x_2\| \leq \tau\},$$

while that for the spherical depth (Elmore, Hettmansperger and Xuan, 2006) is given by

$$Z_\tau^B(x) = \{(x_1, x_2) \in (\mathbb{R}^p)^2 : \|2x - (x_1 + x_2)\| \leq \|x_1 - x_2\| \leq \tau\}.$$

The local set for the β-skeleton depth (Yang and Modarres, 2018) is given by

$$Z_\tau^{K_\beta}(x) = \{(x_1, x_2) \in (\mathbb{R}^p)^2 : \max_{\substack{i,j=1,2 \\ i \neq j}} \|x_i + \left(\frac{2}{\beta} - 1\right)x_j - \frac{2}{\beta}x\| \leq \|x_1 - x_2\| \leq \tau\},$$

while that for the simplicial depth (Liu, 1990) is

$$Z_\tau^{S}(x) = \{(x_1, \ldots, x_{p+1}) \in (\mathbb{R}^p)^{(p+1)} : x \in \triangle[x_1, \ldots, x_{p+1}], \max_{i,j} \|x_i - x_j\| \leq \tau\},$$

where $\triangle[x_1, \ldots, x_{p+1}]$ is the closed simplex with vertices $x_1, \ldots, x_{p+1} \in \mathbb{R}^p$. While the definitions of LLD and LSD (local simplicial depth) were available in the literature, the definitions of LBD (local spherical (ball) depth) and $LK_\beta D$ (local β-skeleton depth), as defined here, seem new. Of course, β-skeleton depth reduces to spherical depth and lens depth for $\beta = 1$ and $\beta = 2$, respectively. Also, when $p = 1$ all of the above four local depths coincide. Finally, taking $\tau = \infty$ (that is, there is no localization) in the above, one obtains the general depth (GD) function

$$GD(x, P) = \int_{(\mathbb{R}^p)^k} h_\infty^{(G)}(x; x_1, \ldots, x_k) dP(x_1) \ldots dP(x_k)$$

studied in Zuo and Serfling (2000a) and referred to as *Type A* DFs. Accordingly, we refer to the class of LDFs above as *Type A* LDFs. When there is no scope for confusion, we suppress P in $GD(x, P)$ and $LGD(x, \tau, P)$ and use the notation $GD(x)$ and $LGD(x, \tau)$.

The LDFs scaled by τ^{-p}, as in Definition 2.2 below, are referred to as τ-approximations. When P is absolutely continuous with respect to (w.r.t.) the Lebesgue measure with density $f(\cdot)$, the τ-approximations converge as $\tau \to 0^+$ to a power of $f(\cdot)$. Under additional conditions, one can also prove the convergence of the derivatives of $f_\tau(\cdot)$ to the derivatives of $f(\cdot)$ which facilitates an inquiry into the modes of the density *via* a gradient system analysis. This, in turn, allows characterization of the related *stable manifolds* paving the way for cluster analysis. Related ideas about clustering appear in Chazal et al. (2013). Our methodology differs from the existing literature in that we take advantage of the τ-approximation $f_\tau(\cdot)$ and its properties, developed in Sections 2 and 3 below. For some discussion on the choice of τ see Remark 2.6 and Subsection 3.3.

Statistical inquiry about local depth requires an investigation into their sample versions, specifically of sample local depth and sample τ-approximation ($S\tau A$), $f_{\tau,n}(\cdot)$. Borrowing tools from empirical process theory, we establish that, when $\mathcal{H}_G$ is a VC-subgraph class, the sample local depth is uniformly consistent. We also obtain a related limit distribution in the class $\mathcal{H}_G$. Additionally, we develop a *Bernstein* type inequality for sample local depth. These results rely on the Hoeffding's decomposition of U-statistics representation of the local depth, which incidentally is a critical component of our analysis. A technical issue to the above development is that the space of bounded functions on $\mathcal{H}_G$ is not separable and it is here that we use the VC-subgraph property of the class $\mathcal{H}_G$. These results are described in Section 2.

We next focus on the application of the above methods to clustering. To this end, we recall from dynamical systems that the stable manifold generated by a mode m of a "smooth" density $f(\cdot)$ is given by

$$C(m) := \{x \in S_f : \lim_{t\to\infty} u_x(t) = m\},$$

where S_f is the interior of the support of $f(\cdot)$ and $u_x(t)$ is the solution at time t of the gradient system

$$u'(t) = \nabla f(u(t))$$

with initial value $u(0) = x$ and $\nabla f(\cdot)$ represents the gradient of $f(\cdot)$. If $m_1, \cdots m_M$ are the modes of $f(\cdot)$, then the clusters associated with $f(\cdot)$ are given by $C(m_1), \ldots, C(m_M)$ (Chacón, 2015). In this paper, we establish the convergence of the clusters derived using $f_\tau(\cdot)$, as $\tau \to 0^+$. This yields consistency at the population level. Next, using $S\tau A$, we also prove consistency of empirical clusters. For this, we require uniform convergence of the empirical finite difference approximations to the appropriate derivatives which we establish using the *Bernstein*-type inequality described previously. These results are in Section 3.

The consistency proof of the clustering method requires additional analyses via the use of discrete Grönwall lemma and subtle arguments involving the density of data points. The use of $S\tau A$s require a specification of τ. While in some cases, such as β-skeleton depths, one can choose τ to be an appropriate quantile, care is required for other DFs. An approach to choosing τ for clustering is via cross-validation as suggested by Wang (2010). We use this idea in Subsection 3.3. The convergence of the clustering algorithm requires a careful "real analysis" argument involving delicate probability bounds and path-tracing of the solution of the gradient system. This is described in Section 4. Numerical results and data analyses related to clustering algorithm are in Subsection 3.4 and 3.5. We end this introduction section with a comment about the notations used in the paper.

While the discussion on clustering focused on values of τ near 0, large and intermediate values of τ are also useful in applications as described in Chandler and Polonik (2021).

We assume that $(\mathbb{R}^p)^{k_G}$ is equipped with $\mathscr{B}((\mathbb{R}^p)^{k_G})$, where $\mathscr{B}(\mathscr{X})$ is the family of Borel subsets of a topological space $\mathscr{X}$. We denote by S^{p-1} the unit sphere in $\mathbb{R}^p$ and by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^p$. For a Borel measure μ on $\mathbb{R}^p$, $\mu^{\otimes k}$ is the k-fold product measure, $\lambda(\cdot)$ the Lebesgue measure on $\mathbb{R}^p$. We use *a.e.* to mean almost everywhere with respect to the Lebesgue measure on $\mathbb{R}^p$ and *a.s.* to mean almost surely with respect to a probability measure P on $\mathbb{R}^p$. The support of a function $g(\cdot)$ and its interior are denoted by $\overline{S}_g$ and S_g, respectively. Finally, we denote by $B_r(x)$ and $\overline{B}_r(x)$ the open and closed ball in $\mathbb{R}^p$ with radius $r \geq 0$ and center $x \in \mathbb{R}^p$.

2. Local depth and extreme localization

2.1. Analytic properties

We begin by describing in detail local notions of *Type A* depth functions studied in Zuo and Serfling (2000a). Let $\mathcal{G}$ denote the class of kernel functions $G(\cdot) : (\mathbb{R}^p)^{k_G} \to [0,\infty)$ satisfying the properties (**P1**)-(**P4**) below. When the kernel $G(\cdot)$ is an indicator of $Z_1^I(0)$, $I = L, B, K_\beta, S$ we obtain the classical depth functions. The sets $Z_\tau^I(x)$ are referred to as local sets. In the sequel when analyzing specific depth functions, we will interchangeably use the notations $G(\cdot)$ and $\mathbf{I}(\cdot \in Z_1^G(0))$. The *Type A* local depth function is defined as follows:

Definition 2.1 *Let $G \in \mathcal{G}$, $\tau \in [0,\infty]$, and let $h_\tau^{(G)} : \mathbb{R}^p \times (\mathbb{R}^p)^{k_G} \to [0,\infty)$ be given by*

$$h_\tau^{(G)}(x; x_1, \ldots, x_{k_G}) \coloneqq \begin{cases} G(\frac{x_1 - x}{\tau}, \ldots, \frac{x_{k_G} - x}{\tau}) & \text{if } \tau \in (0,\infty) \\ \lim_{\tau \to 0^+} G(\frac{x_1 - x}{\tau}, \ldots, \frac{x_{k_G} - x}{\tau}) & \text{if } \tau = 0 \\ \lim_{\tau \to \infty} G(\frac{x_1 - x}{\tau}, \ldots, \frac{x_{k_G} - x}{\tau}) & \text{if } \tau = \infty. \end{cases} \tag{2.1}$$

(i) The local general depth at localization level $\tau \in [0,\infty]$ of a point $x \in \mathbb{R}^p$ with respect to P is given by

$$LGD(x, \tau, P) \coloneqq \int h_\tau^{(G)}(x; x_1, \ldots, x_{k_G}) dP(x_1) \ldots dP(x_{k_G}). \tag{2.2}$$

(ii) The general depth of a point $x \in \mathbb{R}^p$ with respect to a probability measure P is given by

$$GD(x, P) \coloneqq LGD(x, \infty, P). \tag{2.3}$$

Properties of the Kernel $G(\cdot)$:

(**P1**) $G(\cdot)$ is a non-negative and Borel measurable function satisfying

$$\Lambda_1^{(G)} \coloneqq \int G(x_1, \ldots, x_{k_G}) dx_1 \ldots dx_{k_G} < \infty.$$

(**P2**) $G(\cdot)$ is symmetric and non-increasing along any ray from the origin in $(\mathbb{R}^p)^{k_G}$; that is, for any scalar $\alpha \geq 0$ and $v \in (\mathbb{R}^p)^{k_G}$, $G(v) = G(-v)$ and $G(\alpha v)$ is non-increasing in α.

(**P3**) $G(x_1, \ldots, x_{k_G}) \to 0$ as $\max_{i=1,\ldots,k_G} \|x_i\| \to \infty$.

(**P4**) For any $\epsilon > 0$, there exist $0 < \delta \leq \epsilon$ and $c_G > 0$ such that $\lambda((\overline{B}_\delta(0))^{k_G} \cap S_G) > 0$ and $G(\cdot) \geq c_G$ in $(\overline{B}_\delta(0))^{k_G} \cap S_G$.

In typical examples studied in the literature, such as simplicial, lens, and spherical depth, $G(\cdot)$ will have bounded support implying (**P3**); *i.e.*, for some $\rho > 0$,

$$\overline{S}_G \subset (\overline{B}_\rho(0))^k, \tag{2.4}$$

where we have suppressed G in k_G. Frequently, when there is no scope of confusion we will suppress the superscript or subscript G. Additionally we assume, without loss of generality (w.l.o.g.), that $\Lambda_1 = 1$ and functions in $\mathcal{G}$ are permutation invariant (see Appendix A (Francisci et al., 2023) for details). From the discussion in Appendix A it follows that if P is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$ with density $f(\cdot)$, then

$$LGD(x, \tau, P) = (h_\tau^{(G)}(0; \cdot) * f^{\otimes k}(\cdot))(x, \ldots, x), \quad x \in \mathbb{R}^p, \ \tau \in [0, \infty], \tag{2.5}$$

where $*$ is the convolution operator and $f^{\otimes k}(x_1, \ldots x_k) = f(x_1) \ldots f(x_k)$. Since P is fixed, in the following we write $GD(x)$ for $GD(x, P)$ and $LGD(x, \tau)$ for $LGD(x, \tau, P)$. Also, for $j = 1, \ldots, p$, we denote by $\partial_j g(\cdot)$ the partial derivative of the function $g : \mathbb{R}^p \to \mathbb{R}$ with respect to its j^{th} component. Our first proposition summarizes several continuity and differentiability properties of the LDFs.

Proposition 2.1 *(i) For all $x \in \mathbb{R}^p$, $LGD(x, \cdot)$ is monotonically non-decreasing with*

$$\lim_{\tau \to 0^+} LGD(x, \tau) = G(0, \ldots, 0) P^k(\{x\}) \text{ and } \lim_{\tau \to \infty} LGD(x, \tau) = GD(x).$$

(ii) For $\tau \in [0, \infty)$, $\lim_{r \to \infty} \sup_{x \in \mathbb{R}^p \setminus B_r(0)} LGD(x, \tau) = 0$.
(iii) If P is absolutely continuous with respect to the Lebesgue measure, then, for each $\tau \in [0, \infty)$, $LGD(\cdot, \tau)$ is bounded and continuous.
(iv) Under assumption (2.4), if P is absolutely continuous with respect to the Lebesgue measure, with m-times continuously differentiable density $f(\cdot)$, then, for each $\tau \in [0, \infty)$, $LGD(\cdot, \tau)$ is m-times continuously differentiable and, for $i_1, \ldots, i_m \in \{1, \ldots, p\}$,

$$\partial_{i_m} \ldots \partial_{i_1} LGD(x, \tau) = (h_\tau(0; \cdot) * (\partial_{i_m} \ldots \partial_{i_1} f^{\otimes k}(\cdot)))(x, \ldots, x). \tag{2.6}$$

When $\tau = \infty$, part (ii) does not hold in general. For instance, if P is absolutely continuous with respect to the Lebesgue measure with density function $f(\cdot)$, $k = 1$, and $G(\cdot) = \exp(-\|\cdot\|^2/2)$, then $h_\infty(\cdot; \cdot) \equiv 1$ and (ii) holds for $LGD(\cdot, \infty)$ if and only if it holds for $f(\cdot)$ (see also Zuo and Serfling (2000a)).

Our next result is concerned with the convergence of scaled versions of LDFs in spaces of integrable functions, under extreme localization. To this end, let $L^d((\mathbb{R}^p)^k) = L^d((\mathbb{R}^p)^k, \lambda^{\otimes k})$, $1 \le d < \infty$, denote the space of Lebesgue measurable functions $g : (\mathbb{R}^p)^k \to \mathbb{R}$ for which $g^d(\cdot)$ is absolutely integrable, and $L^\infty((\mathbb{R}^p)^k) = L^\infty((\mathbb{R}^p)^k, \lambda^{\otimes k})$ be the space of Lebesgue measurable functions $g : (\mathbb{R}^p)^k \to \mathbb{R}$ that are essentially bounded.

Theorem 2.1 *Let P be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$, with density $f(\cdot)$.*
(i) Under assumption (2.4) at every point of continuity of $f(\cdot)$, it holds that

$$\lim_{\tau \to 0^+} \tau^{-kp} LGD(\cdot, \tau) = f^k(\cdot). \tag{2.7}$$

Furthermore, (2.7) holds uniformly on any set where $f(\cdot)$ is uniformly continuous.

(ii) If $f(\cdot) \in L^\infty(\mathbb{R}^p)$, then (2.7) holds at every point of continuity of $f(\cdot)$ and the convergence in (2.7) is uniform on any set where $f(\cdot)$ is uniformly continuous.
(iii) Let $f(\cdot)$ be twice continuously differentiable. Then, under assumption (2.4), there exists a non-trivial function $R(\cdot)$ such that, for all $x \in S_f$,

$$\lim_{\tau \to 0^+} \tau^{-2} \left(\tau^{-kp} LGD(x, \tau) - f^k(x) \right) = R(x).$$

(iv) If $f^k(\cdot) \in L^d(\mathbb{R}^p)$, $1 \leq d < \infty$, then $\tau^{-kp} LGD(\cdot, \tau)$ converges in $L^d(\mathbb{R}^p)$ to $f^k(\cdot)$.

We observe that (iii) provides the rate of convergence of the local depth to the k^{th} power of the density under extreme localization. An explicit formula for $R(\cdot)$ is provided in Appendix A (Francisci et al., 2023). It is worth noticing that, under the assumption (2.4), for all $x \in \mathbb{R}^p \backslash \overline{S}_f$, $f^k(x) = 0$ and $\frac{1}{\tau^{kp}} LGD(x, \tau) = 0$ for small values of τ.

Using (2.7) one can express $f(\cdot)$ in terms of the limit of LDFs, for a given choice of $G(\cdot)$. This leads to our next definition, namely the τ-approximation.

