Journal of Machine Learning Research 26 (2025) 1-50 Submitted 1/21; Revised 10/25; Published 11/25

Algorithms for ridge estimation with convergence guarantees

Wanli Qiao WQIAO@GMU.EDU
Department of Statistics
George Mason University
4400 University Drive, MS 4A7
Fairfax, VA 22030, USA

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

Editor: Xiaotong Shen

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, by which we mean that the algorithms can asymptotically recover the full ridge set. We consider the new algorithms as alternatives to the Subspace Constrained Mean Shift (SCMS) algorithm for which no such theoretical guarantees are known.

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 (Szymczak et al., 2006), road system (Wang et al., 2015), DNA strands (Backer et al., 2016) or fault lines (Scott et al., 2023) 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 (Novikov et al., 2006). Usually, the first step is the extraction of this structure, and this is the topic discussed below.

Ridges characterize the low-dimensional structures as collection of local maxima of probability density functions in local orthogonal subspace. See Genovese et al. (2014), Chen et al. (2015), Qiao and Polonik (2016), Qiao (2021), and Qiao (2025+) for statistical analyses of ridge estimation.

The estimation of ridges is related to the problem of manifold learning, for which it is assumed that data are observed near a manifold with noise and the task is to recover the manifold. It has been shown in Genovese et al. (2014) that ridges can be used as surrogates for manifold estimation. Some useful references for manifold learning include Niyogi et al. (2008), Genovese et al. (2012a), Fefferman et al. (2020), Yao et al. (2023) and the references therein.

Ridges can also be used for the purpose of (non-linear) dimension reduction. This falls under the more general umbrella of statistical embedding where the goal is to find a low-dimensional representation of the data that is not necessarily in the ambient space but preserves the geometric and/or topological structures in the original data. See Tjøstheim et al. (2023) for a recent survey on this topic.

The literature on the estimation of low-dimensional structures (or filaments) is rich, and different approaches use different geometric ideas. For example, the local principal curves (Einbeck et al., 2005, 2008) are formed by tracking a localized version of the first principal component directions, but it requires selection of good starting points lying on or near the filaments already. The candy model (Stoica et al., 2007) uses possibly connected cylinders (in 3D) of a fixed radius and height to represent the filaments. The medial axis of the data distribution's support (Genovese et al., 2012b) can also be used to estimate filaments, under the assumption that the noise around the filaments is symmetric. The multiscale method developed in Arias-Castro et al. (2006) can be used to detect the presence of a single filament in a noise background. However, it does not provide a low-dimensional filament estimate. Genovese et al. (2009) proposed the concept of the path density to characterize filaments, which does not lend itself to a low-dimensional estimator, either.

One of the most well-known concepts for the extraction of low-dimensional features is principal curves (Hastie and Stuetzle, 1989), which generalized PCA in the nonlinear setting. The principal curve is a smooth curve that passes through the middle of a data set. Any point on a principal curve is defined as the conditional expectation of all the data that project to that point, which is a property called self-consistency. Following this concept, a lot of research work has been generated to investigate the properties and algorithms for principal curves. See, for example, Banfield and Raftery (1992), Tibshirani (1992), Stanford and Raftery (2000) and Verbeek et al. (2002). The fact that principal curves do not always exist, motivated a related line of work on modified principal curves, which started with Kégl et al. (2000). See also Biau and Fischer (2011) and Delattre and Fischer (2020) and references therein.

If the idea of local averaging in the original definition of principal curves is replaced by that of taking a local maximum in the orthogonal subspace, then we obtain the concept of ridges, which first appears in the literature of image analysis. See Eberly (1996). Ridges have a mathematical definition using derivatives up to second order and they come with intuitive geometric interpretation (see Section 2.1). In practice, ridges can be used to estimate filaments with flexible shapes and structures without strong requirements on the starting points of algorithms for ridge extraction. As shown in Ozertem and Erdogmus (2011), the ridge estimators can perform well even when there are loops, bifurcations, and self intersections in data, while these are difficult to handle for the principal curve method.

