Electronic Journal of Statistics
Vol. 16 (2022) 5623–5658
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2073

Space partitioning and regression maxima seeking via a mean-shift-inspired algorithm*

Wanli Qiao

Department of Statistics, George Mason University
e-mail: wqiao@gmu.edu
and

Amarda Shehu

Department of Computer Science, George Mason University
e-mail: ashehu@gmu.edu

Abstract: The mean shift (MS) algorithm is a nonparametric method used to cluster sample points and find the local modes of kernel density estimates, using an idea based on iterative gradient ascent. In this paper we develop a mean-shift-inspired algorithm to estimate the maxima of regression functions and partition the sample points in the input space. We prove convergence of the sequences generated by the algorithm and derive the rates of convergence of the estimated local maxima for the underlying regression model. We also demonstrate the utility of the algorithm for data-enabled discovery through an application on biomolecular structure data.

Keywords and phrases: Gradient ascent, nonparametric regression derivative estimation, maxima hunting, spatial partitioning.

Received December 2021.

1. Introduction

The mean-shift (MS) algorithm is a well-known method to cluster sample points and find the local modes of kernel density estimators (KDE) using a gradient ascent idea. This algorithm was introduced by Fukunaga and Hostetler (1975), and was generalized by Cheng (1995). It finds wide applications in image segmentation (Comaniciu and Meer, 2002) and object tracking (Comaniciu et al., 2003). The algorithm has thus far no counterpart that partitions sample points in a regression setting and estimates the local maxima of regression functions. In this paper we propose a regression mean shift algorithm to fill this gap and study the theoretical properties of our maxima estimators for regression functions.

*This work is partially supported by grants NSF DMS 1821154 and NSF FET 1900061. This material is additionally based upon work by AS supported by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random pair, and $r(x) = \mathbb{E}(Y|X = x)$ be the regression function. Suppose that we observe i.i.d. sample points $(X_1, Y_1), \cdots, (X_n, Y_n)$ that have the same joint distribution as (X, Y). Here the goals are (1) to estimate the set of local maxima of r, and (2) to partition the input space or the points $\{X_1, \cdots, X_n\}$ according to their connection with the estimated local maxima. The plug-in method is a natural way of achieving goal (1). In other words, suppose that we have a good estimator r_n of r, we can use the local maxima of r_n as the estimators of the local maxima of r. In fact, this idea has been used in Müller (1985, 1989) using the Gasser-Müller (GM) kernel regression estimator, and in Ziegler (2002) using the Nadaraya-Watson (NW) regression estimator. However, the plug-in approach does not directly render an algorithm to find the local maxima, which is usually challenging because the local maxima are only implicitly defined through the regression estimators. Related to goal (2), partition-based regression methods include CART (Breiman et al., 1993), MARS (Friedman, 1991), and SUPPORT (Chaudhuri et al., 1994), among others. Different from the above works, the space partitioning idea in our approach is based on the geometric characteristics of regression functions. More specifically, we use the basins of attraction associated with the local maxima of regression functions to define the partition. Input space partitioning for regression functions has many applications, for example, clustering for house price (Liu et al., 2016), segregated homogeneous neighborhoods studied in sociology (Legewie, 2018), and division of disease risk zones in epidemiology (Gaudart et al., 2005). The regression MS algorithm we propose in this paper uses a modal clustering idea and is simultaneously useful for the above two goals.

We briefly describe the idea behind the original MS algorithm, in order to elucidate the main differences and challenges in extending the MS algorithm to the regression setting. Let f be a differentiable density function on $\mathbb{R}^d$. For a fixed $a > 0$, consider a sequence of points, starting from $x_0 \in \mathbb{R}^d$, defined iteratively by

$$x_\ell = x_{\ell-1} + a\frac{\nabla f(x_{\ell-1})}{f(x_{\ell-1})},\ \ell \geq 1. \tag{1.1}$$

Having f in the denominator quickly moves points in low-density regions to higher-density locations. Since $\nabla \log f(x) = \frac{\nabla f(x)}{f(x)}$ for all $x \in \mathbb{R}^d$ such that $f(x) > 0$, the procedure in (1.1) can be understood as a gradient ascent algorithm applied to $\log f$, with $x_\infty := \lim_{\ell\to\infty} x_\ell$, if it exists, as a local mode of f under regularity conditions. With a random sample drawn from f, one can get a KDE $\widehat{f}$ defined in (2.1) below, replace f by $\widehat{f}$ in the above iterative procedure and generate a sequence $\widehat{x}_j$, $j = 0, 1, \cdots$, with $\widehat{x}_0 = x_0$, so that $\widehat{x}_\infty := \lim_{\ell\to\infty} \widehat{x}_\ell$ is used as an estimate of x_∞. The MS algorithm implicitly uses $a \propto h^2$, where h is the bandwidth of the KDE, and groups starting points x_0 to the same cluster (i.e., basin of attraction) if their destination $\widehat{x}_\infty$ is the same. The gradient ascent nature of the MS algorithm has been studied in Arias-Castro et al. (2016). See Fig 1.1 for an illustration.

One appealing feature of the MS algorithm, which is perhaps also why it is so popular, is that, the convergence of the algorithm can be guaranteed under

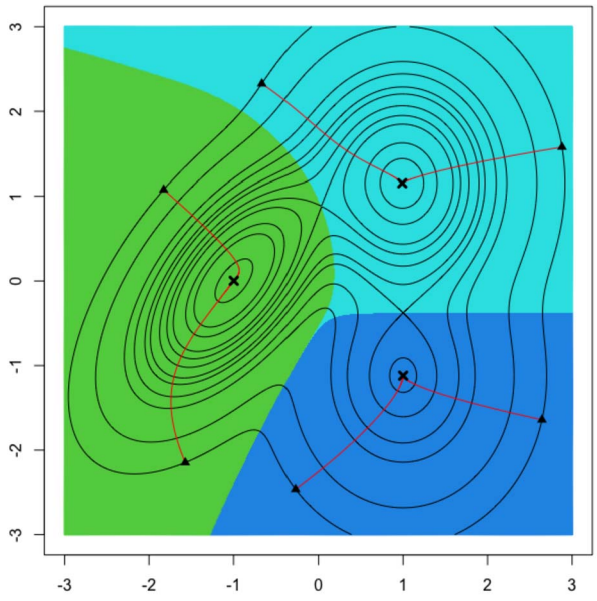

FIG 1.1. *Partition of domain for the trimodal function (K) in Wand and Jones (1993). The black curves are the contour lines of the function; the three local maxima are represented by* $\times$ *symbols; the red curves are gradient integral curves starting from the solid triangles; the domain of the function is partitioned into three parts represented by three different colors, using the modal clustering idea.*

some mild conditions when the kernel function is appropriately chosen. There is no requirement for the step length, i.e., the quantity a in (1.1), which is in fact implicitly determined by h in the MS algorithm. See Ghassabeh (2015), and Yamasaki and Tanaka (2020). When the MS algorithm is applied to modal clustering, the number of clusters does not need to pre-specified, but rather depends on the chosen bandwidth.

Ideally, the MS algorithm can be extended to a nonparametric regression setting, by replacing the density f in (1.1) with a regression function estimator r_n, in order to estimate the local maxima of the regression function r and their associated basins of attraction, which can then be used to naturally partition the input space. However, this extension does not appear straightforward, related to the following aspects:

(1) The regression function r and its estimator r_n are not always non-negative, so that it is not always meaningful to consider $\log r$ or $\log r_n$ directly, while it seems that the logarithm transformation plays a critical role in the convergence property of the MS algorithm without any requirement for the bandwidth choice;
(2) The regression function r has a quotient form as a conditional expectation. The regression estimators that adopt a similar form (such as the NW regression estimator) have more tedious gradients than those of KDE, which

makes the mean-shift implementation using such estimators no longer enjoy the same convergence property as the original MS-type algorithm. See Remark 2.1 below for more discussions.

Briefly speaking, we handle the above two issues in the following way: For (1), we apply a positive transformation to the observed response variable $Y_1, \cdots, Y_n$; For (2), with the transformed response variables we use a regression estimator developed in Mack and Müller (1989), which is a variant of the NW kernel estimator, but enjoys a simpler form of gradients (and higher order derivatives). With this equipment, here is a summary of our contributions in this paper.

1. We present a regression mean shift algorithm that is used to partition the sample points in the input space and estimate the local maxima of r. We prove the convergence of the algorithm under mild conditions, which does not have a requirement for the bandwidth (see Theorem 2.1).
2. We give uniform rates of convergence of the Hausdorff distance of the sets of local maxima between our estimator and the truth (see Theorems 3.2 and 3.3).

Related literature. For clustering and maxima estimation related to regression models, there is a MS-type algorithm called the conditional mean shift (CMS) algorithm, developed by Einbeck and Tutz (2006). The CMS algorithm is used to estimate the local modes of $f(y|x)$, which is the conditional density function of Y given $X = x$. The algorithm searches for local maxima in the space of y, with its output indexed by x, and has been used in nonparametric modal regression studied by Chen et al. (2016). Note that the CMS is still an algorithm of searching for the modes of (conditional) density functions, while the problem we are studying here is to find the local maxima estimators for the regression function $\mathbb{E}(Y|X = x)$ in the space of x. For this reason, we do not view CMS as a competitor of the regression mean shift algorithm studied in this paper. The input space partitioning idea using our regression mean shift can be interpreted based on the Morse theory (see Milnor, 1963), which is also used in the Morse-Smale regression developed in Gerber et al. (2013). Their method is specifically applied to k-nearest neighbor graphs. Estimating the gradient and critical points of regression functions can be also useful for variable selection (see Mukherjee and Zhou, 2006).

We organize the paper as follows. First we present our regression mean shift algorithm in Section 2 with its convergence proved. Section 3 includes theoretical study for the maxima estimators. It is followed by simulation and case studies in Section 4, where in particular we show the application of our algorithm to biomolecular structure datasets. The proofs are given in Section 6.

2. Regression mean shift algorithm

Denote the marginal probability density function of X by f, and let

$$\widehat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \ x \in \mathbb{R}^d \tag{2.1}$$

be the KDE of f, where K is a kernel density function on $\mathbb{R}^d$, and $h > 0$ is a bandwidth. Denote $K_h(x) = K(x/h)$. We have the following kernel regression estimator of $r(x)$, proposed by Mack and Müller (1989):

$$\widehat{r}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \frac{Y_i K_h(x - X_i)}{\widehat{f}(X_i)}. \tag{2.2}$$

Note that if we take derivatives of $\widehat{r}$, the differential operator only needs to be applied to the numerator, which helps avoid the tedious form of the derivatives of, say, the NW regression estimator.

Let K be a spherically symmetric kernel with profile $k : \mathbb{R}_{\geq 0} \to \mathbb{R}$, that is, $K(x) = c_{k,d} k(\|x\|^2)$, where $c_{k,d} > 0$ is a normalization factor such that $c_{k,d}^{-1} = \int_{\mathbb{R}^d} k(\|x\|^2) dx$. Examples of K include the Gaussian kernel and Epanechnikov kernel. Then we can write

$$\widehat{r}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} \frac{Y_i k(\|x - X_i\|^2/h^2)}{\widehat{f}(X_i)}. \tag{2.3}$$

We will transform $Y_i, i = 1, \cdots, n$ by applying a strictly increasing positive function $\xi : \mathbb{R} \to \mathbb{R}_{>0}$. We consider the following two choices of ξ.

- **T1** (Transformation 1): ξ is a deterministic bounded function. For example, $\xi(x) = \text{logistic(x)}$.
- **T2** (Transformation 2): ξ is a random function depending on $Y_{[n]} := \min_i Y_i$ such that $\xi(x) = x + \pi(Y_{[n]})$, where $\pi(Y_{[n]}) = (-Y_{[n]} + c_0)\mathbf{1}(Y_{[n]} < c_0)$ for some positive constant c_0. Note that $\min_i \xi(Y_i) \geq c_0$.

Let $\widetilde{Y} = \xi(Y)$ and $\widetilde{Y}_i = \xi(Y_i) > 0$, $i = 1, \cdots, n$. Define

$$\widehat{r}_*(x) = \widehat{r}_{*,k}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} \frac{\widetilde{Y}_i k(\|x - X_i\|^2/h^2)}{\widehat{f}(X_i)}, \tag{2.4}$$

which is considered as an estimator of $\widetilde{r}(x) := \mathbb{E}(\widetilde{Y}|X = x)$. Define $g(x) = -k'(x)$ for all $x \in [0, \infty)$, assuming that the derivative exists. For any $x \in \mathbb{R}^d$, denote

$$w_i(x) = \frac{g\left(\|x - X_i\|^2/h^2\right)}{\widehat{f}(X_i)}, \tag{2.5}$$

and define

$$\widehat{m}_*(x) = \frac{\sum_{i=1}^{n} w_i(x) \widetilde{Y}_i X_i}{\sum_{i=1}^{n} w_i(x) \widetilde{Y}_i} - x, \tag{2.6}$$

which is called the *regression mean shift*. Note that we have the following relation.

$$\nabla \widehat{r}_*(x) = \nabla \widehat{r}_{*,k}(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} \frac{-\widetilde{Y}_i (x - X_i) g(\|x - X_i\|^2/h^2)}{\widehat{f}(X_i)}$$

$$=\frac{2c_{k,d}}{h^2 c_{g,d}}\widehat{r}_{*,g}(x)\widehat{m}_*(x). \tag{2.7}$$

Here $\widehat{r}_{*,g}$ is defined as in (2.4) for $\widehat{r}_{*,k}$, where we replace k by g. In other words, the regression mean shift $\widehat{m}_*(x)$ is proportional to $\nabla\widehat{r}_{*,k}(x)/\widehat{r}_{*,g}(x)$ up to a constant coefficient, so that the following regression MS algorithm can be understood as a gradient ascent algorithm.

Regression MS algorithm Our regression MS algorithm is as follows. Let z_0 be a starting point in the domain of r (e.g., one of X_i's). Obtain $z_1, z_2, \cdots$, iteratively from

$$z_{j+1} = \widehat{m}_*(z_j) + z_j,\ j = 0, 1, 2, \cdots. \tag{2.8}$$

The limit of the sequence $\{z_0, z_1, \cdots\}$ is considered as an estimator of a local maximum of r. In practice, the algorithm stops when the distance between two consecutive points $\|z_{j+1} - z_j\|$ is less than a pre-specified small threshold.

Lemma 2.1. *If k is convex and strictly decreasing such that $-\infty < k'(x) < 0$ for all $x \geq 0$, then we have*

(1) $\widehat{r}_*(z_j)$ converges,
(2) $\|z_{j+1} - z_j\| \to 0$, and
(3) $\nabla\widehat{r}_*(z_j) \to 0$, as $j \to \infty$.