Definition 2.2 (τ-approximation) *For any $\tau > 0$,*

$$f_\tau^{(G)}(x) \coloneqq \tau^{-p} (LGD(x, \tau))^{1/k}. \tag{2.8}$$

Remark 2.1 *From Proposition 2.1 (iii), it follows that when P has a density $f(\cdot)$ then, $f_\tau^{(G)}(\cdot)$ is continuous. Additionally, Proposition 2.1 (iv) implies that $f_\tau^{(G)}(\cdot)$ is m-times continuously differentiable in $S_{f_\tau^{(G)}}$.*

Remark 2.2 *Evidently, when $k = 1$ the τ-approximation reduces to the classical approximation by convolution in $\mathbb{R}^p$ with kernel scaled by τ. Using (A.5), we see that the same conclusion holds if $k > 1$ and $G(\cdot)$ is the product kernel $K^{\otimes k}(\cdot)$, where $K^{\otimes k}(x_1, \ldots x_k) = K(x_1) \ldots K(x_k)$. See also Remark 2.5 below.*

Our next proposition provides a uniform approximation of the density and its derivatives using the τ-approximation.

Proposition 2.2 *Let P be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$ with density $f(\cdot)$. Then the following hold:*
(i) If $f(\cdot)$ is uniformly continuous, then

$$\lim_{\tau \to 0^+} \sup_{x \in \mathbb{R}^p} |f_\tau^{(G)}(x) - f(x)| = 0. \tag{2.9}$$

(ii) If $f(\cdot)$ is continuous, then for all compact sets $K \subset \mathbb{R}^p$

$$\lim_{\tau \to 0^+} \sup_{x \in K} |f_\tau^{(G)}(x) - f(x)| = 0.$$

In particular, for all $x \in \mathbb{R}^p$, $\lim_{\tau,\ \epsilon \to 0^+} \sup_{y \in \overline{B}_\epsilon(x)} |f_\tau^{(G)}(y) - f(x)| = 0$.
(iii) If $f(\cdot) \in L^{k_G d}(\mathbb{R}^p)$, $d \geq 1$, then $f_\tau^{(G)}(\cdot)$ converges in $L^{k_G d}(\mathbb{R}^p)$ to $f(\cdot)$.

(iv) Suppose (2.4) holds and $f(\cdot)$ is m-times continuously differentiable, then, for all compact sets $K \subset S_f$ and $i_1, \ldots, i_m \in \{1, \ldots, p\}$,

$$\lim_{\tau \to 0^+} \sup_{x \in K} |\partial_{i_m} \ldots \partial_{i_1} f_\tau(x) - \partial_{i_m} \ldots \partial_{i_1} f(x)| = 0.$$

Remark 2.3 *The above proposition implies that the τ-approximation converges uniformly to the density under extreme localization. We also note that continuity is not enough in Proposition 2.2 (i). (iv) of the Proposition provides a uniform approximation to the partial derivatives of the τ-approximation and plays a central role in the properties of clustering investigated in the Section 3.*

2.2. Sample local depth

Let $\{X_1, \ldots, X_n\}$ be independent and identically distributed (i.i.d.) random variables from P on $\mathbb{R}^p$; then the estimate of LGD, called sample local depth, is the U-statistics of order k (Korolyuk and Borovskich, 2013)

$$LGD_n(x, \tau) := \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} h_\tau^{(G)}(x; X_{i_1}, \ldots, X_{i_k}), \tag{2.10}$$

where $x \in \mathbb{R}^p$ and $\tau \in [0, \infty]$. In particular, $GD(x)$ is estimated by setting $GD_n(x) := LGD_n(x, \infty)$. For $1 \le j \le k$, let

$$h_\tau^{(G,j)}(x; x_1, \ldots, x_j) := E[h_\tau^{(G)}(x; x_1, \ldots, x_j, X_{j+1}, \ldots, X_k)] \text{ and}$$
$$\tilde{h}_\tau^{(G,j)}(x; x_1, \ldots, x_j) := h_\tau^{(G,j)}(x; x_1, \ldots, x_j) - LGD(x, \tau).$$

When there is no scope for confusion we also write $h_\tau^{(j)}(\cdot;\cdot)$ for $h_\tau^{(G,j)}(\cdot;\cdot)$ and $\tilde{h}_\tau^{(j)}(\cdot;\cdot)$ for $\tilde{h}_\tau^{(G,j)}(\cdot;\cdot)$. Using (1.1.34) in Korolyuk and Borovskich (2013), we have that

$$Var[LGD_n(x, \tau)] = \binom{n}{k}^{-1} \sum_{j=1}^{k} \binom{k}{j} \binom{n-k}{k-j} E[(\tilde{h}_\tau^{(G,j)}(x; X_1, \ldots, X_j))^2]. \tag{2.11}$$

It follows that, for all $n \in \mathbb{N}$,

$$Var[\sqrt{n} LGD_n(x, \tau)] = nk \binom{n}{k}^{-1} \binom{n-k}{k-1} E[(\tilde{h}_\tau^{(G,1)}(x; X_1))^2] + O\left(\frac{1}{n}\right)$$
$$\xrightarrow[n \to \infty]{} k^2 E[(\tilde{h}_\tau^{(G,1)}(x; X_1))^2].$$

The above calculation yields that LGD_n is a consistent estimator of LGD even though this holds under much weaker conditions on $G(\cdot)$. In typical applications, the choice of x, τ, and G vary and in exploratory analyses, different choices of x, τ and G may be investigated. Our next result shows that the LGD_n is uniformly consistent over x and τ. The proof relies on the size of the class

$\mathcal{H}_G := \{h_\tau^{(G)}(x;\cdot) : x \in \mathbb{R}^p, \tau \in [0,\infty]\}$ which can be characterized using VC-theory. We impose a weak condition on the class $\mathcal{H}_G$, namely that it is a VC-subgraph class (see Definition 3.6.8 of Giné and Nickl (2016)). We show that this assumption holds in several examples studied in the literature. These details are described in Appendix C (Francisci et al., 2023).

Theorem 2.2 *Let $\mathcal{H}_G$ be a VC-subgraph class of functions. Then*

$$\sup_{\substack{x\in\mathbb{R}^p \\ \tau\in[0,\infty]}} |LGD_n(x,\tau) - LGD(x,\tau)| \xrightarrow[n\to\infty]{} 0 \ a.s.$$

In some examples, it is possible that $G =: G_\theta \in \mathcal{G}$ is indexed by a parameter $\theta \in \Theta \subset \mathbb{R}$, as is the case for β-skeletons. In such cases, one can strengthen the above Theorem 2.2 to obtain uniformity in the indexing parameter under additional assumptions as described in the Assumption A.1 in Appendix A (Francisci et al., 2023). That is,

$$\sup_{\theta\in\Theta} \sup_{\substack{x\in\mathbb{R}^p \\ \tau\in[0,\infty]}} |LG_\theta D_n(x,\tau) - LG_\theta D(x,\tau)| \xrightarrow[n\to\infty]{} 0 \text{ a.s.} \tag{2.12}$$

The details for the β-skeleton are also provided in Appendix C (Francisci et al., 2023). Computational issues are addressed in Appendix E.

Next, we turn to the uniform central limit theorem for LGD_n over a suitable subset T of $\mathbb{R}^p \times [0,\infty]$. Let $\ell^\infty(T)$ denote the space of all bounded functions $\bar{g}(\cdot) : T \to \mathbb{R}$. To study the convergence in distribution in $\ell^\infty(T)$, one needs to address the measurability problems that are encountered due to the non-separability of $\ell^\infty(T)$. We do this by establishing that the class $\mathcal{H}_G$ is image admissible Suslin as in Arcones and Giné (1993). For definition of image-admissible Suslin see Dudley (2014). In the following, convergence in distribution in $\ell^\infty(T)$ is in the sense of Hoffmann-Jørgensen (Giné and Nickl, 2016, Definition 3.7.22).

Theorem 2.3 *Let $T \subset \mathbb{R}^p \times [0,\infty]$ such that $E[(\tilde{h}_\tau^{(1)}(x;X_1))^2] > 0$, for all $(x,\tau) \in T$, and suppose that $\mathcal{H}_G$ is a VC-subgraph class of functions. Then*

$$\sqrt{n}(LGD_n(\cdot,\cdot) - LGD(\cdot,\cdot)) \xrightarrow[n\to\infty]{d} kW(\cdot,\cdot) \text{ in } \ell^\infty(T)$$

where $\{W(x,\tau)\}_{(x,\tau)\in T}$ is a centered Gaussian process with covariance function $\gamma : T \times T \to \mathbb{R}$ given by

$$\gamma((x,\tau),(y,\nu)) = \int h_\tau^{(1)}(x;x_1)h_\nu^{(1)}(y;x_1)dP(x_1) - LGD(x,\tau)LGD(y,\nu).$$

Remark 2.4 *Notice that, for $(x,\tau) \in T$, the variance of $W(x,\tau)$ is given by $\gamma((x,\tau),(x,\tau)) = E[(\tilde{h}_\tau^{(1)}(x;X_1))^2] > 0$ and, in the examples, $T \neq \emptyset$. This implies that the U-statistics (2.10) is non-degenerate, i.e. $\tilde{h}_\tau^{(1)}(x;\cdot) \neq 0$ (Korolyuk and Borovskich, 2013). Furthermore, if P is absolutely continuous with respect to the Lebesgue measure, $x \in S_P$ (the interior of the support of P), and $\tau > 0$, then since $E[(\tilde{h}_\tau^{(1)}(x;X_1))^2] > 0$, T can be taken to be "large".*

In the clustering applications discussed below, we will establish the consistency of the sample clustering algorithm. This will involve approximating the τ-approximations of the depth functions and their derivatives via their sample versions. The quality of this approximation will play a critical role in the consistency arguments. Our next result enables this study by establishing the following *Bernstein*-type inequality for local depth functions. Before we state this result, notice that, by Jensen's inequality and (A.1),

$$\sigma_G^2 \coloneqq \sup_{\substack{x\in\mathbb{R}^p \\ \tau\in[0,\infty]}} E[(\tilde{h}_\tau^{(G,1)}(x;X_1))^2] \leq \sup_{\substack{x\in\mathbb{R}^p \\ \tau\in[0,\infty]}} E[(\tilde{h}_\tau^{(G,k)}(x;X_1,\ldots,X_k))^2] \leq l_G^2,$$

where $l_G \coloneqq G(0,\ldots,0)$.

Theorem 2.4 *Let $\mathcal{H}_G$ be a VC-subgraph class of functions. Then, there are constants $1 < C_{G,0}, C_{G,1}, C_{G,2} < \infty$ such that, for all $t \geq \max(2^3\sigma_G, 2^4 C_{G,0})$,*

$$P^{\otimes n}(\sqrt{n} \sup_{\substack{x\in\mathbb{R}^p \\ \tau\in[0,\infty]}} |LGD_n(x,\tau) - LGD(x,\tau)| \geq t) \leq D_G(n,t) \coloneqq \sum_{j=1}^{3} D_{G,j}(n,t), \tag{2.13}$$

where

$$D_{G,1}(n,t) \coloneqq 8\exp\left(-\frac{t^2\sqrt{n}}{2^{15}k_G^2(\sigma_G^2\sqrt{n}+tl_G)}\right),$$

$$D_{G,2}(n,t) \coloneqq 8C_{G,1}^{2C_{G,2}}\left(\sigma_G^2 + \frac{2tl_G}{\sqrt{n}}\right)^{-C_{G,2}} \exp\left(-\left(\frac{n\sigma_G^2}{2l_G^2} + \frac{\sqrt{n}t}{4l_G}\right)\right), \quad \textit{and}$$

$$D_{G,3}(n,t) \coloneqq 2\exp\left(-\frac{t^2\sqrt{n}}{2^{6+k_G}k_G^{k_G+1}l_G C_{G,0}(\sigma_G^2\sqrt{n}+tl_G)}\right).$$

We now turn to the $S\tau A$ for estimating the density. To this end, let P be absolutely continuous with respect to the Lebesgue measure with density $f(\cdot)$. The plug-in estimator of $f_\tau^{(G)}(\cdot)$ is given by

$$f_{\tau,n}^{(G)}(x) \coloneqq \tau^{-p}(LGD_n(x,\tau))^{1/k}, \tag{2.14}$$

where we recall that we have suppressed G in k_G. Our first result uses Proposition 2.2 and Theorem 2.4 to establish the uniform convergence of $f_{\tau,n}^{(G)}(\cdot)$ to $f(\cdot)$.

Proposition 2.3 *Let $\mathcal{H}_G$ be a VC-subgraph class of functions and suppose that P is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$ with density $f(\cdot)$. Let $\{\tau_n\}_{n=1}^\infty$ and $\{\epsilon_n\}_{n=1}^\infty$ be sequences of positive scalars converging to zero with $\lim_{n\to\infty}\frac{n}{\log(n)}\tau_n^{2kp} = \infty$. Then the following hold:*
(i) If $f(\cdot)$ is uniformly continuous, then

$$\lim_{n\to\infty}\sup_{x\in\mathbb{R}^p}|f_{\tau_n,n}^{(G)}(x) - f(x)| = 0 \ \textit{a.s.}$$

(ii) If $f(\cdot)$ is continuous, then for all compact sets $K \subset \mathbb{R}^p$

$$\lim_{n\to\infty} \sup_{x\in K} |f^{(G)}_{\tau_n,n}(x) - f(x)| = 0 \ \ a.s.$$

In particular, for all $x \in \mathbb{R}^p$, $\lim_{n\to\infty} \sup_{y\in\overline{B}_{\epsilon_n}(x)} |f^{(G)}_{\tau_n,n}(y) - f(x)| = 0$ a.s.

Remark 2.5 *Returning to Remark 2.2, under the additional assumption that $G(\cdot) = K^{\otimes k}(\cdot)$ is a product of kernels, one can obtain our estimator as a U-statistic with a product kernel with the same bandwidth, namely*

$$\binom{n}{k}^{-1} \frac{1}{\tau^{kp}} \sum_{1\le i_1<\cdots<i_k\le n} \prod_{l=1}^{k} K\left(\frac{x - X_{i_l}}{\tau}\right),$$

which is the U-statistic estimator of $f^k(x)$ using the KDE

$$f^{(K)}_{\tau_n,n}(x) = \frac{1}{n\tau^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\tau}\right).$$

Thus, LDFs are a natural extension of product kernels, where the function $K^{\otimes k}(\cdot)$ is replaced by $G(\cdot)$.

Remark 2.6 *Under the additional assumption that $\psi^{(G)}(\cdot)$ is integrable in $(\mathbb{R}^p)^k$, where $\psi^{(G)}(w) := \sup_{v\in(\mathbb{R}^p)^k:\|v-w\|\le 1} G(v)$ (norm in $(\mathbb{R}^p)^k$), it follows from Bertrand-Retali (1978) that consistency can be proved under the weaker condition $\lim_{n\to\infty} \frac{n}{\log(n)} \tau_n^{kp} = \infty$. Einmahl and Mason (2005) study uniformity in x and τ when $k = 1$.*

The asymptotic limit distribution of the $S\tau A$ is provided in Appendix B (Francisci et al., 2023). Examples and verification of the VC-subgraph property are provided in Appendix C. We now turn to discuss clustering application. Appendix D.1 contains applications to estimation of upper level sets of the density.

3. Clustering

In this section, we describe a methodology for clustering multivariate data using the theory of dynamical systems, which involves three distinct but connected steps. In the first step, one constructs cluster(s) in the population as stable manifold(s) generated by the mode(s). Next, the behavior of the gradient system when $f(\cdot)$ is replaced by its τ-approximation is studied and its convergence established under extreme localization. Finally, one replaces the τ-approximated density by its $S\tau A$, $f_{\tau_n,n}(\cdot)$, to obtain the empirical clusters and establish their convergence.