In this work, we are concerned with the actual extraction of ridges, i.e., we will construct and analyze algorithms. One such algorithm is the popular SCMS (Subspace Constrained Mean Shift) algorithm developed by Ozertem and Erdogmus (2011), which extracts k-dimensional ridges of a d-dimensional density f from a point cloud sampled from f. The algorithm consists in running a corrected (i.e., subspace constrained) mean shift algorithm starting at a data point. For each data point the algorithm then provides an estimate of a point on the ridge.

However, there do not seem to be theoretical guarantees for the SCMS algorithm to consistently estimate the full ridge set, and, as discussed below (see Section 2.3 and the Appendix), the SCMS algorithm might miss some parts of the ridge, although the pointwise convergence property of SCMS is studied in Zhang and Chen (2023). In other words, it is not entirely clear what the SCMS is actually estimating. Even though it appears that this does not have a serious impact in many practical examples, this theoretical gap provides a motivation for developing alternative ridge finding algorithms that (i) come with theoretical guarantees offering deeper insights to their behavior, and (ii) do not suffer from potentially missing some parts of the ridge. We mention in passing that there exists another ridge estimation algorithm developed in Pulkkinen (2015), which tracks the ridge lines. However, it relies on a starting point that has been on or close to the ridge.

The remaining part of the paper is organized as follows. In Section 2 we introduce the formal definition of ridges. This is followed by our extraction algorithms, whose performance is illustrated using some numerical studies in $\mathbb{R}^2$, where we give an example for which the SCMS algorithm fails to detect a part of the ridge while our algorithms do not miss it. The main theoretical results are given in Section 3, where we give the convergence results of our algorithms. The mathematical framework for the theoretical analyses is provided in Section 4. All the proofs are provided in the appendix.

2. Extraction of filamentary structures

2.1 Definition

Let f denote a density on $\mathbb{R}^d$ from which data will be drawn. We will assume that f is (at least) twice differentiable. The definition of ridge (or filament) points is as follows:

Definition 1 *(Ridge in $\mathbb{R}^d$). Let $(\lambda_i^f(x), V_i^f(x)), i = 1, \ldots, d$ be eigenpairs of the Hessian $\nabla^2 f(x)$ with $\lambda_1^f(x) \geq \cdots \geq \lambda_d^f(x)$. Let $1 \leq k \leq d-1$. With $V_\perp^f(x) = [V_{k+1}^f(x), \cdots, V_d^f(x)] \in \mathbb{R}^{d\times(d-k)}$ the matrix of the trailing $(d-k)$ eigenvectors, we define*

$$\text{Ridge}(f) = \big\{x \in \mathbb{R}^d : \; V_\perp^f(x)^\top \nabla f(x) = \mathbf{0} \textit{ and } \; \lambda_{k+1}^f(x) < 0\big\}. \tag{2.1}$$

The geometric intuition underlying this definition is the following: Since the (first order) directional derivative of $f(x)$ along $V_i^f(x)$ is $\langle \nabla f(x), V_i^f(x)\rangle$ and the second order directional derivative of $f(x)$ along $V_i^f(x)$ is $\lambda_i^f(x)$, the two conditions in (2.1) mean that a point x on the ridge is a local mode in the linear subspace spanned by $V_i^f(x), \; i = k+1, \cdots, d$, for which the density has the largest concavity (see Eberly, 1996).

Given an iid sample $X_1, \cdots, X_n$ from f, we estimate Ridge(f) by

$$\text{Ridge}(\widehat{f}\,) = \big\{x \in \mathbb{R}^d : \; V_\perp^{\widehat{f}}(x)^\top \nabla \widehat{f}(x) = \mathbf{0} \text{ and } \; \lambda_{k+1}^{\widehat{f}}(x) < 0\big\},$$

where $\widehat{f}$ is a kernel density estimate (KDE), $\big(\lambda_i^{\widehat{f}}(x), V_i^{\widehat{f}}(x)\big), i = 1, \ldots, d$ are the eigenpairs of the Hessian $\nabla^2 \widehat{f}(x)$, where we assume the eigenvalues to be sorted as $\lambda_1^{\widehat{f}}(x) \geq \cdots \geq$

