Algorithms for ridge estimation with convergence guarantees

Wanli Qiao
Department of Statistics
George Mason University
4400 University Drive, MS 4A7
Fairfax, VA 22030, USA

Wolfgang Polonik
Department of Statistics
University of California
One Shields Ave.
Davis, CA 95616, USA

Abstract

The extraction of filamentary structure from a point cloud is discussed. The filaments are modeled as ridge lines or higher dimensional ridges of an underlying density. We propose two novel algorithms, and provide theoretical guarantees for their convergences. We consider the new algorithms as alternatives to the Subspace Constraint Mean Shift (SCMS) algorithm that do not suffer from a shortcoming of the SCMS that is also revealed in this paper.

Keywords: Filaments, Ridges, Manifold Learning, Mean Shift, Gradient Ascent

1. Introduction

The geometric interpretation of a ridge in $\mathbb{R}^d$ is that of a low-dimensional structure along which the density is higher than in the surrounding area when moving away from the set in an orthogonal direction. Blood vessels, road system, DNA strands or fault lines appearing in 2D or 3D images can be modeled as filaments, or maybe better as unions of filaments that might intersect, or that have a common starting point. We sometimes call such unions filamentary structures. Another example is provided by the filamentary structure that can be observed in the distribution of galaxies in the universe, the ‘cosmic web’. Cosmologists are interested in finding rigorous geometric and topological descriptions of the filamentary structures. Usually, the first step is the extraction of this structure, and this is the topic discussed below.

Ozertem and Erdogmus (2011) developed the popular SCMS (Subspace Constrained Mean Shift) algorithm for extracting estimates of $k$-dimensional ridges of a $d$-dimensional density $f$ from a point cloud sampled from $f$. The algorithm consists in running a corrected mean shift algorithm (subspace constrained) starting at a data point. For each data point the algorithm then provides an estimate of a point on the ridge. While the algorithm works well in practice, we here reveal a shortcoming of the algorithm: Some parts of the ridges might be missed by the algorithm. This is discussed below in some detail (see subsection 2.3).
Here we propose algorithms that do not suffer from the shortcoming of the SCMS algorithm just mentioned, and we also derive theoretical guarantees for their convergence. So far, no theoretical guarantees for the SCMS algorithm seem to be known.

The heuristic underlying our algorithm is simple: Consider a one-dimensional ridge line of a smooth density $f$. A point $x$ on the ridge line is characterized by the conditions (i) $V_1^\top(x) \nabla f(x) = 0$ and (ii) $\lambda_i(x) < 0$, $i = 2, \ldots, d$. Here $0$ is the zero-vector in $\mathbb{R}^{d-1}$, $\lambda_i(x)$ are the eigenvalues of the Hessian of $f$ at $x$, $\nabla f(x)$ denotes the gradient of $f$, and the columns of the $(d \times (d-1))$-matrix $V_\perp$ are the orthonormal eigenvectors of the Hessian of $f$ corresponding to $\lambda_2, \ldots, \lambda_d$. Observing that (i) is equivalent to $\|V_\perp^\top(x) \nabla f(x)\|^2 = 0$, we can find ridge points by finding the global minimizers of the real-valued function $r(x) = \|V_\perp^\top(x) \nabla f(x)\|^2$ subject to (ii). Given data $X_1, \ldots, X_n$ sampled from $f$, one can simply replace all the population quantities by the ones corresponding to a KDE (kernel density estimator) of $f$. We will devise subspace-constrained type algorithms for finding the entire set of global minimizers of the function $r$ subject to condition (ii), i.e., the ridge lines, and of higher-dimensional analogs.

The literature on ridge (or filament) detection is rich, and different approaches use different geometric ideas. For instance, Qiao and Polonik (2016) use the fact that integral curves driven by the second eigenvectors of the Hessians of the density $f$ can be used to extract ridges. Other approaches include the skeleton (Novikov et al. 2006), local principal curves (Einbeck et al. 2005, 2007), minimum spanning trees (Barrow et al. 1985), the candy model (Stoica et al. 2007), multiscale detection (Arias-Castro et al. 2006) and the path density (Genovese et al. 2009). Other related concepts include principal curves (Hastie et al. 1989) and the medial axis (Genovese et al. 2012). Also see Chen et al. (2015), Pulkinnen (2015), and Qiao (2021).

The remaining part of the paper is organized as follows. In Section 2 we introduce the definition of ridges. This is followed by our extraction algorithm, which is illustrated using a few datasets in $\mathbb{R}^2$. The main results are given in Subsection 2.9, where we show the convergence of our algorithms. All the mathematical details are given in Section 3.

2. Extraction of filamentary structures

2.1 Definition

We use the following definition of a ridge (filament) as given in Eberly (1996).

**Definition 1** (Ridge points in $\mathbb{R}^d$). Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a twice differentiable density function on $\mathbb{R}^d$. Suppose the Hessian $\nabla^2 f(x)$ has eigenvalues $\lambda_1(x) \geq \cdots \geq \lambda_k(x) > \lambda_{k+1}(x) \geq \cdots \geq \lambda_d(x)$ with $V_1(x), \cdots, V_d(x)$ being the corresponding orthonormal eigenvectors. For $1 \leq k \leq d-1$, a point $x$ is said to be a $k$-dimensional ridge point if

$$\langle \nabla f(x), V_i(x) \rangle = 0, \text{ for } i = k+1, \cdots, d,$$

and

$$\lambda_{k+1}(x) < 0.$$ 

The geometric intuition underlying this definition is clear: Since the (first order) directional derivative of $f(x)$ along $V_i(x)$ is $\langle \nabla f(x), V_i(x) \rangle$ and the second order directional derivative
Ridge Estimation Algorithms

of \( f(x) \) along \( V_i(x) \) is \( \lambda_i(x)\|V_i(x)\|^2 \), conditions (2.1) and (2.2) mean that the point \( x \) is a local mode in the linear subspace spanned by \( V_i(x) \), \( i = k + 1, \ldots, d \), for which the density has the largest concavity (see Eberly, 1996).

Let \( V_\perp = [V_{k+1}, \ldots, V_d] \in \mathbb{R}^{d \times (d-k)} \). For a twice differentiable function \( f \), a ridge is the set of ridge points, given by

\[
\text{Ridge}(f) = \{ x \in \mathbb{R}^d : V_\perp(x)^\top \nabla f(x) = 0 \text{ and } \lambda_{k+1}(x) < 0 \}. 
\]

If we have an estimator \( \hat{f} \) for \( f \), then \( \text{Ridge}(\hat{f}) \) is a plug-in estimator for \( \text{Ridge}(f) \). In this work we consider \( f \) to be a density function, and given an i.i.d. sample \( X_1, \ldots, X_n \) from \( f \), we use a kernel density estimator \( \hat{f} \) of the form

\[
\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left( \frac{X_i - x}{h} \right),
\]

where \( h > 0 \) is a bandwidth and \( \int_{\mathbb{R}^d} K(u) du = 1 \). Denote the eigenvalues of the Hessian \( \nabla^2 \hat{f}(x) \) by \( \hat{\lambda}_1(x) \geq \hat{\lambda}_2(x) \geq \cdots \geq \hat{\lambda}_d(x) \), and let \( \hat{V}_1(x), \ldots, \hat{V}_d(x) \) be the corresponding orthonormal eigenvectors. With \( \hat{V}_\perp = [\hat{V}_{k+1}, \ldots, \hat{V}_d] \), we then have

\[
\text{Ridge}(\hat{f}) = \{ x \in \mathbb{R}^d : \hat{V}_\perp(x)^\top \nabla \hat{f}(x) = 0 \text{ and } \hat{\lambda}_{k+1}(x) < 0 \}. 
\]

Of course, this only gives the definition of an estimated ridge. One still needs an algorithm to actually extract the estimated ridges from data.

2.2 Mean shift algorithm and subspace constrained mean shift algorithm

One of the popular algorithms for ridge extraction is called Subspace Constrained Mean Shift (SCMS), proposed in Ozertem and Erdogmus (2011). The algorithm uses the idea of mean shift, which is briefly reviewed first. The mean shift algorithm (Fukanaga and Hostetler, 1975) is essentially tracking non-parametric estimates of gradients. It is being used for mode finding and clustering, e.g. see Cheng (1995) and Comaniciu and Meer (1999, 2002). Considering the above kernel estimator with \( K(x) = \phi(\|x\|) \), the vector

\[
m(x) := \frac{\sum_{i=1}^n X_i \phi\left( \frac{\|x-X_i\|}{h} \right)}{\sum_{i=1}^n \phi\left( \frac{\|x-X_i\|}{h} \right)} - x
\]

is called mean shift. It is well known that the direction of the mean shift is an estimator of the direction of the gradient of \( f \) at \( x \). Given some initial position \( y_0 \), let the points \( y_1, y_2, \cdots \) be defined iteratively by

\[
y_{j+1} = m(y_j) + y_j, \quad j = 0, 1, 2, \cdots.
\]

Successively connecting these points provides an estimate of the integral curve driven by the gradient, starting from \( y_0 \). The endpoint of this iteration (after applying some stopping criterion) is considered to be an estimate of a mode of \( f \). This is the popular mean shift algorithm.

The idea of using a subspace constraint comes in when the target is a ridge. To this end, one modifies the just described hill climbing algorithm by following the direction of the
gradient projected on the subspace spanned by the eigenvectors of the Hessian corresponding to small eigenvalues, thereby exploiting the definition of a ridge point. In this algorithm, the gradient direction is approximated by the mean shift and the Hessian is estimated by the Hessian of the kernel density estimator. More specifically, given a starting point $y_0$, the SCMS generates a sequence

$$y_{j+1} = \hat{V}_\perp(y_j)\hat{V}_\perp(y_j)^T m(y_j) + y_j, \quad j = 0, 1, 2, \ldots$$

where $m(y)$ is the mean shift vector, and $\hat{V}_\perp(y)\hat{V}_\perp^T$ is the projection matrix onto the subspace spanned by the trailing $(d-k)$ eigenvectors of the estimated Hessian evaluated at $y$. Notice that for $y \in \operatorname{Ridge}(\hat{f})$, this space is orthogonal to the ridge (see below for more). This motivates that the endpoint of this iteration (after applying some stopping criterion) is considered to be an estimated ridge point. The SCMS algorithm is very popular, and it gives nice results in practice.

### 2.3 A shortcoming of the SCMS Algorithm

The main idea behind SCMS is to consider ridge points as local maxima on the trajectories along the tangent direction of $V_\perp V_i^T \nabla f$, the projection of $\nabla f$ onto the subspace spanned by the $(d-k)$ trailing eigenvectors of the estimated Hessian evaluated at $y$. This is the ascending direction that SCMS tries to estimate. Indeed, a ridge point $x_0$ is a local mode in the direction given by the columns of $V_\perp(x_0)$. In other words, a ridge point $x_0$ is a local mode in the linear subspace $\{x_0 + V_\perp(x_0)^T t : t \in \mathbb{R}^{d-k}\}$. This is because for $i = 1, \ldots, n$, the sign of $\lambda_i(x_0)$ determines the sign of the second order directional derivative in the direction of $V_i(x_0)$, assuming that the eigenvalues are simple, i.e., without ties, that is,

$$\nabla V_i(x_0) (\nabla V_i(x_0) f(x)) = \lambda_i(x_0) \|V_i(x_0)\|^2,$$

where $\nabla V_i(x_0)$ is the directional derivative along $V_i(x_0)$. Thus, the ridge point $x_0$ yielding a local maximum of $f$ in the directions of $V_i(x_0)$, $i = k+1, \ldots, d$ follows from the condition in (2.1) that $\lambda_i(x_0) < 0$ for $i = k+1, \ldots, d$, and the first order condition in (2.1). However, when running the SCMS algorithm, we of course do not know which point is a ridge point. Therefore, at a given step of the algorithm, we project the gradient at the current point $y_j$, say, on the space spanned by $\{V_i(y_j), i = k+1, \ldots, d\}$. Implicitly this means that we view the $V_i$’s as functions of $x$, and in this case, the derivatives are taken constrained in the corresponding integral surface, and because of this, the signs of the directional derivatives are usually not determined by the signs of the $\lambda_i$’s. As a consequence, ridge points are not necessarily local modes along the corresponding integral curves, but they can also be local minimizers or saddle points, as we explain now for $d = 2$. For a given point $x_0$ near the ridge, we consider an integral curve $x(t)$ defined as

$$\frac{dx(t)}{dt} = V_\perp(x(t)), \quad x(0) = x_0,$$

where $V_\perp$ is the second unit eigenvector of the Hessian. Of course each eigenvector has two opposite orientations, and we assume that the direction of $V_\perp(x)$ is determined such that it varies continuously with $x$. Note that $V_\perp$ is parallel to $V_\perp V_\perp^T \nabla f$ and so $x(t)$ has
the same trajectory as the integral curve driven by \( V_{\perp} V_{\perp}^T \nabla f \). Using \( V_{\perp} \) has the advantage that allows tracking integral curves both forward and backward. Indeed, the vector field \( V_{\perp} V_{\perp}^T \nabla f \) vanishes on the ridge, while \( V_{\perp} \) always has a unit length. Suppose that there exists an interval \((a, b)\) such that \( \{x(t) : t \in (a, b)\} \) intersects with Ridge\((f)\). Then the first and second order derivatives of \( f(x(t)) \) with respect to \( t \) are

\[
\frac{df(x(t))}{dt} = \langle \nabla f(x(t)), V_{\perp}(x(t)) \rangle \tag{2.8}
\]

and

\[
\frac{d^2 f(x(t))}{dt^2} = \nabla \langle \nabla f(x(t)), V_{\perp}(x(t)) \rangle, V_{\perp}(x(t)) \rangle
= \nabla f(x(t)) \nabla V_{\perp}(x(t)) V_{\perp}(x(t)) + \langle \nabla^2 f(x(t)) V_{\perp}(x(t)), V_{\perp}(x(t)) \rangle
= \nabla f(x(t)) \nabla V_{\perp}(x(t)) V_{\perp}(x(t)) + \lambda_2(x(t)) \|V_{\perp}(x(t))\|^2. \tag{2.9}
\]

If \( x(t) \) is a ridge point, then \( \frac{df(x(t))}{dt} = 0 \) in (2.8) by Definition 1, but \( \frac{d^2 f(x(t))}{dt^2} \) is not always negative. In other words, depending on the sign in (2.9), ridge points on the trajectory driven by \( V_{\perp} \) (or equivalently, by \( V_{\perp} V_{\perp}^T \nabla f \)) can be local maxima, local minima or even saddle points. An example is given below, showing that following the direction of \( V_{\perp} V_{\perp}^T \nabla f \), a piece of ridge will be missed if the starting points are not chosen exactly on that piece.

**Example 1** Consider the density function

\[
f(u, v) = \frac{3}{8} (1 - u^2) v, \quad u \in [-1, 1], \quad v \in [0, 2].
\]

Using the ridge point definition, two curves are detected: \( \{u = 0, v \in (0, 2)\} \) and \( \{(u, v) : v = \frac{1 - u^2}{\sqrt{2 + 2u^2}}, u \in (-1, 1)\} \). The intersection of the two curves is at \( (0, \frac{1}{\sqrt{2}}) \). The plots of this function and the ridge are given in Figures 2.1 and 2.2.

**Figure 2.1:** Surface plot of the function \( f \).

**Figure 2.2:** Contour lines (black), ridge lines (red), and integral curves driven by \( V_{\perp} \) (green).
In Figures 2.3 and 2.4 we compare the ideas behind the original SCMS algorithm and our new algorithms. It can be found that a piece of the ridge is failed to be detected using the idea of the original SCMS algorithm in this example.

Figure 2.3: green curves are the trajectories of $V_\perp V_\perp^\top \nabla f$, and the red dots are the limit points.

Figure 2.4: green curves are the trajectories of $-\nabla r$, and the red dots are the limit points.

2.4 Measuring ridgeness

Our new ridge finding algorithms to be described below, are designed to avoid the issue identified above. Their construction is based on the following ridgeness measure (which actually measures ‘non-ridgeness’):

$$\eta(x) = -\frac{1}{2} \| V_\perp(x)^\top \nabla f(x) \|^2.$$  \hfill (2.10)

Since we will assume that $\lambda_k(x) \neq \lambda_{k+1}(x)$ for all $x$ (see assumption (A2) below), $\eta(x)$ is well-defined. According to their definition, ridge points can be described as

$$\eta(x) = 0 \quad \text{s.t.} \quad \lambda_{k+1}(x) < 0.$$  \hfill (2.11)

Since $\eta \leq 0$, ridge points are global maximizers of $\eta$. In other words, ridge points can be described as

$$\argmax_x \eta(x) \quad \text{s.t.} \quad \lambda_{k+1}(x) < 0.$$  

With this formulation, finding ridges means finding global maximizers of a function plus checking the eigenvalue condition. Notice, however, that in this case, the global maximizers of interest are not isolated points, but we rather have an entire continuum of global maximizers - the ridge points.

Of course, $\eta(x)$ depends on population quantities and is thus unknown. A data-dependent version is given by

$$\hat{\eta}(x) = -\frac{1}{2} \| \hat{V}_\perp(x)^\top \nabla \hat{f}(x) \|^2,$$  \hfill (2.12)
and the corresponding empirical optimization problem becomes finding

$$\arg\max_x \hat{\eta}(x) \quad \text{s.t.} \quad \hat{\lambda}_{k+1}(x) < 0.$$ 

### 2.5 Algorithms

To compute the minimizers of the ridgeness function we now consider two algorithms based on \(\hat{\eta}(x)\). Recall that the columns of \(\hat{V}_\perp(x) \in \mathbb{R}^{d \times (d-k)}\) are the orthonormal eigenvectors of the Hessian of \(\hat{f}(x)\) corresponding to its \((d-k)\) smallest eigenvalues. Similarly, let the columns of \(\hat{V}_{k+1}^*(x) \in \mathbb{R}^{d \times (d-k)}\) be the trailing orthonormal eigenvectors \(\hat{V}_{k+1}^*(x), \ldots, \hat{V}_d^*(x)\) of the Hessian \(\nabla^2 \hat{\eta}(x)\), and let \(\hat{\lambda}_1^*(x) \geq \hat{\lambda}_2^*(x) \geq \cdots \geq \hat{\lambda}_d^*(x)\) denote the eigenvalues of \(\nabla^2 \hat{\eta}(x)\). Further, let

\[
\hat{\xi}(x) = \hat{V}_\perp(x)\hat{V}_\perp(x)^\top \nabla \hat{f}(x), \\
\hat{\xi}^*(x) = \hat{V}_{k+1}^*(x)[\hat{V}_{k+1}^*(x)]^\top \nabla \hat{\eta}(x).
\]

In other words, \(\hat{\xi}(x)\) is the projection of \(\nabla \hat{f}(x)\) onto the subspace spanned by the columns of \(\hat{V}_\perp(x)\), and similarly, \(\hat{\xi}^*(x)\) is the projection of \(\nabla \hat{\eta}(x)\) onto the subspace spanned by the columns of \(\hat{V}_{k+1}^*(x)\). Note that with this notation we have

$$\hat{\eta}(x) = -\frac{1}{2}\|\hat{\xi}(x)\|^2.$$ 

We will see below (see Lemma 10) that the ridge of the ridgeness function essentially equals the original ridge of \(f\), and we will use this fact.