The following discussion is reliant on Assumption 3.1 below concerning the smoothness properties of $f(\cdot)$. Recall that the clusters are defined as the stable manifolds generated by the mode and are obtained using the limiting trajectory of the gradient system. Specifically, for any $\mu \in S_f$, the stable manifold

generated by μ is given by

$$C(\mu) \coloneqq \{x \in S_f : \lim_{t\to\infty} u_x(t) = \mu\}, \tag{3.1}$$

where $u_x(t)$ is the solution at time t of the gradient system

$$u'(t) = \nabla f(u(t)) \tag{3.2}$$

with initial value $u(0) = x$. For any choice of μ, it is not required for the stable manifold so-defined to be non-trivial; i.e. the Lebesgue measure of $C(\mu)$ can be zero. However, if μ is chosen as a mode of $f(\cdot)$, then, one can verify that the resulting manifold has a positive Lebesgue measure. We next turn to define the stationary points type, and, in particular, the mode. Before we state the assumption, we introduce one more notation: the Hessian matrix associated with any function $g(\cdot)$ is denoted by H_g and $\langle\cdot,\cdot\rangle$ denotes the inner product on $\mathbb{R}^p$.

Definition 3.1 *A stationary point $\mu \in S_f$ of $f(\cdot)$ is said to be of type l, $0 \le l \le p$, if $H_f(\mu)$ has l negative and $p-l$ positive eigenvalues. In particular, $m \in S_f$ is said to be a mode (resp. an antimode) for $f(\cdot)$ if it is a stationary point of $f(\cdot)$ and $H_f(m)$ has only negative (resp. positive) eigenvalues, that is, m is a local maximum (resp. minimum) for $f(\cdot)$. If $m_1,\dots,m_M$ are the modes of $f(\cdot)$, then the clusters induced by $m_1,\dots,m_M$ are the stable manifolds $C(m_1),\dots,C(m_M)$.*

Let $m_1,\dots,m_M$ be the modes and $\mu_1,\dots,\mu_L$ the other stationary points of $f(\cdot)$. We deduce from dynamical systems and Morse theory literature (Hirsch, Devaney and Smale, 1974; Matsumoto, 2002; Teschl, 2012) and Chacón (2015) that the clusters $C(m_1),\dots,C(m_M)$ are well-defined, non-trivial, disjoint, and

$$S_f = \cup_{i=1}^{M} C(m_i) \cup \cup_{l=1}^{L} C(\mu_l) \tag{3.3}$$

Hence, $C(m_1),\dots,C(m_M),C(\mu_1),\dots,C(\mu_L)$ form a partition of S_f. Additionally, the clusters $C(m_1),\dots,C(m_M)$ are open sets and are separated in S_f by the lower dimensional stable manifolds $C(\mu_1),\dots,C(\mu_L)$. This completes the first step. The second step is described in Subsection 3.1 where we describe step-by-step analytical tools to fill in the gap between local depths and stable manifolds generated by the modes. The third step is described in Subsection 3.2. The algorithm is provided in Appendix E (Francisci et al., 2023).

3.1. Identification of stationary points and convergence of the gradient system under extreme localization

We replace $f(\cdot)$ by $f_\tau(\cdot)$ in (3.2) and consider the gradient system

$$u'(t) = \nabla f_\tau(u(t)). \tag{3.4}$$

The domain of this new system is S_{f_τ}. We summarize the main properties of (3.4) as $\tau \to 0^+$. We begin with the properties of S_{f_τ}.

Lemma 3.1 *For all $0 < \tau_1 \leq \tau_2$, we have that $S_{f_{\tau_1}} \subset S_{f_{\tau_2}}$. Additionally, if $f(\cdot)$ is continuous, then, for all $\tau > 0$, $S_f \subset S_{f_\tau}$ and $\lim_{\tau \to 0^+} S_{f_\tau} \supset S_f$. Under assumption (2.4), $\lim_{\tau \to 0^+} S_{f_\tau} \subset \overline{S}_f$.*

We observe that the assumption (2.4) is essential in the last part of Lemma 3.1. Indeed, if $G(\cdot)$ is the Gaussian kernel, then $S_{f_\tau} = \mathbb{R}^p$, for all $\tau > 0$, implying $\lim_{\tau \to 0^+} S_{f_\tau} = \mathbb{R}^p$. Also, since ∂S_f and S_G have arbitrary shape, it is unclear if $x \in \partial S_f$ belongs to $\lim_{\tau \to 0^+} S_{f_\tau}$ or not. Under Assumption 3.1 below, Proposition 2.2 (iv) shows that the gradient and the Hessian matrix of $f_\tau(\cdot)$ converge to those of $f(\cdot)$. Recall that, by Remark (2.1), if $f(\cdot)$ is m-times continuously differentiable, then, $f_\tau(\cdot)$ is m-times continuously differentiable in S_{f_τ}. Additionally, if $f(\cdot)$ is τ-symmetric about a stationary point μ (that is, $f(\mu+x) = f(\mu-x)$, for all $x \in \mathbb{R}^p$ with $\|x\| \leq \tau$), then it is easy to see that the stationary points of $f(\cdot)$ are also the stationary points of $f_\tau(\cdot)$. However, the assumption of τ-symmetry may be harder to verify in applications. For this reason, we *do not make this assumption in the developments below* even though in Appendix D.2 (Francisci et al., 2023) we provide sufficient conditions under which the stationary points (resp. modes, antimodes) of $f(\cdot)$ are *exactly* the stationary points (resp. modes, antimodes) of $f_\tau(\cdot)$ for $\tau > 0$ when τ-symmetry obtains.

Next, to characterize the stationary points of $f_\tau(\cdot)$ without the τ-symmetry condition, notice that for small τ, the first and second order derivatives of $f_\tau(\cdot)$ are close to those of $f(\cdot)$ (Proposition 2.2). Hence, one can pick a hypercube, centered at the stationary point with directions provided by eigenvectors of Hessian matrix, so that $f(\cdot)$ and $f_\tau(\cdot)$ share similar properties within the hypercube. This idea is made precise in the following theorem.

Theorem 3.1 *Suppose (2.4) holds true. The following hold:*
(i) If $f(\cdot)$ is continuously differentiable in $\overline{B}_{\rho\tau}(\mu) \subset S_f$, $\tau > 0$, then $\nabla f_\tau(\mu) = 0$ if and only if

$$\int h_\tau(0; x_1, \ldots, x_k) \nabla f(\mu + x_1) f(\mu + x_2) \ldots f(\mu + x_k) dx_1 \ldots dx_k = 0, \quad (3.5)$$

where the integral of a vector is the vector of the integrals.
(ii) If $f(\cdot)$ is twice continuously differentiable in $\overline{B}_\delta(\mu) \subset S_f$, $\delta > 0$, and μ is a stationary point of $f(\cdot)$ of type l, then there exists $h^*, \tau^* > 0$ and a closed hypercube $F_{h^*}(\mu) \subset \overline{B}_\delta(\mu)$ with side length $3/2h^*$ such that, for $0 < \tau \leq \tau^*$, $f_\tau(\cdot)$ has a unique stationary point μ_τ in $\mathring{F}_{h^*}(\mu)$ and μ_τ is of type l. Moreover, $\|\mu_\tau - \mu\| \xrightarrow[\tau \to 0^+]{} 0$.
(iii) If $f(\cdot)$ is three times continuously differentiable, then $\|\mu_\tau - \mu\| = O(\tau^2)$.

We now state the main assumptions required for convergence of clusters obtained using (2.8).

Assumption 3.1 *$f(\cdot)$ is a probability density function on $\mathbb{R}^p$ that is twice continuously differentiable with a finite number of stationary points in S_f. Additionally, the Hessian matrix H_f has non-zero eigenvalues at its stationary points. Also, let $R^\alpha := \{x \in \mathbb{R}^p : f(x) \geq \alpha\}$ be a bounded set for every $\alpha > 0$.*

By continuity of $f(\cdot)$, R^α is compact. We notice that R^α is bounded if $f(\cdot)$ vanishes at infinity, that is, $\sup_{x\in\mathbb{R}^p:\|x\|\geq c} f(x) \to 0$ as $c \to \infty$, which is satisfied, for example, if S_f is bounded. We study next the relationship between the gradient systems (3.4) and (3.2) under extreme localization. To this aim, notice that the sets $\{S_{f_\tau}\}_{\tau>0}$ contain S_f by Lemma 3.1. If it exists, we denote by $u_{x,\tau}(t)$ the solution of (3.4) with initial point $u_{x,\tau}(0) = x$. Since $f_\tau(\cdot)$ is continuous, for $\alpha > 0$, the sets $R^\alpha_\tau := \{x \in \mathbb{R}^p : f_\tau(x) \geq \alpha\} = f_\tau^{-1}([\alpha, \infty))$ are closed. Lemma A.6 in Appendix A (Francisci et al., 2023) shows that they are also bounded. Lemma A along with the boundedness of R^α for all $\alpha > 0$, implies that for all $x \in S_f$ $u_{x,\tau}(\cdot)$ exists and is unique in a maximal time interval (a, ∞), for some $-\infty \leq a < 0$. For a stationary point $\mu_\tau \in S_f$ of $f_\tau(\cdot)$, the stable manifold generated by μ_τ is

$$C_\tau(\mu_\tau) := \{x \in S_f : \lim_{t\to\infty} u_{x,\tau}(t) = \mu_\tau\}.$$

We next exploit the differentiability properties of $f_\tau(\cdot)$ to show that the solutions of the gradient system (3.4) converge for $\tau \to 0^+$ to those of the gradient system (3.2). This is described in Appendix A, Proposition A.2. We now turn to the convergence of the clusters $C_\tau(\mu_\tau)$ under extreme localization. To this end, let $N_f := \{m_1, \ldots, m_M, \mu_1, \ldots, \mu_L\}$ denote the set of stationary points of $f(\cdot)$.

Theorem 3.2 *Suppose that (2.4) and Assumption 3.1 hold true, and $f(\cdot)$ is three times continuously differentiable. Let $\{\tau_j\}_{j=1}^\infty$ be a sequence of positive scalars converging to 0. Then, for all $\mu \in N_f$, there exists $\tau^* > 0$ and $\{\mu_{\tau_j}\}_{j=1,\tau_j\leq\tau^*}^\infty$ such that $\|\mu_{\tau_j} - \mu\| = O(\tau_j^2)$, where, for each τ_j, μ_{τ_j} is a stationary points of $f_{\tau_j}(\cdot)$ and is of the same type as μ satisfying $\lim_{j\to\infty} C_{\tau_j}(\mu_{\tau_j}) = C(\mu)$.*

3.2. Algorithm and consistency of empirical clusters

In this section, we describe the algorithm for the numerical approximation of the clusters induced by the system (3.4) and establish its consistency.

Since the sample τ-approximation is, in general, not differentiable in x, we use a finite difference approximation that converges to the directional derivative. The directional derivative of $g(\cdot)$, in the direction of $v \in S^{p-1}$ (the unit sphere in $\mathbb{R}^p$), is denoted by $\nabla_v g(\cdot) = \langle \nabla g(\cdot), v\rangle$. To this end, for $x \in \mathbb{R}^p$, $\tau > 0$, $n \in \mathbb{N}$, $h > 0$ and a unit vector $v \in \mathbb{R}^p$, the finite difference approximations of the directional derivatives of $f_\tau(\cdot)$ and $f_{\tau,n}(\cdot)$ along v are given by

$$\nabla_v^h f_\tau(x) = \frac{f_\tau(x + hv) - f_\tau(x)}{h} \quad \text{and} \quad \nabla_v^h f_{\tau,n}(x) = \frac{f_{\tau,n}(x + hv) - f_{\tau,n}(x)}{h}.$$

Our first result shows that under the condition $\lim_{n\to\infty} nh_n^{2k}\tau_n^{2kp} = \infty$, the finite difference approximation to the directional derivative converges uniformly on compact sets, in probability.

Theorem 3.3 *Suppose* (2.4) *holds true. Let* K *be a compact subset of* S_f, $\{h_n\}_{n=1}^{\infty}$ *and* $\{\tau_n\}_{n=1}^{\infty}$ *sequences of positive scalars converging to* 0 *and* $\{v_n\}_{n=1}^{\infty}$ *be a sequence in* S^{p-1} *converging to* $v \in S^{p-1}$. *(i) If* $f(\cdot)$ *is continuously differentiable, then*

$$\lim_{n\to\infty} \sup_{x\in K} |\nabla_{v_n}^{h_n} f_{\tau_n}(x) - \nabla_v f(x)| = 0.$$

(ii) If, additionally, $\mathcal{H}_G$ *is a VC-subgraph class of functions and* $\lim_{n\to\infty} nh_n^{2k}\tau_n^{2kp} = \infty$, *then, for all* $\epsilon > 0$,

$$\lim_{n\to\infty} P^{\otimes n}\left(\sup_{x\in K} |\nabla_{v_n}^{h_n} f_{\tau_n,n}(x) - \nabla_v f(x)| \geq \epsilon\right) = 0.$$

The first step towards identifying the modes, is finding a local maximum of a function. To this end, we use the steepest ascent or gradient ascent idea; that is, starting from a point in the space, the next point is chosen in the direction given by the gradient of the function at that point. This procedure is repeated until convergence to a local maximum is achieved. When clustering using modes, this procedure is often combined with kernel density estimators to find the modes of the density underlying the given data points, and the clusters associated with them (Fukunaga and Hostetler, 1975; Menardi, 2016). Our methodology does not require existence of gradients, and considers data as potential candidate points for the next move. Similar ideas were also used in Koontz, Narendra and Fukunaga (1976).

Turning to the consistency result, we need arguments that allows one to approximate uniformly the directional derivative of points over (i) a compact set, (ii) the step-size, and (iii) directions. The next lemma addresses this issue and critically uses the *Bernstein*-type inequality developed in Theorem 2.4. Part (iii) of the lemma below also provides a upper bound on the uniform approximation mentioned above. We need the following notation: for $\delta > 0$, $(A)^{+\delta} := \{x \in \mathbb{R}^p : \inf_{y\in A}\|x - y\| \leq \delta\}$ and $(A)^{-\delta} := \mathbb{R}^p \setminus (\mathbb{R}^p \setminus A)^{+\delta} = \{x \in \mathbb{R}^p : \inf_{y\in\mathbb{R}^p\setminus A}\|x - y\| > \delta\}$.

Lemma 3.2 *Suppose* (2.4) *holds true. Let* K *be a compact subset of* S_f *and let* $h^* > 0$ *be such that* $(K)^{+h^*} \subset S_f$. *Also, let* $\{\tau_n\}_{n=1}^{\infty}$ *and* $\{h_n\}_{n=1}^{\infty}$ *be sequences of positive scalars converging to* 0. *Assume also that* $f(\cdot)$ *is three times continuously differentiable. Then*
(i) the finite difference approximation of the directional derivative of $f_\tau(\cdot)$ *converges uniformly to that of* $f(\cdot)$. *That is,*

$$\lim_{n\to\infty} \sup_{h\in[h_n,h^*]} \sup_{v\in S^{p-1}} \sup_{x\in K} |\nabla_v^h f_{\tau_n}(x) - \nabla_v^h f(x)| = 0.$$

(ii) If, additionally, $\mathcal{H}_G$ *is a VC-subgraph class of functions and* $\lim_{n\to\infty} nh_n^{2k}\tau_n^{2kp} = \infty$, *then, for all* $\epsilon > 0$,

$$\lim_{n\to\infty} P^{\otimes n}\left(\sup_{h\in[h_n,h^*]} \sup_{v\in S^{p-1}} \sup_{x\in K} |\nabla_v^h f_{\tau_n,n}(x) - \nabla_v^h f(x)| \geq \epsilon\right) = 0.$$

(iii) Let $\lim_{n\to\infty} \frac{n}{\log(n)} h_n^{2k}\tau_n^{2kp} = \infty$ *and* $\mathcal{H}_G$ *be a VC-subgraph class of functions. Then, for all* $\epsilon > 0$*, there are constants* $0 < \tilde{C} < \infty$ *and* $\tilde{n} \in \mathbb{N}$ *such that, for all* $n \geq \tilde{n}$*,*

$$P^{\otimes n}\left(\sup_{h\in[h_n,h^*]}\sup_{v\in S^{p-1}}\sup_{x\in K}|\nabla_v^h f_{\tau_n,n}(x) - \nabla_v^h f(x)| \geq \epsilon\right) \leq \frac{\tilde{C}}{n^2}.$$

We now describe the clustering algorithm. Let $\mathcal{X}_n := \{X_1, \ldots, X_n\}$ be a sample of i.i.d. random variables from P with density $f(\cdot)$ and $\{h_n\}_{n=1}^{\infty}$ and $\{\tau_n\}_{n=1}^{\infty}$ be sequences of positive scalars converging to zero. For $x \in \mathbb{R}^p$ and $r > 0$, define

$$\mathcal{X}_{n,r}(x) := \{X \in \mathcal{X}_n : h_n \leq \|X - x\| \leq r\},$$

$Y_{n,r,0} := x$ and, recursively, if

$$\max_{X\in\mathcal{X}_{n,r}(Y_{n,r,j})\cup\{Y_{n,r,j}\}} f_{\tau_n,n}(X) - f_{\tau_n,n}(Y_{n,r,j}) > 0, \tag{3.6}$$

then

$$Y_{n,r,j+1} := \operatorname*{argmax}_{X\in\mathcal{X}_{n,r}(Y_{n,r,j})} \frac{f_{\tau_n,n}(X) - f_{\tau_n,n}(Y_{n,r,j})}{\|X - Y_{n,r,j}\|}; \tag{3.7}$$

else stop and let $j^* := j$ and $L_{n,r}(x) := Y_{n,r,j^*}$. It is clear from the above description that $j^* \leq n$, that is, the algorithm ends in at most n steps. Indeed, for all $j = 1, \ldots, j^*$, $Y_{n,r,j} \in \mathcal{X}_n$ and, by (3.6), $Y_{n,r,j} \neq Y_{n,r,l}$ for all $l < j$. The next theorem shows that, for small r, large n, and $x \in C(m_i)$, $L_{n,r}(x)$ is close to m_i with arbitrary large probability.