$\lambda_d^{\widehat{f}}(x)$. Furthermore, let $V_\perp^{\widehat{f}}(x) = \big[V_{k+1}^{\widehat{f}}(x), \cdots, V_d^{\widehat{f}}(x)\big]$ be the matrix of the $(d-k)$ trailing orthonormal eigenvectors of $\nabla^2 \widehat{f}(x)$. We recall that the KDE has the form

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

where $h > 0$ is a bandwidth, $K \geq 0$ and $\int_{\mathbb{R}^d} K(u)du = 1$. The goal now is to construct and study algorithms to extract the estimated ridges from data. It is well known that KDE suffers from the curse of dimensionality and so we do not consider high dimensions in this paper.

Note that the matrices $V_\perp^f(x)$ and $V_\perp^{\widehat{f}}(x)$ (and similar matrices defined below) depend on k (and d). Since k (and d) are held fixed in the theoretical and methodological developments of this paper, this dependence is not indicated in the notation.

2.2 Mean shift algorithm and subspace constrained mean shift algorithm

The popular mean shift algorithm (Fukunaga and Hostetler, 1975) is being used for mode finding and clustering, e.g. see Cheng (1995) and Comaniciu and Meer (1999, 2002). It is tracking non-parametric estimates of gradients of a KDE. Using the KDE with differentiable $K(x) = \phi(\|x\|^2)$, the vector

$$m(x) := \sum_{i=1}^n w_i(x) X_i - x \qquad \text{with} \qquad w_i(x) = \frac{\phi'\big(\big\|\frac{x-X_i}{h}\big\|^2\big)}{\sum_{i=1}^n \phi'\big(\big\|\frac{x-X_i}{h}\big\|^2\big)} \tag{2.2}$$

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, the basic mean shift algorithm iteratively finds a sequence of points $y_1, y_2, \cdots$ by

$$y_{j+1} = m(y_j) + y_j = \sum_{i=1}^n w_i(y_j) X_i, \quad j = 0, 1, 2, \cdots. \tag{2.3}$$

Successively connecting these points provides an estimate of the integral curve driven by the gradient, starting from y_0. See Arias-Castro et al. (2016) and Arias-Castro and Qiao (2025+). The endpoint of this iteration (after applying some stopping criterion) is considered to be an estimate of a mode (local maximum) of f. When running this algorithm repeatedly with each data point as a starting point, clusters can be formed by grouping all the data points for which the mean shift algorithm has the same endpoint.

Subspace constraints come into the picture when the target is a ridge rather than a mode. The construction of the SCMS algorithm modifies the just described hill climbing algorithm by following the direction of the gradient projected on the subspace spanned by the trailing $(d-k)$ eigenvectors of the Hessian of the KDE. The gradient direction is approximated by the mean shift. More specifically, given a starting point y_0, the SCMS generates a sequence

$$y_{j+1} = \Pi^{\widehat{f}}(y_j)\, m(y_j) + y_j, \quad j = 0, 1, 2, \ldots \tag{2.4}$$

where $m(y)$ is the mean shift vector, and $\Pi^{\widehat{f}}(y) = V_\perp^{\widehat{f}}(y)V_\perp^{\widehat{f}}(y)^\top$ is the projection matrix onto the subspace spanned by the trailing $(d-k)$ eigenvectors of the Hessian of the KDE evaluated at y. Notice that for $y \in \mathrm{Ridge}(\widehat{f}\,)$, this space is orthogonal to $\nabla \widehat{f}(y)$. This motivates that the endpoint of this iteration (after applying some stopping criterion) is considered to be an estimated ridge point. We note that in the original SCMS algorithm proposed in Ozertem and Erdogmus (2011), $\Pi^{\widehat{f}}(y_j)$ is replaced by $\Pi^{\log \widehat{f}}(y_j)$, and this corresponds to the ridge estimation of $\log f$. We focus on the ridge estimation of f in this paper, although the analysis can be easily adapted to that of $\log f$. The SCMS algorithm is very popular, and it gives nice results in practice. However, as discussed next (and in the Appendix A), the SCMS algorithm might miss some parts of the ridge.