Remark 2.1. a). A related and alternative method is to use the discretized gradient ascent with a constant step length based on a smooth estimator of the regression function (such as the NW regression estimator). However, this method requires the step length to be chosen sufficiently small; otherwise it is well-known that there is an overshooting problem and the sequence can diverge (see Bertsekas, 1999, Chapter 1). In contrast, the convergence of our regression MS algorithm does not rely on requirements for step length.
b). It is natural to wonder if a MS-type algorithm can be developed based on the gradient of the NW regression estimator, in a way similar to (2.7). The analysis below shows such an algorithm is not effective in general. The NW regression estimator using $\{(X_i, \widetilde{Y}_i), i = 1, \cdots, n\}$ is given by

$$\widehat{r}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n \widetilde{Y}_i k(\|x - X_i\|^2/h^2)}{\sum_{i=1}^n k(\|x - X_i\|^2/h^2)}. \tag{2.9}$$

Let $w_i^k(x) = k(\|x - X_i\|^2/h^2)$ and $w_i^g(x) = g(\|x - X_i\|^2/h^2)$. Let

$$w_i^*(x) = \widetilde{Y}_i w_i^g(x)\Big[\sum_{i=1}^n w_i^k(x)\Big] - w_i^g(x)\Big[\sum_{i=1}^n \widetilde{Y}_i w_i^k(x)\Big].$$

It follows from a straightforward calculation that

$$\nabla\widehat{r}_{\mathrm{NW}}(x) = \frac{2}{h^2}\frac{\sum_{i=1}^n w_i^*(x)X_i - \sum_{i=1}^n w_i^*(x)x}{[\sum_{i=1}^n w_i^k(x)]^2}. \tag{2.10}$$

If $g \propto k$ (which happens, for example, when K is the Gaussian kernel), then $\sum_{i=1}^n w_i^*(x) \equiv 0$, which makes it hopeless to get a mean shift form as given in (2.6). Now suppose a kernel can be chosen such that $\sum_{i=1}^n w_i^*(x) \neq 0$. Then we can write

$$\nabla \widehat{r}_{\mathrm{NW}}(x) = \frac{2}{h^2} \widehat{s}(x) \widehat{m}_{\mathrm{NW}}(x). \tag{2.11}$$

where

$$\widehat{s}(x) = \frac{\sum_{i=1}^n w_i^*(x)}{[\sum_{i=1}^n w_i^k(x)]^2} \text{ and } \widehat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n w_i^*(x) X_i}{\sum_{i=1}^n w_i^*(x)} - x,$$

the latter corresponding to our regression mean shift. Corresponding to (2.8), the mean shift algorithm using $\widehat{m}_{\mathrm{NW}}$ and starting from z_0 is given by

$$z_{j+1} = \widehat{m}_{\mathrm{NW}}(z_j) + z_j = \frac{\sum_{i=1}^n w_i^*(z_j) X_i}{\sum_{i=1}^n w_i^*(z_j)}, \ j = 0, 1, 2, \cdots. \tag{2.12}$$

Note that in general the sign of w_i^* is not always positive, and hence it is not clear if a similar result as given in Lemma 2.1 holds for $\widehat{r}_{\mathrm{NW}}$. In fact, a simulation we ran shows the converge of the sequence generated by $\widehat{m}_{\mathrm{NW}}$ using the Epanechnikov kernel for g is problematic. See Section 4.2. It appears that the mean shift idea and the quotient form of the NW regression estimator are not compatible.

We need to additionally assume that the critical points of $\widehat{r}_*$ are isolated, in order to have the convergence of our regression mean shift algorithm. The proof of the following theorem is similar to that of Theorem 1 in Ghassabeh (2015) for the MS algorithm.

Theorem 2.1. *Suppose that the assumptions in Lemma 2.1 hold. If the critical points of $\widehat{r}_*$ are isolated, then the sequence of z_j converges to one of the critical points of $\widehat{r}_*$ as $j \to \infty$.*

2.1. Basins of attraction for regression functions

The maxima seeking algorithm in (2.8) can be used to partition the input space into basins of attraction. This can be understood using the framework of Morse theory (see Milnor, 1963). A similar perspective has been used to interpret modal clustering using the MS algorithm. See Chácon (2015). Suppose that $\mathcal{X}$ is a compact set of positive volume contained in the support of the density of X. Also suppose that r is a twice differentiable Morse function, meaning that all of its critical points are non-degenerate, that is, the Hessian at each critical point is nonsingular. Let $\mathcal{M}$ be the collection of all local maxima of r, denoted by $x_1, \cdots, x_m$, where m is the cardinality of $\mathcal{M}$. For any $x \in \mathcal{X}$, let $\phi_x : \mathbb{R} \to \mathcal{X}$ be the integral curve driven by the gradient of r, starting from x:

$$\frac{d\phi_x(t)}{dt} = \nabla r(\phi_x(t)), t \in \mathbb{R}; \ \phi_x(0) = x.$$

Then by the Morse theory, for any $x \in \mathcal{X}$, $\phi_x(\infty) := \lim_{t\to\infty} \phi_x(t)$ is one of the critical points of r. In particular, for $j = 1, \cdots, m$, the basins of attraction associated with x_j is

$$C(x_j) := \{x \in \mathcal{X} : \ \phi_x(\infty) = x_j\},$$

which is also called a stable manifold, or ascending manifold in Morse theory. The sets in $\mathcal{C} := \{C(x_j), j = 1, \cdots, m\}$ are disjoint, and their union covers $\mathcal{X}$ except for a set of zero Lebesgue measure.

Let us first consider the deterministic transformation ξ in **T1**. Under regularity conditions, one can show that r and $\widetilde{r}$ have the same ascending manifolds (see Lemma 3.1 below). The regression estimator $\widehat{r}_*$ is used to estimate $\widetilde{r}$, and the sequence (2.8) is viewed as discretized estimation of trajectories of the integral curves driven by $\nabla \log \widetilde{r}$. Let $\widehat{\mathcal{M}}$ be the set of all local maxima of $\widehat{r}_*$, consisting of $\widehat{x}_1, \cdots, \widehat{x}_{\widehat{m}}$, where $\widehat{m}$ is its cardinality. For any $x \in \mathcal{X}$, let $\widehat{\phi}_x(\infty)$ be the limit of the sequence in (2.8) when $z_0 = x$. Define

$$\widehat{C}(\widehat{x}_j) := \{x \in \mathcal{X} : \ \widehat{\phi}_x(\infty) = \widehat{x}_j\}.$$

Then $\widehat{\mathcal{C}} := \{\widehat{C}(\widehat{x}_j) : \ j = 1, \cdots, \widehat{m}\}$ also gives a partition of $\mathcal{X}$ (up to a small set not covered), and can be used to estimate $\mathcal{C}$.

For the transformation ξ given in **T2**, the idea is similar. Using this transformation, the regression estimator $\widehat{r}_*$ is used to estimate

$$\bar{r} := r + \pi(Y_{[n]}). \tag{2.13}$$

Notice that $\bar{r}$ and r has the same ascending manifolds, assuming that $\pi(Y_{[n]})$ is bounded. So again $\widehat{\mathcal{C}}$ gives an approximate partition of $\mathcal{X}$ and can be used to estimate $\mathcal{C}$.

For both transformations, the sample points $X_1, \cdots, X_n$ in the input space can be partitioned based on which basins of attraction they belong to, and this idea is used in the simulation and case studies in Section 4.

3. Theoretical analysis of the maxima estimators

In this section we study the theoretical properties of $\widehat{r}_*$ and its maxima as direct plug-in estimators of the maxima of r. We derive the uniform rate of convergence of $\widehat{r}_*$, which further gives the rate of convergence of its local maxima in Hausdorff distance. The derived rates of convergence for the local maxima estimation match the minimax rate of mode estimation for density functions up to a logarithm factor (see Remark 3.3).

We will use the following notation. For any d-tuple $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}^d$, let $|\alpha| = \alpha_1 + \cdots + \alpha_d$. For an $|\alpha|$ times differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, denote $\partial^\alpha g(x) = \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} g(x), \ x \in \mathbb{R}^d$. For a composition of functions $g_1 \circ g_2$, we write $\partial_x^\alpha g_1(g_2(x)) = \partial^\alpha (g_1 \circ g_2)(x)$. Let ∇g and $\nabla^2 g$ be the gradient and Hessian of g, respectively. For any real numbers a, b, let $a \wedge b = \min(a, b)$ and

$a \vee b = \max(a, b)$. For simplicity of notation, for any $n \geq 1$, $h > 0$, $j \in \mathbb{Z}_+$, we denote $\gamma_{n,h}^{(j)} = (nh^{d+2j})^{-1/2}$.

Throughout the paper $\mathcal{X}$ denotes a compact subset of $\mathbb{R}^d$ with strictly positive volume. For any $\delta > 0$, let $\mathcal{X}^\delta = \{x \in \mathbb{R}^d : \inf_{t \in \mathcal{X}} \|x - t\| \leq \delta\}$. We will use the following assumptions in our theoretical analysis.

Assumption **A1**: The marginal density f of X satisfies $\inf_{x \in \mathcal{X}} f(x) \geq \varepsilon_0$ for a constant $\varepsilon_0 > 0$.

Assumption **A2**: f has three times continuous bounded derivatives on $\mathcal{X}^\delta$ for some $\delta > 0$.

Assumption **A3**: r has three times continuous bounded derivatives on $\mathcal{X}^\delta$ for some $\delta > 0$.

Assumption **K**: The kernel K is a spherically symmetric density function with its support contained in the unit ball of $\mathbb{R}^d$. K has three times continuous bounded derivatives on $\mathbb{R}^d$.

3.1. Transformation 1

We first consider the transformation ξ in **T1**. We can write the regression model as follows:

$$Y_i = r(X_i) + \epsilon_i, \ i = 1, \cdots, n, \tag{3.1}$$

where $\epsilon_i, i = 1, \cdots, n$, are i.i.d random errors with mean zero. Note that this is the regression model before any transformation is applied. We make the following assumptions.

Assumption **E**: For $i = 1, \cdots, n$, each ϵ_i is independent of X_i.

Assumption **T**: ξ is a strictly increasing function on $\mathbb{R}$ with three times continuous bounded derivatives. Assume that there exist constants $0 < C_\ell < C_u < \infty$ such that $\xi(\mathbb{R}) \subset [C_\ell, C_u]$.

For a twice differential function g, the index of a critical point x_{crit} of g is the number of negative eigenvalues of $\nabla^2 g$ at x_{crit}. We first show that the critical points (including the local maxima) of $\widetilde{r}(x)$ and $r(x)$ are the same under the above conditions.

Lemma 3.1. *Assume that r is twice differentiable. For ξ in **T1**, under the assumptions **E** and **T**, the critical points of $\widetilde{r}$ and r are the same with the same indices. If r is a Morse function, then (1) $\widetilde{r}$ is also a Morse function, and (2) the ascending manifolds of r and $\widetilde{r}$ are the same.*

Remark 3.1. The above lemma implies that the local maxima of r can be estimated by the local maxima of $\widehat{r}_*$ as a plug-in approach, because $\widehat{r}_*$ is considered as an estimator of $\widetilde{r}$ as shown in Theorems 3.1 below.

In the following theorem we give the uniform rate of convergence for the difference between $\partial^\alpha \widetilde{r}$ and $\partial^\alpha \widehat{r}_*$ for all $|\alpha| \leq 2$.

Theorem 3.1. *For ξ in* **T1**, *under assumptions* **A1-A3**, **K**, **E**, *and* **T**, *there exist constants $C > 0$, $c > 0$ and $h_0 > 0$ such that for all $|\alpha| \le 2$, $n \ge 1$, $0 < h \le h_0$, $\tau > 1$ satisfying $nh^d \ge c(\tau \vee |\log h|)$ we have*

$$\mathbb{P}^n\Big(\sup_{x\in\mathcal{X}} |\partial^\alpha \widehat{r}_*(x) - \partial^\alpha \widetilde{r}(x)| \le C\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(|\alpha|)} + Ch^{(3-|\alpha|)\wedge 2}\Big) \ge 1 - 3e^{-\tau}. \tag{3.2}$$

Let $\lambda_1(x)$ be the largest eigenvalue of $\nabla^2 r(x), x \in \mathcal{X}$. We can write the set of local maxima of r as $\mathcal{M} = \{x \in \mathcal{X} : \nabla r(x) = 0,\ \lambda_1(x) < 0\}$, which is assumed to be nonempty. Let $\widehat{\mathcal{M}}$ be the set of local maxima of $\widehat{r}_*$. For any two subset $A, B \subset \mathbb{R}^d$, their Hausdorff distance is defined as

$$d_H(A, B) = \max\Big\{\sup_{a\in A}\inf_{b\in B}\|a - b\|, \sup_{b\in B}\inf_{a\in A}\|a - b\|\Big\}.$$

To study $d_H(\mathcal{M}, \widehat{\mathcal{M}})$, we will use the following perturbation result for the set of local maxima.

Lemma 3.2. *Let $\mathcal{R}$ be a compact subset of $\mathbb{R}^d$ with positive volume, $\partial\mathcal{R}$ be its boundary, and $\mathcal{U} \supset \mathcal{R}$ be an open subset of $\mathbb{R}^d$. Suppose that $p : \mathcal{U} \to \mathbb{R}$ be a three times continuously differentiable Morse function. Let $\lambda_1(x)$ be the largest eigenvalue of $\nabla^2 p(x)$, $x \in \mathcal{R}$, and*

$$\mathcal{M} = \{x \in \mathcal{R} : \nabla p(x) = 0, \lambda_1(x) < 0)\} \text{ and } \mathcal{C} = \{x \in \mathcal{R} : \nabla p(x) = 0\},$$

be the sets of local maxima and all critical points, respectively, of p on $\mathcal{R}$. Assume that $\eta := \inf_{x\in\mathcal{C}} d(x, \partial\mathcal{R}) > 0$. Let $\widetilde{p} : \mathcal{U} \to \mathbb{R}$ be a twice differentiable function, and $\widetilde{\mathcal{M}}$ be the set of local maxima of $\widetilde{p}$ on $\mathcal{R}$. There exists a constant $c_0 > 0$ such that if

$$\sup_{x\in\mathcal{R}}\max_{|\alpha|\le 2} |\partial^\alpha p(x) - \partial^\alpha \widetilde{p}(x)| < c_0,$$

then $\widetilde{p}$ has the same number of local maxima as p on $\mathcal{R}$, and

$$d_H(\mathcal{M}, \widetilde{\mathcal{M}}) \le \frac{4}{\lambda_*}\max_{x\in\mathcal{M}} \|\nabla\widetilde{p}(x) - \nabla p(x)\|,$$

where $\lambda_* := -\inf_{x\in\mathcal{M}} \lambda_1(x) > 0$.

Remark 3.2. The Hausdorff distance between the sets of maxima of the true and estimated functions is also studied in Chen et al. (2016, Theorem 1). As a comparison, our result is given under weaker conditions. In particular, we do not require their assumption (M2), which assumes that there exist $\eta_1 > 0$ and $C_3 > 0$ such that $\{x : \|\nabla p(x)\| \le \eta_1, 0 > -\lambda_*/2 \ge \lambda_1(x)\} \subset \mathcal{M}^{\lambda_*/(2dC_3)}$.