Theorem 3.4 *Suppose that* $\mathcal{H}_G$ *is a VC-subgraph class of functions, Assumption 3.1 and (2.4) hold true and* $f(\cdot)$ *is three times continuously differentiable. Let* $\{h_n\}_{n=1}^{\infty}$ *and* $\{\tau_n\}_{n=1}^{\infty}$ *be sequences of positive scalars converging to zero with* $\lim_{n\to\infty} nh_n^{2k}\tau_n^{2kp} = \infty$*,* $0 < \eta \leq 1$*,* $0 < \overline{\alpha} < \min_{i=1,\ldots,M} f(m_i)$*,* $\epsilon > 0$*,* $\xi > 0$*, and* $0 < r \leq r^*$ *for some* r^**. Then, there exists* $n^* \in \mathbb{N}$ *such that, with probability at least* $1 - \eta$*,* $L_{n,r}(x) \in B_\epsilon(m_i)$ *for all* $n \geq n^*$ *and* $x \in R^{\overline{\alpha}} \cap (C(m_i)^{-\xi})$*.*

Using the above theorem, one can estimate the mode using the last iterate, namely, $L_{n,r}(x) = Y_{n,r,j^*}$. The Corollary 3.1 below provides strong consistency of this estimate. Turning to the proof of Theorem 3.4, it is divided into four distinct but connected steps. For the first step, let j^* be a non-negative integer and define $\{y_{r,j}\}$ recursively as follows: let $y_{r,0} = x$ and

$$y_{r,j+1} = y_{r,j} + h_j v_j, \quad 0 \leq j \leq (j^* - 1),$$

where $0 < h_j \leq r$ for some small $r > 0$, and where v_j is "close" to the normalized gradient of $f(\cdot)$ at $y_{r,j}$. We show that the sequence $\{y_{r,j}\}$ is close to the solution $u_x(\cdot)$ of (3.2). This is achieved, using version of the discrete Grönwall lemma (Lemma A.7 in Appendix A (Francisci et al., 2023)). Next, we show that $\{Y_{n,r,j}\}$ in (3.7) behaves like the sequence $\{y_{r,j}\}$ described in Step 1, with probability $(1 - \eta)$. This is achieved in Step 2 using Lemma 3.2. The proof of this step requires the existence of sufficient number of data points in a small neighborhood

of all points in the direction of the normalized gradient. We establish that this is indeed the case using compactness arguments in Step 3. Finally, we apply the results of Step 1 to $\{Y_{n,r,j}\}_{j=0}^{j^*}$ yielding that this sequence is close to the solution $u_x(\cdot)$. Since for all points that are not close to a mode, there exists, by Step 3, data points yielding a positive finite difference approximation of the directional derivative, (3.6) occurs with the desired probability. This observation allows to conclude, in Step 4, that Y_{n,r,j^*} is close to the mode.

We now give a formal definition of empirical clusters. To this end, we add an additional step to the above algorithm in which we merge the last iterates $L_{n,r}(x)$, $x \in \mathbb{R}^p$, that are close to each other. To this end, let

$$\mathcal{L}_{n,r} \coloneqq \{L_{n,r,1}, L_{n,r,2}, \ldots, L_{n,r,N_n}\} = \{L_{n,r}(x) : x \in \mathbb{R}^p\}$$

be the set of all last iterates. For $\delta > 0$ and $L_0 \in \mathcal{L}_{n,r}$ let

$$[L_0]_\delta \coloneqq \{L \in \mathcal{L}_{n,r} : \exists L_1, \ldots, L_l \in \mathcal{L}_{n,r} : \|L_0 - L_1\|, \ldots, \|L_l - L\| \leq \delta\}$$

and fix $m_{n,r,1}, m_{n,r,2}, \ldots, m_{n,r,M_n} \in \mathcal{L}_{n,r}$ such that $\mathcal{L}_{n,r} = \cup_{i=1}^{M_n} [m_{n,r,i}]_\delta$ and $[m_{n,r,i}]_\delta \cap [m_{n,r,j}]_\delta = \emptyset$ for $i \neq j$. Empirical clusters are given by

$$C_{n,r,\delta,i} \coloneqq \cup_{L \in [m_{n,r,i}]_\delta} \{x \in \mathbb{R}^p : L_{n,r}(x) = L\}.$$

We use probability distance and Hausdorff distance to evaluate the distance between two clusterings $\mathcal{C} = \{C_1, \ldots, C_s\}$ and $\mathcal{D} = \{D_1, \ldots, D_t\}$ (see Chacón (2015), for instance). Suppose w.l.o.g. that $s \leq t$ and recall that the symmetric difference between two subsets A and B of $\mathbb{R}^p$ is $A \Delta B = ((\mathbb{R}^p \setminus A) \cap B) \cup (A \cap (\mathbb{R}^p \setminus B))$. The probability distance between $\mathcal{C}$ and $\mathcal{D}$ is given by

$$d_{P,c}(\mathcal{C}, \mathcal{D}) = \frac{1}{2} \min_{\pi \in \mathcal{P}_t} \left(\sum_{i=1}^{s} P(C_i \Delta D_{\pi(i)}) + c \sum_{i=s+1}^{t} P(D_{\pi(i)}) \right),$$

where $\mathcal{P}_t$ is the set of all permutations of $\{1, \ldots, t\}$ and $c \geq 0$ is a penalization coefficient for clusters that do not match with any other. The Hausdorff distance is given by

$$d_H(\mathcal{C}, \mathcal{D}) = \max\Big(\max_{i=1,\ldots,s} \min_{j=1,\ldots,t} P(C_i \Delta D_j), \max_{j=1,\ldots,t} \min_{i=1,\ldots,s} P(C_i \Delta D_j) \Big).$$

We denote by

$$\mathcal{C} \coloneqq \{C(m_1), C(m_2), \ldots, C(m_M)\}$$

the population clustering and by

$$\mathcal{C}_{n,r,\delta} \coloneqq \{C_{n,r,\delta,1}, C_{n,r,\delta,2}, \ldots, C_{n,r,\delta,M_n}\}$$

the empirical clustering. We are now ready to prove consistency of empirical clusters.

Proposition 3.1 *Assume the conditions in Theorem 3.4 hold and* $\lambda(C(\mu_l)) = 0$ *for all* $l = 1, \ldots, L$. *If* $0 < \delta \leq \delta^*$ *and* $0 < r \leq r^*$, *for some* δ^* *and* r^* *that*

depend on $f(\cdot)$ only, then

$$\lim_{n\to\infty} d_{P,c}(\mathcal{C}, \mathcal{C}_{n,r,\delta}) = 0 \ a.s.$$

Next, let $\mathcal{C}_{n,r,\delta,\zeta}$ be the clustering obtained from $\mathcal{C}_{n,r,\delta}$ by removing all clusters with empirical probability not larger than $\zeta > 0$.

Proposition 3.2 *Assume the conditions in Theorem 3.4 hold and $\lambda(C(\mu_l)) = 0$ for all $l = 1, \ldots, L$. If $0 < \delta \leq \delta^*$, $0 < \zeta < \zeta^*$, and $0 < r \leq r^*$, for some δ^*, ζ^*, and r^* that depend on $f(\cdot)$ only, then*

$$\lim_{n\to\infty} d_H(\mathcal{C}, \mathcal{C}_{n,r,\delta,\zeta}) = 0 \ a.s.$$

As an additional consequence of the Theorem 3.4, setting $J_n \coloneqq \mathbf{I}(L_{n,r}(x) \notin B_\epsilon(m_i))$ and $\{\eta_n\}_{n=1}^\infty$ be a sequence of scalars in $(0, 1]$ with $\lim_{n\to\infty} \eta_n = 0$ one can show by Theorem 3.4 that $\lim_{n\to\infty} P^{\otimes n}(J_n = 1) \leq \lim_{n\to\infty} \eta_n = 0$, implying that J_n converges in probability to zero. Since $L_{n,r}(x)$ is the estimate of the mode, we obtain weak consistency of the mode. Furthermore, using (iii) of Lemma 3.2, one can strengthen the conclusion to almost sure convergence. We summarize this observation as a corollary.

Corollary 3.1 *Suppose that $\lim_{n\to\infty} \frac{n}{\log(n)} h_n^{2k} \tau_n^{2kp} = \infty$ and the assumptions of Theorem 3.4 hold. Then $J_n \xrightarrow[n\to\infty]{} 0$ a.s.*

It is important to note that one can weaken some of the conditions in Theorem 3.4. Specifically, in Lemma D.1 in Appendix D.3 (Francisci et al., 2023) we show that, for $p \geq 6k + 1$, the conditions involving $\{h_n\}_{n=1}^\infty$ can be removed provided that the sequence $\{\tau_n\}_{n=1}^\infty$ does not converge to zero "too fast"; for instance, one could choose $\tau_n = n^{-\delta/(2kp)}$) for some $0 < \delta < 1 - \frac{6k}{p}$. To see this, notice from the lemma that h_n can be replaced by $\tilde{h}_n \coloneqq \min_{y,z\in\mathcal{X}_n\cup\{x\}, y\neq z} \|y - z\|$, which implies that $\mathcal{X}_{n,r}(x) = \{X \in \mathcal{X}_n : h_n \leq \|X - x\| \leq r\}$ can be replaced by $\tilde{\mathcal{X}}_{n,r}(x) = \{X \in \mathcal{X}_n : \|X - x\| \leq r, \ X \neq x\}$.

3.3. Choice of τ

A key issue in the use of LDFs for clustering is that it requires a value of τ. Theorem 3.4 suggests that τ should decrease slowly with n, namely $\tau = \tau_n = o(n^{-1/(2kp)})$. However, it does not provide an optimal value of τ when n is small. In this subsection, we develop an alternative data-driven procedure for choosing τ for a fixed sample size n. Wang (2010) proposes to choose among different clustering algorithms the algorithm that maximizes clustering stability in the sense that clusters vary as little as possible when applying the algorithm to different samples. We use adjusted Rand index as measure for clustering stability (Rand, 1971; Hubert and Arabie, 1985). If only one sample is available clustering stability can be evaluated using cross-validation. For this, the dataset is divided into three parts: the first and second part are used to build two different clusterings of the sample and the third part is used for evaluating stability based on the two

clustering. We repeat this procedure multiple times and compute the adjusted Rand index for a given value of τ. In our numerical experiments we draw 100 different samples with sample size $n = 1000$ and choose $C = 100$ subsamples.

As for β-skeleton and simplicial depths the parameter τ can be chosen as a quantile of the distances between the observations. For more details we refer to Appendix E (Francisci et al., 2023).

3.4. Numerical results

In this subsection, we describe several numerical experiments to evaluate the performance of the clustering algorithm. We use as metric empirical probability distance with $c = 1$ and empirical Hausdorff distance. We consider the following distributions studied in the literature (Wand and Jones, 1993; Chacón, 2015) in two dimensions: (H) Bimodal IV and #10 Fountain. We also study the behavior in dimension five (Mult. Quadrimodal) and for circular distributions (Circular Bimodal II), where additional complexities arise for identifying the true clusters.

We next turn to the choice of τ for lens depth. As explained previously, we choose τ to be a quantile. To be more precise, let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a sample of i.i.d. random variables with distribution P. We notice that choosing the parameter τ for LLD is equivalent to choose a quantile q for the pairwise distances $\|X_i - X_j\|$, $i > j$, $i, j \in \{1, 2, \ldots, n\}$. Similar considerations hold for LSD (see Appendix E (Francisci et al., 2023)). Thus, we choose q following the discussion in Subsection 3.3. We now illustrate this idea when P is the Mult. Quadrimodal distribution. Figure 1 shows the median adjusted Rand index and interquartile range as a function of the quantile order q (left). The center plot shows the boxplot of optimal value of q and the right plot displays the number of clusters detected when q is the optimal quantile order. Based on this preliminary analysis we conclude that the optimal value of q for LLD lies between 0.01 and 0.1. Thus, we restrict our numerical experiments to values of q in that range (cf. Appendix F (Francisci et al., 2023)). Additional analysis shows that some circular distributions require values of q higher than 0.1 (see Figure 1 in Appendix F.1).

Next, we compare LLD and LSD using the clustering algorithm in Theorem 3.4 with KDE using both the above clustering algorithm and mean shift algorithm of Fukunaga and Hostetler (1975) abbreviated as KDE-`"ms"`. We emphasize here that by KDE we mean implementation with our algorithm. We also compare with two recent clustering algorithms, which are a combination of mixture model clustering (Fraley and Raftery, 2002) and modal clustering (Chacón, 2015): (i) mixture model modal merging (MMMM) and (ii) mixture model modal clustering (MMMC). For more details we refer to Chacón (2019) or Appendix F.1. Our simulation results are based on a sample size of 1000 and 100 numerical experiments. For more details on the numerical implementation and the experimental setting we refer to Appendices E and F (Francisci et al., 2023).

Based on the results, we notice that our clustering algorithm performs adequately and outperforms in some cases compared with KDE-`"ms"`. As expected, MMMM and MMMC are the best for mixture densities. However, they perform

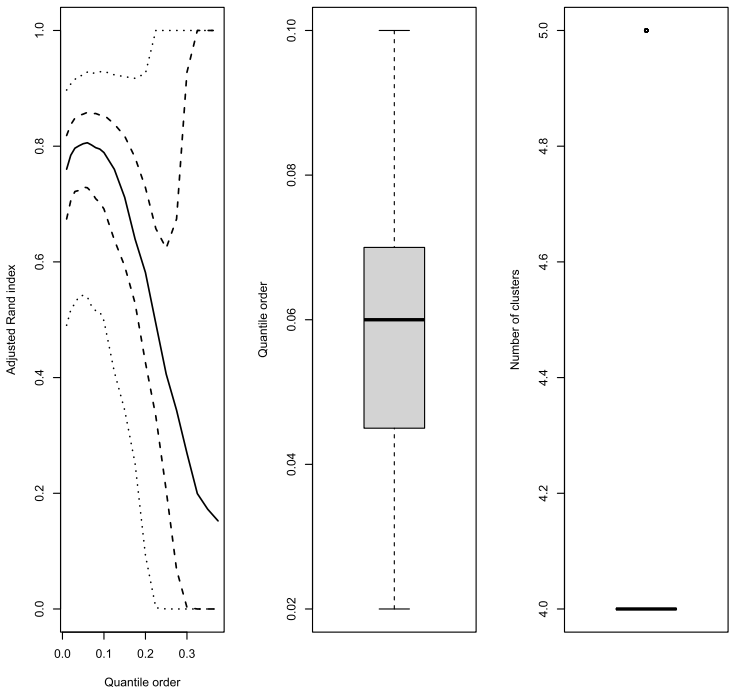

FIG 1. *For* 100 *replications with* $n = 1000$ *samples for the Mult. Quadrimodal distribution and LLD (i) median adjusted Rand index and interquartile range as a function of the quantile order q (left), (ii) boxplot of the optimal quantile q (center), and (iii) boxplot of the number of clusters for the optimal quantile q (right).*

poorly for Circular Bimodal II distribution. More extensive numerical experiments are included in Appendix F (Francisci et al., 2023) and Francisci et al. (2022). Description of the R code used for simulations is included in Appendix F.1.