2.3 The SCMS algorithm might miss some parts of the ridge

By definition, a ridge point x_0 is a local maximum in the directions given by the columns of $V_\perp^f(x_0)$. So the goal of the SCMS algorithm is to stay in this subspace by projecting the mean shift vectors back into this space in each iteration step. Not knowing the ridge points (i.e. not knowing the target subspace), the algorithm projects the gradient at the current iterate y_j on the subspace spanned by the trailing eigenvectors of the Hessian at this point y_j. Thus, the subspace to project on changes with each iteration step. Indeed these subspaces are tangent spaces to the integral curve/surface traced by the SCMS algorithm. It turns out that, because of this, the signs of the directional derivatives taken along these curves/surfaces are not necessarily determined by the signs of the eigenvalues of the Hessian. As a consequence, ridge points are not necessarily local maximizers when traveling along the integral curves/surfaces traced by the SCMS algorithm, but they can also be local minimizers or saddle points, and if they are, they are not identified as a ridge point by the algorithm. More details are provided in Appendix A and an example can be found in Section 2.8.

2.4 Measuring ridgeness

The new ridge finding algorithms proposed in this work are based on the following ridgeness measure (actually measuring 'non-ridgeness'):

$$\eta(x) = -\frac{1}{2}\left\|V_\perp^f(x)^\top \nabla f(x)\right\|^2. \tag{2.5}$$

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

$$\eta(x) = 0 \quad \text{s.t.} \quad \lambda_{k+1}^f(x) < 0, \tag{2.6}$$

or, since $\eta \le 0$, ridge points are global maximizers of η. In other words, ridge points can be described as

$$\underset{x}{\operatorname{argmax}}\, \eta(x) \quad \text{s.t.} \quad \lambda_{k+1}^f(x) < 0.$$

Note that here "argmax" denotes the entire set of maximizers. From this point of view, ridge finding algorithms can be obtained as algorithms maximizing the data-dependent version of $\eta(x)$ given by

$$\text{find } \textit{all} \text{ maximizers of } \quad \widehat{\eta}(x) = -\frac{1}{2}\left\|V_\perp^{\widehat{f}}(x)^\top \nabla \widehat{f}(x)\right\|^2, \quad \text{subject to} \quad \lambda_{k+1}^{\widehat{f}}(x) < 0. \qquad (2.7)$$

The constraint on the eigenvalue will simply be enforced by excluding maximizers violating this condition.

2.5 Notation

Here we collect some important notation used in this work. Let $g, h : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function. Then we use the following notation:

- $\lambda_1^g(x) \geq \cdots \geq \lambda_d^g(x)$: sorted eigenvalues of Hessian of g
- $V_i^g(x)$: eigenvector of Hessian of g corresponding to $\lambda_i^g(x), i = 1, \ldots, d$
- $V_\perp^g(x)$: $(d \times (d-k))$-matrix formed by the $(d-k)$ trailing eigenvectors of Hessian of g
- $\Pi^g(x) = V_\perp^g(x)V_\perp^g(x)^\top$: projection matrix onto linear space spanned by columns of $V_\perp^g(x)$
- $\xi^g(x) = \Pi^g(x)\nabla g(x)$: projection of the gradient of g onto the linear space spanned by $(d-k)$ trailing eigenvectors of Hessian of g
- $\eta(x) = -\frac{1}{2}\left\|V_\perp^f(x)^\top \nabla f(x)\right\|^2 = -\frac{1}{2}\|\xi^f(x)\|^2$: ridgeness function of f
- $\widehat{\eta}(x) = -\frac{1}{2}\left\|V_\perp^{\widehat{f}}(x)^\top \nabla \widehat{f}(x)\right\|^2 = -\frac{1}{2}\|\xi^{\widehat{f}}(x)\|^2$: ridgeness function of KDE $\widehat{f}$
- $\widehat{\eta}_\tau(x)$: smoothed ridgeness function, where τ is a smoothing parameter (see 2.10).
- $$S_\epsilon^{g,h} := \{x \in [0,1]^d : \; g(x) \geq -\epsilon, \lambda_{k+1}^h(x) < 0\}, \qquad (2.8)$$

Recall that $k \in \{0, \ldots, d-1\}$ is fixed throughout the paper (except for the numerical illustrations).