The following theorem gives an upper bound for $d_H(\mathcal{M}, \widehat{\mathcal{M}})$, as a direct consequence of Lemma 3.1, Theorem 3.1, and Lemma 3.2.

Theorem 3.2. *For ξ in* **T1**, *under assumptions* **A1-A3**, **K**, **E**, *and* **T**, *if r is a Morse function and there are no critical points on the boundary of $\mathcal{X}$, then there exist constants $C > 0$, $c > 0$, and $h_0 > 0$ such that for all $n \geq 1$, $0 < h \leq h_0$, $\tau > 1$ satisfying $nh^{d+4} \geq c(\tau \vee |\log h|)$ we have with probability at least $1 - 2(d+2)^2 e^{-\tau}$ that, $\mathcal{M}$ and $\widehat{\mathcal{M}}$ have the same cardinality, and*

$$d_H(\mathcal{M}, \widehat{\mathcal{M}}) \leq C(\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(1)} + h^2).$$

Remark 3.3. a). When $h = O(n^{-\frac{1}{d+6}})$, it is straightforward to show that $d_H(\mathcal{M}, \widehat{\mathcal{M}}) = O(n^{-\frac{2}{d+6}}\sqrt{\log n})$ almost surely, by applying the Borel-Cantelli Lemma to the above result with $\tau = 2\log n$. This matches the minimax rate of convergence up to $\sqrt{\log n}$ of mode estimation for density functions, as given in Tsybakov (1990, Theorem 3) with the smoothness parameter $\beta = 3$ therein. The minimax rate of maxima estimation for regression functions under a similar smoothness assumption is unknown to our best knowledge, and it is expected to be the same as that for density functions. As a side note, in the case of a unique mode, the maxima estimator using the k-NN regression, which is studied in Jiang (2019), matches the minimax rate in Tsybakov (1990) with $\beta = 2$, when k is appropriately chosen.
b). It can be seen from the proof that the constants in this theorem in fact do not depend on the magnitude (e.g. variance) of noise ϵ_i. This is unlike the case for the estimation of regression function itself, because we utilize a bounded transformation ξ and the property in Lemma 3.1.

3.2. Transformation 2

Next we consider the transformation ξ in **T2**. We will replace assumption **E** by the following assumption in our analysis.

Assumption **E′**: There exists a constant $B \in (0, \infty)$ such that $|Y| \leq B$ almost surely.

Under assumption **E′**, we have $0 \leq \pi(Y_{[n]}) \leq B + c_0$ almost surely. We can write $\widehat{r}_*(x) = \widehat{r}(x) + \pi(Y_{[n]})\widehat{t}(x)$, where

$$\widehat{t}(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\frac{K_h(x - X_i)}{\widehat{f}(X_i)}, \tag{3.3}$$

which is an estimator of unity. Here $\widehat{r}_*$ is considered as an estimator of $\bar{r}$, which is given in (2.13). We still denote the set of local maxima of $\widehat{r}_*$ by $\widehat{\mathcal{M}}$, which can be used to estimate $\mathcal{M}$, because the set of maxima of $\bar{r}$ is the same as that of r. The following result is similar to Theorem 3.2.

Theorem 3.3. *For ξ in* **T2**, *under assumptions* **A1-A3**, **K**, *and* **E′**, *if r is a Morse function and there are no critical points on the boundary of $\mathcal{X}$, then there exist constants $C > 0$, $c > 0$, and $h_0 > 0$ such that for all $n \geq 1$,*

$0 < h \le h_0$, $\tau > 1$ satisfying $nh^d \ge c(\tau \vee |\log h|)$ we have with probability at least $1 - 3(d+2)^2 e^{-\tau}$ that, $\mathcal{M}$ and $\widehat{\mathcal{M}}$ have the same cardinality, and

$$d_H(\mathcal{M}, \widehat{\mathcal{M}}) \le C(\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(1)} + h^2).$$

4. Simulations and applications

4.1. Bandwidth selection

The selection of bandwidth is critical for the finite-sample performance of kernel type estimators. In particular, the bandwidth h determines the number of maxima and clusters using the regression mean shift. Since our regression mean shift can be understood as an algorithm tracking the discretized gradient integral curves of the estimated regression function, a bandwidth producing good estimators of the gradient of the regression function is expected to be suitable for our regression MS algorithm. This can also be seen from Lemma 3.2. Based on this observation, below we propose a bandwidth selection strategy for our regression MS algorithm using a cross validation idea, although it is not meant to achieve any optimality. There may be other suitable bandwidth selection strategies that we have not explored, such as those based on the regression function itself or the Hausdorff distance between the estimated and true local maxima (see Zhou and Huang, 2019).

Let $\nabla\widehat{r}_\dagger(x)$ be a nonparametric kernel estimator of the gradient $\nabla\widetilde{r}(x)$, for example, the gradient of the NW regression estimator $\widehat{r}_{\mathrm{NW}}(x)$, or the gradient component using the local linear (LL) regression estimator. For $j = 1, \cdots, n$, let $\nabla\widehat{r}_{*,(-j)}(x)$ be the gradient estimator as given in (2.7), but using the sample points excluding $(X_j, \widetilde{Y}_j)$. The leave-one-out cross-validation error is defined as

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{j=1}^{n} \|\nabla\widehat{r}_\dagger(X_j) - \nabla\widehat{r}_{*,(-j)}(X_j)\|^2, \tag{4.1}$$

which has computational complexity of $O(n^2)$. The least square cross validation bandwidth h_{LSCV} which minimizes $\mathrm{CV}(h)$ is proposed to be used for our regression mean shift algorithm. Note that $\nabla\widehat{r}_\dagger$ itself requires a bandwidth choice. For LL estimator, one can use the gradient-based method as given in Henderson et al. (2015). For NW gradient estimator, one can scale the optimal bandwidth for NW regression estimator by multiplying a factor $n^{1/[(d+4)(d+6)]}$. See Henderson and Parmeter (2015, Chapter 5.5). We adopt the second method in our simulation study.

4.2. Simulation studies

We ran simulations to show the effectiveness of our regression mean shift algorithm (2.8) in partitioning the sample points in the input space and identifying the local maxima of a regression function. We considered the model

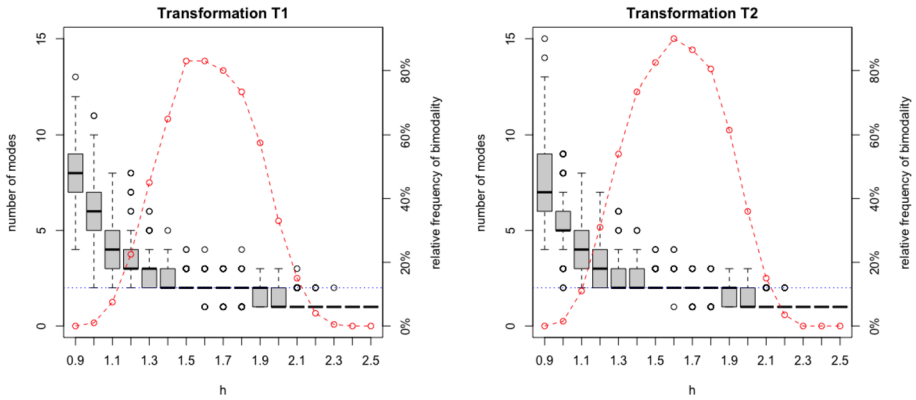

FIG 4.1. *Boxplots of the numbers of maxima/clusters across different bandwidths for transformations T1 (left), and T2 (right) when* $n = 200$ *for 200 samples, with the overlapping red curves representing the relative frequency of detecting two maxima. The number of true maxima is 2, as represented by the horizontal dotted lines.*

$Y_i = r(X_i) + \epsilon_i$ as in (3.1), where r is a bivariate function with two local maxima. Specifically, $r(x) = f_1(x) + f_2(x)$, where f_1 and f_2 are the density functions of $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, respectively, with $\mu_1 = (1,1)^T$, $\Sigma_1 = \text{diag}(0.5, 0.5)$, $\mu_2 = (-1,-1)^T$, $\Sigma_2 = \text{diag}(0.3, 0.9)$; for $i = 1, \cdots, n$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.01)$, and X_i is i.i.d. truncated bivariate normal such that $X_1 \sim \mathcal{N}(\mu_3, \Sigma_3)$, with $\mu_3 = (0,0)^T$ and $\Sigma_3 = \text{diag}(1.5, 1.5)$ conditional on $X_1 \in [-2,2]^2$. In each run $n = 200$ data points were generated from the above model as the input of our algorithm and we repeated the procedure for 200 times. We used the Epanechnikov kernel for g, and $\xi(x) = 1/(1 + \exp(-10x)) + 0.01$ for transformation **T1** and $c_0 = 0.1$ for transformation **T2**. Figure 4.1 shows the boxplots of the number of maxima detected by the algorithm using grid points of bandwidth values, overlapped with the relative frequency of correct bimodal identification (red curves). Using the bandwidth selection strategy in Section 4.1, among the 200 replications the relative frequencies that algorithm can correctly find two maxima are 78% for **T1** and 81% for **T2**, respectively, which are comparable to the peak values in Figure 4.1. When we increased the sample size n to 500, the relative frequencies of correct number of maxima reach 91% and 93.5%, respectively. These numbers are as high as 94.5% and 97.5% when $n = 1000$.

To further evaluate the quality of clustering, we use the adjusted Rand index (ARI) (Hubert and Arabie, 1985) to measure the similarity between the clusters returned by the regression mean shift and its underlying model. The ARI has a range $[-1, 1]$, and a larger value represents a better quality of clustering, with ARI=1 corresponding to a perfect matching. In Table 1 we report the averages and standard deviations (among 200 replications) of the ARI values for the two transformations applied to the same regression model as above. There is a clear trend that ARI approaches to 1 as the sample size n increases. We note that the result in Table 1 also shows the sensitivity to the distance of the two local

maxima in the model, in the sense that the effect of decreasing the sample size is comparable to that of moving data points (and hence the local maxima) closer.

TABLE 1
*The table includes the averages and standard deviations (in parentheses) of the ARI values using regression mean shift for various sample sizes n and the transformations **T1** and **T2**.*

	n =100	n =200	n =500	n =1000	n =2000	n = 5000	n =10000
T1	0.40 (0.27)	0.53 (0.24)	0.69 (0.16)	0.78 (0.10)	0.83 (0.07)	0.88 (0.03)	0.89 (0.02)
T2	0.46 (0.27)	0.59 (0.21)	0.70 (0.14)	0.78 (0.10)	0.83 (0.07)	0.87 (0.04)	0.88 (0.02)

In Figure 4.2 we visualize the outcome of the algorithm using a representative random sample of size 200 based on **T1**, which shows the paths of the estimation sequence in our regression MS algorithm, as well as the impact of different bandwidths on the maxima estimation results. Not surprisingly, when the bandwidth is small, there tend to be more local maxima (or basins), which can also be seen from the boxplots in Figure 4.1. The bandwidth selection strategy given in Section 4.1 works well with this sample.

With the same sample, we also tested the mean shift algorithm (2.12), which is based on the NW regression estimator. Using the Epanechnikov kernel, the sequence generated by (2.12) in not convergent and appears to be "chaotic". In fact, starting from each sample point, the algorithm has to stop after at most 20 iterations, because the sequence jumps to a location where there is no data point within the distance of the bandwidth, so that all the weights become zero in the next iteration. This issue arises because the weights w_i^* are not necessarily positive, and z_{j+1} may not be in a neighborhood of z_j, as argued in Remark 2.1.

4.3. Examples of applications

4.3.1. Partitions of protein energy landscapes

The proposed algorithm can be useful to obtain deep insight about the structure-function relationship in biological molecules (biomolecules). In the application highlighted here, we focus on protein molecules, which are ubiquitous in the cell, and where the three-dimensional structures accessed at equilibrium (under physiological conditions) often regulate a rich set of activities. Figure 4.3 relates the results obtained when the proposed algorithm is employed to organize the three-dimensional structures of the human H-Ras protein by their potential energies.

The structures (data sets) of the human H-Ras protein are obtained via the biophysical methodology described in Maximova et al. (2016) and Maximove et al. (2018). This work obtains structures for different versions of human H-Ras, the naturally-occurring, also referred to as the wildtype (WT) version, and mutated versions, known as variants. In the WT version, the protein accesses groups of structures that regulate its activity between an "on" and "off" state; in the on state, H-Ras instigates cellular reactions that signal the cell to grow; in the off state, such signals stop. In mutated variants, which are found in many

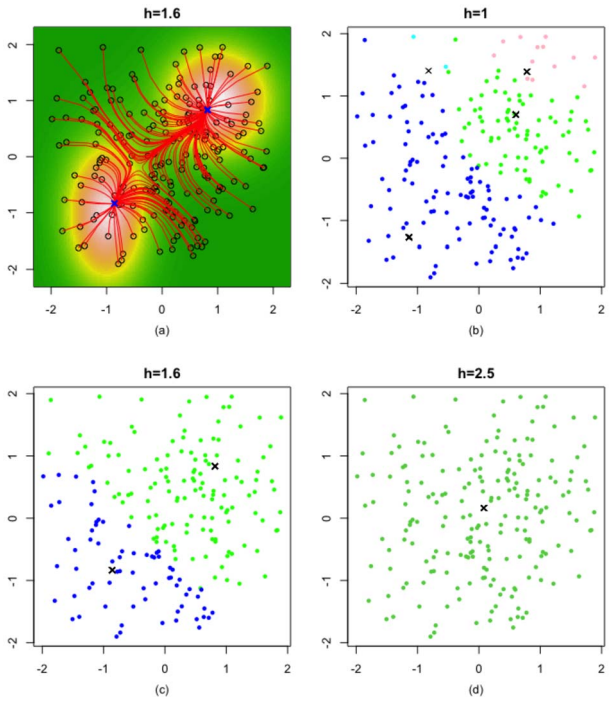

FIG 4.2. *Panel (a): The regression function is represented by the color in the background. Black dots are the sample points. Red curves are lines connecting sequential points generated from the algorithm in (2.8) starting from the sample points. The two* × *symbols are the points of convergence. Using the method in Section 4.1, the selected bandwidth is 1.6. Panels (b)–(d) shows the effect of the bandwidth choice. The sample points are partitioned (shown by different colors) according to their points of convergence (represented by* ×*). There are 4 basins when* $h = 1$*, 2 basins when* $h = 1.6$*, and 1 basin when* $h = 2.5$*.*

disorders, the regulation is disrupted in some manner, but only a view of the space of structures accessible can reveal exactly what, at the structure level, is responsible for dysfunction. The proposed algorithm promises to reveal exactly such organization of structures, as we show here.

From the structures/datasets produced by work in Maximova et al. (2016) and Maximove et al. (2018) for the WT and a common oncogenic variant, Q61L (the naming indicates the position where the naturally-occurring amino acid, "Q", has been replaced with a different amino acid, "L" in this case), we randomly selected 2000 structures for each, WT and Q61L*. Each structure (data point) comes with an associated energy value, which sums the physical interactions among the atoms in a particular structure. These energy values (in the original data sets) are all negative, and we used their absolute values in the analysis, so that our regression mean shift algorithm can cluster data points based on the local minima that they converge to.