3.5. Data analysis

We evaluate the performance of our methodology on Iris dataset and Seeds dataset, both available from the UCI machine learning repository (`http://archive.ics.uci.edu/ml/`). In the Iris dataset the sample size is $n = 150$ and there are three classes (Iris Setosa, Iris Versicolour,and Iris Virginica) with four measurements each (sepal length, sepal width, petal length, and petal width). Our algorithm using LDFs and KDE correctly identifies the true number of clusters with probability and Hausdorff distances of 0.1 and 0.03 for both methods. KDE-`"ms"` overestimates the number of clusters with probability and Hausdorff distances 0.37 and 0.31, respectively. Next, turning to Seeds dataset, the sample size is $n = 210$ and there are three clusters relating to three varieties of wheat (Kama, Rosa and Canadian). The data are in seven dimensions rep-

TABLE 1
Mean of the clustering errors based on distance in probability ($c = 1$) for the densities (H) Bimodal IV, #10 Fountain, Mult. Quadrimodal, and Circular Bimodal II. In parentheses the standard deviation.

Clustering errors (distance in probability)		
	(H) Bimodal IV	#10 Fountain
MMMM	**0.00 (0.00)**	0.23 (0.07)
MMMC	**0.00 (0.00)**	0.10 (0.07)
KDE	**0.00 (0.00)**	**0.06 (0.01)**
KDE-`"ms"`	0.01 (0.07)	0.21 (0.31)
LLD [1]	0.13 (0.28)	**0.06 (0.01)**
LSD [2]	0.12 (0.27)	**0.06 (0.01)**
	Mult. Quadrimodal	Circular Bimodal II
MMMM	**0.01 (0.00)**	0.55 (0.05)
MMMC	**0.01 (0.00)**	0.53 (0.05)
KDE	0.34 (0.37)	0.36 (0.06)
KDE-`"ms"`	0.57 (0.33)	0.38 (0.12)
LLD [1]	0.09 (0.22)	0.43 (0.11)
LSD [3]	0.45 (0.17)	**0.23 (0.17)**

[1] $q = 0.1$. [2] $q = 0.01$. [3] $q = 0.05$.

TABLE 2
Mean of the clustering errors based on Hausdorff distance for the densities (H) Bimodal IV, #10 Fountain, Mult. Quadrimodal, and Circular Bimodal II. In parentheses the standard deviation.

Clustering errors (Hausdorff distance)		
	(H) Bimodal IV	#10 Fountain
MMMM	**0.00 (0.00)**	0.22 (0.03)
MMMC	**0.00 (0.00)**	0.09 (0.02)
KDE	**0.00 (0.00)**	**0.06 (0.01)**
KDE-`"ms"`	**0.00 (0.03)**	0.08 (0.05)
LLD [1]	0.05 (0.11)	**0.06 (0.01)**
LSD [2]	0.05 (0.11)	**0.06 (0.01)**

	Mult. Quadrimodal	Circular Bimodal II
MMMM	**0.01 (0.00)**	0.55 (0.04)
MMMC	**0.01 (0.00)**	0.53 (0.05)
KDE	0.10 (0.08)	0.44 (0.07)
KDE-`"ms"`	0.16 (0.08)	0.44 (0.07)
LLD [1]	0.03 (0.05)	0.47 (0.07)
LSD [3]	0.38 (0.18)	**0.28 (0.20)**

resenting geometric parameters (continuous) of wheat kernels. Our algorithm correctly identified the three clusters with probability and Hausdorff distance 0.1 for LLD, 0.16 for KDE, and 0.17 for LSD. On the other hand, KDE-`"ms"` (with built-in bandwidth) overestimated the number of clusters with probability and Hausdorff distances 0.75 and 0.33, respectively. For more details we refer to Appendix F (Francisci et al., 2023).

4. Proofs

In this section, we provide detailed proofs of Theorems 2.2-2.4 and Theorem 3.4. The proofs of preliminary results and Theorem 2.1 are given in Appendix A (Francisci et al., 2023).

Proof of Theorem 2.2. Recall that $\mathcal{H}_G = \{h_\tau^{(G)}(x;\cdot) : x \in \mathbb{R}^p, \tau \in [0,\infty]\}$ and let $\mathcal{H}_{G,1} \coloneqq \{h_\tau^{(G,1)}(x;\cdot) : x \in \mathbb{R}^p, \tau \in [0,\infty]\}$. We will show that

$$\sup_{h^{(G)}\in\mathcal{H}_G}\left|\int h^{(G)}(x_1,\ldots,x_{k_G})\prod_{i=1}^{k_G}dP(x_i) - \binom{n}{k_G}^{-1}\sum_{1\le i_1<\cdots<i_{k_G}\le n} h^{(G)}(X_{i_1},\ldots,X_{i_{k_G}})\right|$$

converges to 0 with probability one. To this end, we use Corollary 3.3 of Arcones and Giné (1993). Since $\mathcal{H}_G$ is a VC-subgraph class by hypothesis it is enough to verify that (i) $\sup_{h^{(G)}\in\mathcal{H}_G}|h^{(G)}(\cdot)| < \infty$ and $\sup_{h^{(G,1)}\in\mathcal{H}_{G,1}}|h^{(G,1)}(\cdot)| < \infty$ and (ii) $\mathcal{H}_G$ is image admissible Suslin (Dudley, 2014, p. 186). This then shows that $\mathcal{H}_G$ is a measurable class (Arcones and Giné, 1993, p. 1497) with a bounded envelope. To this end, by (A.1), $\sup_{h^{(G)}\in\mathcal{H}_G}|h^{(G)}(\cdot)| \le l_G$, $\sup_{h^{(G,1)}\in\mathcal{H}_{G,1}}|h^{(G,1)}(\cdot)| \le l_G$, and hence (i) holds. Turning to (ii), we show that the function $\mathfrak{i}_G : [0,\infty] \times \mathbb{R}^p \times (\mathbb{R}^p)^{k_G} \to \mathbb{R}$ given by $\mathfrak{i}_G(\tau;x;x_1,\ldots,x_{k_G}) = h_\tau^{(G)}(x;x_1,\ldots,x_{k_G})$ is Borel measurable. To see this, let $F_G : (0,\infty)\times\mathbb{R}^p\times(\mathbb{R}^p)^{k_G} \to (\mathbb{R}^p)^{k_G}$ be given by $F_G(\tau;x;x_1,\ldots,x_{k_G}) = (\frac{x_1-x}{\tau},\ldots,\frac{x_{k_G}-x}{\tau})^\top$. Since $G(\cdot)$ is Borel measurable and $F_G(\cdot)$ is continuous, $h_{(\cdot)}^{(G)}(\cdot;\cdot) = G(F_G(\cdot))$ is Borel measurable. In particular, $h_\tau^{(G)}(\cdot;\cdot)$ is Borel measurable for all $\tau \in (0,\infty)$ and $h_0^{(G)}(\cdot;\cdot)$ and $h_\infty^{(G)}(\cdot;\cdot)$ are Borel measurable because they are limit of Borel measurable functions. It follows that, for all $A \in \mathscr{B}(\mathbb{R})$,

$$\begin{aligned}\mathfrak{i}_G^{-1}(A) =& (F_G^{-1}(G^{-1}(A)) \cup (\{0\}\times(h_0^{(G)})^{-1}(A)) \cup (\{\infty\}\times(h_\infty^{(G)})^{-1}(A))\\ &\in \mathscr{B}([0,\infty])\times\mathscr{B}(\mathbb{R}^p)\times\mathscr{B}((\mathbb{R}^p)^{k_G}) = \mathscr{B}([0,\infty]\times\mathbb{R}^p\times(\mathbb{R}^p)^{k_G}),\end{aligned}$$

that is, $\mathfrak{i}_G(\cdot)$ is Borel measurable. Hence, by Dudley (2014, p. 186), the class $\mathcal{H}_G$ is image admissible Suslin via the onto Borel measurable map $\mathfrak{e}_G : [0,\infty]\times\mathbb{R}^p \to \mathcal{H}_G$ given by $\mathfrak{e}_G(\tau;x) = h_\tau^{(G)}(x;\cdot)$. ∎

Before proving Theorem 2.3 and Theorem 2.4, we recall that, given a pseudometric space $(\mathcal{H}, d)$, the ϵ-covering number of $\mathcal{H}$ w.r.t. the pseudodistance d, $N(\mathcal{H}, d, \epsilon)$, is the minimum number of balls with radius at most ϵ required to cover $\mathcal{H}$.

Proof of Theorem 2.3. To prove Theorem 2.3, we will verify the conditions of Theorem 4.9 in Arcones and Giné (1993). To this end, first let $\mathcal{H}_G^{(T)} \coloneqq \{h_\tau^{(G)}(x;\cdot) : (x,\tau) \in T\} \subset \mathcal{H}_G$ and $\mathcal{H}_{G,1}^{(T)} \coloneqq \{h_\tau^{(G,1)}(x;\cdot) : (x,\tau) \in T\} \subset \mathcal{H}_{G,1}$. As in the proof of Theorem 2.2 where $[0,\infty]\times\mathbb{R}^p$ is replaced by T with the corresponding subspace topology, we see that $\mathcal{H}_G^{(T)}$ is image admissible Suslin (Dudley, 2014, p. 186). Also, using (A.1), it holds that $\sup_{h\in\mathcal{H}_G^{(T)}}|h(\cdot)| \le l_G$ and $\sup_{h^{(1)}\in\mathcal{H}_{G,1}^{(T)}}|h^{(1)}(\cdot)| \le l_G$. This then shows that $\mathcal{H}_G^{(T)}$ is a measurable class with

a bounded envelope and (ii) of Theorem 4.9 in Arcones and Giné (1993) holds. To verify (iii) in Arcones and Giné (1993) we appeal to Lemma 4.4 and (4.2) in Alexander (1987) concerning the covering number $N(\mathcal{H}_G^{(T)}, d_{L^2(\mathcal{H}_G^{(T)},P)}, \cdot)$ of $\mathcal{H}_G^{(T)}$ with respect to the L^2-distance, $d_{L^2(\mathcal{H}_G^{(T)},P)}$, given by

$$d^2_{L^2(\mathcal{H}_G^{(T)},P)}((x,\tau),(y,\nu))=\int (h_\tau^{(G)}(x;x_1,\ldots,x_k)-h_\nu^{(G)}(y;x_1,\ldots,x_k))^2\prod_{i=1}^k dP(x_i).$$

For this, we observe that $\mathcal{H}_G^{(T)}$ is a VC-subgraph class of functions. Thus to complete the proof, we need to verify (i) in Arcones and Giné (1993). To this end, we need to show: (a) the finite dimensional distributions of $\sqrt{n}(LGD_n(x,\tau,P)-LGD(x,\tau,P))$ converge to a multivariate normal distribution and (b) for each (x,τ), the limiting normal random variable $\{W(x,\tau)\}_{(x,\tau)\in T}$ admits a version whose sample paths are all bounded and uniformly continuous with respect to the distance $d^2_{\mathcal{H}_{G,1}^{(T)},P}$ on $\mathcal{H}_{G,1}^{(T)}$ given by

$$\begin{aligned} d^2_{\mathcal{H}_{G,1}^{(T)},P}((x,\tau),(y,\nu)) = &\int (h_\tau^{(G,1)}(x;x_1)-h_\nu^{(G,1)}(y;x_1))^2 dP(x_1) \\ &-(LGD(x,\tau)-LGD(y,\nu))^2, \end{aligned}$$

where we identify a function $h_\tau^{(G,1)}(x;\cdot)$ for $(x,\tau)\in T$ with its parameter (x,τ). In this sense, $d^2_{\mathcal{H}_{G,1}^{(T)},P}$ is a metric on T. Since $W(x,\tau)$ is Gaussian, we can apply Giné and Nickl (2016, Theorem 2.3.7) with $T=T$ and $d=d^2_{\mathcal{H}_{G,1}^{(T)},P}$. First, note that $\{W(x,\tau)\}_{(x,\tau)\in T}$ is a sub-Gaussian process relative to $d^2_{\mathcal{H}_{G,1}^{(T)},P}$. Indeed, using Proposition A.1 in Appendix A (Francisci et al., 2023), for $(x,\tau),(y,\nu)\in T$, $(W(x,\tau),W(y,\nu))^\top$ has a bivariate normal distribution with mean $(0,0)^\top$ and covariance matrix

$$\begin{pmatrix} E[(\tilde{h}_\tau^{(G,1)}(x;X_1))^2] & \gamma((x,\tau),(y,\nu)) \\ \gamma((y,\nu),(x,\tau)) & E[(\tilde{h}_\nu^{(G,1)}(y;X_1))^2]. \end{pmatrix}.$$

It follows that $W(x,\tau)-W(y,\nu)$ is normally distributed with mean 0 and variance

$$E[(\tilde{h}_\tau^{(G,1)}(x;X_1))^2]+E[(\tilde{h}_\nu^{(G,1)}(y;X_1))^2]-2\gamma((x,\tau),(y,\nu))=d^2_{\mathcal{H}_{G,1}^{(T)},P}((x,\tau),(y,\nu)).$$

Therefore, for all $\alpha\in\mathbb{R}$

$$E[\exp(\alpha(W(x,\tau)-W(y,\nu)))]=\exp\left(\frac{\alpha^2}{2}d^2_{\mathcal{H}_{G,1}^{(T)},P}((x,\tau),(y,\nu))\right)$$

and the process $\{W(x,\tau)\}_{(x,\tau)\in T}$ is sub-Gaussian with respect to $d^2_{\mathcal{H}_{G,1}^{(T)},P}$. We next verify the integrability condition for the metric entropy. To this end, notice

that, for $(x,\tau),(y,\nu) \in T$, the L^2-distance on $\mathcal{H}_{G,1}^{(T)}$, $d_{L^2(\mathcal{H}_{G,1}^{(T)},P)}$ is given by

$$d^2_{L^2(\mathcal{H}_{G,1}^{(T)},P)}((x,\tau),(y,\nu)) = \int (h_\tau^{(G,1)}(x;x_1) - h_\nu^{(G,1)}(y;x_1))^2 dP(x_1).$$

Now using yet another application of Lemma 4.4 of Alexander (1987), it follows that there are constants $C_1, C_2 > 1$ such that

$$N(\mathcal{H}_G^{(T)}, d_{L^2(\mathcal{H}_G^{(T)},P)}, \sqrt{\epsilon}) \leq \left(\frac{C_1}{\sqrt{\epsilon}}\right)^{C_2}.$$

By Jensen's inequality, it follows that

$$d_{L^2(\mathcal{H}_{G,1}^{(T)},P)}((x,\tau),(y,\nu)) \leq d_{L^2(\mathcal{H}_G^{(T)},P)}((x,\tau),(y,\nu)),$$

which in turn, implies that

$$N(\mathcal{H}_{G,1}^{(T)}, d_{L^2(\mathcal{H}_{G,1}^{(T)},P)}, \sqrt{\epsilon}) \leq N(\mathcal{H}_G^{(T)}, d_{L^2(\mathcal{H}_G^{(T)},P)}, \sqrt{\epsilon}) \leq \left(\frac{C_1}{\sqrt{\epsilon}}\right)^{C_2}.$$