2.6 Algorithms

To compute the maximizers of the ridgeness function we now consider two algorithms based on $\widehat{\eta}(x)$. As can be seen below, our proposed algorithms target the ridge of the ridgeness function, and we will see below (see Lemma 6) that the ridge of the ridgeness function essentially equals the original ridge of f.

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

$$y^{j+1} = y^j + a\,\xi^{\widehat{\eta}}(y^j)$$

$$\begin{aligned} &= y^j - a\,\Pi^{\widehat{\eta}}(y^j)\,[\nabla\xi^{\widehat{f}}(y^j)]^\top\,\xi^{\widehat{f}}(y^j) \\ &= y^j - a\,\Pi^{\widehat{\eta}}(y^j)\,[\nabla\xi^{\widehat{f}}(y^j)]^\top\Pi^{\widehat{f}}(y^j)\,\nabla\widehat{f}(y^j),\ j = 0,1,2,\cdots \end{aligned}$$

More precisely, the structure of the algorithm is as follows:

```
Input: $y_i^0 = X_i,\ i = 1,\cdots,n,\ a > 0,\ h > 0.$

Update: For $i = 0,1,2,\ldots,n$, for $j = 1,2,\ldots,$
  while $y_i^j \in [0,1]^d$:
      $y_i^{j+1} = y_i^j - a\,\Pi^{\widehat{\eta}}(y_i^j)\,[\nabla\xi^{\widehat{f}}(y_i^j)]^\top\,\Pi^{\widehat{f}}(y_i^j)\,\nabla\widehat{f}(y_i^j),$   (2.9)
  else: discard the entire sequence $y_i^0, y_i^1, \ldots$
Output:   $\{y_i^\infty :\ \widehat{\eta}(y_i^\infty) = 0,\ \lambda_{k+1}^{\widehat{f}}(y_i^\infty) < 0\}.$
```

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.

Basic Algorithm 2: *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 $\widehat{\eta}(x)$ as

$$\widehat{\eta}_\tau(x) = \frac{1}{\tau^d}\int_{\mathbb{R}^d} L\left(\frac{x-u}{\tau}\right)\widehat{\eta}(u)du. \tag{2.10}$$

Our algorithm will approximate the ridge of this smoothed version of the ridgeness function. Using this smoothed version has some computational advantages (see Section 2.7.3 below). Let $V_\perp^{\widehat{\eta}_\tau}(x)$ be the matrix whose columns are the trailing $(d-k)$ orthonormal eigenvectors of $\nabla^2\widehat{\eta}_\tau(x)$, and let $\xi^{\widehat{\eta}_\tau}(x) = \Pi^{\widehat{\eta}_\tau}(x)\ \nabla\widehat{\eta}_\tau(x)$, where $\Pi^{\widehat{\eta}_\tau}(x) = V_\perp^{\widehat{\eta}_\tau}(x)V_\perp^{\widehat{\eta}_\tau}(x)^\top$. With this notation, the structure of the algorithm is as follows:

```
Input: Given $y_i^0 = X_i,\ i = 1,\cdots,n,\ a > 0,\ \tau > 0,\ h > 0.$

Update: For $i = 0,1,2,\ldots,n$, for $j = 1,2,\ldots,$
  while $y_i^j \in [0,1]^d$:
      $y_i^{j+1} = y_i^j + a\,\xi^{\widehat{\eta}_\tau}(y_i^j)$
               $= y_i^j + a\,\Pi^{\widehat{\eta}_\tau}(y_i^j)\,\nabla\widehat{\eta}_\tau(y_i^j),$
  else: discard the entire sequence $y_i^0, y_i^1, \ldots$
Output:   $\{y_i^\infty :\ \widehat{\eta}_\tau(y_i^\infty) = 0,\ \lambda_{r+1}^{\widehat{f}}(y_i^\infty) < 0\}.$
```