*The data sets are downloadable at https://dx.doi.org/10.21227/331n-7019

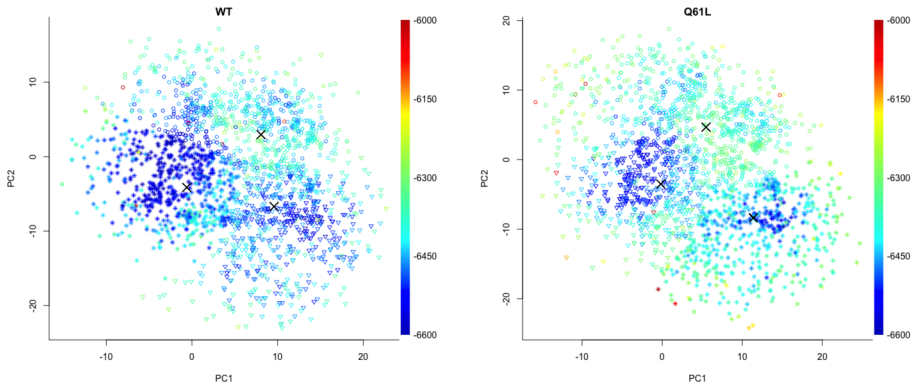

FIG 4.3. *Each panel shows 2000 structures (as symbols ✳, ▽, and ○) of the human H-Ras protein, and each type of these symbols represents a cluster (basin) obtained with the proposed algorithm, for which bandwidths are selected using the gradient-based LSCV method proposed in Section 4.1. The symbols × are the local minima. To aid visualization, structures are embedded onto the top two principal components obtained via Principal Component Analysis. Symbols are color-coded by their potential energies in a red-to-blue color scheme showing high-to-low energies. The left panel organizes structures accessed under physiological conditions by WT H-Ras; the right panel does so for the mutated, oncogenic Q61L H-Ras.*

Each panel in Figure 4.3 organizes the samples by their energies. Each dot corresponds to a three-dimensional structure. The red-to-blue color-coding scheme indicates high-to-low energies. The left panel shows the WT form/variant of the human H-Ras protein; the right panel shows the oncogenic variant known as Q61L.

The proposed algorithm is used to group the structures accessed by each H-Ras variant into local minima (to which we refer as energy basins). Basins with many low-energy structures (blue dots in Figure 4.3) correspond to stable and semi-stable structural states. The left panel in Figure 4.3 shows two such basins, one on the top left and one on the bottom right. More low-energy structures are contained in the basin shown in the top left, which indicates this is a wider and so more stable basin. Blue dots are found in between the basins, which indicate that the protein can transition between the two basins via low-energy structures; that is, an energetically feasible pathways exists to regulate the transition between the basins. Knowledge of the transition between on and off states for WT H-Ras allows us to speculate that the basins correspond to such states, as revealed by the proposed algorithm.

A comparison with the right panel in Figure 4.3, which shows the organization for Q61L H-Ras, shows two major differences with WT H-Ras. First, both basins become narrower; that is, they contain fewer low-energy structures. This suggests that the mutation impacts the structural plasticity of H-Ras. Second, few to no low-energy structures can be found between the basins, which suggests that the energetic pathway between the basins becomes more energetically costly. This in turn suggests that the Q61L mutation directly impacts the tran-

sition between the on and off states and so rigifies H-Ras. Such information is precious, as it allows formulating a detailed structure-based hypothesis that links sequence mutations to dysfunction via changes to structure and structure dynamics.

In addition to the two main basins, a third one is evident in the left panel of Figure 4.3 for the WT H-Ras. This contains fewer structures and is shallower. The right panel of Figure 4.3 indicates that the basin becomes even shallower in Q61L H-Ras. These results are in great agreement with early work in Clausen et al. (2015), where a Conf1 basin was suggested to exist in WT H-RAS and correspond to an unanticipated structural state. Specifically, by analyzing crystallographic structures whose projections over PC1 and PC2 fell on this basin, work in Clausen et al. (2015) suggested that this smaller and shallower basin corresponded to a structural state that was an intermediate between the known on and off states between the GTP- and GDP-bound states of WT H-Ras. In strong agreement with the results presented here, work in Clausen et al. (2015) additionally reported that this basin all but disappeared in Q61L.

More broadly, the shown application suggests that by organizing an energy landscape into the major local minima, the proposed algorithm allows understanding in great detail the impact of a mutation on the structural basin-to-basin dynamics that characterizes flexible biomolecules, such as proteins, and even obtaining an explanation for dysfunction in terms of changes to the underlying energy landscape and the dynamics on it.

4.3.2. Spatial clustering of malaria episodes

We applied our regression mean shift algorithm to a malaria episodes dataset available in the `R` package `SPODT` (Gaudart et al., 2015) and obtained a spatial clustering result, as shown in Figure 4.4. The dataset contains 168 observations, each corresponding to the longitudinal and latitudinal coordinates of a household, and the mean value of the number of malaria episodes per child in the household in Bandiagara, Mali, from November to December 2009. Our algorithm returns three clusters using the automatically selected bandwidth. The estimated maxima represent high-risk locations and different clusters are separated by low-risk valleys. In this example there exist wide regions where there are no data points, but they do not cause any issues to our algorithm, and the generated sequence still converges.

As a comparison, we also show the partitioning result of the CART algorithm (Breiman et al., 1993) in Figure 4.4. The same dataset has also been analyzed using a variant of CART algorithm called spatial oblique decision tree (SpODT). See Gaudart et al. (2005). The shape of clusters found using our regression MS algorithm appear different from that obtained from CART and its variant, which reflects the fundamental difference in the ideas of partitioning: the mathematical models behind the clusters in our regression MS are the ascending manifolds of the regression functions (see Section 2.1), while CART and its variants can be viewed as piecewise constant approximation of the regression function through their leaf nodes.

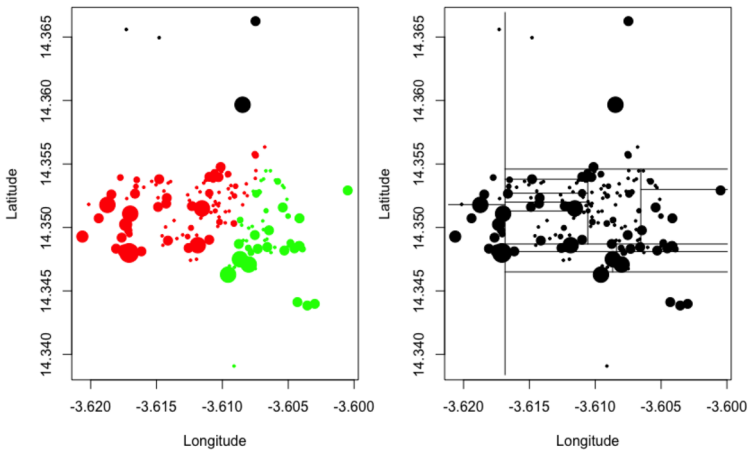

FIG 4.4. *The graphs shows the spatial partitioning results for the malaria episodes dataset available in the R package* **SPODT**, *using our regression mean shift algorithm (left panel), and CART (right panel). The mean value of the malaria episodes at each location is represented by the size of the dots. The bandwidth in regression MS algorithm is selected using the method as given in Section 4.1, and the three clusters are represented by different colors (black, red, and green).*

5. Discussions

In this paper we develop a regression mean shift algorithm to partition the input space and estimate the local maxima of regression functions. The algorithm is shown to be convergent and we give the rates of convergence for the local maxima estimators. Our algorithm is shown to be effective in simulations and real data applications. We note that our mean shift algorithm can also be used to estimate the local minima of regression functions, by simply replacing Y_i with $-Y_i$ for all $i = 1, \cdots, n$, and applying one of the two transformations **T1** and **T2** to $-Y_i$'s, as has been done in Section 4.3.1.

Between the two transformations, **T2** is linear, which is relatively easy to determine but requires the boundedness of the response (assumption $\mathbf{E}'$) in our theoretical analysis; **T1** includes a family of nonlinear transformations, which can potentially sharpen the local maxima of regression functions, and improve the performance of our algorithm, if ξ is carefully selected. In practice, one can first obtain the regression function estimator $\widehat{r}(x)$ and then choose ξ by assessing how transformations in **T1** or **T2** affect the landscape of $\widehat{r}$, in particular, the sharpness of its local maxima.

The idea of using regression MS to find local maxima can be extended to extract ridges of regression functions. Ridges are low-dimensional geometric features where the function values are local maximum in a subspace, which generalizes the concepts of local maxima and can be used to model filamentary structures. An algorithm called Subspace Constrained Mean Shift (SCMS) was developed in Ozertem and Erdogmus (2011) to extract ridges of KDEs. Some

theoretical analysis of this algorithm can be found in Genovese et al. (2014) and Qiao and Polonik (2016). We leave the extension of our regression mean shift algorithm to its subspace constrained version as a future work.

6. Proofs

This section contains the proofs of theoretical results in Sections 2 and 3. Note that the proof of Theorem 2.1 is very similar to that of Theorem 1 in Ghassabeh (2015) and is hence omitted. In the proofs we use C to denote a constant that may change its value depending on where it appears.

6.1. Proof of Lemma 2.1

Proof. Using the expression in (2.3), we have

$$\widehat{r}_*(z_{j+1}) - \widehat{r}_*(z_j) = \frac{c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} \frac{\widetilde{Y}_i}{\widehat{f}(X_i)} \Big[k\Big(\| \frac{z_{j+1} - X_i}{h} \|^2 \Big) - k\Big(\| \frac{z_j - X_i}{h} \|^2 \Big) \Big].$$

The convexity assumption of k implies that $k(x_2) - k(x_1) \geq g(x_1)(x_1 - x_2)$ for all $x_1, x_2 \in [0, \infty)$ and $x_1 \neq x_2$. Then using (2.5) we have

$$\begin{aligned}
&\widehat{r}_*(z_{j+1}) - \widehat{r}_*(z_j) \\
\geq &\frac{c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} \widetilde{Y}_i w_i(z_j) [2(z_{j+1} - z_j)^T X_i + \|z_j\|^2 - \|z_{j+1}\|^2] \\
= &\frac{c_{k,d}}{nh^{d+2}} \Big[2(z_{j+1} - z_j)^T \sum_{i=1}^{n} \widetilde{Y}_i w_i(z_j) X_i + (\|z_j\|^2 - \|z_{j+1}\|^2) \sum_{i=1}^{n} \widetilde{Y}_i w_i(z_j) \Big] \\
= &\frac{c_{k,d}}{nh^{d+2}} \|z_{j+1} - z_j\|^2 \sum_{i=1}^{n} \widetilde{Y}_i w_i(z_j) \\
\geq &\frac{c_{k,d}}{nh^{d+2}} \|z_{j+1} - z_j\|^2 \inf_{z \in \mathcal{C}} \sum_{i=1}^{n} \widetilde{Y}_i w_i(z),
\end{aligned}$$

where $\mathcal{C}$ is the convex hull of $\{X_1, \cdots, X_n\}$. Notice that $\inf_{z \in \mathcal{C}} \sum_{i=1}^{n} \widetilde{Y}_i w_i(z) > 0$, which implies that $\widehat{r}_*(z_{j+1}) - \widehat{r}_*(z_j) > 0$ as long as $z_{j+1} \neq z_j$. Since $\widehat{r}$ is upper bounded, the sequence $\widehat{r}(z_j)$ converges, and it follows that $\|z_{j+1} - z_j\| \to 0$ as $j \to \infty$. Since $\widehat{m}_*(z_j) = z_{j+1} - z_j$, using (2.7) we then get $\nabla \widehat{r}_*(z_j) \to 0$. □

6.2. Proof of Lemma 3.1

Proof. Let ξ' and ξ'' be the first two derivatives of ξ, respectively, and define

$$\rho_1(x) = \mathbb{E}\xi'(r(x) + \epsilon_1), \text{ and } \rho_2(x) = \mathbb{E}\xi''(r(x) + \epsilon_1). \tag{6.1}$$

We have that $\rho_1(x) > 0$, and both $\rho_1(x)$ and $\rho_2(x)$ are bounded. Note that for $\alpha \in \mathbb{N}^d$ with $|\alpha| = 1$,

$$\partial^\alpha \widetilde{r}(x) = \rho_1(x)\partial^\alpha r(x), \tag{6.2}$$

and for $\alpha \in \mathbb{N}^d$ with $|\alpha| = 2$ such that $\alpha = \alpha_1 + \alpha_2$, where $|\alpha_1| = |\alpha_2| = 1$,

$$\partial^\alpha \widetilde{r}(x) = \rho_1(x)\partial^\alpha r(x) + \rho_2(x)\partial^{\alpha_1} r(x)\partial^{\alpha_2} r(x). \tag{6.3}$$

In the matrix form, we get

$$\begin{aligned}
&\nabla \widetilde{r}(x) = \rho_1(x)\nabla r(x),\\
&\nabla^2 \widetilde{r}(x) = \rho_1(x)\nabla^2 r(x) + \rho_2(x)\nabla r(x)[\nabla r(x)]^T.
\end{aligned}$$

Hence $\nabla^2 \widetilde{r}(x) = \rho_1(x)\nabla^2 r(x)$ for all x such that $\nabla \widetilde{r}(x) = 0$. Since $\rho_1(x) > 0$, the critical points of $\widetilde{r}(x)$ and $r(x)$ are the same with the same indices.