Thus, for any $0 < \epsilon \leq 1$,

$$\begin{aligned} N(\mathcal{H}_{G,1}^{(T)}, d^2_{\mathcal{H}_{G,1}^{(T)},P}, \epsilon) &\leq N(\mathcal{H}_{G,1}^{(T)}, d^2_{L^2(\mathcal{H}_{G,1}^{(T)},P)}, \epsilon) = N(\mathcal{H}_{G,1}^{(T)}, d_{L^2(\mathcal{H}_{G,1}^{(T)},P)}, \sqrt{\epsilon}) \\ &\leq \left(\frac{C_1}{\sqrt{\epsilon}}\right)^{C_2} \leq \left(\frac{C_1}{\epsilon}\right)^{C_2}. \end{aligned} \tag{4.1}$$

It follows that

$$\int_0^1 \sqrt{\log(N(\mathcal{H}_{G,1}^{(T)}, d^2_{\mathcal{H}_{G,1}^{(T)},P}, \epsilon))} d\epsilon \leq \sqrt{C_2} \int_0^1 \sqrt{\log(C_1) - \log(\epsilon)} d\epsilon$$

Now, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, it follows that the left hand side (LHS) is bounded above by $\sqrt{C_2}$ times

$$\begin{aligned} &\sqrt{\log(C_1)} + \int_0^{e^{-1}} \sqrt{-\log(\epsilon)} d\epsilon + \int_{e^{-1}}^1 \sqrt{-\log(\epsilon)} d\epsilon \\ \leq &\sqrt{\log(C_1)} - \int_0^{e^{-1}} \log(\epsilon) d\epsilon + 1 - e^{-1} \\ = &\sqrt{\log(C_1)} + e^{-1} + 1 < \infty. \end{aligned}$$

The proof finally follows from Proposition A.1 in Appendix A (Francisci et al., 2023). ∎

Proof of Theorem 2.4. We will show that there are constants $1 < C_{G,0}, C_{G,1}, C_{G,2} < \infty$ such that

$$P^{\otimes n}(\sqrt{n} M_{G,n} \geq t) \leq D_G(n,t),$$

where $M_{G,n}$ is defined to be

$$\sup_{h^{(G)}\in\mathcal{H}_G}\left|\int h^{(G)}(x_1,\ldots,x_{k_G})\prod_{j=1}^{k_G}dP(x_j)-\binom{n}{k_G}^{-1}_{1\le i_1<\cdots<i_{k_G}\le n}\sum h^{(G)}(X_{i_1},\ldots,X_{i_{k_G}})\right|.$$

To this end, we verify the conditions of Theorem 5 in Arcones (1995). By (i)-(ii) in the proof of Theorem 2.2, it follows that $\mathcal{H}_G$ is a uniformly bounded, measurable (Arcones and Giné, 1993, p. 1497), VC-subgraph class, where the bounding constant is l_G. We show that this implies conditions (i)-(iii) of Theorem 5 in Arcones (1995). Condition (i) is clear. Let

$$d^2_{L^2(\mathcal{H}_G,P)}((x,\tau),(y,\nu))=\int(h^{(G)}_\tau(x;x_1,\ldots,x_{k_G})-h^{(G)}_\nu(y;x_1,\ldots,x_{k_G}))^2\prod_{j=1}^{k_G}dP(x_j)$$

and

$$d^2_{L^2(\mathcal{H}_{G,1},P)}((x,\tau),(y,\nu))=\int(h^{(G,1)}_\tau(x;x_1)-h^{(G,1)}_\nu(y;x_1))^2dP(x_1).$$

By Jensen's inequality, it holds that

$$d_{L^2(\mathcal{H}_{G,1},P)}((x,\tau),(y,\nu))\le d_{L^2(\mathcal{H}_G,P)}((x,\tau),(y,\nu)).$$

Using Lemma 4.4 and (4.2) in Alexander (1987), we see that there are constants $C_{G,1},C_{G,2}>1$ such that

$$N(\mathcal{H}_{G,1},d_{L^2(\mathcal{H}^{(T)}_{G,1},P)},\epsilon)\le N(\mathcal{H}_G,d_{\mathcal{H}_G},\epsilon)\le\left(\frac{C_{G,1}}{\epsilon}\right)^{C_{G,2}}. \tag{4.2}$$

Thus, condition (ii) holds true. Finally, (4.2) and (Arcones, 1995, (3.3) and p. 245) imply that there is a constant $C_{G,0}$ such that (iii) holds true. ∎

Proof of Theorem 3.4. The proof is divided into four steps. In the first step below we introduce few notations and preliminary calculations. We recall that the distance between two sets $A,B\subset\mathbb{R}^p$ is $\operatorname{dist}(A,B):=\inf_{y\in A,z\in B}\|y-z\|$.

Step 0. We define

$$A_{\overline{\alpha},\xi,i}:=R^{\overline{\alpha}}\cap(C(m_i)^{-\xi})\text{ and }K_{\overline{\alpha},\xi,i}:=\overline{(\cup_{x\in A_{\overline{\alpha},\xi,i}}G_x)},$$

where $G_x:=\{u_x(t):t\in[0,\infty)\}$. We notice that $K_{\overline{\alpha},\xi,i}$ is a closed subset of $R^{\overline{\alpha}}$ and $C(m_i)$, which is open, and for $0<\tilde{\alpha}<\overline{\alpha}$, $0<\tilde{\xi}<\xi$, and $K_i:=K_{\tilde{\alpha},\tilde{\xi},i}$

$$\delta:=\min_{i=1,\ldots,M}\operatorname{dist}(K_{\overline{\alpha},\xi,i},\mathbb{R}^p\setminus K_i)>0,$$

yielding that $(G_x)^{+\delta}\subset K_i$ for all $x\in A_{\overline{\alpha},\xi,i}$. Also, let $K:=\cup_{i=1}^M K_i$. At the cost of possibly reducing $\tilde{\alpha}$ and $\tilde{\xi}$ (and thus δ), we let $0<\alpha_i<\tilde{\alpha}_i<f(m_i)$ such that

$$R^{\tilde{\alpha}_i}\cap C(m_i)\subset\overline{B}_\delta(m_i), \tag{4.3}$$

$$\overline{B}_{4\delta}(m_i) \subset R^{\alpha_i} \cap C(m_i) \subset B_\epsilon(m_i), \text{ and} \tag{4.4}$$

$$\overline{B}_{4\delta}(m_i) \subset K_{\overline{\alpha},\xi,i} \subset K_i. \tag{4.5}$$

For $z \in \mathbb{R}^p$ with $\nabla f(z) \neq 0$, let

$$w(z) \coloneqq \nabla f(z)/\|\nabla f(z)\| \tag{4.6}$$

and, for $0 < r \leq \xi$ and $j^* \geq 0$ let

$$\mathcal{G}_{x,r} \coloneqq \Bigg\{ \{y_{r,j}\}_{j=0}^{j^*} : j^* \geq 0,\, y_{r,0} = x \text{ and, recursively, } y_{r,j+1} = y_{r,j} + h_j v_j \\ \text{for some } (h_j, v_j) \in (0,r] \times (S^{p-1} \cap \overline{B}_r(w(y_{r,j}))) \Bigg\}. \tag{4.7}$$

Step 1. We show that, for small r, every sequence $\{y_{r,j}\}_{j=0}^{j^*} \in \mathcal{G}_{x,r}$, $x \in \cup_{i=1}^M K_{\overline{\alpha},\xi,i}$, either remains in $(G_x)^{+\delta} \setminus B_\delta(m_i)$ or, for some $j \in \{0,\ldots,j^*\}$, $y_{r,j} \in \overline{B}_{4\delta}(m_i)$. To this end, we suppose w.l.o.g. that $x \in A_{\overline{\alpha},\xi} \coloneqq \cup_{i=1}^M (K_{\overline{\alpha},\xi,i} \setminus \overline{B}_{2\delta}(m_i))$. Indeed, if $x \in \overline{B}_{2\delta}(m_i)$, then $y_{r,0} = x \in \overline{B}_{4\delta}(m_i)$. We now define some quantities that are used in the proof of this fact. Specifically, let $t_0 \coloneqq 0$ and, recursively, $t_{j+1} = \sum_{l=0}^{j} h_l/\|\nabla f(y_{r,l})\|$. Also, let

$$\tilde{t}(x) \coloneqq \inf\{t \in [0,\infty) : u_x(t) \in \overline{B}_{2\delta}(m_i)\},$$
$$\tilde{j}^* \coloneqq \max\{j \in \{0,\ldots,j^*\} : t_j \leq \tilde{t}(x)\},$$

$\tilde{K}_i \coloneqq K_i \setminus \mathring{R}^{\tilde{\alpha}_i}$, $\tilde{K} \coloneqq \cup_{i=1}^M \tilde{K}_i$, $\underline{\nu} \coloneqq \min_{y\in\tilde{K}} \|\nabla f(y)\|$, and $\overline{\nu} \coloneqq \max_{y\in\tilde{K}} \|\nabla f(y)\|$. Notice that by (4.3) $(G_x)^{+\delta} \setminus B_\delta(m_i) \subset \tilde{K}_i$ and $\underline{\nu}, \overline{\nu} > 0$ as $\tilde{K} \cap N_f = \emptyset$. First, we show that

$$t^* \coloneqq \sup_{x \in A_{\overline{\alpha},\xi}} \tilde{t}(x) < \infty. \tag{4.8}$$

To this end, suppose by contradiction that $t^* = \infty$ and let $\{x_l\}_{l=1}^\infty$ in $A_{\overline{\alpha},\xi}$ such that $\lim_{l\to\infty} \tilde{t}(x_l) = \infty$. $\{x_l\}_{l=1}^\infty$ has a convergent subsequence $\{x_{j_l}\}_{l=1}^\infty$ in $\overline{A}_{\overline{\alpha},\xi}$, that is, $x \coloneqq \lim_{l\to\infty} x_{j_l} \in \overline{A}_{\overline{\alpha},\xi}$. It is clear that

$$\overline{t}(x) \coloneqq \inf\{t \in [0,\infty) : u_x(t) \in \overline{B}_\delta(m_i)\} < \infty.$$

Since $\nabla f(\cdot)$ is differentiable, it is Lipschitz in $\tilde{K}$. Denote by L the Lipschitz constant. By continuity of solutions of ordinary differential equations with respect to the initial value (see Theorem 2.8 and (2.43) in Teschl (2012)), for all $t \geq 0$, it holds that

$$\|u_{x_{j_l}}(t) - u_x(t)\| \leq \|x_{j_l} - x\| e^{Lt}.$$

Fix $0 < \zeta < \delta$ and let $l^* \in \mathbb{N}$ such that

$$\|x_{j_l} - x\| \leq \zeta e^{-L\overline{t}(x)} \text{ for all } l \leq l^*.$$

It follows that

$$\|u_{x_{j_l}}(t) - u_x(t)\| \leq \zeta \text{ for all } 0 \leq t \leq \overline{t}(x).$$

Since $x_{j_l} \in A_{\overline{\alpha},\xi}$ and $u_{x_{j_l}}(\overline{t}(x)) \in B_{2\delta}(m_i)$, it obtains $0 \leq \tilde{t}(x_{j_l}) \leq \overline{t}(x)$ for all $l \geq l^*$. (4.8) follows. Next, we show that $u_{(\cdot)}(\cdot)$ is jointly continuous in $\tilde{K} \times [0, t^*]$. Let $(x_l, t_l) \to (x, t) \in \tilde{K} \times [0, t^*]$. If $\|x_l - x\| \leq e^{-Lt^*}\zeta/2$ and $\|u_x(t_l) - u_x(t)\| \leq \zeta/2$ for all $l \geq l^*$, then, using again continuity w.r.t. the initial value, we obtain

$$\begin{aligned}\|u_{x_l}(t_l) - u_x(t)\| &\leq \|u_{x_l}(t_l) - u_x(t_l)\| + \|u_x(t_l) - u_x(t)\| \\ &\leq \|x_l - x\| e^{Lt^*} + \zeta/2 \leq \zeta.\end{aligned}$$

Since $u''_x(t) = H_f(u_x(t))\nabla f(u_x(t))$ and $f(\cdot)$ is three times continuously differentiable, $u''_{(\cdot)}(\cdot)$ is also uniformly continuous in $\tilde{K} \times [0, t^*]$. Using (4.8) and uniform continuity of $u_{(\cdot)}(\cdot)$, $u''_{(\cdot)}(\cdot)$, let $0 < r_1 \leq \delta$, such that

$$r_1 t^* \left(\overline{\nu} + \sup_{x \in \tilde{K}} \sup_{t \in [0,t^*]} \|u''_x(t)\|/(2\underline{\nu}) \right) \exp(Lt^*) \leq \delta \tag{4.9}$$

and, for all $0 < r \leq r_1$ and $x \in A_{\overline{\alpha},\xi}$,

$$\|u_x(\tilde{t}(x) - r/\underline{\nu}) - u_x(\tilde{t}(x))\| \leq \delta. \tag{4.10}$$

We show that, for all $j = 0, \ldots, \tilde{j}^*$ and $0 < r \leq r_1$, $y_{r,j} \in (G_x)^{+\delta} \setminus B_\delta(m_i)$. We recall that, by (4.3), $(G_x)^{+\delta} \setminus B_\delta(m_i) \subset \tilde{K}$. First, notice that $u_x(t_0) = x$ and, since $\|x - m_i\| > 2\delta$, it holds that $y_{r,0} = x \in (G_x)^{+\delta} \setminus B_\delta(m_i)$. We now suppose by induction that, for $j \geq 1$, $y_{r,j-1} \in (G_x)^{+\delta} \setminus B_\delta(m_i)$ and show that $y_{r,j} \in (G_x)^{+\delta} \setminus B_\delta(m_i)$. Since $u'_x(t) = \nabla f(u_x(t))$ and $f(\cdot)$ is three times continuously differentiable, then so is $u_x(\cdot)$. By Taylor theorem with Lagrange's form of remainder, there exists $t_{j-1} \leq \tilde{t}_{j-1} \leq t_j$ such that

$$u_x(t_j) = u_x(t_{j-1}) + \frac{h_{j-1}}{\|\nabla f(y_{r,j-1})\|} \nabla f(u_x(t_{j-1})) + \frac{h_{j-1}^2}{2\|\nabla f(y_{r,j-1})\|^2} u''_x(\tilde{t}_{j-1}).$$

It follows that

$$\begin{aligned}(y_{r,j} - u_x(t_j)) &= (y_{r,j-1} - u_x(t_{j-1})) + h_{j-1}(v_{j-1} - w(y_{r,j-1})) \\ &\quad + \frac{h_{j-1}}{\|\nabla f(y_{r,j-1})\|} \left(\nabla f(y_{r,j-1}) - \nabla f(u_x(t_{j-1})) \right) \\ &\quad - \frac{h_{j-1}^2}{2\|\nabla f(y_{r,j-1})\|^2} u''_x(\tilde{t}_{j-1}).\end{aligned}$$

Now, we use the Lipschitz property of $\nabla f(\cdot)$ and get

$$\begin{aligned}\|y_{r,j} - u_x(t_j)\| &\leq \left(1 + \frac{h_{j-1}L}{\|\nabla f(y_{r,j-1})\|}\right) \|y_{r,j-1} - u_x(t_{j-1})\| + r_1 h_{j-1} \\ &\quad + \frac{h_{j-1}^2}{2\|\nabla f(y_{r,j-1})\|^2} \sup_{t \in [0, \tilde{t}(x)]} \|u''_x(t)\|.\end{aligned}$$

We now apply Lemma A.7 in Appendix A (Francisci et al., 2023) with $a_j = \|y_{r,j} - u_x(t_j)\|$,

$$b_j = r_1 h_j + \frac{h_j^2}{2\|\nabla f(y_{r,j})\|^2} \sup_{t\in[0,\tilde{t}(x)]} \|u_x''(t)\|,$$

$c_j = \frac{h_j L}{\|\nabla f(y_{r,j})\|}$ and, using (4.9) and $t_j \le \tilde{t}(x)$, we get that $\|y_{r,j} - u_x(t_j)\|$ is bounded above by

$$\begin{aligned}&\left(r_1 \sum_{l=0}^{j-1} h_l + \sum_{l=0}^{j-1} \frac{h_j^2}{2\|\nabla f(y_{r,l})\|^2} \sup_{t\in[0,\tilde{t}(x)]} \|u_x''(t)\|\right) \exp\left(L \sum_{l=1}^{j-1} \frac{h_l}{\|\nabla f(y_{r,j})\|}\right)\\ \le& r_1 t_j \left(\overline{\nu} + \sup_{t\in[0,\tilde{t}(x)]} \|u_x''(t)\|/(2\underline{\nu})\right) \exp(L t_j) \le \delta.\end{aligned}$$