Remark 2 *While the algorithms above are defined by using the data as starting points, we do have other options. What we need is a set of starting points that becomes dense in the support $[0,1]^d$, such as a fine grid, where, for the theory, the grid size would tend to zero. The theory presented in this work is using a continuous set. It seems possible to extend this theory to the case of the data being the starting points, but we do not pursue this case here.*

Remark 3 *The proposed algorithms have time complexity $O(kn^2d^3)$, where k is the number of iterations, n is the number of the sample points, and d is the space dimension, because these are gradient-ascent type algorithms (which have time complexity of $O(kn^2)$), and the eigen-decomposition required by the algorithms at each step has complexity $O(d^3)$).*

2.7 Practical implementation and illustration of the algorithms

For practical implementation, the basic algorithms given above require a stopping criterion, choice of tuning parameters, and 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 the above two basic algorithms, for each i, we stop the iterations when $\|y_i^j - y_i^{j-1}\| \leq \varepsilon_{\text{tol}}$ for some small tolerance $\varepsilon_{\text{tol}} > 0$, and we use y_i^j as the final point, denoted y_i^*, for the sequence starting from y_i^0. Here ε_{tol} can be chosen as small as possible, with smaller ε_{tol} giving better accuracy and heavier computation cost.

The bandwidth h in the kernel estimators impacts the rate of convergence for ridge estimation. Since the ridge is determined by up to the second order derivatives of f, it is recommended that the optimal bandwidth for the estimation of the Hessian of f is used for ridge estimation. However, if the density on the ridge is completely flat, the ridge becomes a set of degenerate local maxima and then the optimal bandwidth for gradient estimation should be used. See Qiao (2025+). Plug-in and cross-validation approaches are most commonly used for selecting such bandwidths (see Chacón and Duong, 2013).

The smoothing parameter τ used in the Basic Algorithm 2 can be chosen as small as possible. In fact, our theoretical results in Theorem 9 suggests that τ will not change the rate of convergence for ridge estimation as along as $\tau = O(h) \to 0$. However, if the goal is to approximate $\text{Ridge}(\widehat{f}\,)$ through the Basic Algorithm 2, then τ should be much smaller than h. In practice, the computation of the integral in (2.10) can be done using numerical methods. For example, one can evaluate $\widehat{\eta}$ on a grid, and based on this, (2.10) can be approximated by a Riemann sum (see (2.11)). If the kernel L has bounded support, then τ determines the number of grid points near x effectively used in the approximation. The interplay between the grid size and τ again leads to a question between balancing accuracy and computation cost.

The step size a plays a similar role as the learning rate in gradient descent. It is well-known that if such a hyper-parameter is too large, then a overshooting problem can occur. Usually it is safer to use a small a to achieve convergence of the algorithms, with the downside of slow convergence speed. To find a balance and determine a good value of a, one can use trial and error, which is in fact a commonly used approach for choosing the learning rate for gradient descent (Bengio, 2012).

In practice our algorithms can encounter 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. We note that this challenge is faced by all ridge estimation algorithms, due to the fact that ridges are local features which may arise in any low-estimated-density regions as long as the density is positive, even when true ridges do not exist there. A second challenge is possible local (but non-global) modes of our ridgeness function η, which again might lead to spurious ridge points. This challenge is relatively easy to handle, because the global maximum of η is known, which is zero. This known maximum provides a way to distinguish between local and global modes of η. We address these challenges by introducing the following pre-processing and post-processing steps.