In the following, we assume that all the functions considered are defined on \([0,1]^d\).

**Basic Algorithm 1:** *Alternative SCMS approach using an estimated ridgeness function.*

Observing that \(\nabla \hat{\eta}(x) = -[\nabla \hat{\xi}(x)]^\top \hat{\xi}(x)\), we have the following SCMS-type algorithm: Given \(a > 0\) (step size) and a starting point \(y_0\), we update through

\[
y_{j+1} = y_j + a \hat{\xi}^*(y_j) \\
= y_j - a \hat{V}_{k+1}^*(y_j)[\hat{V}_{k+1}^*(y_j)]^\top [\nabla \hat{\xi}(y_j)]^\top \hat{\xi}(y_j) \\
= y_j - a \hat{V}_{k+1}^*(y_j)[\hat{V}_{k+1}^*(y_j)]^\top [\nabla \hat{\xi}(y_j)]^\top \hat{V}_\perp(y_j) \hat{V}_\perp(y_j)^\top \nabla \hat{f}(y_j), \quad j = 0, 1, 2, \ldots
\]

More precisely, the algorithm is as follows:

**Input:** \(y_i^0 = X_i, \ i = 1, \ldots, n, \ a > 0, \ \varepsilon_{\text{tol}} > 0\).

**Update:** For \(i = 0, 1, 2, \ldots, n\), if \(y_i^j \in [0,1]^d\), run

\[
y_{j+1} = y_j - a \hat{V}_{k+1}^*(y_j)[\hat{V}_{k+1}^*(y_j)]^\top [\nabla \hat{\xi}(y_j)]^\top \hat{V}_\perp(y_j) \hat{V}_\perp(y_j)^\top \nabla \hat{f}(y_j)
\]

until \(\|y_{j+1} - y_j^j\| \leq \varepsilon_{\text{tol}}\); otherwise discard the sequence. Let \(y_i^*\) be the last point before stopping.
\textbf{Output:} \( \{ y_i^* : \lambda_{k+1}(y_i^*) < 0 \} \).

In other words, in the output step, we remove points that do not comply with the condition for eigenvalues in the definition of ridges, because this condition is not being taken into account when constructing the iterations of the algorithm.

\textbf{Basic Algorithm 2:} \textit{Alternative SCMS approach using a smoothed estimated ridgeness function.}

Let \( \tau > 0 \) be another bandwidth and \( L : \mathbb{R}^d \to \mathbb{R}_{\geq 0} \) be a twice differentiable kernel function. Define a smoothed version of the ridgeness function \( \hat{\eta}(x) \) as

\[
\hat{\eta}_\tau(x) = \frac{1}{\tau^d} \int_{\mathbb{R}^d} L \left( \frac{x - u}{\tau} \right) \hat{\eta}(u) du. \tag{2.14}
\]

Our algorithm will approximate the ridge of this smoothed version of the ridgeness function. Using this smoothed version has some computational advantages (see subsection 2.6.3 below). Let \( \hat{V}_\tau^T(x) \) be the matrix whose columns are the last \((d-k)\) orthonormal eigenvectors of \( \nabla^2 \hat{\eta}_\tau(x) \), and let \( \hat{\xi}_\tau^r(x) = \hat{V}_\tau^T(x)[\hat{V}_\tau^T(x)]^\top \nabla \hat{\eta}_\tau(x) \). With this notation, the algorithm is as follows:

\textbf{Input:} Given \( y_i^0 = X_i, i = 1, \cdots, n, a > 0, \tau > 0, \varepsilon_{\text{tol}} > 0 \).

\textbf{Update:} For \( i = 0, 1, 2, \cdots, n \), if \( y_i \in [0, 1]^d \) run

\[
y_i^{j+1} = y_i^j + a \hat{\xi}_\tau^r(y_i^j) = y_i^j + a \hat{V}_\tau^T(y_i^j)[\hat{V}_\tau^T(y_i^j)]^\top \nabla \hat{\eta}_\tau(y_i^j),
\]

until \( \|y_i^{j+1} - y_i^j\| \leq \varepsilon_{\text{tol}} \); otherwise discard the sequence. Let \( y_i^* \) be the last point before stopping.

\textbf{Output:} \( \{ y_i^* : \lambda_{r+1}(y_i^*) < 0 \} \).

\subsection*{2.6 Practical implementation and illustration of the algorithms}

For practical implementation, the basic algorithms given above require some additional pre- and post-processing. This, along with some other aspects that are of some importance in the actual implementation of the algorithms, are discussed in the following.

In practice our algorithms can suffer from the following two challenges. The first is posed by low density regions, where the estimated density tends to be flat leading to possible spurious ridge points identified by the algorithms. A second challenge is possible local (but non-global) modes of our ridgeness function \( \eta \), which again might lead to spurious ridge points. We address these challenges by introducing the following pre-processing and post-processing steps.
2.6.1 Pre-processing

While in our basic algorithms (and also in the theory presented below), we assume that
all the ridges considered are defined on $[0,1]^d$, which is supposed to be contained in the
support of $f$. In the actual implementation we are replacing $[0,1]^d$ by the set
\[ \{ x \in \mathbb{R}^d : \hat{f}(x) \geq \varepsilon_f \} \]
for a given threshold $\varepsilon_f > 0$. We choose $\varepsilon_f$ as an $\alpha$-quantile of the distribution
\[ \{ \hat{f}(X_1), \ldots, \hat{f}(X_n) \} \].
In our theoretical investigations, we assume that the density
at the ridge points is larger than $2\varepsilon_f$, say. Using a consistent data-dependent estimate for $\varepsilon_f$
could also be dealt with theoretically, even though we are not explicitly considering this
in the theory section.

2.6.2 Post-processing

Low density points are now excluded by our pre-processing step.

To address possible local maxima of the functions $\hat{\eta}(x)$ and $\hat{\eta}_\tau(x)$, respectively, we remove
an output point $y_i^*$, if $\hat{\eta}(y_i^*) \leq -\varepsilon_\eta$, or $\hat{\eta}_\tau(y_i^*) \leq -\varepsilon_\eta$, respectively, for some $\varepsilon_\eta > 0$. Here we
utilize the fact that $\eta(x) \leq 0$ and all the ridge points of $\eta(x)$ satisfy $\eta(x) = 0$.

A practical choice of $\varepsilon_\eta$ is as follows. Consider the quantile function of the (one-
dimensional, discrete) distribution given by the values $\{ \hat{\eta}(X_1), \ldots, \hat{\eta}(X_n) \}$, and choose $\varepsilon_\eta$ as the location of the ‘last significant jump’ of the quantile function. In other words, we are
removing ‘outliers’ in this distribution. In our numerical experiments, this first significant
jump was clearly visible. An example is given in Figures 2.5 and 2.6 below, which are based
on Algorithm 2 with step length $a = 0.005$, while Algorithm 1 gives very similar results.

Figure 2.5: X-cross: 200 data points. Black solid dots are data points; Blue circles are the points
removed by pre-processing; Red dots are point removed by post-processing; Green dots are the final
output of the algorithm.
Figure 2.6: Sorted ridgeness values of the output of Algorithm 2 run on the 200 data points with x-cross as ridges without imposing the threshold $\varepsilon_\eta$. A clear jump is visible between the ridgeness values close to zero and those of non-global modes. This observation then informed our choice of $\varepsilon_\eta = -0.1$. The points below this threshold correspond to the red dots in the left panel of Figure 2.5.

Our algorithms are also implemented using two additional data sets, as shown Figures 2.7 and 2.8, for which we only present the final results after the pre-processing and post-processing steps, with the black and green dots having the same meaning as in Figure 2.5.

Figure 2.7: Circle: 200 data points.  
Figure 2.8: Spiral: 200 data points.
2.6.3 Other more practical aspects

In Algorithm 1, we need to compute $\nabla \hat{\xi}(x)$ and eigenvalues and eigenvectors of $\nabla^2 \hat{\eta}(x)$. In the following we present the formulas for them. Let $\text{vec}$ be the matrix vectorization operator such that $\text{vec}(A)$ stacks all the columns of a matrix $A$ into a vector.

As for $\nabla \hat{\xi}(x)$, we have the following by using the product rule for matrix calculus:

$$
\nabla \hat{\xi}(x) = (\nabla \hat{f}(x)^\top \otimes \mathbf{I}_d) \nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top] + \hat{V}_\perp(x) \hat{V}_\perp(x)^\top \nabla^2 \hat{f}(x) \\
= (\nabla \hat{f}(x)^\top \otimes \mathbf{I}_d) \nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top] + \hat{V}_\perp(x) \hat{A}_\perp(x) \hat{V}_\perp(x)^\top,
$$

(2.15)

where $\hat{A}_\perp = \text{diag}(\hat{\lambda}_{r+1}, \ldots, \hat{\lambda}_d)$, $\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top] = \frac{d \text{vec}[\hat{V}_\perp(x) \hat{V}_\perp(x)^\top]}{dx} \in \mathbb{R}^{d^2 \times d}$, and $\otimes$ denotes Kronecker product.

The Hessian $\nabla^2 \hat{\eta}(x)$ is given by

$$
\nabla^2 \hat{\eta}(x) = - (\mathbf{I}_d \otimes \xi(x)^\top) \nabla (\nabla \hat{\xi}(x)) - \nabla \hat{\xi}(x)^\top \nabla \hat{\xi}(x),
$$

(2.16)

where $\nabla (\nabla \hat{\xi}(x))$ can be found by using the product rule:

$$
\nabla (\nabla \hat{\xi}(x)) = [(\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top])^\top \otimes \mathbf{I}_d] \nabla (\nabla \hat{f}(x)^\top \otimes \mathbf{I}_d) \\
+ (\mathbf{I}_d \otimes (\nabla \hat{f}(x)^\top \otimes \mathbf{I}_d)) \nabla (\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top]) \\
+ (\nabla^2 \hat{f}(x) \otimes \mathbf{I}_d) \nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top] \\
+ (\mathbf{I}_d \otimes [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top]) \nabla (\nabla^2 \hat{f}(x)).
$$

Here $\nabla (\nabla \hat{f}(x)^\top \otimes \mathbf{I}_d) = (\mathbf{I}_d \otimes \text{vec}(\mathbf{I}_d)) \nabla^2 \hat{f}(x)$. Furthermore, we can find the expressions of $\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top]$ and $\nabla (\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top])$, which by the chain rule involves the third and fourth derivatives of $\hat{f}$, respectively. For example, for $\ell = 1, \ldots, d$, the $\ell$th column of $\nabla [\hat{V}_\perp(x) \hat{V}_\perp(x)^\top]$ is given by the vectorization of

$$
\sum_{i=1}^{k} \sum_{j=k+1}^{d} \frac{1}{\hat{\lambda}_j(x) - \hat{\lambda}_i(x)} [\hat{V}_j(x) \hat{V}_j(x)^\top \hat{H}_\ell(x) \hat{V}_i(x) \hat{V}_i(x)^\top + \hat{V}_i(x) \hat{V}_i(x)^\top \hat{H}_\ell(x) \hat{V}_j(x) \hat{V}_j(x)^\top],
$$

(2.17)

where $\hat{H}_\ell(x)$ is the partial derivative of $\nabla^2 \hat{f}(x)$ with respect to the $\ell$th component of $x$.

See Qiao (2020) for some relevant calculations.

Algorithm 2 avoids direct computation of these third and fourth derivatives of $\hat{f}$, and thus has some computational advantages in practice. Indeed the connection between the two algorithms can be understood in such a way that the symbolic computation of $\nabla \hat{\eta}$ and $\nabla^2 \hat{\eta}$ in Algorithm 1 are replaced by their numerical approximations ($\nabla \hat{\eta}_r$ and $\nabla^2 \hat{\eta}_r$) based on the evaluations of $\hat{\eta}$ in Algorithm 2. In practice, we implement the computation of $\nabla \hat{\eta}_r$ and $\nabla^2 \hat{\eta}_r$ as follows. Let $\{x_i, i = 1, 2, \ldots\}$ be a grid over $\mathbb{R}^d$ with grid length $\rho < \tau$. Then $\nabla \hat{\eta}_r$ is approximated by

$$
\frac{\rho^d}{\tau^d} \sum_i \nabla \left[L \left(\frac{x - x_i}{\tau}\right)\right] \hat{\eta}(x_i),
$$

and $\nabla^2 \hat{\eta}_r$ can be approximated in a similar way. The kernel $L$ is often chosen with bounded support or as the Gaussian kernel with truncation so that there are only a limited number of grid points involved in the above summation.
2.7 Other variants

Here we briefly discuss two other variants of our algorithm.

The first variant is a SCMS type algorithm. The original mean shift algorithm proposed by Fukunaga and Hostetler (1975) is a gradient ascent algorithm using the logarithm transformation of kernel density estimates, which is analyzed in Arias-Castro et al. (2016). More specifically, for any $x_0 \in \mathbb{R}$, they consider the sequence $x_{j+1} = x_j + a \nabla \log \hat{f}(x_j)$, $j = 0, 1, 2, \cdots$, whose limit is a local mode of $\hat{f}$, provided some regularity assumptions hold.

A similar idea can be applied to our algorithms as well. Let $g : \mathbb{R}_{\leq 0} \to \mathbb{R}_{> 0}$ be a (known) twice differentiable increasing positive function. In Algorithm 1, we can replace the derivatives (and the induced eigenvalues and eigenvectors) of $\hat{\eta}$ by those of $\log(g(\hat{\eta}))$. Note that $\hat{\eta}$ by definition is non-positive and this explains the need for a positive transformation function $g$. More specifically, let $\hat{V}_{g, \perp}(x)$ be a $d \times (d - k)$ matrix whose columns consist of the orthonormal eigenvectors of the Hessian of $\log(g(\hat{\eta}))$ corresponding to its smallest $(d - k)$ eigenvalues. Let $\hat{\xi}^*_{g}(x) = \hat{V}_{g, \perp}(x)[\hat{V}_{g, \perp}(x)]^\top \nabla \log(g(\hat{\eta}(x)))$. Then the update step (2.13) in Algorithm 1 can be replaced by

$$y_{i}^{j+1} = y_{i}^{j} + a \hat{\xi}^*_{g}(y_{i}^{j}),$$

which is a SCMS type algorithm. Algorithm 2 can be modified in a similar way.

The second variant is based on modified objective functions in the optimization. For some $q > 0$, and assuming that $\hat{f}(x) > 0$, write

$$\hat{s}_q(x) = \frac{\hat{\eta}(x)}{|\hat{f}(x)|^q}.$$

We have that $\hat{s}_q \leq 0$ and $\hat{s}_q = 0$ on Ridge($\hat{f}$), and finding minimizers of $\hat{s}_q$ and $\hat{\eta}$ is equivalent. Thus we can replace the derivatives (and the induced eigenvalues and eigenvectors) of $\hat{\eta}$ by those of $\hat{s}_q$. Recall that our basic algorithms can possibly return non-global modes of the ridgeness function depending on the start points, and thus requires a pruning process (see subsection 2.6). The numerator in (2.18) is a penalization for low density. The heuristic underlying the introduction of this penalization is to sharpen the ridgeness function near the ridge, so as to potentially accelerate the algorithm and enlarge the set of starting points whose corresponding limit points are global maxima.

The theoretical analyses of these two variants are not explicitly given in this paper, although they are straightforward extensions by following the same procedure as presented for Algorithms 1 and 2 below.

2.8 Notation, assumptions and some technical implications

Recall that $\hat{\xi}^*(x)$ and $\hat{\xi}^*_\tau(x)$ denotes the projections of the gradients of $\hat{\eta}(x)$ and $\hat{\eta}_\tau(x)$, respectively, onto the subspace spanned by the trailing $(d - k)$ eigenvectors of $\nabla^2 \hat{\eta}$ and $\nabla^2 \hat{\eta}_\tau$, respectively. Similarly, we define

$$\xi(x) = V_\perp(x) V_\perp(x)^\top \nabla f(x),$$

$$\xi^*(x) = V_\perp^*(x) [V_\perp^*(x)]^\top \nabla \eta(x),$$

$$\xi^*_\tau(x) = V_\perp^*_\tau(x) [V_\perp^*_\tau(x)]^\top \nabla \eta_\tau(x).$$
where $V_\perp = [V_{k+1}, \ldots, V_d]$ and $V_\perp^* = [V_{k+1}^*, \ldots, V_d^*]$ denote the $d \times (d - k)$ matrices formed by the trailing $(d - k)$ orthonormal eigenvectors of $\nabla^2 f(x)$ and $\nabla^2 \eta(x)$, respectively. The eigenvalues of the Hessians $\nabla^2 f(x)$ and $\nabla^2 \eta(x)$ are denoted by $\lambda_1 \geq \cdots \geq \lambda_d$ and $\lambda_1^* \geq \cdots \geq \lambda_d^*$, respectively. For $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$, let $|\alpha| = \alpha_1 + \cdots + \alpha_d$. For a function $g : \mathbb{R}^d \to \mathbb{R}$ with partial derivatives of order $|\alpha|$, define
\[
\partial^{(\alpha)} g(x) = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} g(x), \quad x \in \mathbb{R}^d.
\]

With slight abuse of notation, we use $\text{Ridge}(g)$ to denote the ridge of any twice differentiable function $g$ restricted to $[0, 1]^d$.

In the formulations of our theoretical results, we will use the following assumptions. Let $m$ be a positive integer (where our results will require $m \geq 4$).

(A1)$_{f,m}$ $f > 0$ is a density on $\mathbb{R}^d$ such that, for some $\varepsilon_0 > 0$, $[-\varepsilon_0, 1 + \varepsilon_0]^d$ is contained in the support of $f$; the partial derivatives of $f$ up to order $m$ exist and are bounded and continuous.

For $\delta > 0$ and $A \subset \mathbb{R}^d$, we define $A^\delta = \bigcup_{x \in A} B(x, \delta)$, where $B(x, \delta)$ denotes the (open) ball with midpoint $x$ and radius $\delta$.

(A2)$_f$ There exist $\beta, \delta > 0$ such that: $|\lambda_{k+1}(x)| > \beta$ for all $x \in \{t \in [0, 1]^d : \eta(t) = 0\}^\delta$, and $\lambda_k(x) - \lambda_{k+1}(x) < \beta$, and for all $x \in [0, 1]^d$. Furthermore, $\text{Ridge}(f)^\delta \subset [0, 1]^d$.