When r is a Morse function, the above analysis implies that $\widetilde{r}$ is also a Morse function. To show that the ascending manifolds of r and $\widetilde{r}$ are the same, we will prove that the trajectories of integral curves driven by ∇r and $\nabla \widetilde{r}$ are the same when the starting points are the same. To this end, we show that there exists a reparameterization function $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\widetilde{\phi}_x(t) = \phi_x(\eta(t)), t \geq 0$, where ϕ_x and $\widetilde{\phi}_x$ are integral curves driven by ∇r and $\nabla \widetilde{r} = \rho_1 \nabla r$, respectively, defined as the solutions of

$$\begin{aligned}
&\phi_x'(t) = \nabla r(\phi_x(t)),\ t \geq 0;\ \phi_x(0) = x;\\
&\widetilde{\phi}_x'(t) = \nabla \widetilde{r}(\widetilde{\phi}_x(t)),\ t \geq 0;\ \widetilde{\phi}_x(0) = x.
\end{aligned}$$

Here η is the solution of the ODE $\eta'(t) = \rho_1(\phi_x(\eta(t)));\ \eta(0) = 0$. Then we have $\widetilde{\phi}_x(0) = \phi_x(\eta(0)) = x$ and for $t \geq 0$,

$$(\phi_x \circ \eta)'(t) = \phi_x'(\eta(t))\ \eta'(t) = \nabla r(\phi_x \circ \eta(t))\, \rho_1(\phi_x \circ \eta(t)) = \nabla \widetilde{r}(\phi_x \circ \eta(t)).$$

Hence $\widetilde{\phi}_x(t) = \phi_x(\eta(t)), t \geq 0$. So the conclusion of this lemma follows. □

6.3. Proof of Theorem 3.1

We use empirical process theory in the proofs. Let $\mathbb{P}$ be the probability measure of (X, Y), and $\mathbb{P}_n$ be the empirical probability measure with respect to $\{(X_i, Y_i) : i = 1, \cdots, n\}$ such that we write $\mathbb{P}(g) = \mathbb{E}g(X, Y)$, and $\mathbb{P}_n(g) = n^{-1}\sum_{i=1}^n g(X_i, Y_i)$, for any measurable function $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. Let

$$\mathbb{G}_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\mathbb{P}_n(g) - \mathbb{P}(g)]. \tag{6.4}$$

Let $\mathcal{G}$ be a set of measurable functions from $\mathbb{R}^{d+1}$ to $\mathbb{R}$. $\mathcal{G}$ is called a uniformly bounded VC-class if there exists a constant $B > 0$ such that $\sup_{x \in \mathbb{R}^{d+1}} |g(x)| \leq$

B for all $g \in \mathcal{G}$, and there exist positive numbers A and ν (called the characteristics) such that for every $0 < \epsilon \leq B$,

$$\sup_Q \mathcal{N}(\mathcal{G}, L_2(Q), \epsilon) \leq \Big(\frac{AB}{\epsilon}\Big)^\nu,$$

where the covering number $\mathcal{N}(\mathcal{G}, L_2(Q), \epsilon)$ denotes the smallest number of $L_2(Q)$-balls of radius at most ϵ needed to cover $\mathcal{G}$ and the supremum is taken over all probability measures Q on $\mathbb{R}^{d+1}$.

The following proposition generalizes Sriperumbudur and Steinwart (2012, Proposition A.5), based on Talagrand's inequality. Also see Giné and Guillou (2002) and Einmahl and Mason (2000). Its proof is given Section 6.6.

Proposition 6.1. *Let M is a real-valued function on $\mathbb{R}^d$ with bounded support $\mathcal{S}$ such that $M \in L_\infty(\mathbb{R}^d) \cap L_2(\mathbb{R}^d)$. Suppose that the marginal density f of X is uniformly bounded on $\mathcal{X}^{\eta h_0}$ for some constant $h_0 > 0$, where $\eta = \sup_{x \in \mathcal{S}} \|x\|$. Suppose that*

$$\mathcal{F} := \{\mathbb{R}^d \times \mathbb{R} \ni (u, v) \mapsto M(x - u) : \ x \in \mathbb{R}^d\} \tag{6.5}$$

is a uniformly bounded VC-class with characteristics (V, ν). For $h > 0$, let $\zeta_h : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ a function indexed by h such that

$$\sup_{0<h\leq h_0} \sup_{x,y \in \mathcal{X}\times\mathbb{R}} |\zeta_h(x, y)| \leq L$$

for some constant $L \in (0, \infty)$. Denote $\mathcal{G}_h = \{g_{x,h}(\cdot) : x \in \mathcal{X}\}$ where for $(u, v) \in \mathcal{X} \times \mathbb{R}$ and $h > 0$,

$$g_{x,h}(u, v) = \frac{1}{h^d} \zeta_h(u, v) M\Big(\frac{x - u}{h}\Big).$$

Then, there exists a positive constant C only depending on L, M, f, A and ν such that, for all $n \geq 1$, $0 < h < h_0$, and $\tau > 0$ we have

$$\mathbb{P}^n\Big(\frac{1}{\sqrt{n}} \sup_{g \in \mathcal{G}_h} |\mathbb{G}_n(g)| \leq \frac{C}{nh^d} \log \frac{C}{h} + \sqrt{\frac{C}{nh^d} \log \frac{C}{h}} + \frac{\tau C}{nh^d} + \frac{C\sqrt{\tau}}{\sqrt{nh^d}}\Big) \geq 1 - e^{-\tau}.$$

The proof of Theorem 3.1 needs an intermediate estimator, as define below. Denote $f_h = \mathbb{E}\widehat{f}$. Let

$$\widehat{r}_0(x) = \frac{1}{nh^d} \sum_{i=1}^n \frac{\widetilde{Y}_i K_h(x - X_i)}{f_h(X_i)}, \ x \in \mathcal{X}.$$

Note that

$$\partial^\alpha \widehat{r}_0(x) = \frac{1}{nh^{d+|\alpha|}} \sum_{i=1}^n \frac{\widetilde{Y}_i \partial^\alpha K((x - X_i)/h)}{f_h(X_i)}. \tag{6.6}$$

Then for $x \in \mathcal{X}$ we can write $\partial^\alpha \widehat{r}_*(x) - \partial^\alpha \widetilde{r}(x) = \mathrm{I}_n(x) + \mathrm{II}_n(x) + \mathrm{III}_n(x)$, where

$$\begin{aligned}
\mathrm{I}_n(x) &= \partial^\alpha \widehat{r}_*(x) - \partial^\alpha \widehat{r}_0(x), \\
\mathrm{II}_n(x) &= \partial^\alpha \widehat{r}_0(x) - \mathbb{E}\partial^\alpha \widehat{r}_0(x), \\
\mathrm{III}_n(x) &= \mathbb{E}\partial^\alpha \widehat{r}_0(x) - \partial^\alpha \widetilde{r}(x).
\end{aligned}$$

The conclusion in Theorem 3.1 is a direct consequence of Propositions 6.2, 6.3, 6.4 given in the sequel, which are used to analyze $\sup_{x\in\mathcal{X}} |\mathrm{I}_n(x)|$, $\sup_{x\in\mathcal{X}} |\mathrm{II}_n(x)|$, and $\sup_{x\in\mathcal{X}} |\mathrm{III}_n(x)|$, respectively. In particular, $\sup_{x\in\mathcal{X}} |\mathrm{I}_n(x)|$, $\sup_{x\in\mathcal{X}} |\mathrm{II}_n(x)|$ are stochastic terms that are analyzed using Proposition 6.1. We first consider $\sup_{x\in\mathcal{X}} |\mathrm{I}_n(x)|$.

Proposition 6.2. *Under the same assumptions as in Theorem 3.1, there exist constants $C > 0$, $c > 0$, and $h_0 > 0$ such that for all $|\alpha| \le 2$, $n \ge 1$, $0 < h \le h_0$, $\tau > 1$ satisfying $nh^d \ge c(\tau \vee |\log h|)$ we have*

$$\mathbb{P}^n\Big(\sup_{x\in\mathcal{X}} |\partial^\alpha \widehat{r}_*(x) - \partial^\alpha \widehat{r}_0(x)| < C\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(|\alpha|)} \Big) \ge 1 - 2e^{-\tau}. \tag{6.7}$$

Proof. Assume $h \le \delta$. Using Taylor expansion and the assumption that K is spherically symmetric with its support contained in the unit ball, we have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} |f_h(x) - f(x)| \\
=&\sup_{x\in\mathcal{X}} \Big| \int K(u) f(x - hu) du - f(x) \Big| \\
\le& h^2 d \int K(u)\|u\|^2 du \sup_{x\in\mathcal{X}^\delta} \max_{|\alpha|=2} |\partial^\alpha f(x)|.
\end{aligned} \tag{6.8}$$

Hence there exists $h_0 \in (0, \delta]$ such that for all $0 < h \le h_0$,

$$\sup_{x\in\mathcal{X}} |f_h(x) - f(x)| \le \frac{1}{3}\varepsilon_0,$$

where $0 < \varepsilon_0 \le \inf_{x\in\mathcal{X}} f(x)$ is given in assumption **A1**. This implies that $\inf_{x\in\mathcal{X}} f_h(x) \ge \frac{2}{3}\varepsilon_0$. Below we always assume that $0 < h \le h_0$. Notice that

$$|\partial^\alpha \widehat{r}_*(x) - \partial^\alpha \widehat{r}_0(x)| \le \sup_{x\in\mathcal{X}} |s_n(x)| \sup_{x\in\mathcal{X}} \widehat{r}_+^\alpha(x), \tag{6.9}$$

where $s_n(x) = [\widehat{f}(x)]^{-1} - [f_h(x)]^{-1}$ and

$$\widehat{r}_+^\alpha(x) = \frac{1}{nh^{d+|\alpha|}} \sum_{i=1}^n |\widetilde{Y}_i \partial^\alpha K_h(x - X_i)|. \tag{6.10}$$

Notice that

$$s_n(X_i) = \frac{f_h(X_i) - \widehat{f}(X_i)}{f_h(X_i)^2} + \frac{(f_h(X_i) - \widehat{f}(X_i))^2}{f_h(X_i)^2 \widehat{f}(X_i)}. \tag{6.11}$$

For $|\alpha| \le 2$, consider the class of functions $\mathcal{K}_\alpha = \{\partial^\alpha K(x - \cdot) : x \in \mathbb{R}^d\}$. Then $\mathcal{K}_\alpha$ is uniformly bounded under assumption **K**. Note that $K(x-\cdot) = k(\|x-\cdot\|^2)$, where k and its first two derivatives have bounded variation. It is known from Nolan and Pollard (1987) that in general $\mathcal{F}$ in (6.5) is a VC-class if $M(x) = \phi(p(x))$, where p is a polynomial and ϕ is a bounded real function of bounded variation. When $|\alpha| = 0$, it is clear that $\mathcal{K}_\alpha$ is a VC-class. When $|\alpha| = 1$, we have $\partial^\alpha K(x - \cdot) = k'(\|x - \cdot\|^2)[2\alpha^T(x - \cdot)]$. Noice that both $\{2\alpha^T(x - \cdot) : x \in \mathbb{R}^d\}$ and $\{k'(\|x-\cdot\|^2) : x \in \mathbb{R}^d\}$ are VC-classes. We then apply Chernozhukov et al. (2013, Lemma A.6) to conclude that $\mathcal{K}_\alpha$ is also a VC-class. A similar argument also applies to $|\alpha| = 2$.

For $u \in \mathbb{R}^d$, let $g_{x,h}(u) = \frac{1}{h^d} K((u - x)/h)$ and $\mathcal{G}_h = \{g_{x,h}(\cdot) : x \in \mathcal{X}\}$. Then notice that

$$\sup_{x\in\mathcal{X}} |\widehat{f}(x) - \mathbb{E}\widehat{f}(x)| = \frac{1}{\sqrt{n}} \sup_{g\in\mathcal{G}_h} |\mathbb{G}_n(g)|.$$

Applying Proposition 6.1 we get that there exists a constant $C_0 > 0$ such that for all $n \ge 1$, $h \in (0, 1)$, and $\tau > 1$ satisfying $nh^d \ge \tau$ and $nh^d \ge |\log h|$, with probability at least $1 - e^{-\tau}$,

$$\sup_{x\in\mathcal{X}} |\widehat{f}(x) - f_h(x)| < C_0\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(0)}. \tag{6.12}$$

Suppose that $C_0\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(0)} < \frac{1}{3}\varepsilon_0$. On the event in (6.12), we have

$$\sup_{x\in\mathcal{X}} |\widehat{f}(x) - f_h(x)| < \frac{1}{3}\varepsilon_0 \text{ and } \inf_{x\in\mathcal{X}} \widehat{f}(x) \ge \frac{1}{3}\varepsilon_0.$$

Therefore $\sup_{x\in\mathcal{X}} |s_n(x)| \le 5\varepsilon_0^{-2} \sup_{x\in\mathcal{X}} |\widehat{f}(x) - f_h(x)|$ and with probability at least $1 - e^{-\tau}$,

$$\sup_{x\in\mathcal{X}} |s_n(x)| < 5\varepsilon_0^{-2} C_0\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(0)}. \tag{6.13}$$

Here for all $x \in \mathcal{X}$,

$$\begin{aligned}
0 \le& \mathbb{E}\widehat{r}_+^\alpha(x) \\
=& \frac{1}{h^{d+|\alpha|}} \mathbb{E} \int_{\mathbb{R}^d} \left|\xi(r(x) + \epsilon_1)\partial^\alpha K\Big(\frac{x - u}{h}\Big)\right| f(u)du \\
=& \frac{1}{h^{|\alpha|}} \mathbb{E} \int_{\mathbb{R}^d} |\xi(r(x) + \epsilon_1)\partial^\alpha K(w)| f(x - hw)dw \\
\le& \frac{1}{h^{|\alpha|}} C_u \|\partial^\alpha K\|_1 \sup_{x\in\mathcal{X}^\delta} f(x) =: C_1 \frac{1}{h^{|\alpha|}},
\end{aligned} \tag{6.14}$$

where $\|\cdot\|_1$ is the L_1 norm.

For $(u, v) \in \mathcal{X} \times \mathbb{R}$, let $g_{x,h}^\alpha(u, v) = \frac{1}{h^d}|\xi(v)\partial^\alpha K((u-x)/h)|$. Then notice that we can write $\widehat{r}_+^\alpha(x) - \mathbb{E}\widehat{r}_+^\alpha(x) = \frac{1}{\sqrt{n}h^{|\alpha|}}\mathbb{G}_n(g_{x,h}^\alpha)$ and so that

$$\sup_{x\in\mathcal{X}} |\widehat{r}_+^\alpha(x) - \mathbb{E}\widehat{r}_+^\alpha(x)| = \frac{1}{\sqrt{n}h^{|\alpha|}} \sup_{g\in\mathcal{G}_{h,\alpha}} |\mathbb{G}_n(g)|,$$

where $\mathcal{G}_{h,\alpha} = \{g^{\alpha}_{x,h}(\cdot) : x \in \mathcal{X}\}$. Applying Proposition 6.1 we get that for $n \geq 1$, $h \in (0,1)$, and $\tau > 1$ satisfying $nh^d \geq \tau$ and $nh^d \geq |\log h|$,

$$\mathbb{P}^n\Big(\sup_{x\in\mathcal{X}} |\widehat{r}^{\alpha}_{+}(x) - \mathbb{E}\widehat{r}^{\alpha}_{+}(x)| < C_2\sqrt{\tau \vee |\log h|}\gamma^{(|\alpha|)}_{n,h}\Big) \geq 1 - e^{-\tau}, \tag{6.15}$$

for some constant $C_2 > 0$. So it follows from (6.14) and (6.15) that for $C_3 = C_1 \vee C_2$,

$$\mathbb{P}^n\Big(\sup_{x\in\mathcal{X}} \widehat{r}^{\alpha}_{+}(x) < C_3(h^{-|\alpha|} + \sqrt{\tau \vee |\log h|}\gamma^{(|\alpha|)}_{n,h})\Big) \geq 1 - e^{-\tau}. \tag{6.16}$$

Combing (6.9), (6.13) and (6.16), we then get the conclusion of this proposition. □

Next we consider $\sup_{x\in\mathcal{X}} |\mathrm{II}_n(x)|$.