It follows that $y_{r,j} \in (G_x)^{+\delta}$. Moreover, $t_j \le \tilde{t}(x)$ implies that $\|m_i - u_x(t_j)\| \ge 2\delta$. Hence,

$$\|m_i - y_{r,j}\| \ge \|m_i - u_x(t_j)\| - \|u_x(t_j) - y_{r,j}\| \ge \delta,$$

that is, $y_{r,j} \notin B_\delta(m_i)$. In particular, if $\tilde{j}^* = j^*$, then $y_{r,j} \in (G_x)^{+\delta} \setminus B_\delta(m_i)$ for all $j = 0, \dots, j^*$. Next, we show that, if $\tilde{j}^* < j^*$, then $y_{r,\tilde{j}^*} \in \overline{B}_{4\delta}(m_i)$. Since $\tilde{t}(x) - r_1/\underline{\nu} < t_{\tilde{j}^*+1} - r_1/\underline{\nu} \le t_{\tilde{j}^*} \le \tilde{t}(x)$, by (4.10) it holds that $\|u_x(t_{\tilde{j}^*}) - u_x(\tilde{t}(x))\| \le \delta$. Since $u_x(\tilde{t}(x)) \in \partial B_{2\delta}(m_i)$, we conclude that

$$\|y_{r,\tilde{j}^*} - m_i\| \le \|y_{r,\tilde{j}^*} - u_x(t_{\tilde{j}^*})\| + \|u_x(t_{\tilde{j}^*}) - u_x(\tilde{t}(x))\| + \|u_x(\tilde{t}(x)) - m_i\| \le 4\delta.$$

Step 2. We apply Lemma A.8 in Appendix A (Francisci et al., 2023) with $K = \tilde{K}$ and get constants $r^* := \min(r_1, r(\tilde{K})) > 0$ and $c^* := c(\tilde{K}) > 0$ such that, for all $x \in \tilde{K}$ and $(h, v) \in (0, r^*] \times (S^{p-1} \cap \overline{B}_{r^*}(w(x)))$

$$\nabla_v^h f(x) \ge c^*. \tag{4.11}$$

For $X \in \mathcal{X}_n$ and $x \in S_f$ let $h_{X,x} := \|X - x\|$ and $v_{X,x} := (X - x)/h_{X,x}$. We show the existence of $0 < r_2 \le r^*$ such that for all $0 < r \le r_2$ there exist $n_1, n_2 \in \mathbb{N}$ such that, with probability at least $1 - \eta$, for $n \ge \max(n_1, n_2)$ and $x \in \tilde{K}$, we have that $\mathcal{X}_{n,r}(x) \ne \emptyset$,

$$\max_{X\in\mathcal{X}_{n,r}(x)\cup\{x\}} f_{\tau_n,n}(X) - f_{\tau_n,n}(x) > 0 \tag{4.12}$$

and

$$X^*(x) := \operatorname*{argmax}_{X\in\mathcal{X}_{n,r}(x)} \frac{f_{\tau_n,n}(X) - f_{\tau_n,n}(x)}{\|X - x\|} = \operatorname*{argmax}_{X\in\mathcal{X}_{n,r}(x)} \nabla_{v_{X^*(x),x}}^{h_{X^*(x),x}} f_{\tau_n,n}(x)$$

satisfies

$$(h_{X^*(x),x}, v_{X^*(x),x}) \in [h_n, r] \times (S^{p-1} \cap \overline{B}_{r^*}(w(x)). \tag{4.13}$$

To this end, suppose w.l.o.g. that $r^* \leq 1$. Let

$$d(r^*) \coloneqq \inf_{y \in \tilde{K}} \inf_{v \in S^{p-1} \setminus \overline{B}_{r^*}(w(y))} \langle w(y) - v, \nabla f(y) \rangle > 0$$

and $0 < d^* < d(r^*)/(5\overline{\nu})$. Notice that, since $d(r^*) \leq \underline{\nu} r^*$, $d^* < r^*/5 \leq 1/5$ and $\tilde{d}(x) \coloneqq (1 - 3d^*)\|\nabla f(x)\| > 0$ for all $x \in \tilde{K}$. By the mean value theorem, there exists $0 \leq c \leq 1$ such that

$$\nabla_v^h f(x) = \langle v, \nabla f(x + chv) \rangle.$$

Next, by the uniform continuity of $\nabla f(\cdot)$ over compact sets, we have that $\nabla_v^h f(x)$ converges to $\nabla_v f(x)$ uniformly over $v \in S^{p-1}$ and $x \in \tilde{K}$. Let $r_3 > 0$ be such that for all $h \in (0, r_3]$, $v \in S^{p-1}$, and $x \in \tilde{K}$

$$|\nabla_v^h f(x) - \nabla_v f(x)| \leq \underline{\nu} d^*.$$

Then, for all $x \in \tilde{K}$ and $v \in S^{p-1} \cap \overline{B}_{d^*}(w(x))$, it holds that

$$\nabla_v f(x) \geq \|\nabla f(x)\|(1 - \|w(x) - v\|) \geq \|\nabla f(x)\|(1 - d^*), \tag{4.14}$$

which implies that for all $x \in \tilde{K}$, $h \in (0, r_3]$, and $v \in S^{p-1} \cap \overline{B}_{d^*}(w(x))$

$$\nabla_v^h f(x) \geq \nabla_v f(x) - \underline{\nu} d^* \geq \|\nabla f(x)\|(1 - 2d^*). \tag{4.15}$$

On the other hand, by definition of d^*, we have that, for all $x \in \tilde{K}$, and $v \in S^{p-1} \setminus \overline{B}_{r^*}(w(x))$,

$$\nabla_v f(x) \leq (1 - 5d^*)\|\nabla f(x)\|,$$

which implies that, for all $x \in \tilde{K}$, $h \in (0, r_3]$, and $v \in S^{p-1} \setminus \overline{B}_{r^*}(w(x))$,

$$\nabla_v^h f(x) \leq \nabla_v f(x) + \underline{\nu} d^* \leq \|\nabla f(x)\|(1 - 4d^*).$$

Now, let $r_2 \coloneqq \min(r_3, d^*) < r^*$ and $0 < r \leq r_2$. Notice that $(K)^{+r} \subset (K)^{+r^*} \subset (K)^{+r(\tilde{K})} \subset S_f$. Using Lemma 3.2 (ii) with $K = \tilde{K}$ and $h^* = r$, we choose $n_2 \in \mathbb{N}$ such that for all $n \geq n_2$, with probability at least $1 - \eta/2$,

$$\sup_{h \in [h_n, r]} \sup_{v \in S^{p-1}} \sup_{x \in \tilde{K}} |\nabla_v^h f_{\tau_n, n}(x) - \nabla_v^h f(x)| < d^* \underline{\nu}. \tag{4.16}$$

It follows from (4.15), (4.14), and (4.16) that, with probability at least $1 - \eta/2$, for all $x \in \tilde{K}$, $h \in [h_n, r]$, and $v \in S^{p-1} \cap \overline{B}_r(w(x))$

$$\nabla_v^h f_{\tau_n, n}(x) > (1 - 2d^*)\|\nabla f(x)\| - d^* \underline{\nu} \geq \tilde{d}(x), \tag{4.17}$$

and, for all $x \in \tilde{K}$, $h \in [h_n, r]$, and $v \in S^{p-1} \setminus \overline{B}_{r^*}(w(x))$,

$$\nabla_v^h f_{\tau_n, n}(x) < (1 - 4d^*)\|\nabla f(x)\| + d^* \underline{\nu} \leq \tilde{d}(x). \tag{4.18}$$

We show in **Step 3** below that there exists $n_1 \in \mathbb{N}$ such that, with probability at least $1 - \eta/2$, for all $x \in \tilde{K}$ and $n \geq n_1$ there exists $X \in \mathcal{X}_{n,r}(x)$ such that

$$(h_{X,x}, v_{X,x}) \in [h_n, r] \times (S^{p-1} \cap \overline{B}_r(w(x))). \tag{4.19}$$

In particular, (4.16) and (4.19) hold simultaneously with probability at least $1-\eta$. It follows from (4.17) and (4.18) that, with probability at least $1-\eta$, for all $x \in \tilde{K}$ and $n \geq \max(n_1, n_2)$

$$\sup_{(h,v)\in[h_n,r]\times S^{p-1}\setminus\overline{B}_{r^*}(w(x))} \nabla_v^h f_{\tau_n,n}(x) \leq \tilde{d}(x) < \nabla_{v_{X,x}}^{h_{X,x}} f_{\tau_n,n}(x).$$

Thus, we have shown that the finite difference approximation of $f_{\tau_n,n}(\cdot)$ with step $h_{X,x}$ and direction $v_{X,x}$ is larger than all finite difference approximations with step $h \in [h_n, r]$ and directions $v \in S^{p-1} \setminus \overline{B}_{r^*}(w(x))$. Since $\tilde{d}(x) > 0$, (4.12) and (4.13) follow.

Step 3. We show (4.19). To this end, let $0 < s_1 < s_2 < r < 1$ and $n_3 \in \mathbb{N}$ be such that $h_n < s_1$ for all $n \geq n_3$. It is enough to show that there exists $n_1 \geq n_3$ such that, for all $n \geq n_1$,

$$P^{\otimes n}([\mathcal{X}_n \cap D_{s_1,s_2}(x) \neq \emptyset \; \forall x \in \tilde{K}]) \geq 1 - \eta/2,$$

where $D_{s_1,s_2}(x) \coloneqq A_{s_1,s_2}(x) \cap C_{s_2}(x)$, $A_{s_1,s_2}(x) \coloneqq \overline{B}_{s_2}(x) \setminus B_{s_1}(x)$, and

$$C_{s_2}(x) \coloneqq \left\{ y \in \mathbb{R}^p \setminus \{x\} : \| \frac{y-x}{\|y-x\|} - w(x) \| \leq s_2 \right\}.$$

Let $0 < \epsilon_1 < \frac{s_2 - s_1}{2}$. We first notice that

$$A_{s_1+\epsilon_1, s_2-\epsilon_1}(x) \subset \cap_{z \in B_{\epsilon_1}(x)} A_{s_1,s_2}(z). \tag{4.20}$$

Indeed, $y \in A_{s_1+\epsilon_1,s_2-\epsilon_1}(x)$ satisfies $s_1 + \epsilon_1 \leq \|y-x\| \leq s_2 - \epsilon_1$. Therefore, for all $z \in B_{\epsilon_1}(x)$, it holds that

$$s_1 \leq \|y-x\| - \|x-z\| \leq \|y-z\| \leq \|y-x\| + \|x-z\| \leq s_2,$$

that is, $y \in A_{s_1,s_2}(z)$. Now, let $h^* > 0$ such that $(\tilde{K})^{+h^*}$ does not contain stationary points of $f(\cdot)$. Since $w(\cdot)$ is uniformly continuous in $(\tilde{K})^{+h^*}$, there exists $\epsilon_2 \in (0, h^*]$ such that, for all $x \in \tilde{K}$,

$$\sup_{y \in B_{\epsilon_2}(x)} \|w(x) - w(y)\| \leq \epsilon_1/2. \tag{4.21}$$

Suppose w.l.o.g. that $\epsilon_2 \leq \frac{s_1+\epsilon_1}{4}\epsilon_1$. We show that

$$D_{s_1+\epsilon_1, s_2-\epsilon_1}(x) \subset \cap_{z \in B_{\epsilon_2}(x)} D_{s_1,s_2}(z). \tag{4.22}$$

To this end, let $y \in D_{s_1+\epsilon_1,s_2-\epsilon_1}(z)$. By (4.20), it holds that $y \in \cap_{z\in B_{\epsilon_2}(x)} A_{s_1,s_2}(z)$. We need to show that $y \in \cap_{z\in B_{\epsilon_2}(x)} C_{s_2}(x)$. Since for all $z \in B_{\epsilon_2}(x)$

$$\| \frac{y-z}{\|y-z\|} - \frac{y-x}{\|y-x\|} \| \leq 2\frac{\|z-x\|}{\|y-z\|} \leq \frac{2\epsilon_2}{s_1+\epsilon_1} \leq \epsilon_1/2,$$

using the triangle inequality and (4.21), we have that

$$\| \frac{y-z}{\|y-z\|} - w(z) \| \leq \| \frac{y-z}{\|y-z\|} - \frac{y-x}{\|y-x\|} \| + \| \frac{y-x}{\|y-x\|} - w(x) \| + \|w(x) - w(z)\| \leq s_2.$$

(4.22) follows. Notice that, for all $x \in \tilde{K}$,

$$\lambda(D_{s_1+\epsilon_1,s_2-\epsilon_1}(x)) = \lambda(D_{s_1+\epsilon_1,s_2-\epsilon_1}(0)) =: \tilde{\Lambda} > 0.$$

Now, by the compactness of $\tilde{K} \subset \cup_{x\in\tilde{K}} B_{\epsilon_2}(x)$, there exist $x_1, \dots, x_q \in \tilde{K}$ such that $\tilde{K} \subset \cup_{l=1}^{q} B_{\epsilon_2}(x_l)$. It follows from (4.22) that, for all $z \in \tilde{K}$, there exists x_l such that $z \in B_{\epsilon_2}(x_l)$ and $D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) \subset D_{s_1,s_2}(z)$. Therefore, it is enough to show that there exists $n_1 \geq n_3$ such that for all $n \geq n_1$

$$P^{\otimes n}([\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) \neq \emptyset \ \forall l \in \{1,\dots,q\}]) \geq 1 - \eta/2.$$

To this end, notice that $\cup_{l=1}^{q} D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) \subset (\tilde{K})^{+r} \subset S_f$ and let $\underline{\alpha} := \min_{y\in(\tilde{K})^{+r}} f(y)$. Then, $p_l := P(D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l)) \geq \underline{\alpha}\tilde{\Lambda} > 0$. Observe that

$$\begin{aligned} &P^{\otimes n}(\cap_{l=1}^{q}[\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) \neq \emptyset]) \\ =&1 - P^{\otimes n}(\cup_{l=1}^{q}[\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) = \emptyset]) \\ \geq&1 - \sum_{l=1}^{q} P^{\otimes n}([\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) = \emptyset]). \end{aligned}$$

Let G_l have the geometric distribution with parameter p_l. Since $\{X_l\}$ are independent, it holds that

$$P^{\otimes n}([\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) = \emptyset]) = P(G_l > n) = \sum_{j=n}^{\infty}(1-p_l)^j p_l = (1-p_l)^n,$$

which implies that

$$P^{\otimes n}(\cap_{l=1}^{q}[\mathcal{X}_n \cap D_{s_1+\epsilon_1,s_2-\epsilon_1}(x_l) \neq \emptyset]) \geq 1 - \sum_{l=1}^{q}(1-p_l)^n \geq 1 - q(1-\underline{\alpha}\tilde{\Lambda})^n. \quad (4.23)$$

The statement follows by taking $n_1 \geq n_3$ such that $\eta/2 \geq q(1-\underline{\alpha}\tilde{\Lambda})^{n_1}$.