(A3)$_f$ For all $x \in \text{Ridge}(f)$, $\nabla \xi(x) \in \mathbb{R}^{d \times d}$ has rank $d - k$.

(K)$_m$ The kernel function $K : \mathbb{R}^d \to [0, \infty)$ is spherically symmetric and integrates to 1. All the partial derivatives up to order $m$ exist and are continuous and bounded. Moreover, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq m$, the class of functions
\[
\{ x \mapsto \partial^{(\alpha)} K\left(\frac{x - y}{h}\right) : y \in \mathbb{R}^d, h > 0 \}
\]
is a VC-class (see van der Vaart and Wellner, 1996). Also assume $\int_{\mathbb{R}^d} K(x)||x||^2 dx < \infty$.

(L) The kernel function $L : \mathbb{R}^d \to [0, \infty)$ is spherically symmetric with bounded support and integrates to 1. The partial derivatives of $L$ up to order four exist and are continuous.

Discussion of the assumptions: (i) Assumption (A1)$_{f,m}$ is made to avoid boundary issues of kernel density estimation on $[0, 1]^d$. The unit cube can of course be replaced by any compact set in $\mathbb{R}^d$. (ii) Assumption (A2)$_f$ is typical in the literature of ridge estimation (see, e.g., Assumption A1 in Genovese et al., 2014). It avoids spurious ridge points under small perturbation. (iii) Assumption (A3)$_f$ implies that $\text{Ridge}(f)$ is a manifold (without boundary). E.g., see Theorem 5.12 in Lee (2013). In particular, this prevents the existence of intersections of the ridge. For ridges with boundary or intersections, our results (except Theorem 16 (iv)) still apply to the part on the ridges strictly bounded away from the
boundary and intersections. (v) The VC-class assumption in (K)_m holds if K is of the form
\( K(x) = \phi(p(x)) \) with \( \phi : \mathbb{R} \to \mathbb{R}_{\geq 0} \) bounded and of bounded variation, and \( p(x) \) a polynomial (see Nolan and Pollard, 1987). In particular this is the case for the Gaussian kernel. (vi) Under the above assumptions, \( \eta(x) \) is twice differentiable and the second derivatives are Lipschitz continuous. In particular, the Hessian \( \nabla^2 \eta(x) \) is well-defined, and we have the following properties:

**Lemma 2** Under (A1)_f, (A2)_f, and (A3)_f, we have \( \lambda_1^*(x) = \cdots = \lambda_k^*(x) = 0 \) for all \( x \in \text{Ridge}(f) \), and there exist constants \( A, \alpha, \delta > 0 \), such that for all \( x \in \text{Ridge}(f)^\delta \),

\[
-\alpha \geq \lambda_{k+1}^*(x) \cdots \geq \lambda_d^*(x) \geq -A. \tag{2.19}
\]

Moreover, the columns of \( V^*_\perp(x) \) span the normal space of \( \text{Ridge}(f) \) at \( x \in \text{Ridge}(f) \).

**Proof** First we show that \( \text{Ridge}(f) \) is a compact set. Notice that from assumption (A2)_f,

\[
\text{Ridge}(f) = \{ x \in [0,1]^d : \xi(x) = 0, \lambda_{k+1}(x) \leq 0 \},
\]

because the set \{ \( x \in [0,1]^d : \xi(x) = 0, \lambda_{k+1}(x) = 0 \) \} is empty. If we treat \( V^*_\perp V^\top_\perp \) as a function of \( \nabla^2 \eta \), due to the assumed positive gap between \( \lambda_k(x) \) and \( \lambda_{k+1}(x) \) in assumption (A2)_f, \( V^*_\perp V^\top_\perp \) is an analytical matrix-valued function on the space of real symmetric matrices by the classical matrix perturbation theory (see Kato, 1995), which further implies that \( V^*_\perp V^\top_\perp \) is a Lipschitz function of \( x \) by assumption (A1)_f. Since both \( \xi(x) \) and \( \lambda_{k+1}(x) \) are continuous functions, \( \text{Ridge}(f) \) is closed, and hence compact because it is defined on the compact set \([0,1]^d\).

We have \( \nabla^2 \eta(x) = -(I_d \otimes \xi(x)^\top) \nabla(\nabla \xi(x)) - \nabla \xi(x)^\top \nabla \xi(x) \). Thus, for \( x \in \text{Ridge}(f) \), \( \nabla^2 \eta(x) = -\nabla \xi(x)^\top \nabla \xi(x) \), which, by using assumption (A3)_f, implies that rank of \( \nabla^2 \eta(x) \) is \( d - k \). Hence we have for \( x \in \text{Ridge}(f) \),

\[
0 = \lambda_1^*(x) = \cdots = \lambda_k^*(x) > \lambda_{k+1}^*(x) \cdots \geq \lambda_d^*(x).
\]

Since \( \xi(x) = 0 \) for all \( x \in \text{Ridge}(f) \), under assumption (A1)_f, there exists a constant \( c_0 > 0 \) such that for \( \delta' > 0 \) small enough, \( \| (I_d \otimes \xi(x)^\top) \nabla(\nabla \xi(x)) \|_F \leq c_0 \delta' \) for all \( x \in \text{Ridge}(f)^{\delta'} \). Also there exists a constant \( c_1 > 0 \) such that for all \( x, y \in \text{Ridge}(f)^{\delta'} \), \( \| \nabla \xi(x)^\top \nabla \xi(x) - \nabla \xi(y)^\top \nabla \xi(y) \|_F \leq c_1 \| x - y \| \). By using Weyl’s inequality (see page 15 of Serre, 2002), we have \( |\lambda_j^*(x) - \lambda_j^*(y)| \leq 2c_0 \delta' + c_1 \| x - y \| \), for all \( x, y \in \text{Ridge}(f)^{\delta'} \). Therefore (2.19) holds for \( \delta > 0 \) small enough.

Since \( \xi(x) = 0 \) for all \( x \in \text{Ridge}(f) \), the row vectors of \( \nabla \xi(x) \) span the normal space of \( \text{Ridge}(f) \) at \( x \), denoted by \( N(x) \). Note that \( N(x) = \{ \nabla \xi(x)^\top a : a \in \mathbb{R}^d \} = \{ \nabla \xi(x)^\top \nabla \xi(x) a : a \in \mathbb{R}^d \} \), which is the same as the space spanned by the eigenvectors of \( \nabla^2 \eta(x) \) corresponding to the non-zero eigenvalues, i.e., the space spanned by the columns of \( V^*_\perp(x) \).
2.9 Main results

We will need the following notation:

\[ S_\epsilon := \{ x \in [0,1]^d : \eta(x) \geq -\epsilon, \lambda_{k+1}(x) < 0 \}. \]  

(2.20)

Similarly, we define \( \hat{S}_\epsilon \) and \( \hat{S}_{r,\epsilon} \) by replacing \( \eta(x) \) with \( \hat{\eta}(x) \) and \( \hat{\eta}_r(x) \), respectively, and by also replacing \( \lambda_{k+1}(x) \) with \( \hat{\lambda}_{k+1}(x) \).

For any subset \( A \subset \mathbb{R}^d \), and \( x \in \mathbb{R}^d \), let \( d(x,A) = \inf_{y \in A} \| x - y \| \). For another subset \( B \subset \mathbb{R}^d \), the Hausdorff distance between \( A \) and \( B \) is defined as

\[ d_H(A,B) = \max\{ \sup_{x \in A} d(x,B), \sup_{x \in B} d(x,A) \}. \]

For any \( c_1, c_2 \in \mathbb{R} \), let \( c_1 \lor c_2 = \max(c_1,c_2) \) and \( c_1 \land c_2 = \min(c_1,c_2) \). Finally, we will use \( \| A \|_F \) to denote Frobenius-norm of a matrix \( A \).

**Theorem 3** Assume \((\text{A1})_{f,A}, (\text{A2})_f, (\text{A3})_f, (\text{K})_A, h = o(1), \) and \( \frac{\log n}{n h^{d+3}} = o(1) \) as \( n \rightarrow \infty \). Then, for any \( B > 0 \) and \( n \) large enough, with probability at least \( 1 - n^{-B} \):

\[ d_H(\text{Ridge}(f), \text{Ridge}(\hat{f})) = O\left( \left( \frac{\log n}{n h^{d+4}} \right)^{1/2} + h^2 \right). \]

(2.21)

**Proof** Using Talagrand’s inequality (see Proposition A.5 in Sriperumbudur and Steinwart, 2012), there exists a constant \( C > 0 \) such that, for all \( n \geq 1, h \in (0,1), b > 1 \) and \( |\alpha| \leq 4 \) with \( n h^{d+2|\alpha|} \geq (b \lor |\log h|) \), we have

\[ \mathbb{P}\left( \sup_{x \in [0,1]^d} |\partial^{(\alpha)} \hat{f}(x) - \mathbb{E} \partial^{(\alpha)} \hat{f}(x)| < C \sqrt{\frac{b \lor |\log h|}{n h^{d+2|\alpha|}}} \right) \geq 1 - e^{-b}. \]

(2.22)

It follows from standard calculation for kernel density estimation (see, e.g., Lemma 2 of Arias-Castro et al., 2016) that for all \( |\alpha| \leq 4 \),

\[ \sup_{x \in [0,1]^d} |\partial^{(\alpha)} f(x) - \mathbb{E} \partial^{(\alpha)} \hat{f}(x)| = O(h^{(4-|\alpha|) \land 2}). \]

(2.23)

Then (2.22) and (2.23) imply that for any \( B > 0 \), on a set \( A_n \) with probability at least \( 1 - n^{-B} \),

\[ \sup_{x \in [0,1]^d} |\partial^{(\alpha)} f(x) - \mathbb{E} \partial^{(\alpha)} \hat{f}(x)| = O\left( \left( \frac{\log n}{n h^{d+1}} \right)^{1/2} + h^{(4-|\alpha|) \land 2} \right), \]

for all \( |\alpha| \leq 4 \). Observe that, since the eigenvalues of a symmetric matrix are Lipschitz continuous functions, we obtain from that \( \sup_{x \in [0,1]^d} |\lambda_{k+1}(x) - \lambda_{k+1}(x)| = o(1) \) on \( A_n \) for \( n \) large enough. By assumption \((\text{A2})_f\), this implies that on \( A_n \) we have for all \( \epsilon \geq 0 \) small enough, that \( \hat{S}_\epsilon = \{ x \in [0,1]^d : \hat{\eta}(x) \geq -\epsilon, \lambda_{k+1}(x) < 0 \} \) for \( n \) large enough. Denote \( \rho_n = \sup_{x \in [0,1]^d} |\sqrt{-\hat{\eta}(x)} - \sqrt{-\eta(x)}| \). It follows that on \( A_n \), we have that for \( n \) large enough,

\[ \text{Ridge}(\hat{f}) = \hat{S}_0 \subset \{ x \in [0,1]^d : \eta(x) \geq -\rho_n^2, \lambda_{k+1}(x) < 0 \} = S_{\rho_n^2}. \]
We then have on $A_n$,

$$
\sup_{x \in \text{Ridge}(f)} d(x, \text{Ridge}(f)) = O\left(\left(\frac{\log n}{nh^{d+4}}\right)^{1/2} + h^2\right), \quad (2.24)
$$

which follows from (3.14) below, by using that $\rho_n = O\left((\frac{\log n}{nh^{d+4}})^{1/2} + h^2\right)$. Indeed, note that

$$
\sqrt{2}\rho_n \leq \sup_{x \in [0,1]^d} \|\widehat{V}_\perp(x)\widenot{\hat{V}}_\perp(x)\nabla \widehat{f}_\perp(x) - V_\perp(x)V_\perp^\top(x)\nabla f(x)\|

\leq \sup_{x \in [0,1]^d} \|\widehat{V}_\perp(x)\widenot{\hat{V}}_\perp(x)\nabla \widehat{f}_\perp(x)\|_F \sup_{x \in [0,1]^d} \|\nabla f(x)\| + \sup_{x \in [0,1]^d} \|\nabla \widehat{f}_\perp(x) - \nabla f(x)\|

\leq \frac{2\sqrt{2}}{\beta} \sup_{x \in [0,1]^d} \|\nabla^2 \widehat{f}_\perp(x) - \nabla^2 f(x)\|_F \sup_{x \in [0,1]^d} \|\nabla f(x)\| + \sup_{x \in [0,1]^d} \|\nabla \widehat{f}_\perp(x) - \nabla f(x)\|
$$

where the last step follows from the Davis-Kahan theorem (see Yu et al., 2015). Using (2.22) and (2.23) gives the asserted rate in (2.24). The result in Theorem 5(i) allows us to swap the roles of $f$ and $\widehat{f}$ in (2.24) and get on $A_n$,

$$
\sup_{x \in \text{Ridge}(\widehat{f})} d(x, \text{Ridge}(\widehat{f})) = O\left(\left(\frac{\log n}{nh^{d+4}}\right)^{1/2} + h^2\right). \quad (2.25)
$$

We conclude the proof by combining (2.24) and (2.25). 

**Remark 4** Genovese et al. (2014) also gives the same rate of convergence for ridge estimation using the Hausdorff distance. However, their assumptions and methods of proof are different from ours. In particular, in their assumption (A2) they require that $\|I_d - V_\perp(x)V_\perp^\top(x)\nabla f(x)\|_{\text{max},3}(x) \leq \frac{\beta^2}{20}$ for all $x \in \text{Ridge}(f)$, where $f_{\text{max},3}(x)$ is the maximum of the absolute values of all the third partial derivatives of $f$ at $x$, and $\beta$ is essentially the same one as given in our (A2)$_{f}$. Instead of this assumption, we use (A3)$_{f}$, which is weaker (see Lemma 2 in Chen et al., 2015).

### 2.9.1 Continuous versions of the algorithms

**Further notation.** Let $U : \mathbb{R}^d \to \mathbb{R}^d$ be a vector field, and $x_0 \in \mathbb{R}^d$. Consider the general autonomous system of ODEs

$$
\frac{dx(t)}{dt} = U(x(t)), \quad t \in \mathbb{R},
$$

where $x : \mathbb{R} \to \mathbb{R}^d$ with $x(0) = x_0$. This generates a flow $\pi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ such that $\pi(x, 0) = x$, $\pi(x, t + s) = \pi(x, s)$ and $\frac{d}{dt} \pi(x, t) = U(\pi(x, t))$, for $s, t \in \mathbb{R}$. A trajectory, or an integral curve, with a starting point $x_0 \in \mathbb{R}^d$ is denoted by $\pi(x_0, t)$. Let $\gamma^*(x, t)$, $\hat{\gamma}^*(x, t)$ and $\tilde{\gamma}^*(x, t)$ denote the flows generated by $\xi^*$, $\hat{\xi}^*$ and $\tilde{\xi}^*$, respectively.
Theorem 5 Assume (A1)\textsubscript{f.A}, (A2)\textsubscript{f}, (A3)\textsubscript{f}, (K)\textsubscript{4}, h = o(1) and $\frac{\log n}{nh^{\alpha_{f}+\beta}} = o(1)$ as $n \to \infty$. Then, for any $B > 0$ and $n$ large enough, with probability at least $1 - n^{-B}$:

(i) $\hat{f}$ satisfies assumptions (A1)\textsubscript{f.A}, (A2)\textsubscript{f}, and (A3)\textsubscript{f};

(ii) Continuous version of Algorithm 1: there exists $\epsilon_0 > 0$ such that, for all $0 \leq \epsilon \leq \epsilon_0$,
\[
\text{Ridge}(\hat{f}) = \{\lim_{t \to \infty} \tilde{\gamma}^*(x,t), x \in \partial \tilde{S}_\epsilon\} = \text{Ridge}(\hat{\eta}) \cap \tilde{S}_\epsilon.
\]

(iii) Continuous version of Algorithm 2: under the additional assumptions (A1)\textsubscript{f.6}, (K)\textsubscript{6}, (L) and $\frac{\log n}{nh^{\alpha_{f}+\beta}} = o(1)$, and for $\epsilon > 0$ small enough,
\[
\begin{align*}
&\quad a) \text{Ridge}(\hat{\eta}) \cap \tilde{S}_{\tau,\epsilon} = \{\lim_{t \to \infty} \tilde{\gamma}^*_\tau(x,t), x \in \partial \tilde{S}_{\tau,\epsilon}\} \text{ for } \tau > 0 \text{ small enough; }
&\quad b) d_H(\text{Ridge}(\hat{\eta}) \cap \tilde{S}_{\tau,\epsilon}, \text{Ridge}(\hat{f})) = O(\tau^2) \text{ as } \tau \to 0;
&\quad c) d_H(\text{Ridge}(\hat{\eta}) \cap \tilde{S}_{\tau,\epsilon}, \text{Ridge}(f)) = O((\frac{\log n}{nh^{\alpha_{f}+\beta}})^{1/2} + h^2 + \tau^2) \text{ as } \tau \to 0.
\end{align*}
\]

Proof (i) Using (2.22) and (2.23), properties (A1)\textsubscript{f.A} and (A2)\textsubscript{f} are consequences of the Lipschitz continuity of the eigenvalues as functions of symmetric matrices. Let $\lambda_{d-k}^+(A)$ be the $(d - k)$th largest eigenvalue of a symmetric matrix $A$. To show (A3)\textsubscript{f}, notice that assumption (A3)\textsubscript{f} implies that for $\delta > 0$ small enough, there exists a constant $a_0 > 0$ such that
\[
\inf_{x \in \text{Ridge}(f)} \lambda_{d-k}^+(\nabla \xi(x)\nabla \xi(x)) \geq a_0.
\]

Due to the perturbation stability of $\lambda_{d-k}^+(A)$, we have $\inf_{x \in \text{Ridge}(f)} \lambda_{d-k}^+(\nabla \xi(x)\nabla \xi(x)) \geq \frac{1}{2}a_0$ with probability at least $1 - n^{-B}$ when $n$ is large. This then implies the rank of $\nabla \xi(x)$ is at least $d - k$. Furthermore, using the expression of $\nabla \xi(x)$ in (2.15) and the calculation in (2.17), it can be seen that $\tilde{V}_i(x)^\top \nabla \xi(x) = 0$, for all $i = 1, \ldots, k$, which implies that the rank of $\nabla \xi(x)$ is at most $d - k$. Hence (A3)\textsubscript{f} is satisfied.

(ii). Given (i), this is a consequence of Lemma 10 and Theorem 11.