Proposition 6.3. *Under the same assumptions as in Theorem 3.1, there exist constants $C > 0$, $c > 0$, and $h_0 > 0$ such that for all $|\alpha| \leq 2$, $n \geq 1$, $0 < h \leq h_0$, $\tau > 1$ satisfying $nh^d \geq c(\tau \vee |\log h|)$ we have*

$$\mathbb{P}^n\Big(\sup_{x\in\mathcal{X}} |\partial^{\alpha}\widehat{r}_0(x) - \mathbb{E}\partial^{\alpha}\widehat{r}_0(x)| < C\sqrt{\tau \vee |\log h|}\gamma^{(|\alpha|)}_{n,h}\Big) \geq 1 - e^{-\tau}. \tag{6.17}$$

Proof. For $(u,v) \in \mathcal{X} \times \mathbb{R}$, let $\zeta_h(u,v) = \xi(v)/f_h(u)$ and

$$p^{\alpha}_{x,h}(u,v) = \frac{1}{h^d}\zeta_h(u,v)\partial^{\alpha}K((x-u)/h).$$

Then notice that we can write $\partial^{\alpha}\widehat{r}_0(x) - \mathbb{E}\partial^{\alpha}\widehat{r}_0(x) = \frac{1}{\sqrt{n}h^{|\alpha|}}\mathbb{G}_n(p^{\alpha}_{x,h})$ and so that

$$\sup_{x\in\mathcal{X}} |\partial^{\alpha}\widehat{r}_0(x) - \mathbb{E}\partial^{\alpha}\widehat{r}_0(x)| = \frac{1}{\sqrt{n}h^{|\alpha|}} \sup_{p\in\mathcal{P}_{h,\alpha}} |\mathbb{G}_n(p)|.$$

where $\mathcal{P}_{h,\alpha} = \{p^{\alpha}_{x,h}(\cdot) : x \in \mathcal{X}\}$. Note that using the same h_0 in the proof of Proposition 6.9, we have $\sup_{u\in\mathcal{X}}\sup_{v\in\mathbb{R}} |\zeta_h(u,v)| \leq 2\varepsilon_0^{-1}C_u$. Applying Proposition 6.1 we then get (6.17). □

Next we consider $\sup_{x\in\mathcal{X}} |\mathrm{III}_n(x)|$.

Proposition 6.4. *Under the same assumptions as in Theorem 3.1, there exist constants $C > 0$ and $h_0 > 0$ such that for all $0 < h \leq h_0$, and $|\alpha| \leq 2$ we have*

$$\sup_{x\in\mathcal{X}} |\mathbb{E}\partial^{\alpha}\widehat{r}_0(x) - \partial^{\alpha}\widetilde{r}(x)| \leq Ch^{(3-|\alpha|)\wedge 2}. \tag{6.18}$$

Proof. Let

$$c_f = \sup_{x\in\mathcal{X}^{\delta}} \sup_{|\beta|\leq 3} |\partial^{\beta} f(x)|,$$

$$c_K = \Big[\max\Big\{\int_{\mathbb{R}^d} K(u)\|u\|^j du : j = 1, 2\Big\}\Big] \vee 1,$$
$$c_\xi = \Big[\sup_{x\in\mathbb{R}} \max(\xi(x), \xi'(x), \xi''(x), \xi'''(x))\Big] \vee 1,$$
$$c_r = \Big[\sup_{x\in\mathcal{X}^\delta} \sup_{|\beta|\le 3} |\partial^\beta r(x)|\Big] \vee 1.$$

Below we take $0 < h \le \frac{1}{2}\delta$ so that $(\mathcal{X}^h)^h \subset \mathcal{X}^\delta$. For $|\alpha| = 0, 1$, using a Taylor expansion of order 2 and assumption **K**, we have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^h} |\partial^\alpha f_h(x) - \partial^\alpha f(x)| \\
&= \sup_{x\in\mathcal{X}^h} |\int_{\mathbb{R}^d} K(u)\partial^\alpha f(x - hu)du - \partial^\alpha f(x)| \\
&\le h^2 \int_{\mathbb{R}^d} K(u)\|u\|^2 du \sup_{x\in\mathcal{X}^\delta} \max_{|\beta|=|\alpha|+2} |\partial^\beta f(x)| \\
&\le dc_f c_K h^2. \qquad (6.19)
\end{aligned}$$

Similarly for $|\alpha| = 2$, using a Taylor expansion of order 1 we have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}^h} |\partial^\alpha f_h(x) - \partial^\alpha f(x)| \\
&= \sup_{x\in\mathcal{X}^h} |\int_{\mathbb{R}^d} K(u)\partial^\alpha f(x - hu)du - \partial^\alpha f(x)| \\
&\le h \int K(u)\|u\| du \sup_{x\in\mathcal{X}^\delta} \max_{|\beta|=3} |\partial^\beta f(x)| \\
&\le \sqrt{d} c_f c_K h. \qquad (6.20)
\end{aligned}$$

Let $\eta_0 = \sup\{\eta \in (0, \delta] : \inf_{x\in\mathcal{X}^\eta} f(x) \ge \frac{1}{2}\varepsilon_0\}$. Under assumptions **A1** and **A2**, we have $\eta_0 > 0$. Let

$$h_0 = \min\Big\{\eta_0, \frac{1}{2}\delta, \frac{1}{\sqrt{d}c_K}, \Big(\frac{\varepsilon_0}{2dc_f c_K}\Big)^{1/2}\Big\},$$

and we take $h \in (0, h_0]$ below. Using (6.19) and (6.20) we have for all $|\alpha| = 1, 2$,

$$\sup_{x\in\mathcal{X}^h} |f(x) - f_h(x)| \le c_f \wedge (\frac{1}{2}\varepsilon_0),$$
$$\sup_{x\in\mathcal{X}^h} |\partial^\alpha f_h(x) - \partial^\alpha f(x)| \le c_f.$$

This implies that for all $x \in \mathcal{X}^h$ we have $\frac{1}{2}\varepsilon_0 \le f_h(x) \le 2c_f$, and $|\partial^\alpha f_h(x)| \le 2c_f$ for all $|\alpha| = 1, 2$.

Let $q_h(x) = f(x)/f_h(x)$. Using (6.19) we have for all $x \in \mathcal{X}^h$,

$$|q_h(x) - 1| \le \frac{2c_f c_K h^2}{\varepsilon_0} := C_{q,0} h^2. \qquad (6.21)$$

When $|\alpha| = 1$,

$$\partial^\alpha q_h(x) = \frac{\partial^\alpha f(x) - \partial^\alpha f_h(x)}{f_h(x)} - \frac{[f(x) - f_h(x)]\partial^\alpha f_h(x)}{f_h(x)^2}.$$

Hence using (6.19) we have that for all $x \in \mathcal{X}^h$,

$$|\partial^\alpha q_h(x)| \leq \frac{2c_f c_K h^2}{\varepsilon_0} + \frac{8c_f^2 c_K h^2}{\varepsilon_0^2} := C_{q,1} h^2. \tag{6.22}$$

When $|\alpha| = 2$, suppose that $\alpha = \alpha_1 + \alpha_2$, where $|\alpha_1| = |\alpha_2| = 1$. We have

$$\begin{aligned}\partial^\alpha q_h(x) =& \frac{\partial^\alpha f(x) - \partial^\alpha f_h(x)}{f_h(x)} \\ &- \frac{[\partial^{\alpha_1} f(x) - \partial^{\alpha_1} f_h(x)]\partial^{\alpha_2} f_h(x) + [\partial^{\alpha_2} f(x) - \partial^{\alpha_2} f_h(x)]\partial^{\alpha_1} f_h(x)}{[f_h(x)]^2} \\ &+ [f_h(x) - f(x)]\Big[\frac{2\partial^{\alpha_1} f_h(x)\partial^{\alpha_2} f_h(x)}{f_h(x)^4} - \frac{\partial^\alpha f_h(x)}{f_h(x)^2}\Big].\end{aligned}$$

Hence using (6.19) and (6.20) we have that for all $x \in \mathcal{X}^h$,

$$\begin{aligned}|\partial^\alpha q_h(x)| &\leq \frac{2c_f c_K h}{\varepsilon_0} + \frac{24c_f^2 c_K h^2}{\varepsilon_0^2} + \frac{128c_f^3 c_K h^2}{\varepsilon_0^4} \\ &\leq \frac{2c_f c_K h}{\varepsilon_0} + \frac{12c_f^2 c_K \delta h}{\varepsilon_0^2} + \frac{64c_f^3 c_K \delta h}{\varepsilon_0^4} := C_{q,2} h.\end{aligned} \tag{6.23}$$

We can write for $|\alpha| = 0, 1, 2$,

$$\begin{aligned}&\mathbb{E}\partial^\alpha \widehat{r}_0(x) \\ =& \frac{1}{h^{d+|\alpha|}} \mathbb{E} \frac{\widetilde{Y}_1 \partial^\alpha K(\frac{x - X_1}{h})}{f_h(X_1)} \\ =& \frac{1}{h^{d+|\alpha|}} \mathbb{E} \frac{\xi(r(X_1) + \epsilon_1)\partial^\alpha K(\frac{x - X_1}{h})}{f_h(X_1)} \\ =& \frac{1}{h^{d+|\alpha|}} \mathbb{E} \int_{\mathbb{R}^d} \xi(r(u) + \epsilon_1)\partial^\alpha K\Big(\frac{x - u}{h}\Big) q_h(u) du \\ =& \frac{1}{h^{|\alpha|}} \mathbb{E} \int_{\mathbb{R}^d} \xi(r(x - hw) + \epsilon_1) q_h(x - hw)\partial^\alpha K(w) dw \\ =& \frac{1}{h^{|\alpha|}} \mathbb{E} \int_{\mathbb{R}^d} \partial_w^\alpha [\xi(r(x - hw) + \epsilon_1) q_h(x - hw)] K(w) dw.\end{aligned} \tag{6.24}$$

Here for $|\alpha| = 1$,

$$\begin{aligned}\partial_w^\alpha [\xi(r(x + hw) + \epsilon_1) q_h(x + hw)] =& h\xi'(r(x + hw) + \epsilon_1)\partial^\alpha r(x + hw) q_h(x + hw) \\ &+ h\xi(r(x + hw) + \epsilon_1)\partial^\alpha q_h(x + hw).\end{aligned} \tag{6.25}$$

For $|\alpha| = 2$, suppose that $\alpha = \alpha_1 + \alpha_2$, where $|\alpha_1| = |\alpha_2| = 1$. We have

$$\begin{aligned}
&\partial_w^\alpha[\xi(r(x+hw)+\epsilon_1)q_h(x+hw)] \\
=&h^2\xi''(r(x+hw)+\epsilon_1)\partial^{\alpha_1}r(x+hw)\partial^{\alpha_2}r(x+hw)q_h(x+hw) \\
&+h^2\xi'(r(x+hw)+\epsilon_1)\partial^{\alpha}r(x+hw)q_h(x+hw) \\
&+h^2\xi'(r(x+hw)+\epsilon_1)\partial^{\alpha_1}r(x+hw)\partial^{\alpha_2}q_h(x+hw) \\
&+h^2\xi'(r(x+hw)+\epsilon_1)\partial^{\alpha_2}r(x+hw)\partial^{\alpha_1}q_h(x+hw) \\
&+h^2\xi(r(x+hw)+\epsilon_1)\partial^{\alpha}q_h(x+hw). \qquad (6.26)
\end{aligned}$$

Using a Taylor expansion of order 2, we have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}|\xi(r(x+hw)+\epsilon_1)-\xi(r(x)+\epsilon_1)-h\xi'(r(x)+\epsilon_1)w^T\nabla r(x)| \\
\le&h^2dc_\xi c_r\|w\|^2, \qquad (6.27)
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}|\xi'(r(x+hw)+\epsilon_1)-\xi'(r(x)+\epsilon_1)-h\xi''(r(x)+\epsilon_1)w^T\nabla r(x)| \\
\le&h^2dc_\xi c_r\|w\|^2. \qquad (6.28)
\end{aligned}$$

Using a Taylor expansion of order 1, we have

$$\sup_{x\in\mathcal{X}}|\xi''(r(x+hw)+\epsilon_1)-\xi''(r(x)+\epsilon_1)| \le h\sqrt{d}c_\xi c_r\|w\|. \qquad (6.29)$$

For $|\alpha| = 1$, using a Taylor expansion of order 2 we have

$$\sup_{x\in\mathcal{X}}|\partial^\alpha r(x+hw)-\partial^\alpha r(x)+hw^T\nabla\partial^\alpha r(x)| \le h^2dc_r\|w\|^2. \qquad (6.30)$$

For $|\alpha| = 2$, using a Taylor expansion of order 1 we have

$$\sup_{x\in\mathcal{X}}|\partial^\alpha r(x+hw)-\partial^\alpha r(x)| \le h\sqrt{d}c_r\|w\|. \qquad (6.31)$$

Therefore it follows from (6.24), (6.21), and (6.27) that

$$\sup_{x\in\mathcal{X}}|\mathbb{E}\widehat{r}_0(x)-\widetilde{r}_0(x)| \le (C_{q,0}c_\xi + dc_\xi c_r c_K)h^2 := C_{r,0}h^2. \qquad (6.32)$$

For $|\alpha| = 1$, the calculations in (6.24), (6.21), (6.22), (6.25), (6.27), (6.28) and (6.30) yield

$$\sup_{x\in\mathcal{X}}|\mathbb{E}\partial^\alpha\widehat{r}_0(x)-\partial^\alpha\widetilde{r}_0(x)| \le (C_{q,1}c_\xi + C_{q,0}c_\xi c_r + 3dc_r^2c_Kc_\xi)h^2 := C_{r,1}h^2. \qquad (6.33)$$

For $|\alpha| = 2$, using 6.24), (6.21) – (6.23), and (6.26) – (6.31) we get

$$\sup_{x\in\mathcal{X}}|\mathbb{E}\partial^\alpha\widehat{r}_0(x)-\partial^\alpha\widetilde{r}_0(x)|$$

$$\begin{aligned}&\leq (C_{q,2}c_\xi + C_{q,1}c_\xi c_r + 3C_{q,0}c_\xi c_r^2 + 4dc_\xi c_r^2 c_K)h^2 + 2\sqrt{d}c_\xi c_r^3 c_K h \\ &\leq [\delta(C_{q,2}c_\xi + C_{q,1}c_\xi c_r + 3C_{q,0}c_\xi c_r^2 + 4dc_\xi c_r^2 c_K) + 2\sqrt{d}c_\xi c_r^3 c_K]h := C_{r,2}h.\end{aligned} \tag{6.34}$$

The proof is completed with a constant $C = \max\{C_{r,0}, C_{r,1}, C_{r,2}\}$. □

6.4. *Proofs of Lemma 3.2 and Theorem 3.2*

Theorem 3.2 is a direct consequence of the application of Lemma 3.2 and Theorem 3.1, so we only give the proof of Lemma 3.2 below.