2.7.1 Pre-processing

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, however, we are replacing $[0,1]^d$ by the set $\{x \in \mathbb{R}^d : \widehat{f}(x) \geq \varepsilon_f\}$, for a given threshold $\varepsilon_f \geq 0$. We can choose ε_f as an α-quantile of the distribution of $\{\widehat{f}(X_1), \ldots, \widehat{f}(X_n)\}$. In our numerical studies, we used $\alpha = 5\%$ unless otherwise noted. A similar idea of using ε_f has been suggested in Genovese et al. (2014). Notice that under some mild assumptions, the upper-level set $\{x \in \mathbb{R}^d : \widehat{f}(x) \geq \varepsilon_f\}$ is known to be a consistent estimate of $\{x \in \mathbb{R}^d : f(x) \geq \varepsilon_f\}$ (see, e.g., Qiao and Polonik (2019)), and in our theoretical investigations, we could replace the compact set $[0,1]^d$ by $\{x \in \mathbb{R}^d : f(x) \geq \varepsilon_f\}$, and assume that the density at the ridge points is larger than $2\varepsilon_f$, say. Using a consistent data-dependent estimate for ε_f could also be dealt with theoretically, even though we are not explicitly considering this in the theory section. Note that the estimated upper level set plays the role of $[0,1]^d$ in the algorithms, so that if a sequence moves out of this region, it will be discarded. Alternatively, one can also use an estimate of the density support, for example, by using the alpha-convex hull (Rodríguez-Casal, 2007).

2.7.2 Post-processing

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

To address the problem of possible local maxima of the functions $\widehat{\eta}$ and $\widehat{\eta}_\tau$, we remove an output point y_i^*, if $\widehat{\eta}(y_i^*) < -\varepsilon_\eta$, and $\widehat{\eta}_\tau(y_i^*) < -\varepsilon_\eta$, respectively, for some small $\varepsilon_\eta > 0$. Note that in an ideal scenario where we can obtain y_i^∞ as in the two basic algorithms, we can choose $\varepsilon_\eta = 0$, because these two algorithms converge to the estimated ridge as $j \to \infty$, as shown in our theoretical results in Theorem 11. However, in practice, we impose a stopping criterion with tolerance ε_{tol}, which prevents the algorithms from reaching the estimated ridge points exactly. This also explains the necessity of $\varepsilon_\eta > 0$, which intrinsically depends on ε_{tol}. If ε_{tol} is chosen very small, so should be ε_η. However, we do not know a fixed relation between these two parameters so that the latter can automatically chosen based on the former.

A practical choice of ε_η is as follows. Consider the quantile function of the (one-dimensional, discrete) distribution given by the values $\{\widehat{\eta}(y_1^*), \ldots, \widehat{\eta}(y_n^*)\}$ assuming that all the y_i^* are included in the final output, and choose ε_η as the location of the "last significant jump" of the quantile function. In other words, we are removing "outliers" in this distribution. It is possible to automatically identify such jumps by using change point detection techniques, however we do not pursue this further. In our numerical experiments, this first significant jump was clearly visible. An example is given in Figures 2.1 and 2.2

below, which are based on Algorithm 2 with step length $a = 0.005$, while Algorithm 1 gives very similar results.

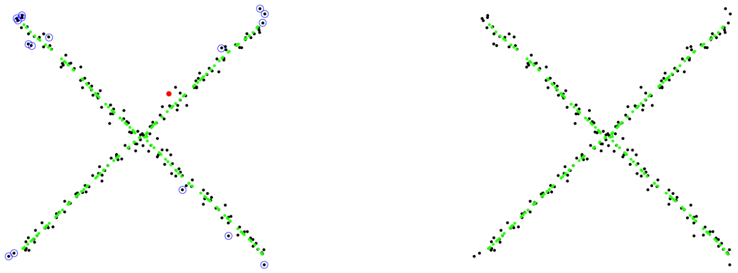


Figure 2.1: *X-cross example with 200 data points; black solid dots are data points; blue circles are points removed by pre-processing; red dot is the point removed by post-processing; green dots are the final output of the algorithm; right panel shows the final result with red dot and blue circles removed from the left panel.*

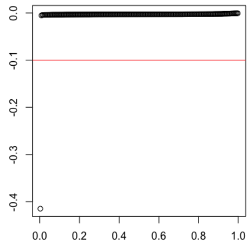


Figure 2.2: *Sorted ridgeness values of the output of Algorithm 2 run on 200 data points with X-cross as ridges without imposing the threshold ε_η. A clear jump is visible between the ridgeness values close to zero and those of spurious ridge points. This observation then informed our choice of $\varepsilon_\eta = 0.1$ (red line). The point below this threshold corresponds to the red dot in the left panel of Figure 2.1.*

Our algorithms are also implemented using two additional data sets, as shown in Figures 2.3 and 2.4, 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 2.1.