(iii). Given (i), we have the following: For some $\delta > 0$, we have for $|\alpha| = 0, 1, 2$,
\[
\sup_{x \in [0,1]^d} |\partial^{(\alpha)}\hat{\eta}(x) - \partial^{(\alpha)}\hat{\eta}_r(x)| = O(\tau^2).
\]  

This can be shown by using standard techniques for calculating the rate of the bias in kernel density estimation (see, e.g., proof of Lemma 2 in Arias Castro et al., 2016). For $\tau$ small, we have $\text{Ridge}(\hat{\eta}_r) = \{\lim_{t \to \infty} \tilde{\gamma}^*_\tau(x,t), x \in \partial \tilde{S}_{\tau,\epsilon}\}$, as a result of Theorem 14(iii). This is a). Part b) is a consequence of Theorem 14 (iv). Part c) immediately follows from Theorem 3. $\blacksquare$

Remark 6 Without the additional assumptions in part (iii) of the above theorem, we can still show that with probability at least $1 - n^{-B}$, $\sup_{x \in [0,1]^d} |\partial^{(\alpha)}\hat{\eta}(x) - \partial^{(\alpha)}\hat{\eta}_r(x)| = o(1)$ for all $|\alpha| = 0, 1, 2$, and therefore $d_H(\text{Ridge}(\hat{\eta}_r) \cap \tilde{S}_{\tau,\epsilon}, \text{Ridge}(f)) = o(1)$ as $\tau \to 0$. 

17
2.9.2 Discrete approximation: the Euler method

Here we study discrete versions of the algorithms given above. To this end, recall that 
\[ \hat{\xi}^* = \hat{V}_i^* [\hat{V}_i^*]^{-1} \nabla \hat{\eta}. \] 
For a constant \( a > 0 \), let 
\[ \hat{\gamma}^*_a(x, \ell) = \hat{\gamma}^*_a(x, 0) + \sum_{\ell=1}^\infty a \hat{\xi}^*(\hat{\gamma}^*_a(x, \ell)), \quad \ell = 1, 2, \ldots \]
This is a discrete approximation to \( \gamma^*(x, t), t \geq 0 \) using Euler’s method. Similarly, we define the discrete approximation \( \hat{\gamma}^*_{\tau, a}(x, \ell) \) of \( \hat{\gamma}^*_{\tau, \epsilon}(x, t) \) by replacing \( \hat{\xi}^* \) by \( \hat{\xi}^*_{\tau} \) (or, in other words, by replacing \( \hat{\eta} \) by \( \hat{\eta}_{\tau} \)). The following result says that these discretized approximations can be used to recover the corresponding ridges.

**Theorem 7** Assume \((A1), (A2), (A3), (K), (L), h = o(1)\) and \( \frac{\log n}{nh^{d+1}} = o(1) \) as \( n \to \infty \). Then, for any \( B > 0 \) and \( n \) large enough, with probability at least \( 1 - n^{-B} \):

(i) Algorithm 1: With \( R_a(\hat{f}) = \{ \lim_{\ell \to \infty} \hat{\gamma}^*_a(x, \ell), x \in \partial \hat{S}_\epsilon \} \), we have for \( \epsilon \geq 0 \) small enough
\[ d_H(\text{Ridge}(\hat{f}), R_a(\hat{f})) = o(1) \quad \text{as} \quad a \to 0. \] (2.27)

(ii) Algorithm 2: With \( R_{\tau, a}(\hat{f}) = \{ \lim_{\ell \to \infty} \hat{\gamma}^*_{\tau, a}(x, \ell), x \in \partial \hat{S}_{\tau, \epsilon} \} \), and \( \epsilon \geq 0 \) small enough,
we have
\[ d_H(\text{Ridge}(\hat{f}), R_{\tau, a}(\hat{f})) = o(1) \quad \text{as} \quad a \to 0 \quad \text{and} \quad \tau \to 0. \] (2.28)

**Proof** The result in (i) follows from Theorem 5(i) and Theorem 16. The result in (ii) is a consequence of Theorem 5 and Corollary 18.

**Remark 8** By using Theorem 16, one can also obtain rates of convergence in (2.27), and one can show that \( \text{Ridge}(\hat{f}) = R_a(\hat{f}) \) and \( \text{Ridge}(\hat{\eta}_{\tau}) \cap \hat{S}_{\tau, \epsilon} = R_{\tau, a}(\hat{f}) \) for \( d = 1 \), with probability at least \( 1 - n^{-B} \) for any \( B > 0 \) and \( n \) large enough.

3. The mathematical framework for the ODE-based algorithms of ridge extraction

The SCMS type algorithms discussed above can be considered as numerical approximations of trajectories of Ordinary Differential Equations (ODE). These corresponding ODEs (the mathematical models) for our algorithm are analyzed in the following. For the original mean shift algorithm, such theoretical work can be found in Arias-Castro et al. (2016). For the Subspace Constraint Mean Shift algorithm, see Genovese et al. (2014) and Qiao and Polonik (2016) for the theoretical analysis of the corresponding ODE models.

In the following \( f \) denotes a generic positive function on \([0, 1]^d\). While we apply the below results to densities (e.g. to our kernel density estimates), \( f \) does not have to integrate to 1.
3.1 Some useful background knowledge of ODEs

A reference for the following material is Wiggins (2003). As above consider

\[ \frac{dx(t)}{dt} = U(x(t)), \quad t \in \mathbb{R}, \]

where \( x : \mathbb{R} \rightarrow \mathbb{R}^d \) with \( x(0) = x_0 \in \mathbb{R}^d \), and \( U : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is a vector field. Let \( \pi : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}^d \) denote the corresponding flow.

A compact set \( S \subset \mathbb{R}^d \) is called a positively invariant set under the above vector field if for any \( x_0 \in S \) we have \( \pi(x_0, t) \in S \) for all \( t \geq 0 \). We assume that the boundary of \( S \) is a \( C^1 \) manifold. A point \( \bar{x} \) is called an equilibrium point, or fixed point, if \( U(\bar{x}) = 0 \). Let \( A \subset S \) be the set of all equilibrium points in \( S \). A continuously differentiable scalar-valued function \( V \) defined on \( S \) is called a Lyapunov function if it satisfies: \( \frac{dV(x(t))}{dt} \leq 0 \) for all \( x(t) \in S \). We also define two sets

\[ E = \left\{ x \in S : \frac{dV(x(t))}{dt} = 0 \right\}, \]

\[ M = \bigcup_{x_0 \in E} \left\{ \pi(x_0, t), t > 0 : \pi(x_0, t) \in E \text{ for all } t > 0 \right\}. \]

Note that \( A \subset M \subset E \). We will use LaSalle’s Invariance Principle (see Theorem 8.3.1, Wiggins, 2003), which states:

**Theorem 9** For all \( x \in S \), \( \pi(x, t) \to M \) as \( t \to \infty \).

Later LaSalle’s Theorem will be applied with \( M = E = \text{Ridge}(f) \).

3.2 ODE theory for ridge extraction

Recall that \( \eta(x) = -\frac{1}{2}||V_\perp(x)\nabla f(x)||^2 = -\frac{1}{2}||\xi(x)||^2 \). Also recall the definition of \( S_\epsilon \) given in (2.20). The following lemma states that in a neighborhood of the ridge of \( f \), the ridge of the ridgeness function \( \eta \) equals the ridge of \( f \).

**Lemma 10** Assuming \( (A1)_f, (A2)_f \), and \( (A3)_f \), there exists an \( \epsilon_0 > 0 \) such that

\[ \text{Ridge}(f) = \text{Ridge}(\eta) \cap S_\epsilon \quad \forall 0 < \epsilon < \epsilon_0. \]

**Proof** First we show that \( \text{Ridge}(f) \subset \text{Ridge}(\eta) \cap S_\epsilon \). We have \( \nabla \eta(x) = -\nabla \xi(x)^\top \xi(x) \) for \( x \in [0, 1]^d \). Thus, for \( x \in \text{Ridge}(f) \) (meaning that \( \xi(x) = 0 \)), we have \( \nabla \eta(x) = 0 \), and hence \( \xi^*(x) = 0 \). Then \( \text{Ridge}(f) \subset \text{Ridge}(\eta) \cap S_\epsilon \) is a direct consequence of Lemma 2.

Next we show that \( \text{Ridge}(\eta) \cap S_\epsilon \subset \text{Ridge}(f) \) for \( \epsilon \) small enough. To this end we show that

\[ ||\xi^*(x)|| > 0 \quad \text{for all } x \in S_\epsilon \setminus \text{Ridge}(f), \]

which implies the result. Note that \( S_\epsilon \subset \text{Ridge}(f)^\delta \) when \( \epsilon \) is small enough. See (3.14) below. For any \( x \in S_\epsilon \setminus \text{Ridge}(f) \), let \( x_0 \in \text{Ridge}(f) \) be the projection of \( x \) onto \( \text{Ridge}(f) \), that is, there exists a \( \delta_x \in (0, \delta) \), such that we can write \( x = x_0 + \delta_x u \), where \( u = u(x) \in N(x_0) \).
is a unit vector with \(N(x_0)\) the normal space to the ridge at \(x_0\). Notice that by using Lemma 2, \(N(x_0) = \{V_+^*(x_0)a : a \in \mathbb{R}^{d-k}\}\). We will show that there exists an \(\epsilon > 0\), such that \(\|\nabla \eta(x)\| > 0\) for all \(x \in S_\epsilon \setminus \text{Ridge}(f)\). Using Lipschitz continuity of the fourth partial derivatives of \(f\), there exists a constant \(\kappa_1 > 0\) such that \(\|\nabla^2 \eta(x_0) - \nabla^2 \eta(x_1)\|_F \leq \kappa_1\|x_0 - x_1\|\) for all \(x_1 \in B(x_0, \delta)\). Then we can write

\[
\nabla \eta(x) = \nabla \eta(x_0) + \nabla^2 \eta(x_0)(x - x_0) + R_1(x)
= \delta_x \nabla^2 \eta(x_0)u + R_1(x),
\]

where \(\|R_1(x)\| \leq \kappa_1\delta_x^2\). Note that \(\nabla^2 \eta(x_0) = \sum_{j=k+1}^d \lambda_j^*(x_0)V_j^*(x_0)V_j^*(x_0)^\top\) by Lemma 2 and \(u\) is a unit vector in the space spanned by \(V_{k+1}^*(x_0), \ldots, V_d^*(x_0)\). Hence \(A \geq |\lambda_d^*(x_0)| \geq \|\nabla^2 \eta(x_0)u\| \geq |\lambda_{k+1}^*(x_0)| \geq \alpha\), where \(\alpha\) and \(A\) are positive constants given in Lemma 2. Thus, for \(\epsilon > 0\) small enough (and hence \(\delta_x\) is small), we have \(\|\nabla \eta(x)\| > 0\).

Next we show that \(\frac{|\langle \nabla \eta(x), u(x) \rangle|}{\|\nabla \eta(x)\|}\) is bounded from below. Using (3.4) we have that

\[
|\langle \nabla \eta(x), u(x) \rangle| \geq \frac{\delta_x |\lambda_{k+1}^*(x_0)| - \kappa_1 \delta_x^2}{\|\nabla \eta(x)\|}. \tag{3.4}
\]

For \(\epsilon\) (and hence \(\delta_x\)) small enough, the right-hand side is bounded by a positive constant from below, say,

\[
\frac{|\langle \nabla \eta(x), u(x) \rangle|}{\|\nabla \eta(x)\|} \geq \frac{1}{2} \frac{|\lambda_{k+1}^*(x_0)|}{|\lambda_d^*(x_0)|}.
\]

Since \(u\) can be written as \(V_+^*(x_0)a\) with \(a \in \mathbb{R}^{d-k}\) and \(\|a\| = 1\), we have by using the Cauchy-Schwarz inequality that \(|\langle \nabla \eta(x), u(x) \rangle| = |\langle \nabla \eta(x)^\top V_+^*(x_0)a \rangle| \leq \|V_+^*(x_0)^\top \nabla \eta(x)\| = \|V_+^*(x_0)V_+^*(x_0)^\top \nabla \eta(x)\|\), and thus

\[
\frac{\|V_+^*(x_0)V_+^*(x_0)^\top \nabla \eta(x)\|}{\|\nabla \eta(x)\|} \geq \frac{1}{2} \frac{|\lambda_{k+1}^*(x_0)|}{|\lambda_d^*(x_0)|}. \tag{3.5}
\]

On the other hand, the angle between the eigenspace spanned by \(V_+^*(x_0)\) and \(V_+^*(x)\) should be small if \(\delta_x\) is small because \(\nabla^2 \eta\) is a continuous function of \(x\). This can be seen by using the Davis-Kahan inequality:

\[
\|V_+^*(x_0)V_+^*(x_0)^\top - V_+^*(x)V_+^*(x)^\top\|_F \leq \frac{2\sqrt{2}\|\nabla^2 \eta(x_0) - \nabla^2 \eta(x)\|_F}{|\lambda_{k+1}^*(x_0)|} \leq \frac{2\sqrt{2}\kappa_1 \delta_x}{|\lambda_{k+1}^*(x_0)|}. \tag{3.6}
\]

Hence using (3.5) and (3.6),

\[
\frac{\|\xi^*(x)\|}{\|\nabla \eta(x)\|} = \frac{\|V_+^*(x)V_+^*(x)^\top \nabla \eta(x)\|}{\|\nabla \eta(x)\|} \geq \frac{1}{2} \frac{|\lambda_{k+1}^*(x_0)|}{|\lambda_d^*(x_0)|} - \frac{2\sqrt{2}\kappa_1 \delta_x}{|\lambda_{k+1}^*(x_0)|},
\]

which can be bounded from below by a positive constant when \(\epsilon\) (and hence \(\delta_x\)) is small enough. This is (3.3).
Recall that \( \frac{\partial \gamma^*(x,t)}{\partial t} = \xi^*(\gamma^*(x,t)) \), \( t \in \mathbb{R} \), which is the mathematical model of Algorithm 1. There exists a positive \( \epsilon > 0 \) such that \( S_\epsilon \) is a positively invariant set corresponding to this flow. This can be seen as follows. Consider the derivative of \( \eta(\gamma^*(x,t)) \) with respect to \( t \):

\[
\frac{\partial \eta(\gamma^*(x,t))}{\partial t} = [\nabla \eta(\gamma^*(x,t))]^T V^*_\perp(\gamma^*(x,t))[V^*_\perp(\gamma^*(x,t))]^T \nabla \eta(\gamma^*(x,t)) = \|\xi^*(\gamma^*(x,t))\|^2 \geq 0.
\]

(3.7)

In other words, \( \eta \) is always non-decreasing along the trajectories of the flow. Indeed, notice that for \( \epsilon > 0 \) small enough, \( \|\xi^*(\gamma^*(x,t))\|^2 > 0 \) for all \( x \in S_\epsilon \setminus \text{Ridge}(f) \) (or when \( t < \infty \)) based on Lemma 10. In other words, \( \xi^* \) is a gradient-like vector field with respect to \( \eta \) (see page 45, Nicolaescu, 2011), and it follows that \( S_\epsilon \) is a positively invariant set as we have claimed. Note that \( \text{Ridge}(f) \) is the set of all the equilibrium points in \( S_\epsilon \). A natural choice for a Lyapunov function \( V \) is

\[ V(x) = -\eta(x). \]

Since

\[
\frac{\partial V(\gamma^*(x,t))}{\partial t} = -\langle \nabla \eta(\gamma^*(x,t)), \xi^*(\gamma^*(x,t)) \rangle = -\|\xi^*(\gamma^*(x,t))\|^2 \leq 0,
\]

this indeed is a Lyapunov function. Notice further that this derivative is equal to zero only when \( x \in \text{Ridge}(f) \). Therefore, for the sets \( E \) and \( M \) given in (3.1) and (3.2), we have \( E = M = \text{Ridge}(f) \).

**Theorem 11** Assume that (A1)$_f$A), (A2)$_f$, and (A3)$_f$ hold. There exists \( \epsilon > 0 \) such that:

(i) For each \( x \in S_\epsilon \setminus \text{Ridge}(f) \), there is a unique path \( \gamma^* \) passing through \( x \).

(ii) For each \( x \in S_\epsilon \) as a starting point, the path \( \gamma^*(x,t) \) converges to a point on \( \text{Ridge}(f) \), as \( t \to +\infty \).

(iii) \( \text{Ridge}(f) = \{\lim_{t\to\infty} \gamma^*(x,t) : x \in \partial S_\epsilon\} \), where \( \partial S_\epsilon \) is the boundary of \( S_\epsilon \).

**Proof** (i) is very similar to Lemma 2 in Genovese et al. (2014) and the proof is similar, too. Details are omitted. (ii) follows from the above discussion by an application of LaSalle’s Invariance Principle. Next we prove (iii), for which we will use the technique in the proof of Theorem 2.6 in Nicolaescu (2011). Recall that the vector field generated by \( \xi^* \) is gradient-like. Let

\[
\zeta := \frac{\xi^*}{\|\xi^*\|^2},
\]

(3.8)

and let \( \gamma^\dagger \) by the integral curve generated by this rescaled vector field, that is,

\[
\frac{\partial \gamma^\dagger(x,t)}{\partial t} = \zeta(\gamma^\dagger(x,t)), \quad \gamma^\dagger(x,0) = x.
\]

(3.9)

We first show (iii) with \( \gamma^* \) replaced by \( \gamma^\dagger \). Observe that (see (3.7))

\[
\frac{\partial \eta(\gamma^\dagger(x,t))}{\partial t} = 1.
\]

(3.10)
In other words, the level of η can be used to parametrize the integral curves γ⁺(x, t), so that for any ϵ₁, ϵ₂ ∈ (0, ε) with ϵ₁ > ϵ₂, we have

\[ \partial S_{ϵ₁} = \{ γ⁺(x, ϵ₂ - ϵ₁) : x ∈ ∂S_{ϵ₂} \}, \quad (3.11) \]
\[ \partial S_{ϵ₂} = \{ γ⁺(x, ϵ₁ - ϵ₁) : x ∈ ∂S_{ϵ₁} \}. \quad (3.12) \]

Both (3.11) and (3.12) are similar to Theorem 2.6 in Nicolaescu (2011). We only show (3.11). Using (3.10), it is clear that for any \( x \in ∂S_{ϵ₂} \), \( γ⁺(x, ϵ₂ - ϵ₁) \in ∂S_{ϵ₁} \). This means \( \{ γ⁺(x, ϵ₂ - ϵ₁) : x ∈ ∂S_{ϵ₂} \} ⊂ ∂S_{ϵ₁} \). Now define \( \tilde{γ}⁺ \) as the solution of

\[ \frac{∂\tilde{γ}⁺(x, t)}{∂t} = -ζ(\tilde{γ}⁺(x, t)), \quad \tilde{γ}⁺(x, 0) = x. \quad (3.13) \]

Then, similar to (3.10), we have \( \frac{∂η(γ⁺(x, t))}{∂t} = -1 \). Therefore for any \( x ∈ ∂S_{ϵ₁} \), there exists \( \tilde{x} = \tilde{γ}⁺(x, ϵ₂ - ϵ₁) \) so that \( x = γ⁺(\tilde{x}, ϵ₂ - ϵ₁) \). This means that \( ∂S_{ϵ₁} \subset \{ γ⁺(x, ϵ₂ - ϵ₁) : x ∈ ∂S_{ϵ₂} \} \). Hence (3.11) is verified.