Step 4. Let $x \in A_{\overline{\alpha},\xi,i}$ and $n \geq n^* := \max(n_1, n_2)$. Notice that, by **Step 2**, $\{Y_{n,r,j}\}_{j=0}^{j^*} \in \mathcal{G}_{x,r}$ with probability at least $1-\eta$. Since, $r \leq r^* \leq r_1 \leq \delta$, by **Step 1**, either (i) $\{Y_{n,r,j}\}_{j=0}^{j^*}$ remains in $(G_x)^{+\delta} \setminus B_\delta(m_i)$ or (ii) $Y_{n,r^*,j} \in \overline{B}_{4\delta}(m_i)$ for some $j \in \{0,\dots,j^*\}$. We show that (i) is not possible. Indeed, if $Y_{n,r,j^*} \in (G_x)^{+\delta} \setminus B_\delta(m_i) \subset \tilde{K}_i$, then, by (4.12), there exists $X^*(Y_{n,r,j^*}) \in \mathcal{X}_{n,r}(Y_{n,r,j^*})$ such that $f_{\tau_n,n}(X^*(Y_{n,r,j^*})) > f_{\tau_n,n}(Y_{n,r,j^*})$. However, since j^* is the last iterate by (3.6) it holds that $f_{\tau_n,n}(Y_{n,r,j^*}) \geq \max_{X\in\mathcal{X}_{n,r}(Y_{n,r,j^*})\cup\{Y_{n,r,j^*}\}} f_{\tau_n,n}(X)$. Let $j_0 = \min\{j \in \{0,\dots,j^*\} : Y_{n,r,j} \in \overline{B}_{4\delta}(m_i)\}$. By (4.4), $Y_{n,r,j_0} \in R^{\alpha_i} \cap C(m_i)$. We show by induction that $Y_{n,r,j} \in R^{\alpha_i} \cap C(m_i)$ for all $j_0 \leq j \leq j^*$. Since $R^{\alpha_i} \cap C(m_i) \subset B_\epsilon(m_i)$ the statement follows. First, notice that, if $Y_{n,r,j} \in R^{\tilde{\alpha}_i} \cap C(m_i)$, then, using (4.3) and (4.4), $Y_{n,r,j+1} \in B_{\delta+r}(m_i) \subset B_{2\delta}(m_i) \subset R^{\alpha_i} \cap C(m_i)$. Second, if $Y_{n,r,j} \in B_{4\delta}(m_i) \setminus (R^{\tilde{\alpha}_i} \cap C(m_i))$, then by (4.5) $Y_{n,r,j} \in \tilde{K}_i$ and by (4.11) it holds that $f(Y_{n,r,j+1}) > f(Y_{n,r,j})$. Using the induction hypothesis, we conclude that $Y_{n,r,j+1} \in R^{\alpha_i} \cap C(m_i)$. ∎

The proof of Propositions 3.1 and 3.2 is based on the following lemma. We recall that $A_{\overline{\alpha},\xi,i} = R^{\overline{\alpha}} \cap (C(m_i))^{-\xi})$.

Lemma 4.1 *Assume the conditions in Proposition 3.1. Let* $\tilde{m}_{n,r,i} \coloneqq L_{n,r}(z)$ *for some* $r > 0$ *and* $z \in A_{\overline{\alpha},\xi,i}$ *and* $\tilde{C}_{n,r,\delta,i} \coloneqq \cup_{L\in[\tilde{m}_{n,r,i}]_\delta}\{x \in \mathbb{R}^p : L_{n,r}(x) = L\}$. *Given* $\eta > 0$, $0 < \delta \le \delta^*$, *and* $0 < r \le r^*$ *there exist* $n^* \in \mathbb{N}$ *such that* $P^{\otimes n}(E_{n,r,\delta}) \ge 1 - \eta$ *for all* $n \ge n^*$, *where* $E_{n,r,\delta} \coloneqq \{\tilde{C}_{n,r,\delta,i}$ *are distinct and* $A_{\overline{\alpha},\xi,i} \subset \tilde{C}_{n,r,\delta,i}\}$.

Proof of Lemma 4.1. Let $\delta^* > 0$ such that $\overline{B}_{2\delta^*}(m_i) \subset A_{\overline{\alpha},\xi,i}$ for all $i = 1,\dots,M$. Using Theorem 3.4, there are $0 < \epsilon \le \frac{\delta}{2} \le \frac{\delta^*}{2}$, $0 < r \le r^*$, and $n^* \in \mathbb{N}$ such that, with probability at least $1-\eta$, $L_{n,r}(x) \in B_\epsilon(m_i)$ for all $x \in A_{\overline{\alpha},\xi,i}$ and $n \ge n^*$. Since $A_{\overline{\alpha},\xi,i} \supset \overline{B}_{2\delta}(m_i) \setminus \overline{B}_\epsilon(m_i)$ and $\|y - z\| \le 2\epsilon \le \delta$ for all $y, z \in \overline{B}_\epsilon(m_i)$, we obtain that

$$\cup_{x\in A_{\overline{\alpha},\xi,i}}\{L_{n,r}(x)\} = [\tilde{m}_{n,r,i}]_\delta \subset \overline{B}_\epsilon(m_i).$$

As $\overline{B}_\epsilon(m_i)$ are disjoint, $[\tilde{m}_{n,r,i}]_\delta \cap [\tilde{m}_{n,r,j}]_\delta = \emptyset$ for $i \ne j$. It follows that $\tilde{C}_{n,r,\delta,i}$ are distinct and $A_{\overline{\alpha},\xi,i} \subset \tilde{C}_{n,r,\delta,i}$. ∎

It follows from the above lemma that, on $E_{n,r,\delta}$, $\tilde{C}_{n,r,\delta,i}$ $i = 1,\dots,M$ are empirical clusters. In particular, $M_n \ge M$ on $E_{n,r,\delta}$.

Proof of Proposition 3.1. Fix $\epsilon, \eta > 0$. Using (3.3) and $P(C(\mu_l)) = 0$ for all $l = 1,\dots,L$, let $0 < \overline{\alpha} < \min_{i=1,\dots,M} f(m_i)$ and $\xi > 0$ such that

$$P(\cup_{i=1}^M A_{\overline{\alpha},\xi,i}) \ge 1 - \frac{2\epsilon}{1 + \max(1,c)}. \tag{4.24}$$

Using Lemma 4.1, let $0 < \delta \le \delta^*$, $0 < r \le r^*$, and n^* such that

$$P^{\otimes n}(E_{n,r,\delta}) \ge 1 - \eta \text{ for all } n \ge n^*. \tag{4.25}$$

Suppose w.l.o.g. that $C_{n,r,\delta,i} = \tilde{C}_{n,r,\delta,i}$ $i = 1,\dots,M$ on $E_{n,r,\delta}$. On the event $E_{n,r,\delta}$ it holds that

$$2d_{P,c}(\mathcal{C}, \mathcal{C}_{n,r,\delta}) \le \sum_{i=1}^M P^{\otimes n}(C(m_i)\Delta C_{n,r,\delta,i}) + c\sum_{i=M+1}^{M_n} P^{\otimes n}(C_{n,r,\delta,i}).$$

Since $\{C(m_i)\}_{i=1}^M$ and $\{C_{n,r,\delta,i)}\}_{i=1}^{M_n}$ are disjoint and $A_{\overline{\alpha},\xi,i} \subset C_{n,r,\delta,i}$ on $E_{n,r,\delta}$, $2d_{P,c}(\mathcal{C}, \mathcal{C}_{n,r,\delta})$ is bounded above by

$$P(\mathbb{R}^p \setminus (\cup_{i=1}^M A_{\overline{\alpha},\xi,i}) + \max(1,c)P(\mathbb{R}^p \setminus (\cup_{i=1}^M A_{\overline{\alpha},\xi,i})).$$

Using (4.24) we conclude that $d_{P,c}(\mathcal{C}, \mathcal{C}_{n,r,\delta}) \le \epsilon$ on $E_{n,r,\delta}$ for all $n \ge n^*$. ∎

Proof of Proposition 3.2. Fix $\epsilon, \eta > 0$ and let $0 < \overline{\alpha} < \min_{i=1,\dots,M} f(m_i)$ and $\xi > 0$ such that $P(\cup_{i=1}^M A_{\overline{\alpha},\xi,i}) \ge 1 - \epsilon/2$. Denote by $\hat{P}_n$ the empirical measure and recall that $C_{n,r,\delta,i} \in \mathcal{C}_{n,r,\delta}$ belongs to $\mathcal{C}_{n,r,\delta,\zeta}$ if and only if $\hat{P}_n(C_{n,r,\delta,i}) > \zeta$. We let $0 < \zeta < \zeta^* \coloneqq \min_{i=1,\dots,M} P(A_{\overline{\alpha},\xi,i}))$ and assume w.l.o.g. that $\epsilon/2 < \zeta$ yielding that $P(\mathbb{R}^p \setminus (\cup_{i=1}^M A_{\overline{\alpha},\xi,i})) < \zeta$. Let δ, δ^*, r, r^*, and n^* as in (4.25) and take $C_{n,r,\delta,i} = \tilde{C}_{n,r,\delta,i}$ $i = 1,\dots,M$ and $\tilde{\mathcal{C}}_{n,r,\delta} \coloneqq \{C_{n,r,\delta,1},\dots,C_{n,r,\delta,M}\}$ on $E_{n,r,\delta}$. On the event $E_{n,r,\delta}$ it holds

$$d_H(\mathcal{C}, \mathcal{C}_{n,r,\delta,\zeta}) \le d_H(\mathcal{C}, \tilde{\mathcal{C}}_{n,r,\delta}\}) + \mathbf{I}_{\cup_{i=1}^M\{\hat{P}_n(C_{n,r,\delta,i})\le\zeta\}\cup\cup_{i=M+1}^{M_n}\{\hat{P}_n(C_{n,r,\delta,i})>\zeta\}}. \tag{4.26}$$

Using $\cup_{i=M+1}^{M_n}\{\hat{P}_n(C_{n,r,\delta,i}) > \zeta\} \subset \{\hat{P}_n(\cup_{i=M+1}^{M_n} C_{n,r,\delta,i}) > \zeta\}$, $A_{\overline{\alpha},\xi,i} \subset \tilde{C}_{n,r,\delta,i}$, and $\cup_{i=M+1}^{M_n} C_{n,r,\delta,i} \subset \mathbb{R}^p \setminus (\cup_{i=1}^{M} A_{\overline{\alpha},\xi,i})$ on $E_{n,r,\delta}$, the indicator in (4.26) is bounded above by

$$\sum_{i=1}^{M} \mathbf{I}_{\{\hat{P}_n(A_{\overline{\alpha},\xi,i}) \le \zeta\}} + \mathbf{I}_{\{\hat{P}_n(\mathbb{R}^p \setminus (\cup_{i=1}^{M} A_{\overline{\alpha},\xi,i})) > \zeta\}}.$$

Since $\zeta < \zeta^*$ and $P(\mathbb{R}^p \setminus (\cup_{i=1}^{M} A_{\overline{\alpha},\xi,i})) < \zeta$, the law of large numbers yields

$$\lim_{n\to\infty} \mathbf{I}_{\cup_{i=1}^{M}\{\hat{P}_n(C_{n,r,\delta,i}) \le \zeta\} \cup \cup_{i=M+1}^{M_n}\{\hat{P}_n(C_{n,r,\delta,i}) > \zeta\}} = 0 \text{ a.s.} \tag{4.27}$$

Using again $A_{\overline{\alpha},\xi,i} \subset \tilde{C}_{n,r,\delta,i}$ on $E_{n,r,\delta}$ and $P(\mathbb{R}^p \setminus (\cup_{i=1}^{M} A_{\overline{\alpha},\xi,i})) \le \epsilon/2$, we obtain

$$d_H(\mathcal{C}, \tilde{\mathcal{C}}_{n,r,\delta}) \le \max_{i=1,\dots,M} P(C(m_i) \Delta C_{n,r,\delta,i}) \le \epsilon. \tag{4.28}$$

Using (4.27) and (4.28) in (4.26) we conclude that on the event $E_{n,\delta,r}$

$$\lim_{n\to\infty} d_H(\mathcal{C}, \mathcal{C}_{n,r,\delta,\zeta}) \le \epsilon.$$

■

5. Concluding remarks

In this paper, we developed the notions of local depth for general *Type A* DFs and established its analytic and statistical properties. Specifically, we established the uniform convergence of sample local depth and related asymptotic limit distribution in $\ell^\infty(T)$ spaces. These results are then used to derive new approaches to clustering, mode estimation, and upper level set estimation. Specifically, we developed a modal clustering approach (via a gradient system) where the density is replaced by the population approximation. Convergence results show that the approximated approach provides, in the limit, the same clusters as those given by the true density. In particular, we have shown that, our approximated approach correctly detects the true modes. We proposed an algorithm for the numerical computation of the clusters at sample level and established its consistency.

Acknowledgments

The authors thank the anonymous reviewers for careful reading of the manuscript and for valuable suggestions that improved the paper.

Supplementary Material

Appendices to: Analytical and statistical properties of local depth functions motivated by clustering applications
(doi: 10.1214/23-EJS2110SUPP; .pdf). The supplementary material consists of proofs of several results needed for proving the main results. Also, it contains additional technical results, a detailed description of the clustering algorithm, and extensive simulations and data analysis. For more details see also Francisci et al. (2022).

References

Agostinelli, C. and Romanazzi, M. (2011). Local depth. *Journal of Statistical Planning and Inference* **141** 817-830. MR2732952

Alexander, K. S. (1987). The central limit theorem for empirical processes on Vapnik-Cervonenkis classes. *The Annals of Probability* **15** 178-203. MR0877597

Arcones, M. A. (1995). A Bernstein-type inequality for U-statistics and U-processes. *Statistics & probability letters* **22** 239-247. MR1323145

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *The Annals of Probability* **21** 1494-1542. MR1235426

Bertrand-Retali, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode de noyau. *Revue Roumaine de Mathématiques Pures et Appliquées* **23** 361-385. MR0494658

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science* **30** 518–532. MR3432839

Chacón, J. E. (2019). Mixture Model Modal Clustering. *Adv. Data Anal. Classif.* **13** 379–404. MR3954514

Chandler, G. and Polonik, W. (2021). Multiscale geometric feature extraction for high-dimensional and non-Euclidean data with applications. *The Annals of Statistics* **49** 988–1010. MR4255116

Chazal, F., Guibas, L. J., Oudot, S. Y. and Skraba, P. (2013). Persistence-Based Clustering in Riemannian Manifolds. *Journal of the ACM* **60**. MR3144911

Dudley, R. M. (2014). *Uniform central limit theorems* **142**. Cambridge university press. MR3445285

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics* **33** 1380–1403. MR2195639

Elmore, R. T., Hettmansperger, T. P. and Xuan, F. (2006). Spherical data depth and a multivariate median. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72** 87. MR2343115

Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97** 611-631. MR1951635

Francisci, G., Agostinelli, C., Nieto-Reyes, A. and Vidyashankar, A. N. (2022). Analytical and statistical properties of local depth functions motivated by clustering applications. *arXiv preprint arXiv:2008.11957*. MR2673155

Francisci, G., Agostinelli, C., Nieto-Reyes, A. and Vidyashankar, A. N. (2023). Appendices to: Analytical and statistical properties of local depth functions motivated by clustering applications. *DOI: 10.1214/23-EJS2110SUPP*.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory* **21** 32–40. MR0388638

Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press. MR3588285

Hirsch, M. W., Devaney, R. L. and Smale, S. (1974). *Differential equations, dynamical systems, and linear algebra* **60**. Academic press. MR0486784

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.

Koontz, W. L. G., Narendra, P. M. and Fukunaga, K. (1976). A graph-theoretic approach to nonparametric cluster analysis. *Institute of Electrical and Electronics Engineers. Transactions on Computers* **C-25** 936–944. MR0489093

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics* **273**. Springer Science & Business Media. MR1472486

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics* **18** 405–414. MR1041400

Liu, Z. and Modarres, R. (2011). Lens data depth and median. *Journal of Nonparametric Statistics* **23** 1063–1074. MR2854255

Matsumoto, Y. (2002). *An introduction to Morse theory* **208**. American Mathematical Soc. MR1873233

Menardi, G. (2016). A review on modal clustering. *International Statistical Review* **84** 413–433. MR3580423

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66** 846–850.

Teschl, G. (2012). *Ordinary differential equations and dynamical systems* **140**. American Mathematical Society. MR2961944

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association* **88** 520–528. MR1224377

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97** 893–904. MR2746159

Yang, M. and Modarres, R. (2018). β-Skeleton depth functions and medians. *Communications in Statistics-Theory and Methods* **47** 5127–5143. MR3833887

Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *Annals of statistics* 461–482. MR1790005

Zuo, Y. and Serfling, R. (2000b). Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics* **28** 483–499. MR1790006