Proof. Let $c_p = \sup_{x\in\mathcal{R}} \sup_{|\beta|\leq 3} |\partial^\beta p(x)|$ and $\lambda_\dagger = \inf_{x\in\mathcal{C}} |\lambda_1(x)|$. Since p is a Morse function and $\mathcal{R}$ is a compact set, we have $\lambda_* \geq \lambda_\dagger > 0$. Let $\kappa = \frac{\lambda_\dagger}{2dc_p} \wedge \eta$. Then $\mathcal{M}^\kappa \subset \mathcal{R}$. Let $\mathcal{C}_\circ^\kappa = \{y \in \mathcal{R} : \inf_{x\in\mathcal{C}} \|x - y\| < \kappa\}$ be the interior of $\mathcal{C}^\kappa$, and $\mathcal{T} = \mathcal{R}\backslash\mathcal{C}_\circ^\kappa$. Let $\theta = \inf_{x\in\mathcal{T}} \max_{|\alpha|=1} |\partial^\alpha p(x)|$. Note that $\theta > 0$ when $\mathcal{T} \neq \emptyset$, because $\max_{|\alpha|=1} |\partial^\alpha p|$ is a continuous function on $\mathcal{R}$ and $\mathcal{T}$ is a compact set. We will show the result in this lemma holds when the following three conditions are satisfied.

$$\widetilde{\delta}_0 := \sup_{x\in\mathcal{R}} |p(x) - \widetilde{p}(x)| < \frac{1}{8}\lambda_*\kappa^2, \tag{6.35}$$

$$\widetilde{\delta}_2 := \sup_{x\in\mathcal{R}} \max_{|\alpha|=2} |\partial^\alpha p(x) - \partial^\alpha \widetilde{p}(x)| \leq \frac{\lambda_\dagger}{4d}, \tag{6.36}$$

$$\widetilde{\delta}_1 := \sup_{x\in\mathcal{R}} \max_{|\alpha|=1} |\partial^\alpha p(x) - \partial^\alpha \widetilde{p}(x)| \leq \frac{1}{2}\theta, \quad \text{when } \mathcal{T} \neq \emptyset. \tag{6.37}$$

Step 1. For any $x \in \mathcal{M}$, consider any $y \in \mathcal{B}_x^\kappa := \{y \in \mathcal{R} : \|x - y\| \leq \kappa\}$, and using Weyl's inequality (see Serre, 2002, page 15) we have

$$\begin{aligned}|\lambda_1(y) - \lambda_1(x)| &\leq d \sup_{|\beta|=2} |\partial^\beta p(x) - \partial^\beta p(y)| \\ &\leq d \sup_{z\in\mathcal{R}} \sup_{|\beta|=3} |\partial^\beta p(z)| \|x - y\| \leq dc_p\|x - y\|.\end{aligned}$$

Therefore for all $y \in \mathcal{B}_x^\kappa$,

$$\lambda_1(y) \leq -\lambda_* + dc_p\kappa \leq -\frac{1}{2}\lambda_*. \tag{6.38}$$

In other words, $\mathcal{M}^\kappa \subset \mathcal{A} := \{x \in \mathcal{R} : \lambda_1(x) \leq -\frac{1}{2}\lambda_*\}$. For all $x \in \mathcal{M}$ and all $y \in \mathcal{B}_x^\kappa$, using a Taylor expansion we have

$$p(y) \leq p(x) + \frac{1}{2} \sup_{z\in\mathcal{M}^\kappa} \lambda_1(z)\|x - y\|^2 \leq p(x) - \frac{1}{4}\lambda_*\|x - y\|^2. \tag{6.39}$$

Then by using (6.35) we must have for all $x \in \mathcal{M}$ and all $y \in \mathcal{R}$ such that $\|x - y\| = \kappa$,

$$\widetilde{p}(y) < p(y) + \frac{1}{8}\lambda_*\kappa^2 \leq p(x) - \frac{1}{8}\lambda_*\kappa^2 < \widetilde{p}(x). \tag{6.40}$$

Therefore there must exist at least one local maximum of $\widetilde{p}$ on $\mathcal{B}_x^\kappa$ for each $x \in \mathcal{M}$.

Step 2a. Let $\mathcal{S} = \mathcal{C} \backslash \mathcal{M}$ the set of critical points of p on $\mathcal{R}$ excluding local maxima. Suppose $\mathcal{S} \neq \emptyset$. Then following a similar calculation in (6.38), we have that for all $y \in \mathcal{S}^\kappa$, $\lambda_1(y) \geq \frac{1}{2}\lambda^\dagger > 0$. For any $x \in \mathcal{R}$, let $\widetilde{\lambda}_1(x)$ be the largest eigenvalue of $\nabla^2 \widetilde{p}(x)$. For all $y \in \mathcal{S}^\kappa$, by using (6.36) and Weyl's inequality we have

$$\widetilde{\lambda}_1(y) \geq \lambda_1(y) - d \sup_{|\beta|=2} |\partial^\beta \widetilde{p}(y) - \partial^\beta p(y)| \geq \frac{\lambda_\dagger}{4}. \tag{6.41}$$

So there are no local maxima of $\widetilde{p}$ on $\mathcal{S}^\kappa$. The same statement is trivially true when $\mathcal{S} = \emptyset$ because $\mathcal{S}^\kappa = \emptyset$ in such a case.

Step 2b. If $\mathcal{T} = \emptyset$, then we must have $\widetilde{\mathcal{M}} \subset \mathcal{M}^\kappa$ based on the arguments in **Step 1** and **Step 2a**, since $\mathcal{R} = \mathcal{M}^\kappa \cup \mathcal{S}^\kappa \cup \mathcal{T}$. Otherwise for all $y \in \mathcal{T}$, by using (6.37) we have that

$$\max_{|\alpha|=1} |\partial^\alpha \widetilde{p}(y)| \geq \theta - \widetilde{\delta}_1 \geq \frac{1}{2}\theta > 0. \tag{6.42}$$

This means that there are no local maxima of $\widetilde{p}$ on $\mathcal{T}$, and hence $\widetilde{\mathcal{M}} \subset \mathcal{M}^\kappa$.

Step 2c. Suppose that there exists $x \in \mathcal{M}$ such that there are at least two different local maxima $\widetilde{x}_1$ and $\widetilde{x}_2$ of $\widetilde{p}$ within $\mathcal{B}_x^\kappa$. For any $y \in \mathcal{B}_x^\kappa$, by using (6.38), (6.36) and Weyl's inequality, we have that

$$\widetilde{\lambda}_1(y) \leq \lambda_1(y) + d \sup_{|\beta|=2} |\partial^\beta \widetilde{p}(y) - \partial^\beta p(y)| \leq -\frac{\lambda_*}{4}. \tag{6.43}$$

Using a Taylor expansion we have

$$0 = (\widetilde{x}_1 - \widetilde{x}_2)^T [\nabla \widetilde{p}(\widetilde{x}_1) - \nabla \widetilde{p}(\widetilde{x}_2)] \leq \sup_{y \in \mathcal{B}_x^\kappa} \widetilde{\lambda}_1(y) \|\widetilde{x}_1 - \widetilde{x}_2\|^2, \tag{6.44}$$

which leads to a contradiction with (6.43). Hence there exists only one local maximum $\widetilde{x}$ of $\widetilde{p}$ in $\mathcal{B}_x^\kappa$ for each $x \in \mathcal{M}$. For the same reason, using (6.38) it can be seen that there exists only one local maximum x of p in $\mathcal{B}_x^\kappa$. In other words, we have that the number of maxima of p and $\widetilde{p}$ are the same and can be matched in such a way that

$$d_H(\mathcal{M}, \widetilde{\mathcal{M}}) = \max_{x \in \mathcal{M}} \|\widetilde{x} - x\|. \tag{6.45}$$

Step 3. Let us consider any local maximum of p, denoted by x and its corresponding local maximum $\widetilde{x}$ of $\widetilde{p}$ in $\mathcal{B}_x^\kappa$. Let $|\cdot|_{\max}$ and $\|\cdot\|_{\text{op}}$be the element-wise maximum and the operator norm of a matrix, respectively. Since $\nabla p(x) = \nabla \widetilde{p}(\widetilde{x}) = 0$, using a Taylor expansion, we have

$$\nabla p(x) - \nabla \widetilde{p}(x) = \nabla \widetilde{p}(\widetilde{x}) - \nabla \widetilde{p}(x) = [\nabla^2 p(x) + \Delta(\widetilde{x}, x)](\widetilde{x} - x), \tag{6.46}$$

where $\Delta(\widetilde{x},x)$ is a $d\times d$ symmetric matrix such that $|\Delta(\widetilde{x},x)|_{\max}\leq\widetilde{\delta}_2+c_p|\widetilde{x}-x|_{\max}$. Therefore

$$\begin{aligned}\|\nabla p(x)-\nabla\widetilde{p}(x)\| &\geq \|\nabla^2 p(x)(\widetilde{x}-x)\|-\|\Delta(\widetilde{x},x)(\widetilde{x}-x)\|\\ &\geq \lambda_*\|\widetilde{x}-x\|-\|\Delta(\widetilde{x},x)\|_{\mathrm{op}}\|\widetilde{x}-x\|\\ &\geq \lambda_*\|\widetilde{x}-x\|-d[\widetilde{\delta}_2+c_p|\widetilde{x}-x|_{\max}]\|\widetilde{x}-x\|\\ &\geq \frac{1}{2}\lambda_*\|\widetilde{x}-x\|-d\widetilde{\delta}_2\|\widetilde{x}-x\|\\ &\geq \frac{1}{4}\lambda_*\|\widetilde{x}-x\|,\end{aligned}$$

where in the last step we use (6.36). The conclusion of this lemma follows by noticing (6.45). □

6.5. *Proof of Theorem 3.3*

Proof. First of all, similar to Theorem 3.1, there exist constants $C_1>0$, $c_1>0$ and $h_1>0$ such that for all $|\alpha|\leq 2$, $n\geq 1$, $0<h\leq h_1$, $\tau>1$ satisfying $nh^d\geq c_1(\tau\vee|\log h|)$ we have with probability at least $1-3e^{-\tau}$,

$$\sup_{x\in\mathcal{X}}|\partial^\alpha\widehat{r}(x)-\partial^\alpha r(x)|<C_1(\sqrt{\tau\vee|\log h|}\gamma_{n,h}^{(|\alpha|)}+h^{(3-|\alpha|)\wedge 2}). \tag{6.47}$$

Recall $\widehat{t}$ and $\bar{r}$ that have been defined in (3.7) and (2.8), respectively. Note that $\widehat{t}(x)$ corresponds to $\widehat{r}(x)$ in the case of $\widetilde{Y}_i=1$, for all $i=1,\cdots,n$. Let

$$\widehat{t}_0(x)=\frac{1}{nh^d}\sum_{i=1}^n\frac{K_h(x-X_i)}{f_h(X_i)}.$$

Then similar to Proposition 5.2, there exist constants $C_2>0$, $c_2>0$, and $h_2>0$ such that for all $|\alpha|\leq 2$, $n\geq 1$, $0<h\leq h_2$, $\tau>1$ satisfying $nh^d\geq c_2(\tau\vee|\log h|)$ we have with probability at least $1-2e^{-\tau}$,

$$\sup_{x\in\mathcal{X}}|\partial^\alpha\widehat{t}(x)-\partial^\alpha\widehat{t}_0(x)|<C_2\sqrt{\tau\vee|\log h|}\gamma_{n,h}^{(|\alpha|)}. \tag{6.48}$$

Similar to Proposition 5.3, there exist constants $C_3>0$, $c_3>0$, and $h_3>0$ such that for all $|\alpha|\leq 2$, $n\geq 1$, $0<h\leq h_3$, $\tau>1$ satisfying $nh^d\geq c_3(\tau\vee|\log h|)$ we have with probability at least $1-e^{-\tau}$,

$$\sup_{x\in\mathcal{X}}|\partial^\alpha\widehat{t}_0(x)-\mathbb{E}\partial^\alpha\widehat{t}_0(x)|<C_3\sqrt{\tau\vee|\log h|}\gamma_{n,h}^{(|\alpha|)}. \tag{6.49}$$

Recall that $q_h=f/f_h$. Note that

$$\mathbb{E}\partial^\alpha\widehat{t}_0(x)=\int_{\mathbb{R}^d}\partial^\alpha q_h(x-hw)K(w)dw$$

Similar to Proposition 5.4, there exist constants $C_4 > 0$ and $h_4 > 0$ such that for all $0 < h \le h_4$, and $|\alpha| \le 2$ we have

$$\sup_{x\in\mathcal{X}} |\mathbb{E}\partial^\alpha \widehat{t}_0(x) - b_{|\alpha|}| \le C_4 h^{(3-|\alpha|)\wedge 2} \tag{6.50}$$

where $b_{|\alpha|} = 0$ when $|\alpha| = 0$ and $b_{|\alpha|} = 1$ when $|\alpha| = 1, 2$. Hence combing (6.48), (6.49) and (6.50), for all $|\alpha| \le 2$ we get with probability at least $1 - 3e^{-\tau}$ that

$$\sup_{x\in\mathcal{X}} |\partial^\alpha \widehat{t}(x) - b_{|\alpha|}| < C_5(\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(|\alpha|)} + h^{(3-|\alpha|)\wedge 2}), \tag{6.51}$$

where $C_5 = (C_2 + C_3) \vee C_4$. Due to the almost sure boundedness of $\pi(Y_{[n]})$, using (6.47), (6.51), and the relations $\widehat{r}_*(x) = \widehat{r}(x) + \pi(Y_{[n]})\widehat{t}(x)$ and $\partial^\alpha \bar{r} = \partial^\alpha r + b_{|\alpha|}\pi(Y_{[n]})$, for all $|\alpha| \le 2$ we have with probability at least $1 - 6e^{-\tau}$ that

$$\sup_{x\in\mathcal{X}} |\partial^\alpha \widehat{r}_*(x) - \partial^\alpha \bar{r}(x)| < C_6(\sqrt{\tau \vee |\log h|}\gamma_{n,h}^{(|\alpha|)} + h^{(3-|\alpha|)\wedge 2}),$$

where $C_6 = C_1 + C_5(B \vee 1)$, where B is given in assumption $\mathbf{E}'$. Then the conclusion of the theorem follows from the application of Lemma 3.2. □

6.6. Proof of Proposition 6.1

Proof. For any measurable function g and probability measure Q on $\mathbb{R}^{d+1}$, let $\|g\|_{L_2(Q)} = [\int_{\mathbb{R}^{d+1}} |g(u)|^2 dQ(u)]^{1/2}$ be the $L_2(Q)$-norm of g. We first show that $\widetilde{\mathcal{G}}_h := \{h^d(g - \mathbb{P}g) : g \in \mathcal{G}_h\}$ is a uniformly bounded VC-class, where $\mathbb{P}g = \mathbb{E}g(X, Y)$. For any $x \in \mathbb{R}^d$, let $m_{x,h}(u, v) = M((x-u)/h)$ for all $(u, v) \in \mathbb{R}^d \times \mathbb{R}$. Let $B \in (0, \infty)$ be a constant envelope of $\mathcal{F}$ such that $\sup_{g\in\mathcal{F}} \sup_{x\in\mathbb{R}^{d+1}} |g(x)| \le B$. Define $\mathcal{F}_h = \{m_{x,h}(\cdot) : x \in \mathbb{R}^d\}$, and $\mathcal{F}_{h,\mathcal{X}} = \{m_{x,h}(\cdot) : x \in \mathcal{X}\}$ for all $h > 0$. Using Sriperumbudur and Steinwart (2012, Lemma A.3), we obtain that for all $h > 0$, and $\epsilon \in (0, B]$,

$$\sup_Q \mathcal{N}(\mathcal{F}_{h,\mathcal{X}}, L_2(Q), \epsilon) \le \sup_Q \mathcal{N}(\mathcal{F}_h, L_2(Q), \epsilon) = \sup_Q \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon) \le \Big(\frac{AB}{\epsilon}\Big)^\nu, \tag{6.52}$$

where the supremum is taken over all the probability measures Q on $\mathbb{R}^{d+1}$.