Furthermore, as will be shown below, one can find constants \( L₁, L₂ > 0 \) such that for all \( ϵ' ≥ 0 \) small enough

\[ L₁\sqrt{ε'} ≤ d_{H}(∂S_{ϵ'}, \text{Ridge}(f)) = d_{H}(∂S_{ϵ'}, S₀) ≤ L₂\sqrt{ε'}. \quad (3.14) \]

where \( d_{H} \) is the Hausdorff distance. By using (3.11) and (3.12), this means that, for all \( ϵ' > 0 \) small enough

\[ L₁\sqrt{ε'} ≤ d_{H}(\{ γ⁺(x, ϵ - ϵ') : x ∈ ∂S_{ϵ} \}, \text{Ridge}(f)) ≤ L₂\sqrt{ε'}. \quad (3.15) \]

Next we show that

\[ \text{Ridge}(f) = \{ \lim_{t→ε} γ⁺(x, t) : x ∈ ∂S_{ϵ} \}. \quad (3.16) \]

Recall that \( \text{Ridge}(f) \) is a compact set (see proof of Lemma 2). The set \( \{ \lim_{t→ε} γ⁺(x, t) : x ∈ ∂S_{ϵ} \} \) is also compact. This is because \( \lim_{t→ε} γ⁺(x, t) \) is a continuous function of \( x \) (see Fact 6 below) and \( ∂S_{ϵ} \) is a compact set. Suppose that \( \text{Ridge}(f) \neq \{ \lim_{t→ε} γ⁺(x, t) : x ∈ ∂S_{ϵ} \} \). Then there must exist a constant \( δ₀ > 0 \) such that \( δ₀ ≤ d_{H}(\{ \lim_{t→ε} γ⁺(x, t) : x ∈ ∂S_{ϵ} \}, \text{Ridge}(f)) \), which further implies that there exists \( x₀ ∈ ∂S_{ϵ} \) and \( ϵ₀ > 0 \) such that for all \( 0 < ϵ' ≤ ϵ₀ \) we have \( d_{H}(γ⁺(x, ϵ - ϵ'), \text{Ridge}(f)) ≥ δ₀ \). This contradicts (3.15) and therefore (3.16) has to be true.

Next we show that \( γ⁺(x, \cdot) \) and \( γ⁺(x, \cdot) \) have the same trajectories, which shows (iii). To this end we show a reparameterization relation between the two flows. For each \( x ∈ ∂S_{ϵ} \), let

\[ s(t) = \int_{0}^{t} \| ξ⁺(x, u) \|^2 du \]
\[ = \int_{0}^{t} \left[ \frac{∂η(γ⁺(x, u))}{∂u} \right] du \]
\[ = η(γ⁺(x, t)) - η(γ⁺(x, 0)) \]
\[ = η(γ⁺(x, t)) - η(x). \]
Note that we have suppressed the dependence of $s(t)$ on $x$ in the notation. Let $t(s)$ be the inverse of $s(t)$. Then $t(0) = 0$, and

$$
\frac{dt(s)}{ds} = \frac{1}{\|\xi^*(x,s)\|^2}. \tag{3.17}
$$

We obtain $\gamma^t(x,s) = \gamma^*(x,t(s))$, because

$$
\frac{\partial \gamma^*(x,t(s))}{\partial s} = \zeta(\gamma^t(x,s)), \quad \gamma^*(x,t(0)) = x. \tag{3.18}
$$

Note that as $t \to \infty$, we have $s(t) \to \eta(\gamma^*(x,\infty)) - \eta(x) = \epsilon$ for all $x \in S_\epsilon$, because $\gamma^*(x,\infty) \in \text{Ridge}(f)$. Hence $\lim_{t \to \infty} \gamma^*(x,t) = \lim_{s \to \epsilon} \gamma^t(x,s)$ for all $x \in S_\epsilon$.

It remains to prove (3.14). For any $b \in \mathbb{R}^d$, consider

$$
M_b := \{x \in [0, 1]^d : \xi(x) = b, \lambda_{k+1}(x) < 0\},
$$

and note that $S_{\epsilon'} = \bigcup_{\|b\| \leq \epsilon'} M_b$ for all $\epsilon' \geq 0$. We will show that for all $\|b\|$ small enough, $C_1\|b\| \leq d_H(M_b, \text{Ridge}(f)) \leq C_2\|b\|$ for some positive constants $C_1, C_2$. For any $x \in M_b$, let $x_0$ be its projection point on $\text{Ridge}(f)$. There exists a unit vector $u = u(x) \in N(x_0)$ and $\delta_x > 0$ such that $x = x_0 + \delta_xu$. Using a Taylor expansion, we get

$$
\xi(x) = \xi(x_0) + \nabla \xi(x_0)(x-x_0) + R_1(x) = \delta_x \nabla \xi(x_0)u + R_2(x),
$$

where $\|R_2(x)\| \leq \kappa_2\delta_x^2$, for a constant $\kappa_2 > 0$ for all $x \in M_b$. Therefore

$$
\|b\|^2 = \|\xi(x)\|^2 = \delta_x^2 u^T \nabla \xi(x_0)^T \nabla \xi(x_0)u + R_3(x),
$$

where $|R_3(x)| \leq \kappa_3\delta_x^2$, for a constant $\kappa_3 > 0$ for all $x \in M_b$. Let $\omega_{\max}(x)$ and $\omega_{\min}(x)$ be the largest and the $(d-k)$th largest eigenvalues of $\nabla \xi(x)^T \nabla \xi(x)$, respectively. For $x_0 \in \text{Ridge}(f)$, we have $\omega_{\min}(x_0) = -\lambda_{k+1}^*(x_0)$ and $\omega_{\max}(x_0) = -\lambda_k^*(x_0)$ because $\nabla^2 \eta(x_0) = -\nabla \xi(x_0)^T \nabla \xi(x_0)$. Observe that the first $k$ eigenvalues $\lambda^*_i(x_0)$, $i = 1, \ldots, k$ of $\nabla \xi(x_0)^T \nabla \xi(x_0)$ are zero, and that $\lambda^*_k(x_0) < 0$, $i = k+1, \ldots, d$. Moreover, $u$ is in the space spanned by $V_{k+1}^*_0(x_0), \ldots, V_d^*(x_0)$. So when $\|b\|$ is small (and hence $\delta_x$ is small), we have that,

$$
2|\lambda^*_d(x_0)|\delta_x^2 \geq 2\omega_{\max}(x_0)\delta_x^2 \geq \|b\|^2 \geq \frac{\omega_{\min}(x_0)}{2}\delta_x^2 = \frac{|\lambda^*_{k+1}(x_0)|}{2}\delta_x^2. \tag{3.19}
$$

Since $\delta_x = d(x, \text{Ridge}(f)) = \inf_{y \in \text{Ridge}(f)}\|x-y\|$, we have for all $b \in \mathbb{R}^d$ such that $\frac{1}{2}\|b\|^2 = \epsilon'$ is small enough,

$$
\sup_{x \in M_b} d(x, \text{Ridge}(f)) \leq \sqrt{\frac{2}{|\lambda^*_{k+1}(x_0)|}}\|b\| \leq \sqrt{\frac{2}{|\lambda^*_{k+1}(x_0)|}}\sqrt{2\epsilon'} \leq \sqrt{\frac{4}{\alpha}}\sqrt{\epsilon'}, \tag{3.20}
$$

$$
\sup_{x \in M_b} d(x, \text{Ridge}(f)) \geq \sqrt{\frac{1}{2|\lambda^*_d(x_0)|}}\|b\| \geq \sqrt{\frac{1}{2|\lambda^*_d(x_0)|}}\sqrt{2\epsilon'} \geq \sqrt{\frac{1}{A}}\sqrt{\epsilon'}, \tag{3.21}
$$

where $A \geq \alpha > 0$ are given in Lemma 2.
The other direction can be proved in a similar way, as given as follows. Now let \( x \) be a point on \( \text{Ridge}(f) \), and \( x_0 \) be its projection on \( M_b \), such that \( x = x_0 + \delta_x u \), where \( \delta_x = \| x - x_0 \| > 0 \), and \( u \in N(x_0) \) is a unit normal vector of \( M_b \) at \( x_0 \). Following the above analysis (in particular (3.19)), when \( \| b \| \) is small enough, we still have \( 2\omega_{\max}(x_0)\delta_x^2 \geq \| b \| \geq \frac{1}{2}\omega_{\min}(x_0)\delta_x^2 \). Since \( \omega \) is a continuous function in a neighborhood of \( \text{Ridge}(f) \), we have that 
\[
\frac{4}{\sqrt{A}} \| b \| \leq \frac{1}{\sqrt{2}} \alpha \delta_x^2 \sqrt{\epsilon'} \leq \sup_{x \in \text{Ridge}(f)} d(x, M_b) \leq \sqrt{8|\alpha|} \sqrt{\epsilon'}.
\]
(3.22)

Combining (3.20), (3.21) and (3.22), we get
\[
\sqrt{\frac{1}{2A}} \sqrt{\epsilon'} \leq d_H(\partial S_{\epsilon'}, \text{Ridge}(f)) \leq \sqrt{\frac{8}{|\alpha|}} \sqrt{\epsilon'}.
\]
(3.23)

Hence we get (3.14) and the proof is completed.

**Remark 12** Let \( \gamma(x, t) \) be the flow corresponding to \( \frac{dx(t)}{dt} = \xi(x(t)), t \in \mathbb{R} \), which is the model for the original SCMS. We can now see the major difference between using \( \gamma(x, t) \) and using the flow \( \gamma^*(x, t) \) considered in our approach. For \( \gamma(x, t), \) we have
\[
\frac{df(\gamma(x, t))}{dt} = [\nabla f(\gamma(x, t))]^\top V_\perp(\gamma(x, t))[V_\perp(\gamma(x, t))]^\top \nabla f(\gamma(x, t)) \geq 0.
\]
(3.24)

In other words, it is the height of \( f \) that increases along the path of \( \gamma(t, x) \) as \( t \) increases. In contrast to that, it is the ridgeness \( \eta(x) \) that increases along the path of \( \gamma^*(x, t) \). In general, while the height and ridgeness are closed related, they are two different quantities. This also provides a different point of view to the shortcoming of the original SCMS discussed above, which is that there is no guarantee for the entire ridge to be covered by the limit points of the SCMS paths (unless the starting points already cover the entire ridge).

**Remark 13** Part (iii) of Theorem 11 can be generalized as follows. Let \( \mathcal{A} \) be a set such that for any \( x \in \partial S_\epsilon \), there exists a point \( y \in \mathcal{A} \) such that there exists a finite \( t_x \in (-\infty, \infty) \) such that \( y = \gamma^*(x, t_x) \). Then by (iii) in Theorem 11, we have \( \text{Ridge}(f) = \{ \lim_{t \to \infty} \gamma^*(x, t) : x \in \mathcal{A} \} \).

### 3.3 Stability of the flows

Now suppose we have a perturbed ridgeness function \( \tilde{\eta} \), which we assume is twice differentiable. We measure the perturbation by the following quantities.

\[
\delta_0 = \sup_{x \in [0,1]^d} |\eta(x) - \tilde{\eta}(x)|, \quad (3.25)
\]
\[
\delta_1 = \sup_{x \in [0,1]^d} \| \nabla \eta(x) - \nabla \tilde{\eta}(x) \|. \quad (3.26)
\]
\[ \delta_2 = \sup_{x \in [0,1]^d} \|\nabla^2 \eta(x) - \nabla^2 \tilde{\eta}(x)\|_F. \] (3.27)

Denote the eigenvalues and eigenvectors of the Hessian \( \nabla^2 \tilde{\eta} \) by \( \tilde{\lambda}_1^* \geq \tilde{\lambda}_2^* \geq \cdots \geq \tilde{\lambda}_d^* \) and \( \tilde{V}_1^*, \cdots, \tilde{V}_d^* \), respectively, and let \( \tilde{V}_+ = [\tilde{V}_{k+1}^*, \cdots, \tilde{V}_d^*] \). With \( \xi^*(x) = \tilde{V}_+^*(x)[\tilde{V}_+^*(x)]^T \nabla \eta(x) \), for \( x \in [0,1]^d \), let \( \tilde{\gamma}^*(x, t) \) be the flow satisfying
\[ \frac{d\tilde{\gamma}^*(x, t)}{dt} = \tilde{\xi}^*(\tilde{\gamma}^*(x, t)), \quad t \in \mathbb{R}; \quad \tilde{\gamma}^*(x, 0) = x. \] (3.28)

Let \( \tilde{S}_\epsilon = \{x \in [0,1]^d : \tilde{\eta}(x) \geq -\epsilon, \lambda_{k+1}(x) < 0\} \). By definition \( \text{Ridge}(\tilde{\eta}) = \{x \in [0,1]^d : \tilde{\xi}^*(x) = 0, \lambda_{k+1}^*(x) < 0\} \).

**Theorem 14** Suppose that (A1)\(_f\), (A1)\(_\tilde{f}\), (A2)\(_f\), and (A3)\(_f\) hold. There exists \( \epsilon_0 > 0 \) such that for any \( \epsilon \in (0, \epsilon_0] \), if \( \max(\delta_0, \delta_1, \delta_2) \) is small enough,

(i) For each \( x \in \tilde{S}_\epsilon \setminus \text{Ridge}(\tilde{\eta}) \), there is a unique path \( \tilde{\gamma}^* \) passing through \( x \).

(ii) For each \( x \in \tilde{S}_\epsilon \) as a starting point, the path \( \tilde{\gamma}^*(x, t) \) converges to a point on \( \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon \), as \( t \rightarrow +\infty \).

(iii) \( \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = \{\lim_{t \to \infty} \tilde{\gamma}^*(x, t) : x \in \partial \tilde{S}_\epsilon\} \), where \( \partial \tilde{S}_\epsilon = \{x \in [0,1]^d : \tilde{\eta}(x) = -\epsilon, \lambda_{k+1} < 0\} \) is the boundary of \( \tilde{S}_\epsilon \).

(iv) There exists a constant \( C > 0 \) such that \( d_H(\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon, \text{Ridge}(f)) \leq C \max(\delta_0, \delta_1, \delta_2) \).

**Remark 15** This theorem is applied to \( f, \tilde{\eta} \) being \( \tilde{f} \) and \( \tilde{\eta}_r \), respectively, for which small enough values of \( \delta_0, \delta_1, \) and \( \delta_2 \) can be found on a set with high probability (see the proof of Theorem 5).

**Proof** First we show that
\[ \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = R(\tilde{\eta}) \cap \tilde{S}_\epsilon, \] (3.29)
where \( R(\tilde{\eta}) = \{x \in [0,1]^d : \tilde{\xi}^*_+ = 0\} \). Using Lemma 2 and the continuity of eigenvalues as functions of symmetric matrices, when \( \max(\delta_0, \delta_1, \delta_2) \) is small enough, we have that for all \( x \in \text{Ridge}(f) \),
\[ -\frac{1}{2} \alpha \geq \tilde{\lambda}_{k+1}^*(x) \geq \cdots \geq \tilde{\lambda}_d^*(x) \geq -2A. \] (3.30)

Then for any \( \epsilon \) small enough we can choose \( \delta_0 \in [0, \epsilon) \) small enough such that
\[ S_{\epsilon-\delta_0} \subset \tilde{S}_\epsilon \subset S_{\epsilon+\delta_0} \subset \text{Ridge}(f) \] (3.31)
and hence we get (3.29). Then (i) and (ii) follows from similar arguments for Theorem 11 (i) and (ii).

To prove (iii) we need to prove the result analogue to (3.14). Let \( \tilde{\eta}_r(x) = -\frac{1}{2}||\tilde{\xi}_+^*(x)||^2 \) and \( \tilde{S}_r^* = \{x \in [0,1]^d : \tilde{\eta}_r(x) \geq -\epsilon, \lambda_{k+1}(x) < 0\} \). Then following a similar argument as in the proof of (3.14), we can show that for all \( \epsilon' > 0 \) small enough
\[ \tilde{L}_1 \sqrt{\epsilon'} \leq d_H(\partial \tilde{S}_r^* \cap \tilde{S}_\epsilon, \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon) \leq \tilde{L}_2 \sqrt{\epsilon'}, \] (3.32)
Using (3.34) and (3.35) we have
\[
\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = \{ \lim_{t \to \infty} \tilde{\gamma}^*(x, t) : x \in \partial \tilde{S}_\epsilon^* \cap \tilde{S}_\epsilon \}. \tag{3.33}
\]

Using similar arguments given in the proof of Theorem 3, we can show that there exists a constant $C \geq 0$ such that when $\max(\delta_0, \delta_1, \delta_2)$ and $\epsilon > 0$ are small enough,
\[
d_H(\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon, \text{Ridge}(\eta) \cap \tilde{S}_\epsilon) \leq C \max(\delta_1, \delta_2). \tag{3.34}
\]

Using (3.34) and Lemma 10, we have for some constants $L_1, L_2 > 0$, and all $\epsilon'' > 0$ small enough,
\[
L_1 \sqrt{\epsilon''} \leq d_H(\partial \tilde{S}_\epsilon^*, \text{Ridge}(\eta) \cap \tilde{S}_\epsilon) \leq L_2 \sqrt{\epsilon''}, \tag{3.35}
\]

Suppose that $\delta_0$ is small enough so that $\epsilon - 2\delta_0 > 0$, and for all $\epsilon'' \in (0, \epsilon - 2\delta_0]$ we have
\[
\tilde{S}_{\epsilon'' - \delta_0} \subset \tilde{S}_{\epsilon''} \subset \tilde{S}_{\epsilon'' + \delta_0} \subset \tilde{S}_\epsilon. \tag{3.36}
\]

Using (3.34) and (3.35) we have
\[
L_1 \sqrt{\epsilon''} - \sqrt{\delta_0} - C \max(\delta_1, \delta_2) \leq d_H(\partial \tilde{S}_\epsilon^*, \text{Ridge}(\eta) \cap \tilde{S}_\epsilon) \leq L_2 \sqrt{\epsilon''} + \sqrt{\delta_0} + C \max(\delta_1, \delta_2), \tag{3.37}
\]

When $\max(\delta_0, \delta_1, \delta_2)$ is small enough, we can further adjust $\epsilon'$ and $\epsilon''$ and choose $\epsilon''' > 0$ such that
\[
L_2 \sqrt{\epsilon''} + \sqrt{\delta_0} + C \max\{\delta_1, \delta_2\} \leq \tilde{L}_2 \sqrt{\epsilon'}, \tag{3.38}
\]
\[
L_1 \sqrt{\epsilon''} - \sqrt{\delta_0} - C \max\{\delta_1, \delta_2\} \geq \tilde{L}_1 \sqrt{\epsilon''}. \tag{3.39}
\]

Then it follows from (3.32) that
\[
(\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon) \subset (\tilde{S}_\epsilon^{**} \cap \tilde{S}_\epsilon) \subset \tilde{S}_{\epsilon''} \subset (\tilde{S}_\epsilon^* \cap \tilde{S}_\epsilon). \tag{3.40}
\]

Therefore, by the results in (i) and (ii), for each $x \in \partial \tilde{S}_\epsilon^* \cap \tilde{S}_\epsilon$, there exists a $t'_x \in (0, \infty)$ such that $\tilde{\gamma}^*(x, t'_x) \in \partial \tilde{S}_\epsilon^{**}$; and for each $x \in \partial \tilde{S}_\epsilon^{**}$, there exists a $t''_x \in (0, \infty)$ such that $\tilde{\gamma}^*(x, t''_x) \in \tilde{S}_\epsilon^{**} \cap \tilde{S}_\epsilon$. Using (3.33) and noticing $\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon^*$, we get
\[
\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = \{ \lim_{t \to \infty} \tilde{\gamma}^*(x, t) : x \in \partial \tilde{S}_\epsilon^* \}, \tag{3.41}
\]

and hence (iii) is proved (where we take $\epsilon$ as $\epsilon''$). The result in (iv) follows from (3.34) and Lemma 10.

\[\blacksquare\]

### 3.4 Euler’s method

Recall that $\xi = \nabla \tilde{\gamma}^*(x, \ell)$. For a fixed step size $\epsilon > 0$, now consider the sequence $\gamma^*_a(x, \ell)$, $\ell = 0, 1, \cdots$, defined as $\gamma^*_a(x, 0) = x$, and
\[
\gamma^*_a(x, \ell + 1) = \gamma^*_a(x, \ell) + \epsilon \xi^*(\gamma^*_a(x, \ell)).
\]

In this section, we will prove the following result about this discretization of the flow $\gamma^*$.
Theorem 16 Suppose that assumptions \((A1)\), \((A2)\), and \((A3)\) hold. Let \(R_a(f) = \{\lim_{\ell \to \infty} \gamma^*_a(x, \ell), x \in \partial S_\epsilon\}\). There exists \(\delta_a > 0\) such that when \(0 < a \leq \delta_a\), we have that

(i) \(R_a(f) \subset \text{Ridge}(f)\), and

(ii) there exists a constant \(C > 0\) such that \(d_H(\text{Ridge}(f), R_a(f)) \leq C a^{1-\sigma_0}\), where \(0 < \sigma_0 < 1\) is given in (3.73).

Moreover,

(iii) When \(k = 1\) (1-dimensional ridge), we have

\[
\text{Ridge}(f) = R_a(f).
\]

Remark 17 The last assertion says that in the case of one-dimensional ridges we still recover the entire ridge of \(f\) (as we do in the continuous case - see Theorem 11(iii)), even when using the discretized algorithm, provided the step-size is small enough. We conjecture that the result also holds true for ridges of dimension larger than 1 (see Remark 19).

Recall that we have analyzed the perturbed flow \(\tilde{\gamma}^*_a\) and its convergence in subsection 3.3. Using the same notation as defined in that subsection, consider the sequence \(\tilde{\gamma}^*_a(x, \ell)\), \(\ell = 0, 1, \ldots\), defined as \(\tilde{\gamma}^*_a(x, 0) = x\), and

\[
\tilde{\gamma}^*_a(x, \ell + 1) = \tilde{\gamma}^*_a(x, \ell) + a \tilde{\xi}^*(\gamma^*_a(x, \ell)).
\]

Using Theorem 14 and following a very similar proof of Theorem 16, we get the following discretization results of the perturbed flows and ridges.

Corollary 18 Suppose that assumptions \((A1)\), \((A1)\), \((A2)\), and \((A3)\) hold. Let \(\tilde{R}_a(\tilde{f}) = \{\lim_{\ell \to \infty} \tilde{\gamma}^*_a(x, \ell), x \in \partial \tilde{S}_\epsilon\}\). When \(\max(\delta_0, \delta_1, \delta_2)\) is small enough, there exist \(\epsilon_0 > 0\) and \(\delta_0 > 0\) such that for all \(0 < \epsilon \leq \epsilon_0\), and \(0 < a \leq \delta_a\), we have that

(i) \(\tilde{R}_a(f) \subset \text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon\), and

(ii) there exists a constant \(C > 0\) such that \(d_H(\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon, \tilde{R}_a(\tilde{f})) \leq C a^{1-\sigma_1}\), for some \(\sigma_1 \in (0, 1)\).

Moreover,

(iii) When \(k = 1\) (1-dimensional ridge), we have

\[
\text{Ridge}(\tilde{\eta}) \cap \tilde{S}_\epsilon = \tilde{R}_a(\tilde{f}).
\]

We will first prove several facts that then lead to Theorem 16. The proof of Corollary 18 is omitted because of its similarity to the proof of Theorem 16. Throughout this section we assume without further mention that assumptions \((A1)\), \((A2)\), and \((A3)\) hold.

Fact 1: If \(\epsilon, a > 0\) are small enough, then, for any starting point \(x \in \partial S_\epsilon\), the sequence \(\gamma^*_a(x, \ell)\) converges to a ridge point in \(S_\epsilon\).
As a consequence, the maximal length of the discretized paths with starting points in all of a matrix. By (3.23), we can choose $\epsilon$ small enough such that $S_* \subset \text{Ridge}(f)\left(\frac{1}{2}\delta\right)$, where $\delta$ is given in Lemma 2. Denote $S_{*,\delta} = S_*\left(\frac{1}{2}\delta\right)$. Under assumption (A2) we have

$$
\sup_{x \in S_{*,\delta}} \lambda_{k+1}(x) \leq -\beta < 0. \tag{3.41}
$$

Since $S_*$ is a compact set when $\epsilon$ is small enough and $\|\xi^*\|$ is a continuous function on $S_*$, we can choose $\epsilon$ small enough such that $\sup_{x \in S_*} \|\xi^*(x)\| < \frac{1}{2}\delta$. From Lemma 2 it is known that $0 < \alpha < \kappa_2(x) < A < \infty$, for all $x \in S_*$. Suppose $0 < a < A^{-1} \land 1$ and $\xi^*(x) \neq 0$ (that is, $x$ is not a ridge point). Using a Taylor expansion, we have

$$
\eta(x + a\xi^*(x)) = \eta(x) + a \nabla \eta(x)^\top \xi^*(x) + R(x, a) = \eta(x) + a\|\xi^*(x)\|^2 + R(x, a),
$$

where $|R(x, a)| \leq \frac{1}{2}a^2\kappa(x)\|\xi^*(x)\|^2 \leq \frac{1}{2}a\|\xi^*(x)\|^2$. Therefore

$$
\eta(x + a\xi^*(x)) > \eta(x).
$$

Thus, for $a$ small enough, the sequence $\eta(\gamma^*_\alpha(x, \ell))$ is bounded and increasing, and therefore convergent. Using (3.41) we can see that $\gamma^*_\alpha(x, \ell) \in S_*$ for all $\ell \geq 0$. Moreover, since

$$
\eta(\gamma^*_\alpha(x, \ell + 1)) - \eta(\gamma^*_\alpha(x, \ell)) \geq \frac{1}{2}a\|\xi^*(\gamma^*_\alpha(x, \ell))\|^2,
$$

we have that

$$
\lim_{\ell \to \infty} \|\xi^*(\gamma^*_\alpha(x, \ell))\| \to 0.
$$

Observe further that by using Fact 2, it can be seen that $\{\gamma^*_\alpha(x, \ell)\}_\ell$ is a Cauchy sequence, and from what we have just shown, its limit has to be a point on the ridge (and also in $S_*$).

**Fact 2:** There exists a constant $c_1 > 0$ such that for $a, \epsilon > 0$ small enough, we have for all $x \in S_*$ and $\ell = 0, 1, 2, \ldots$,

$$
\|\gamma^*_\alpha(x, \ell + 2) - \gamma^*_\alpha(x, \ell + 1)\| \leq (1 - c_1a)\|\gamma^*_\alpha(x, \ell + 1) - \gamma^*_\alpha(x, \ell)\|. \tag{3.42}
$$

As a consequence, the maximal length of the discretized paths with starting points in $S_*$ is finite, i.e. $\sup_{x \in S_*} \sum_{\ell = 0}^\infty \|\gamma^*_\alpha(x, \ell + 1) - \gamma^*_\alpha(x, \ell)\| < C < \infty$ for a constant $C > 0$ not depending on $a$. In other words, for small enough values of $a$ and $\epsilon$, the total length of the discretized path based on Euler’s method is bounded uniformly over starting points in $S_*$. 

**Proof** First notice that if there exists $\ell_0 \geq 1$ such that $\xi^*(\gamma^*_\alpha(x, \ell_0)) = 0$, then $\gamma^*_\alpha(x, \ell_0) = \gamma^*_\alpha(x, \ell_0 + 1) = \gamma^*_\alpha(x, \ell_0 + 2) = \cdots$, and the conclusion of this fact is valid. Below we assume that $\xi^*(\gamma^*_\alpha(x, \ell)) \neq 0$ for all $\ell \geq 1$. We have the following Taylor expansion

$$
\|\xi^*(\gamma^*_\alpha(x, \ell + 1))\|^2 = \|\xi^*(\gamma^*_\alpha(x, \ell))\|^2 + \langle \nabla \|\xi^*(y)\|^2 |_{y = \gamma^*_\alpha(x, \ell)} \rangle \gamma^*_\alpha(x, \ell + 1) - \gamma^*_\alpha(x, \ell)\rangle.
$$

28
\[\frac{1}{2} [\gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell)]^\top \{\nabla^2 \| \xi^*(y) \|^2 | y = \tilde{y}_\delta \} [\gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell)] \]

where \( \tilde{y}_\delta = \delta \gamma^*_a(x, \ell) + (1 - \delta) \gamma^*_a(x, \ell + 1) \) for some \( \delta \in (0, 1) \). From the proof of Fact 1, we see that \( a \) and \( \epsilon \) can be chosen small enough such that \( \{ \tilde{y}_\delta : \delta \in [0, 1] \} \subset S_\ell \) for all \( \ell \geq 0 \), which will also be assumed for the remaining proofs.

Let \( \nu(y) = \lambda_{\text{max}} [\nabla^2 \| \xi^*(y) \|^2] \), where \( \lambda_{\text{max}}(B) \) denotes the maximum eigenvalue of the symmetric matrix \( B \). For each \( y \in \text{Ridge}(f) \), we can write \( \nabla^2 \| \xi^*(y) \|^2 = 2[\nabla \xi^*(y)]^\top \nabla \xi^*(y) = 2\nabla^2 \eta(y) V^*_1(y) [V^*_1(y)]^\top \nabla^2 \eta(y) \). Hence \( \nu(y) = 2\lambda^*_a(y)^2 \) for \( y \in \text{Ridge}(f) \). Since \( \nu \) is a continuous function on \( \text{Ridge}(f)^\delta \), we can find an \( \epsilon > 0 \) small enough such that \( 0 < \lambda_{\text{max}} := \sup_{y \in S} \nu(y) < \infty \). We thus obtain

\[
\begin{align*}
\| \xi^*(\gamma^*_a(x, \ell + 1)) \|^2 \\
\leq \| \xi^*(\gamma^*_a(x, \ell)) \|^2 + 2a [\xi^*(\gamma^*_a(x, \ell))^\top [\nabla \xi^*(\gamma^*_a(x, \ell))] [\xi^*(\gamma^*_a(x, \ell))] + \frac{1}{2} a^2 [\xi^*(\gamma^*_a(x, \ell))]^2 \lambda_{\text{max}}.
\end{align*}
\]

(3.43)

Hence

\[
\begin{align*}
\frac{\| \gamma^*_a(x, \ell + 2) - \gamma^*_a(x, \ell + 1) \|^2}{\| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell) \|^2} &= \frac{\| \xi^*(\gamma^*_a(x, \ell + 1)) \|^2}{\| \xi^*(\gamma^*_a(x, \ell)) \|^2} \\
&\leq 1 + 2a \frac{[\xi^*(\gamma^*_a(x, \ell))^\top [\nabla \xi^*(\gamma^*_a(x, \ell))] [\xi^*(\gamma^*_a(x, \ell))] + \frac{1}{2} a^2 \lambda_{\text{max}}.}{\| \xi^*(\gamma^*_a(x, \ell)) \|^2}
\end{align*}
\]

(3.44)

Let \( N(x) = \{ V^*_1(x) [V^*_1(x)]^\top u : u \in \mathbb{R}^d \} \), which is the subspace spanned by \( V^*_1(x), \ldots, V^*_d(x) \), and \( N_0(x) = A(x) \cap S^{d-1} \), where \( S^{d-1} \) denotes the \((d-1)\)-dimensional unit sphere in \( \mathbb{R}^d \).

Note that \( N_0(x) \) is a \((d-k-1)\)-dimensional unit sphere in \( \mathbb{R}^d \). Define

\[
\tilde{\beta}(x) = \sup_{v \in N(x) \setminus \{0\}} \frac{v^\top \nabla \xi^*(x) w}{\| v \|^2} = \sup_{w \in N_0(x)} w^\top \nabla \xi^*(x) w.
\]

(3.45)

It is clear that the second term on the right-hand side of (3.44) can be upper bounded by \( 2a \tilde{\beta}(x) \). Note that for \( x \in \text{Ridge}(f) \),

\[
\nabla \xi^*(x) = V^*_1(x) [V^*_1(x)]^\top \nabla^2 \eta(x) = \sum_{j=k+1}^d \lambda^*_j(x) V^*_j(x) V^*_j(x)^\top.
\]

Therefore for any \( x \in \text{Ridge}(f) \),

\[
\tilde{\beta}(x) = \lambda^*_{k+1}(x) < 0.
\]

(3.46)

Next we will show that \( \tilde{\beta} \) is continuous in \( \text{Ridge}(f)^\delta \). For any \( x, x_0 \in \text{Ridge}(f)^\delta \), we will find a bijective map \( q_{x,x_0} : N_0(x) \to N_0(x_0) \). Let \( \pi/2 \geq \theta_1 \geq \cdots \geq \theta_{d-k} \geq 0 \) be the principal angles between \( N(x) \) and \( N(x_0) \), and suppose that \( u_1(\bar{x}), \ldots, u_k(\bar{x}) \) are the associated principal vectors for \( N(\bar{x}) \), where \( \bar{x} \in \{x, x_0\} \), respectively. In other words, if the singular value decomposition of \( V^*_1(x)^\top V^*_1(x) \) is given by \( P \Sigma R^\top \), where \( P \) and \( R \)
are \((d-k) \times (d-k)\) orthogonal matrix and \(\Sigma\) is a \((d-k) \times (d-k)\) diagonal matrix, then \([u_1(x), \ldots, u_k(x)] = V^*_x(x)P, \Sigma = \cos \Theta, \) and \([u_1(x_0), \ldots, u_k(x_0)] = V^*_x(x_0)R, \) where \(\Theta = \text{diag}(\theta_1, \ldots, \theta_{d-k})\). Using the Davis-Kahan theorem and Lemma 2, we have

\[
\| \sin \Theta \|_F \leq \frac{2}{\beta} \| \nabla^2 \eta(x) - \nabla^2 \eta(x_0) \|_F. \tag{3.47}
\]

We choose \(\|x - x_0\|\) small enough such that \(\theta_1 \leq 2\pi/3\) (so that \(\sin \theta_1 \geq \sin(\theta_i/2)\) for \(i = 1, \ldots, d-k\)). Let \(U(x, x_0) = [\bar{u}_1(x, x_0), \ldots, \bar{u}_{d-k}(x, x_0)]\), where \(\bar{u}_i(x, x_0) = \frac{1}{2(1 + \cos \theta_1)} [u_i(x) + u_i(x_0)], i = 1, \ldots, d-k, \) and \(N(x, x_0)\) be the range of \(U(x, x_0)\). Note that \(U(x, x_0)\) is an orthogonal matrix and \(\Sigma\) is symmetric. Therefore the images of \(N_o(x)\) and \(N_o(x_0)\) under the projection map \(\Omega_{x,x_0} := U(x, x_0)[U(x, x_0)]^\top\) are the same. For \(w \in N_o(x)\), define \(q_{x,x_0}(w)\) in such a way that \(\Omega_{x,x_0} q_{x,x_0}(w) = \Omega_{x,x_0} w\). Here \(q_{x,x_0}(w)\) is uniquely defined because \(\Omega_{x,x_0}\) is bijective from either \(N(x)\) or \(N(x_0)\) to \(N(x, x_0)\). For \(w \in N_o(x)\), using (3.47) we have

\[
\|q_{x,x_0}(w) - w\| = 2\sqrt{\sin^2\left(\frac{\theta_1}{2}\right) + \cdots + \sin^2\left(\frac{\theta_{d-k}}{2}\right)} \leq 2\| \sin \Theta \|_F \leq \frac{4}{\beta} \| \nabla^2 \eta(x) - \nabla^2 \eta(x_0) \|_F. \tag{3.48}
\]

Then we can write

\[
|\bar{\beta}(x) - \bar{\beta}(x_0)| \leq \left| \sup_{w \in N_o(x)} w^\top \nabla \xi^*(x)w - \sup_{w \in N_o(x)} w^\top \nabla \xi^*(x_0)w \right| \\
+ \left| \sup_{w \in N_o(x)} w^\top \nabla \xi^*(x_0)w - \sup_{w \in N_o(x)} w^\top \nabla \xi^*(x_0)w \right| \\
\leq \| \nabla \xi^*(x) - \nabla \xi^*(x_0) \|_F + 2\| \nabla \xi^*(x_0) \|_F \sup_{w \in N_o(x)} \| w - q_{x,x_0}(w) \| \\
\leq \| \nabla \xi^*(x) - \nabla \xi^*(x_0) \|_F + \frac{8\| \nabla \xi^*(x_0) \|_F}{\beta} \| \nabla^2 \eta(x) - \nabla^2 \eta(x_0) \|_F. \tag{3.49}
\]

Using the boundedness of \(\| \nabla \xi^* \|_F\), we can then show that \(\bar{\beta}(x)\) is a uniformly continuous function on \(\text{Ridge}(f)^d\). Because the ridge is a compact set, using (3.46) we are able to find \(\epsilon > 0\) such that

\[
\beta_0 := \sup_{x \in S_\epsilon} \bar{\beta}(x) < 0, \tag{3.50}
\]

and from (3.44) we have that for all \(\ell \geq 0,\)

\[
\frac{\| \gamma^*_a(x, \ell + 2) - \gamma^*_a(x, \ell + 1) \|}{\| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell) \|} \leq 1 + a\bar{\beta}(x) + \frac{1}{4}a^2\lambda_{\text{max}}^* \leq 1 + a\beta_0 + \frac{1}{4}a^2\lambda_{\text{max}}^*.
\]

We require that \(a \in (0, -\frac{2\beta_0}{\lambda_{\text{max}}^*})\). Then

\[
\frac{\| \gamma^*_a(x, \ell + 2) - \gamma^*_a(x, \ell + 1) \|}{\| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell) \|} \leq 1 + \frac{1}{2}a\beta_0.
\]
This is (3.42) with \( c_1 = -\beta_0 \). Let
\[
\kappa^\dagger = \sup_{x \in S_\epsilon} \| \xi^*(x) \|. \tag{3.51}
\]
Notice that \( 0 < 1 + a\beta_0 < 1 \). The quantity \( T_a(x) = \sum_{\ell=0}^{\infty} \| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell) \| \) can be upper bounded by
\[
T_a(x) ≤ \sum_{i=0}^{\infty} a\| \xi^*(x) \| (1 + \frac{1}{2}a\beta_0)^i ≤ \frac{2\kappa^\dagger}{a\beta_0}, \tag{3.52}
\]
with which we conclude the proof with the constant \( C = \frac{2\kappa^\dagger}{a\beta_0} < \infty \).