Let

$$\mathcal{F}_{h,\mathcal{X}}^{(1)} = \{\mathcal{X} \times \mathbb{R} \ni (u, v) \mapsto \zeta_h(u, v)m_{x,h}(u, v) : \ x \in \mathcal{X}\}.$$

Note that $B^{(1)} := LB$ is a constant envelope of $\mathcal{F}_{h,\mathcal{X}}^{(1)}$ such that

$$\sup_{g\in\mathcal{F}_h^{(1)}} \sup_{x\in\mathbb{R}^{d+1}} |g(x)| \le B^{(1)}.$$

It follows from (6.52) that, for any given probability measure Q on $\mathbb{R}^{d+1}$ and any $\epsilon \in (0, B^{(1)}]$, there exist $x_1, \cdots, x_{N_\epsilon} \in \mathcal{X}$ with $N_\epsilon \le (\frac{AB^{(1)}}{\epsilon})^\nu$ such that

$\{m_{x_j,h} : j = 1, \cdots, N_\epsilon\}$ is an $(\frac{\epsilon}{L})$-covering of $\mathcal{F}_{h,\mathcal{X}}$ with respect to the $L_2(Q)$-norm. In other words, for any $x \in \mathcal{X}$, there exists $j \in \{1, \cdots, N_\epsilon\}$ such that $\|m_{x,h} - m_{x_j,h}\|_{L(Q)} \leq \frac{\epsilon}{L}$. Hence for any $m_{x,h} \in \mathcal{F}^{(1)}_{x,\mathcal{X}}$,

$$\|\zeta_h m_{x,h} - \zeta_h m_{x_j,h}\|_{L(Q)} \leq L\|m_{x,h} - m_{x_j,h}\|_{L(Q)} \leq \epsilon.$$

This means that $\{\zeta_h m_{x_j,h} : j = 1, \cdots, N_\epsilon\}$ is an ϵ-covering of $\mathcal{F}^{(1)}_{h,\mathcal{X}}$ with respect to the $L_2(Q)$-norm. Hence for any $m_{x,h} \in \mathcal{F}_{x,\mathcal{X}}$ and any $\epsilon \in (0, B^{(1)}]$,

$$\sup_Q \mathcal{N}(\mathcal{F}^{(1)}_{h,\mathcal{X}}, L_2(Q), \epsilon) \leq \Big(\frac{AB^{(1)}}{\epsilon}\Big)^\nu,$$

which implies that, for any given probability measure Q on $\mathbb{R}^{d+1}$ and any $\epsilon \in (0, 2B^{(1)}]$, there exist $x'_1, \cdots, x'_{N_\epsilon} \in \mathcal{X}$ with $N_\epsilon \leq (\frac{2AB^{(1)}}{\epsilon})^\nu$ such that $\{\zeta_h m_{x'_j,h} : j = 1, \cdots, N_\epsilon\}$ is a $(\frac{1}{2}\epsilon)$-covering of $\mathcal{F}^{(1)}_{h,\mathcal{X}}$ with respect to the $L_2(Q)$-norm.

Consider the interval $[-B^{(1)}, B^{(1)}]$. For any $\epsilon > 0$, there exist $b_1, \cdots, b_{N_\epsilon}$ with $N_\epsilon \leq \lceil 2B^{(1)}/\epsilon \rceil$ such that $b_1, \cdots, b_{N_\epsilon}$ is a $(\frac{1}{2}\epsilon)$-covering of $[-B^{(1)}, B^{(1)}]$, where $\lceil \cdot \rceil$ is the ceiling function. Let

$$\mathcal{F}^{(2)}_h = \{g(\cdot) - b : g \in \mathcal{F}^{(1)}_{h,\mathcal{X}}, |b| \leq B^{(1)}\}.$$

For any $g \in \mathcal{F}^{(1)}_{h,\mathcal{X}}$ and $|b| \leq B^{(1)}$, there exist $m_{x'_i,h}$ and b_j such that $\|m_{x,h} - m_{x'_i,h}\|_{L(Q)} \leq \frac{1}{2}\epsilon$ and $|b - b_j| \leq \frac{1}{2}\epsilon$. Hence

$$\|(m_{x,h} - b) - (m_{x'_i,h} - b_j)\|_{L_2(Q)} \leq \|m_{x,h} - m_{x'_i,h}\|_{L_2(Q)} + |b - b_j| \leq \epsilon. \quad (6.53)$$

Therefore with $A^{(2)} = 2(A \vee 1)$ and $B^{(2)} = 2B^{(1)}$ we have

$$\sup_Q \mathcal{N}(\mathcal{F}^{(2)}_{h,\mathcal{X}}, L_2(Q), \epsilon) \leq \Big(\frac{2AB^{(1)}}{\epsilon}\Big)^\nu \lceil 2B^{(1)}/\epsilon \rceil \leq \Big(\frac{A^{(2)}B^{(2)}}{\epsilon}\Big)^{\nu+1}.$$

Note that $\sup_{g \in \mathcal{F}^{(2)}_h} \sup_{x \in \mathbb{R}^{d+1}} |g(x)| \leq B^{(2)}$. Since $\widetilde{\mathcal{G}}_h \subset \mathcal{F}^{(2)}_{h,\mathcal{X}}$, we have

$$\sup_Q \mathcal{N}(\widetilde{\mathcal{G}}_h, L_2(Q), \epsilon) \leq \sup_Q \mathcal{N}(\mathcal{F}^{(2)}_{h,\mathcal{X}}, L_2(Q), \epsilon) \leq \Big(\frac{A^{(2)}B^{(2)}}{\epsilon}\Big)^{\nu+1}. \quad (6.54)$$

This then shows that $\widetilde{\mathcal{G}}_h$ is a VC class with characteristics $A^{(2)}$ and $\nu + 1$, and is uniformly bounded by a constant envelope $B^{(2)}$. For any $g \in \widetilde{\mathcal{G}}_h$, for all $h \in (0, h_0]$, we have

$$\begin{aligned}
\mathbb{P}g^2 = \mathbb{E}g^2(X, Y) &\leq \mathbb{E}\Big\{\Big[\zeta_h(X, Y)M\Big(\frac{x - X}{h}\Big)\Big]^2\Big\} \\
&\leq L^2\mathbb{E}\Big\{\Big[M\Big(\frac{x - X}{h}\Big)\Big]^2\Big\} \\
&\leq h^d L^2 \int_{\mathbb{R}^d} [M(w)]^2 f(x - hw)dw
\end{aligned}$$

$$\leq h^d L^2 \sup_{x \in \mathcal{X}^{\eta h_0}} |f(x)| \|M\|_2^2$$
$$:= h^d \sigma_0^2.$$

Applying Sriperumbudur and Steinwart (2012, Theorems A.1 and A.2) we have that for all $h \in (0, h_0]$, $n \geq 1$ and $\tau > 0$, with probability at least $1 - e^{-\tau}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sup_{g \in \mathcal{G}_h} |\mathbb{G}_n(g)| \\
\leq & 4 \frac{1}{\sqrt{n}} \mathbb{E} \sup_{g \in \mathcal{G}_h} |\mathbb{G}_n(g)| + \sqrt{\frac{2\tau\sigma_0^2}{nh^d}} + \frac{\tau B^{(2)}}{nh^d} \\
\leq & 4C \Big[\frac{(\nu+1)B^{(2)}}{nh^d} \log \frac{A^{(2)}B^{(2)}}{\sqrt{h^d\sigma_0^2}} + \sqrt{\frac{(\nu+1)\sigma_0^2}{nh^d}} \log \frac{A^{(2)}B^{(2)}}{\sqrt{h^d\sigma_0^2}} \Big] + \sqrt{\frac{2\tau\sigma_0^2}{nh^d}} + \frac{\tau B^{(2)}}{nh^d},
\end{aligned}$$

where C is a universal constant that is given in Sriperumbudur and Steinwart (2012, Theorem A.2). □

Acknowledgments

We are grateful to two anonymous referees for their helpful comments.

References

E. Arias-Castro, D. Mason, and B. Pelletier (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, **17**: 1–28. MR3491137

D.P. Bertsekas (1999) *Nonlinear Programming*, 2nd edition, Athena Scientific, Belmont, Massachusetts. MR3444832

L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone (1993). *Classification and Regression Trees*. Chapman and Hall. MR0726392

J.E. Chácon (2015). A population background for nonparametric density-based clustering. *Statistical Science*, **30**(4): 518–532. MR3432839

P. Chaudhuri, M.-C. Huang, W.-Y. Loh, R. Yao (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, **4**: 143–167. MR1347613

Y.-C. Chen, C. R. Genovese, J. Tibshirani, and L. Wasserman (2016). Nonparametric modal regression. *Ann. Statist.* **44**(2): 489–514. MR3476607

Y.-C. Chen, C. R. Genovese, and L. Wasserman (2016). A comprehensive approach to mode clustering. *Electron. J. Statist.* **10**(1): 210–241. MR3466181

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

V. Chernozhukov, D. Chetverikov, and K. Kato (2013). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42**(4): 1564–1597. MR3262461

R. Clausen, B. Ma, R. Nussinov, and A. Shehu (2015). Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Computational Biology* **11**(9).

D. Comaniciu and P. Meer (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5):1–18.

D. Comaniciu, V. Ramesh, and P. Meer (2003). Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5): 564–577.

J. Einbeck and G. Tutz (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *J. Roy. Statistical Soc.: Series C (Appl. Statist.)* **55**(4): 461–475. MR2242274

U. Einmahl and D.M. Mason (2000). Uniform in bandwidth consistency of kernel-type function estimators. *Annals of Statistics* **33**: 1380–1403. MR2195639

J.H. Friedman (1991). Multivariate adaptive regression splines. *Annals of Statistics* **19**(1):1–141. MR1091842

K. Fukunaga and L. D. Hostetler (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *EEE Transactions on Information Theory* **21**(1):32–40. MR0388638

J. Gaudart, N. Graffeo, G. Barbet, S. Rebaudet, N. Dessay, O. Doumbo, and R. Giorgi (2015). SPODT: An R Package to Perform Spatial Partitioning. *Journal of Statistical Software*, **63**(16).

J. Gaudart, B. Poudiougou, S. Ranque, and O. Doumbo (2005). Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk. *BMC Medical Research Methodology*, **5**(1), 1–11.

C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman (2014). Nonparametric ridge estimation. *Annals of Statistics*, **42**(4), 1511–1545. MR3262459

S. Gerber, O Rübel, P.T. Bremer, V. Pascucci, R.T. Whitaker (2013). Morse-Smale Regression. *J Comput Graph Stat*, **22**(1):193–214. MR3044330

Y. A. Ghassabeh (2015). A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel. *Journal of Multivariate Analysis* **135**: 1–10. MR3306422

E. Giné and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annals of the Institute Henri Poincaré: Probability and Statistics*, **38**: 907–921. MR1955344

D.J. Henderson, Q. Li, C.F. Parmeter, and S. Yao (2015). Gradient-based smoothing parameter selection for nonparametric regression estimation. *Journal of Econometrics* **184**: 233–241. MR3291000

D.J. Henderson, and C.F. Parmeter (2015). *Applied Nonparametric Econometrics*, Cambridge University Press.

L. Hubert and P. Arabie (1985). Comparing partitions, *Journal of classification*, **2**(1):193–218.

H. Jiang (2019). Non-asymptotic uniform rates of consistency for k-NN regression. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**(1): 3999–4006.

Y. P. Mack, and H.-G. Müller (1989). Derivative estimation in nonparametric regression with random predictor variables. *Sankhya* **51**:59–72, Ser. A. MR1065559

T. Maximova, E. Plaku, and A. Shehu (2016). Structure-guided protein transition modeling with a probabilistic roadmap algorithm. *IEEE/ACM transactions on computational biology and bioinformatics*, **15**(6), 1783–1796.

T. Maximova, Z. Zhang, D. B. Carr, E. Plaku, and A. Shehu (2018). Sample-based models of protein energy landscapes and slow structural rearrangements. *Journal of Computational Biology*, **25**(1): 33–50.

J. Legewie (2018). Living on the edge: neighborhood boundaries and the spatial dynamics of violent crime. *Demography*, **55**(5), 1957–1977.

B. Liu, B. Mavrin, D. Niu, and L. Kong (2016). House price modeling over heterogeneous regions with hierarchical spatial functional analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1047-1052.

J. Milnor (1963). *Morse Theory*, Princeton University Press. MR0163331

H.-G. Müller (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian Journal of Statistics* **12**(3): 221–232. MR0817940

H.-G. Müller (1989). Adaptive nonparametric peak estimation. *Annals of Statistics* **17**(3): 1053–1069. MR1015137

S. Mukherjee and D.X. Zhou (2006). Learning coordinate covariances via gradients. *Journal of Machine Learning Research* **7**(3), 519–549. MR2274377

D. Nolan and D. Pollard (1987). U-processes: rates of convergence. *Annals of Statistics* **15**(2): 780–799. MR0888439

U. Ozertem, and D. Erdogmus, (2011). Locally defined principal curves and surfaces. *The Journal of Machine Learning Research*, **12**, 1249–1286. MR2804600

W. Qiao, and W. Polonik (2016). Theoretical analysis of nonparametric filament estimation. *Annals of Statistics*, **44**(3), 1269–1297. MR3485960

D. Serre (2002). *Matrices: Theory and Applications*. Springer-Verlag, New York. MR1923507

B. Sriperumbudur, I. Steinwart (2012). Consistency and rates for clustering with DBSCAN. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, PMLR* **22**: 1090–1098.

A.B. Tsybakov (1990). Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii*, **26**(1), 38–45. MR1051586

M. P. Wand and M. C. Jones (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, **88**(422), 520–528. MR1224377

R. Yamasaki and T. Tanaka (2020). Properties of mean shift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(9): 2273–2286.

K. Ziegler (2002). On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric*

Statistics **14**(6): 749–774. MR1941713

H. Zhou and X. Huang (2019). Bandwidth selection for nonparametric modal regression. Communications in Statistics – Simulation and Computation, **48**(4), 968–984. MR3957561