Next we will show the continuity of \( \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \) as a function of \( x \).

**Fact 3:** The following holds when \( \epsilon \) is small enough: Let \( x, x_0 \in \partial S_\epsilon \) be two starting points. For any \( \eta > 0 \), there exists \( \delta_\eta > 0 \) such that when \( \| x - x_0 \| ≤ \delta_\eta \), we have
\[
\| \lim_{\ell \to \infty} \gamma^*_a(x, \ell) - \lim_{\ell \to \infty} \gamma^*_a(x_0, \ell) \| ≤ \eta.
\]

**Proof** Note that for \( \bar{x} \in \{ x, x_0 \} \),
\[
\gamma^*_a(\bar{x}, \ell + 1) = \bar{x} + \sum_{i=1}^{\ell} \gamma^*_a(\bar{x}, i + 1) - \gamma^*_a(\bar{x}, i) = \bar{x} + a \times \sum_{i=1}^{\ell} \xi^*(\gamma^*_a(\bar{x}, i)).
\]
Therefore using a Taylor expansion we have
\[
\gamma^*_a(x, \ell + 1) - \gamma^*_a(x_0, \ell + 1) = (x - x_0) + a \sum_{i=1}^{\ell} \{ \xi^*(\gamma^*_a(x, i)) - \xi^*(\gamma^*_a(x_0, i)) \}
\]
\[
= (x - x_0) + a \sum_{i=1}^{\ell} \int_0^1 \nabla \xi^*(\gamma^*_a(t, x, x_0)) dt \{ \gamma^*_a(x, i) - \gamma^*_a(x_0, i) \},
\]
where \( \gamma^*_a(t, x, x_0) = t \gamma^*_a(x, i) + (1 - t) \gamma^*_a(x_0, i) \). As in the proof of Fact 1, we can choose \( \epsilon \) small enough such that \( S_\epsilon \subset \text{Ridge}(f^{\frac{1}{2}}(x)) \). Recalling that \( \| \xi^*(x) \| = 0 \) for \( x \in \text{Ridge}(f) \), we obtain from (3.52) that \( \epsilon \) and \( a \) can be chosen small enough such that \( \max \{ T_a(x), T_a(x_0) \} \leq \frac{1}{5} \delta \) for all \( x, x_0 \in \partial S_\epsilon \). Suppose \( \| x - x_0 \| ≤ \frac{1}{6} \delta \) so that it follows from elementary geometry that \( \gamma^*_a(t, x, x_0) \in \text{Ridge}(f)^{\delta} \) for all \( i \geq 0 \), and \( t \in [0, 1] \). Hence with
\[
\kappa^* := \sup_{x \in \text{Ridge}(f)^{\delta}} \| \nabla \xi^*(x) \|_F < \infty, \tag{3.53}
\]
we have
\[
\| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x_0, \ell + 1) \| ≤ \| x - x_0 \| + a\kappa^* \sum_{i=1}^{\ell} \| \gamma^*_a(x, i) - \gamma^*_a(x_0, i) \|. \tag{3.54}
\]
We use the following discrete Gronwall’s inequality (see Holte, 2009): Let \( \{y_n\} \) and \( \{g_n\} \) be nonnegative sequences and \( c \) a nonnegative constant. If
\[
y_n \leq c + \sum_{0 \leq k < n} g_k y_k, \quad n \geq 0,
\]
then
\[
y_n \leq c \exp \left( \sum_{0 \leq j < n} g_j \right), \quad n \geq 0.
\]

Applying this inequality to (3.54), we get
\[
\| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x_0, \ell + 1) \| \leq \| x - x_0 \| \exp(ac*\ell).
\] (3.55)

Recall \( \kappa^\dagger \) given in (3.51). Using the argument in the proof of Fact 2 (in particular, see (3.52)), we have that for any positive integer \( N \),
\[
\| \gamma^*_a(x, N) - \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \| \leq \sum_{\ell = N}^{\infty} \| \gamma^*_a(x, \ell + 1) - \gamma^*_a(x, \ell) \|
\]
\[
\leq a \kappa^\dagger \sum_{\ell = N}^{\infty} (1 - ac_1)^\ell
\]
\[
= \kappa^\dagger \frac{(1 - ac_1)^N}{c_1},
\] (3.56)

where \( 0 < 1 - ac_1 < 1 \). When \( N \geq N_\eta := \frac{\log (\frac{1}{ac_1})}{\log (1 - ac_1)} \), we have \( \| \gamma^*_a(x, N) - \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \| \leq \frac{1}{3} \eta \), for all \( x \in \partial S_e \). Hence for any \( x, x_0 \in \partial S_e \),
\[
\| \gamma^*_a(x, N_\eta + 1) - \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \| \leq \frac{\eta}{3},
\] (3.57)
\[
\| \gamma^*_a(x_0, N_\eta + 1) - \lim_{\ell \to \infty} \gamma^*_a(x_0, \ell) \| \leq \frac{\eta}{3}.
\] (3.58)

Now
\[
\exp(ac* N_\eta) = \exp \left( \frac{-ac_1}{\log (1 - ac_1)} \left[ -\frac{\kappa^*}{c_1} \log (\frac{c_1 \eta}{3\kappa^\dagger}) \right] \right).
\]

Using that \( \frac{t}{1+t} \leq \log (1 + t) \leq t \), for all \( t > -1 \), we get
\[
1 - ac_1 \leq \frac{-ac_1}{\log (1 - ac_1)} \leq 1,
\]
and consequently
\[
\exp(ac* N_\eta) \leq \exp \left( -\frac{\kappa^*}{c_1} \log (\frac{c_1 \eta}{3\kappa^\dagger}) \right).
\]
Combining this with (3.57), we choose $\delta_\eta > 0$ small enough (independent of $a$ and $x$ once they are within a bound) such that when $\|x - x_0\| \leq \delta_\eta$ we have

$$\|\gamma_a^\ast(x, N_\eta + 1) - \gamma_a^\ast(x_0, N_\eta + 1)\| \leq \|x - x_0\| \exp(a \kappa^* N_\eta)$$

$$\leq \delta_\eta \exp \left( - \frac{\kappa^*}{c_1} \log \left( \frac{c_1 \eta}{3 \kappa^*} \right) \right) \leq \frac{\eta}{3},$$

Combining this with (3.57) and (3.58), we have

$$\lim_{\ell \to \infty} \gamma_a^\ast(x, \ell) - \gamma_a^\ast(x_0, \ell) = \lim_{\ell \to \infty} \gamma_a^\ast(x, N_\eta + 1) - \gamma_a^\ast(x_0, N_\eta + 1)$$

$$+ \|\gamma_a^\ast(x_0, N_\eta + 1) - \gamma_a^\ast(x_0, \ell)\| \leq \eta.$$

This proves the assertion. Note that the set $\partial S_f$ is a compact set, so the continuity of a function on this set is equivalent to uniform continuity.

We are going to show the map $\lim_{\ell \to \infty} \gamma_a^\ast(x, \ell)$ is surjective (onto) from $\partial S_f$ to $\text{Ridge}(f)$ when $d = 1$. The continuity of this map still cannot guarantee this property. We need some more facts. First we show a continuous version of Fact 2.

**Fact 4:** When $\epsilon > 0$ is small enough, the total lengths of the paths $\gamma^\ast(x, t)$ are bounded uniformly over $x \in \partial S_f$, i.e. $\sup_{x \in \partial S_f} \int_0^\infty \|\xi^\ast(\gamma^\ast(x, t))\| dt < \infty$.

**Proof** The proof is similar to that of Fact 2, but more involved. Let $t_\ell = a \ell$. Then notice that

$$\int_0^\infty \|\xi^\ast(\gamma^\ast(x, t))\| dt = \sum_{\ell = 0}^\infty \int_{t_\ell}^{t_{\ell + 1}} \|\xi^\ast(\gamma^\ast(x, t))\| dt.$$ 

For each $\ell \geq 0$, using a Taylor expansion for $\lambda(t_{\ell + 1})$ where $\lambda(s) = \int_{t_\ell}^s \|\xi^\ast(\gamma^\ast(x, t))\| dt$, we have

$$\int_{t_\ell}^{t_{\ell + 1}} \|\xi^\ast(\gamma^\ast(x, t))\| dt = a \|\xi^\ast(\gamma^\ast(x, t_\ell))\| + \frac{1}{2} a^2 \frac{d}{dt} \|\xi^\ast(\gamma^\ast(x, t))\| \bigg|_{t = t_\ell},$$

(3.59)

where $t_\ell = (1 - \delta_\ell) t_\ell + \delta_\ell t_{\ell + 1} = t_\ell + \delta_\ell \times a$ for some $\delta_\ell \in (0, 1)$. Here

$$\frac{d}{dt} \|\xi^\ast(\gamma^\ast(x, t))\| = \left\{ [\nabla \|\xi^\ast(y)\|] \xi^\ast(y) \right\} \bigg|_{y = \gamma^\ast(x, t)} = \{\|\xi^\ast(y)\| \theta(y)\} \bigg|_{y = \gamma^\ast(x, t)},$$

(3.60)

where $\theta(y) = \frac{\xi^\ast(y)^\top \nabla \xi^\ast(y) \xi^\ast(y)}{[\xi^\ast(y)]^2}$. Recall that in the proof of Fact 2 we have shown that when $\epsilon > 0$ is small enough,

$$\sup_{y \in S_f \setminus \text{Ridge}(f)} \theta(y) \leq \sup_{y \in S_f} \bar{\beta}(y) = \beta_0 < 0.$$  

(3.61)
Corresponding to the definition of $\beta$ in (3.45), define

$$\beta(x) = \inf_{w \in \ker(x) \setminus \{0\}} \frac{v^T \nabla \xi^*(x) w}{\|v\|^2} = \inf_{w \in \ker(x)} w^T \nabla \xi^*(x) w. \quad (3.62)$$

Using an argument similar to (3.48) and (3.49), we can show that $\beta$ is a continuous function on $\text{Ridge}(f)$. Similar to (3.46), we have that for any $x \in \text{Ridge}(f)$, $\beta(x) = \lambda^*_2(x) < 0$. Therefore we can find $\epsilon > 0$ small enough such that $\beta_1 := \inf_{x \in S_\epsilon} \beta(x) > -\infty$. Note that $\beta_1 \leq \beta_0 < 0$. Then $\inf_{y \in S_\epsilon \setminus \text{Ridge}(f)} \theta(y) \geq \beta_1$. The calculation in (3.60) means that the length of the vector field along the trajectory is strictly decreasing. So from (3.59) we can write

$$\int_{t_{\ell}}^{t_{\ell+1}} \|\xi^*(\gamma^*(x,t))\| dt = a\|\xi^*(\gamma^*(x,t_\ell))\| + \frac{1}{2} a^2 \|\xi^*(\gamma^*(x,\bar{t}_\ell))\| \theta(\gamma^*(x,\bar{t}_\ell)). \quad (3.63)$$

Let $\rho_\ell(x) = \gamma^*(x,t_{\ell+1}) - \gamma^*(x,t_\ell)$. Similar to (3.43) we have

$$\|\xi^*(\gamma^*(x,t_{\ell+1}))\|^2 \leq \|\xi^*(\gamma^*(x,t_\ell))\|^2 + 2[\rho_\ell(x)]^T \nabla \xi^*(\gamma^*(x,t_\ell)) \xi^*(\gamma^*(x,t_\ell)) + \frac{1}{2} \|\rho_\ell(x)\|^2 \lambda^*_\max. \quad (3.64)$$

We have the following Taylor expansion

$$\rho_\ell(x) = \int_{t_\ell}^{t_{\ell+1}} \xi^*(\gamma^*(x,t)) dt = a\xi^*(\gamma^*(x,t_\ell)) + a^2 \nabla \xi^*(y) \xi^*(y)|_{y = \gamma^*(x,\bar{t}_\ell)},$$

where $\bar{t}_\ell = t_\ell + \tilde{\delta}_\ell \times a$ for some $\tilde{\delta}_\ell \in (0, 1)$. Note that we have used the same $\bar{t}_\ell$ in the Taylor expansion of each element of the vector $\rho_\ell$ for the simplicity of notation. Recall $\kappa^*$ given in (3.51). Then we have

$$\left\|\nabla \xi^*(y) \xi^*(y)|_{y = \gamma^*(x,\bar{t}_\ell)}\right\| \leq \kappa^* \|\xi^*(\gamma^*(x,\bar{t}_\ell))\| \leq \kappa^* \|\xi^*(\gamma^*(x,t_\ell))\|,$$

where we have used the fact that $\|\xi^*(\gamma^*(x,t))\|$ is a decreasing function of $t$, as we proved above, and hence

$$\|\rho_\ell(x)\| \leq a\|\xi^*(\gamma^*(x,t_\ell))\| + a^2 \kappa^* \|\xi^*(\gamma^*(x,t_\ell))\|.$$

Then from (3.64) we have

$$\|\xi^*(\gamma^*(x,t_{\ell+1}))\|^2 \leq \|\xi^*(\gamma^*(x,t_\ell))\|^2 + 2a[\xi^*(\gamma^*(x,t_\ell))]^T \nabla \xi^*(\gamma^*(x,t_\ell)) \xi^*(\gamma^*(x,t_\ell))$$

$$+ 2a^2 \nabla \xi^*(y) \xi^*(y)|_{y = \gamma^*(x,\bar{t}_\ell)}^T \nabla \xi^*(\gamma^*(x,t_\ell)) \xi^*(\gamma^*(x,t_\ell)) + \frac{1}{2} \|\rho_\ell(x)\|^2 \lambda^*_\max$$

$$\leq \|\xi^*(\gamma^*(x,t_\ell))\|^2 + 2a[\xi^*(\gamma^*(x,t_\ell))]^T \nabla \xi^*(\gamma^*(x,t_\ell)) \xi^*(\gamma^*(x,t_\ell))$$

$$+ 2a^2 (\kappa^*)^2 \|\xi^*(\gamma^*(x,t_\ell))\|^2 + \frac{1}{2} (a + a^2 \kappa^*)^2 \lambda^*_\max \|\xi^*(\gamma^*(x,t_\ell))\|^2.$$
This leads to
\[
\frac{\|\xi^*(\gamma^*(x, t_{\ell+1}))\|^2}{\|\xi^*(\gamma^*(x, t_\ell))\|^2} \\
\leq 1 + 2a \left[ \frac{\xi^*(\gamma^*(x, t_{\ell}))^\top \nabla \xi^*(\gamma^*(x, t_{\ell}))}{\|\xi^*(\gamma^*(x, t_{\ell}))\|^2} \right] \xi^*(\gamma^*(x, t_{\ell})) + 2a^2(\kappa^*)^2 + \frac{1}{2}(a + a^2\kappa^*)^2 \lambda^*_\text{max} \\
= 1 + 2a \theta(\gamma^*(x, t_\ell)) + 2a^2(\kappa^*)^2 + \frac{1}{2}(a + a^2\kappa^*)^2 \lambda^*_\text{max}.
\]

Therefore
\[
\frac{\|\xi^*(\gamma^*(x, t_{\ell+1}))\|}{\|\xi^*(\gamma^*(x, t_\ell))\|} \leq 1 + a \theta(\gamma^*(x, t_\ell)) + a^2(\kappa^*)^2 + \frac{1}{4}(a + a^2\kappa^*)^2 \lambda^*_\text{max}.
\] (3.65)

From (3.63) we have
\[
\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi^*(\gamma^*(x, t))\| \, dt = a\|\xi^*(\gamma^*(x, t_{\ell+1}))\| + \frac{1}{2}a^2\|\xi^*(\gamma^*(x, t_{\ell+1}))\| \theta(\gamma^*(x, t_{\ell+1})).
\]

As we have shown in (3.61), \( \theta(\gamma^*(x, t_{\ell+1})) < 0, \) for all \( \ell \geq 0. \) Hence by (3.65),
\[
\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi^*(\gamma^*(x, t))\| \, dt \leq a\|\xi^*(\gamma^*(x, t_{\ell+1}))\| \left[ 1 + a \theta(\gamma^*(x, t_\ell)) + a^2(\kappa^*)^2 + \frac{1}{4}(a + a^2\kappa^*)^2 \lambda^*_\text{max} \right].
\] (3.66)

Now we turn back to \( \int_{t_\ell}^{t_{\ell+1}} \|\xi^*(\gamma^*(x, t))\| \, dt. \) Using (3.63) and the facts that \( \xi^*(\gamma^*(x, t)) \) is a decreasing function of \( t \) and that \( \theta(y) \) is a negative function of \( y, \) we have
\[
\int_{t_\ell}^{t_{\ell+1}} \|\xi^*(\gamma^*(x, t))\| \, dt \geq a\|\xi^*(\gamma^*(x, t_\ell))\| + \frac{1}{2}a^2\|\xi^*(\gamma^*(x, t_\ell))\| \theta(\gamma^*(x, t_\ell)).
\] (3.67)

We do a Taylor expansion for \( \theta(\gamma^*(x, t_\ell)), \) and have
\[
\theta(\gamma^*(x, t_\ell)) = \theta(\gamma^*(x, t_\ell)) + (\delta t) \frac{d}{dt} \theta(\gamma^*(x, t)) \bigg|_{t=t_\ell},
\] (3.68)

where \( t_\ell^* = (1 - \delta t^*) t_\ell + \delta t^* \ell_{\ell+1} = t_\ell + \delta t^* \times a \) for some \( \delta t^* \in (0, \delta \ell). \) Here
\[
\frac{d}{dt} \theta(\gamma^*(x, t)) = \frac{d}{dt} \left[ \frac{\xi^*(\gamma^*(x, t))^\top \nabla \xi^*(\gamma^*(x, t))}{\|\xi^*(\gamma^*(x, t))\|^2} \right] \xi^*(\gamma^*(x, t)) + \frac{2}{\|\xi^*(\gamma^*(x, t))\|^4} \left[ (\xi^*(\gamma^*(x, t))^\top \nabla \xi^*(\gamma^*(x, t))) \xi^*(\gamma^*(x, t)) \right]^2
\]
\[
= \pi(\gamma^*(x, t)),
\]

35
where with $\xi^*_\theta(y) = \xi^*(y)/\|\xi^*(y)\|$, 
\[ \pi(y) = \xi^*_\theta(y)\{[\nabla\xi^*(y)]^T\nabla\xi^*(y) - \nabla\xi^*(y)\xi^*_\theta(y) + \xi^*_\theta(y)[\xi^*(y)^T \otimes I_d] \nabla(\nabla\xi^*(y))\xi^*_\theta(y) \}
- 2[\xi^*_\theta(y)\nabla\xi^*(y)\xi^*_\theta(y)]^2.\]

Hence when $\epsilon > 0$ is small enough, we have that
\[ \kappa^\dagger := \sup_{y \in S_\epsilon \setminus \text{Ridge}(f)} |\pi(y)| < \infty.\]

Then from (3.68) we have
\[ \theta(\gamma^*(x,\ell)) \geq \theta(\gamma^*(x,t_\ell)) - a\kappa^\dagger.\]

Plugging this result into (3.67) we get
\[ \int_{t_\ell}^{t_{\ell+1}} \|\xi^*(\gamma^*(x,t))\|dt \geq a\|\xi^*(\gamma^*(x,t_\ell))\| \left\{ 1 + \frac{1}{2}a[\theta(\gamma^*(x,t_\ell)) - a\kappa^\dagger] \right\}. \tag{3.69} \]

Here we require $a < \min\{1, \frac{1}{\kappa^\dagger - \beta_1}\}$ so that the right-hand side of (3.69) is positive. Now combining (3.66) and (3.69) we have
\[ \frac{\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi^*(\gamma^*(x,t))\|dt}{\int_{t_\ell}^{t_{\ell+1}} \|\xi^*(\gamma^*(x,t))\|dt} \leq 1 + \frac{1}{4}a[\theta(\gamma^*(x,t_\ell)) - a\kappa^\dagger]. \tag{3.70} \]

Using (3.61), we can show that if further $a \leq \frac{-\beta_0}{4(\kappa^*)^2 + (1 + \kappa^*)^2\lambda_{\max}^* + 2\kappa^\dagger}$, then
\[ \frac{\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi^*(\gamma^*(x,t))\|dt}{\int_{t_\ell}^{t_{\ell+1}} \|\xi^*(\gamma^*(x,t))\|dt} \leq \frac{1}{4}a\beta_0. \tag{3.71} \]

Then
\[ \int_0^\infty \|\xi^*(\gamma^*(x,\ell))\|dt \leq \int_0^{t_1} \|\xi^*(\gamma^*(x,t))\|dt \sum_{i=0}^{\infty} \left( 1 + \frac{1}{4}a\beta_0 \right)^i \leq \frac{\int_0^{t_1} \|\xi^*(\gamma^*(x,t))\|dt}{1 - (1 + \frac{1}{4}a\beta_0)} \leq \frac{4\kappa^\dagger}{-\beta_0}, \tag{3.72} \]

where $\kappa^\dagger$ is given in (3.51). This completes the proof.

The next fact concerns the comparison of two sequences: $\{\gamma^*(x,t_\ell), \ell \geq 0\}$ and $\{\gamma^*_a(x,\ell), \ell \geq 0\}$, where as above, $t_\ell = at_\ell$.

**Fact 5:** When $\epsilon > 0$ is small enough, there exists $a_0 > 0$ such that when $a < a_0$, for all $x \in \partial S_\epsilon$ we have
\[ \| \lim_{\ell \to \infty} \gamma^*(x,t_\ell) - \lim_{\ell \to \infty} \gamma^*_a(x,\ell) \| \leq Ca^{1-a_0}. \]
for some constant $C > 0$, where $0 < \sigma_0 < 1$ is given in (3.73).

**Proof** We will use some similar arguments as in the proof of Theorem 1 in Arias-Castro et al. (2016). Let $e_\ell = \gamma^*(x, t_\ell) - \gamma_a^*(x, \ell)$. Then

$$
e_\ell+1 = \gamma^*(x, t_{\ell+1}) - \gamma_a^*(x, \ell + 1) = e_\ell + [\gamma^*(x, t_{\ell+1}) - \gamma^*(x, t_\ell)] - [\gamma_a^*(x, \ell + 1) - \gamma_a^*(x, \ell)] = e_\ell + [\gamma^*(x, t_{\ell+1}) - \gamma^*(x, t_\ell)] - a \xi^*(\gamma^*(x, t_\ell))] + a[\xi^*(\gamma^*(x, t_\ell)) - \xi^*(\gamma_a^*(x, \ell))].$$

Recall $\kappa^*$ defined in (3.53). We have

$$||\xi^*(\gamma^*(x, t_\ell)) - \xi^*(\gamma_a^*(x, \ell))|| \leq \kappa^*||\gamma^*(x, t_\ell) - \gamma_a^*(x, \ell)|| = \kappa^*||e_\ell||.$$  

Also we can write

$$\gamma^*(x, t_{\ell+1}) - \gamma^*(x, t_\ell) - a \xi^*(\gamma^*(x, t_\ell)) = \int_{t_\ell}^{t_{\ell+1}} [\xi^*(\gamma^*(x, t)) - \xi^*(\gamma^*(x, t_\ell))] dt.$$  

Hence we have

$$||\gamma^*(x, t_{\ell+1}) - \gamma^*(x, t_\ell) - a \xi^*(\gamma^*(x, t_\ell))|| \leq \int_{t_\ell}^{t_{\ell+1}} ||\xi^*(\gamma^*(x, t)) - \xi^*(\gamma^*(x, t_\ell))|| dt \leq \kappa^* \int_{t_\ell}^{t_{\ell+1}} ||\gamma^*(x, t) - \gamma^*(x, t_\ell)|| dt \leq \kappa^* \kappa^\dagger \int_{t_\ell}^{t_{\ell+1}} |t - t_\ell| dt = \frac{1}{2} \kappa^* \kappa^\dagger (t_{\ell+1} - t_\ell)^2 = \frac{1}{2} a^2 \kappa^* \kappa^\dagger,$$

where $\kappa^\dagger$ is given in (3.51). Therefore

$$||e_{\ell+1}|| \leq (1 + a \kappa^*)||e_\ell|| + \frac{1}{2} a^2 \kappa^* \kappa^\dagger.$$  

Using the Gronwall’s inequality (see Lemma 4 in Arias-Castro et al., 2016), we have

$$||\gamma^*(x, t_\ell) - \gamma_a^*(x, \ell)|| \leq \frac{1}{2} [e^{a \kappa^*} - 1] \kappa^\dagger a.$$  

From (3.56) in the proof of Fact 3 we know

$$||\gamma_a^*(x, \ell) - \lim_{\ell \to \infty} \gamma_a^*(x, \ell)|| \leq 2 \kappa^\dagger \frac{[1 + \frac{1}{2} a \beta_0] \ell}{-\beta_0} \leq \frac{2 \kappa^\dagger}{-\beta_0} e^{\frac{1}{2} a \beta_0}.$$  

Using (3.71) in the proof of Fact 4, the same arguments as in the derivation of (3.56) give that

$$||\gamma^*(x, t_\ell) - \lim_{\ell \to \infty} \gamma^*(x, t_\ell)|| \leq 4 \kappa^\dagger \frac{[1 + \frac{1}{4} a \beta_0] \ell}{-\beta_0} \leq \frac{4 \kappa^\dagger}{-\beta_0} e^{\frac{1}{4} a \beta_0}.$$  

37
Hence combining the above three inequalities, we get
\[ \| \lim_{\ell \to \infty} \gamma^*(x, t_\ell) - \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \| \leq \min_{\ell \geq 0} \psi(\ell), \]
where
\[ \psi(\ell) = \frac{1}{2} \kappa^\dagger a e^{a\kappa^*} + \frac{6\kappa^\dagger}{-\beta_0} e^{\frac{1}{2} a \beta_0}. \]

We consider \( \ell \) in the form of \( \ell_a(\sigma) = \lceil -\frac{\sigma}{a \kappa^*} \log a \rceil \), where \( \sigma > 0 \) is a parameter to be determined. Denote the upper bound we have imposed on \( a \) by \( a_0 \). Choosing \( \sigma = \sigma_0 = 1 + \frac{\beta_0}{4\kappa^*} \), (3.73)
we have
\[ \| \lim_{\ell \to \infty} \gamma^*(x, t_\ell) - \lim_{\ell \to \infty} \gamma^*_a(x, \ell) \| \leq \psi(\ell_a(\sigma_0)) \leq \kappa^\dagger \left( \frac{1}{2} e^{a_0 \kappa^*} + \frac{6}{-\beta_0} \right) a^{1-\sigma_0}. \]

(3.74)

We have the following continuous version of Fact 3.

**Fact 6:** The following holds when \( \epsilon \) is small enough: Let \( x, x_0 \in \partial S_\epsilon \) be two starting points. For any \( \eta > 0 \), there exists \( \delta_\eta > 0 \) such that when \( \| x - x_0 \| \leq \delta_\eta \), we have
\[ \| \lim_{t \to \infty} \gamma^*(x, t) - \lim_{t \to \infty} \gamma^*(x_0, t) \| \leq \eta. \]

**Proof:** Note that for \( \bar{x} \in \{x, x_0\} \in \partial S_\epsilon \),
\[ \gamma^*(\bar{x}, t) = \bar{x} + \int_0^t \xi^*(\gamma^*(\bar{x}, s)) ds. \]

When \( \epsilon \) is small enough, we obtain for any \( t \geq 0 \),
\[ \| \gamma^*(x, t) - \gamma^*(x_0, t) \| \leq \| x - x_0 \| + \int_0^t \| \xi^*(\gamma^*(x, s)) - \xi^*(\gamma^*(x_0, s)) \| ds \]
\[ \leq \| x - x_0 \| + \kappa^* \int_0^t \| \gamma^*(x, s) - \gamma^*(x_0, s) \| ds. \]

Using Gronwall's inequality, we have
\[ \| \gamma^*(x, t) - \gamma^*(x_0, t) \| \leq \| x - x_0 \| e^{\kappa^* t}. \]
This is similar to (3.55). The rest of the proof is similar to the part below (3.55) in the proof of Fact 3, where we use Fact 4 to replace Fact 2. Details are omitted.

Now we are ready to prove the main result in this subsection, Theorem 16.

**Proof** Part (i) follows from Fact 1. Part (ii) is a consequence of Theorem 11 and Fact 5.

To show (iii), we need to show that \( \text{Ridge}(f) \subset R_a(f) \) when \( k = 1 \). In other words, for any \( \bar{x} \in \text{Ridge}(f) \), we want to show that there exists \( x \in \partial S_\epsilon \) such that \( \bar{x} = \lim_{\ell \to \infty} \gamma_a^*(x, \ell) \).
Note that Ridge($f$) is a 1-dimensional closed curve when $k = 1$ under our assumptions (see the discussion of the assumptions). We can parametrize this curve by $\zeta : [0, 1] \rightarrow \text{Ridge}(f)$ with $\zeta(0) = \zeta(1)$. Without loss of generality, we set $\bar{x} = \zeta(0)$. Also we assume that the total length of ridge is 1 and the parametrization is based on the arclength. Due to Fact 5, for a fixed (small) $\epsilon_0 \in (0, 1/8)$, when $a$ and $\epsilon$ are small enough, there exist $x_1, x_2 \in \partial S_\epsilon$ such that

$$\lim_{\ell \to \infty} \gamma_a^*(x_1, \ell) \in \{ \zeta(\alpha) : \alpha \in (0, \epsilon_0) \},$$

$$\lim_{\ell \to \infty} \gamma_a^*(x_2, \ell) \in \{ \zeta(\alpha) : \alpha \in [1 - \epsilon_0, 1) \}.$$ 

In other words, the limit points of $\gamma_a^*(x_1, \ell)$ and $\gamma_a^*(x_2, \ell)$ are not too far away from $\bar{x}$ and they are located on two sides of $\bar{x}$. Let $\alpha_1 \in (0, \epsilon_0]$ and $\alpha_2 \in [1 - \epsilon_0, 1)$ be such that $\lim_{\ell \to \infty} \gamma_a^*(x_1, \ell) = \zeta(\alpha_1)$ and $\lim_{\ell \to \infty} \gamma_a^*(x_2, \ell) = \zeta(\alpha_2)$. Let $\bar{x}_1 \bar{x}_2$ be the shortest curve (a connected set) in $\partial S_\epsilon$ connecting $x_1$ and $x_2$. When $a$ is small enough, we have

$$\{ \lim_{\ell \to \infty} \gamma^*(x, \ell) : x \in \bar{x}_1 \bar{x}_2 \} \subset \{ \zeta(\alpha) : \alpha \in [0, 2\epsilon_0] \cup [1 - 2\epsilon_0, 1) \}. \quad (3.76)$$

Since the image of the continuous map from a connected set is a connected set, the set

$$\gamma_a^*(\bar{x}_1 \bar{x}_2, \infty) := \{ \lim_{\ell \to \infty} \gamma_a^*(x, \ell) : x \in \bar{x}_1 \bar{x}_2 \}$$

is also connected. Since $\gamma_a^*(\bar{x}_1 \bar{x}_2, \infty) \subset \text{Ridge}(f)$, we must have $\bar{x} \in \gamma_a^*(\bar{x}_1 \bar{x}_2, \infty)$, because otherwise we have

$$\{ \zeta(\alpha) : \alpha \in [\alpha_1, \alpha_2] \} \subset \gamma_a^*(\bar{x}_1 \bar{x}_2, \infty),$$

which, however, contradicts (3.76) and Fact 5. Thus there exists $\bar{x}_0 \in \bar{x}_1 \bar{x}_2$ such that $\lim_{\ell \to \infty} \gamma_a^*(\bar{x}_0, \ell) = \bar{x}$, and we complete the proof. $\blacksquare$

**Remark 19** The arguments in the above proof will fail if we consider $k$-dimensional ridges ($k \geq 2$), because the connected sets in $\mathbb{R}^d$ may have “holes” when $d \geq 2$.

To make it work for $k \geq 2$, one could extend $\bar{x}_1 \bar{x}_2$ to a simply connected set (which is a set without holes, like a disk). But for this we need to show that $\lim_{\ell \to \infty} \gamma_a^* \ell$ is injective on this small disk (that is, limit points are different as long as starting points are different). If this could be shown, then Brouwer’s domain invariance theorem (see Brouwer, 1911) could be used to obtain a proof that follows similar ideas as in the one-dimensional case.

4. Reference

Arias-Castro, E., Donoho, D.L. and Huo, X. (2006). Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.*, **34** 326-349.

Arias-Castro, E., Mason, D., and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, **17**(1), 1487-1514.
Barrow, J. D., Bhavsar, S. P., and Sonoda, D. H. (1985). Minimal spanning trees, filaments and galaxy clustering. *Monthly Notices of the Royal Astronomical Society, 216*(1), 17-35.

Brouwer, L.E.J. (1911). Beweis der invarianz desn-dimensionalen gebiets. *Mathematische Annalen, 71* 305-313.

Chen, Y-C., Genovese, C. R., and Wasserman, L. (2015). Asymptotic theory for density ridges. *Ann. Statist., 43*, 1896-1928.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence, 17*(8), 790-799.

Comaniciu, D. and Meer, P. (1999). Mean shift analysis and applications. *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 2 1197-1203.

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

Eberly, D. (1996). Ridges in image and data analysis. Published by Boston, Mass. : Kluwer Academic Publishers. Computational imaging and vision ; v.7

Einbeck, J., Tutz, G. and Evers, L. (2005). Local principal curves. *Statistics and Computing, 15* 301-313.

Einbeck, J., Evers, L. and Bailer-Jones, C. (2007). Representing complex data using localized principal components with application to astronomical data. Lecture Notes in Computational Science and Engineering, Springer, 2007, pp. 180-204.

Fukunaga, K., Hostetler, L.D. (1975). The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Transactions on Information Theory 21*, 32-40.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2009). On the path density of a gradient field. *Ann. Statist. 37* 3236-3271.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). The geometry of nonparametric filament estimation. *Journal of the American Statistical Association, 107*(498), 788-799.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2014). Nonparametric ridge estimation. *Ann. Statist. 42*(4), 1511-1545.

Hastie, T., and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association, Vol. 84, No. 406 (Jun., 1989), pp. 502-516.

Holte, J.M. (2009). Discrete Gronwall lemma and applications, MAA-NCS Meeting at the University of North Dakota, 24 October, 2009.
Kato, T. (1976). *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin.

Lee, J.M. (2013). Introduction To Smooth Manifolds. 2nd Edition. Springer, New York.

Nicolaescu, L. (2011). *An Invitation To Morse Theory* (2nd ed.), Springer.

Nolan D. and Pollard, D. (1987). U-processes: rates of convergence, *Ann. Statist.*, 15, 780-799.

Novikov, D., Colombi, S. and Doré, O. (2006). Skeleton as a probe of the cosmic web: the two-dimensional case. *Mon. Not. R. Astron. Soc.*, 366, 1201-1216.

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

Pulkinnen, S. (2015). Ridge-based method for finding curvilinear structures from noisy data. *Computational Statistics and Data Analysis*, 82, 89-109.

Qiao, W. (2020). Confidence regions for filamentary structures. Technical Report.

Qiao, W. (2021). Asymptotic confidence regions for density ridges. arXiv: 2004.11354. To appear in Bernoulli.

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

Qiao, W. and Polonik, W. (2019). Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electronic Journal of Statistics*, 13, 985-1030.

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

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

Stoica, R.S. Martínez, V.J. and Saar E. (2007). A three-dimensional object point process for detection of cosmic filaments. *Appl. Statist.* 56, 459-477.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wiggins, S. (2003). *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. 2nd edition. Springer-Verlag, New York.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2), 315-323.