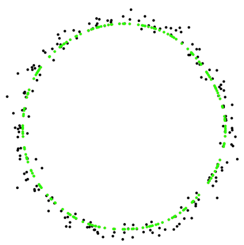


Figure 2.3: *Circle: 200 data points.*

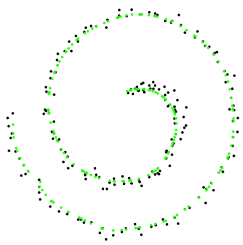


Figure 2.4: *Spiral: 200 data points.*

2.7.3 Other more practical aspects

In Algorithm 1, we need to compute $\nabla\widehat{\xi}(x)$ and eigenvalues and eigenvectors of $\nabla^2\widehat{\eta}(x)$. This requires some matrix algebra, and we present some explicit formulas in Appendix A.

Algorithm 2 avoids direct evaluation of these formulas, 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 $\nabla\widehat{\eta}$ and $\nabla^2\widehat{\eta}$ in Algorithm 1 are replaced by their numerical approximations ($\nabla\widehat{\eta}_\tau$ and $\nabla^2\widehat{\eta}_\tau$) based on the evaluations of $\widehat{\eta}$ in Algorithm 2. In practice, we implement the computation of $\nabla\widehat{\eta}_\tau$ and $\nabla^2\widehat{\eta}_\tau$ as follows. Let $\{x_i, i = 1, 2, \cdots\}$ be a grid over $\mathbb{R}^d$ with grid length $\rho < \tau$. Then $\nabla\widehat{\eta}_\tau$ is approximated by

$$\frac{\rho^d}{\tau^d}\sum_i \nabla\Big[L\left(\frac{x - x_i}{\tau}\right)\Big]\widehat{\eta}(x_i), \tag{2.11}$$

and $\nabla^2\widehat{\eta}_\tau$ 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.

2.8 Simulation study

We conducted a small simulation study to compare the performance of our algorithms 1 and 2 with SCMS. We first consider the ridge set of the density f of a bivariate random vector $X = Y + Z$, where Y has a distribution restricted on a circle $\mathcal{C}$ with center at the origin and unit radius, and $Z \sim N(0, 0.05^2)$. Two models depending on the distributions of $Y = (\cos(\Theta), \sin(\Theta))$ were used: Θ has a uniform distribution on $[0, 2\pi]$ for Model 1 and Θ has a density $[\sin(\theta) + 2]/4\pi$ for $\theta \in [0, 2\pi]$ for Model 2. Note that the distribution of X is a convolution of those of Y and Z, and therefore ridge(f) slightly deviates from $\mathcal{C}$. We use the Hausdorff distance (see (3.1) below for the definition) between the algorithm outputs and the true ridge ridge(f) as the error in the estimation. We compare the performance of the our algorithms and SCMS using this error based on 200 random samples from each of the two models. The bandwidth h used in the kernel estimates was set to be 0.2 for sample size $n = 500$ and $\tau = 0.001$ for Algorithm 2. No density estimate cutoff was used, i.e., $\varepsilon_f = 0$. All the three algorithms used the same starting points for each model, which are grid points near the ridges. The results are shown in Figure 2.5 and Table 1, where we see that in this example all the three algorithms are able to estimate the true ridges well and their performances are almost identical.

	Alg 1	Alg 2	SCMS
Model 1	0.0360 (0.0043)	0.0360 (0.0043)	0.0359 (0.0043)
Model 2	0.0369 (0.0058)	0.0369 (0.0058)	0.0369 (0.0057)

Table 1: *This table contains the means and standard deviations (in parentheses) of the errors shown in Figure 2.5.*

Next we provide an example for which SCMS fails to detect a part of a ridge while our algorithms are able to capture it. Consider the density function

$$f(u, v) = \frac{3}{8}(1 - p)(1 - u^2)v + \frac{1}{4}p, \quad u \in [-1, 1], \quad v \in [0, 2], \tag{2.12}$$

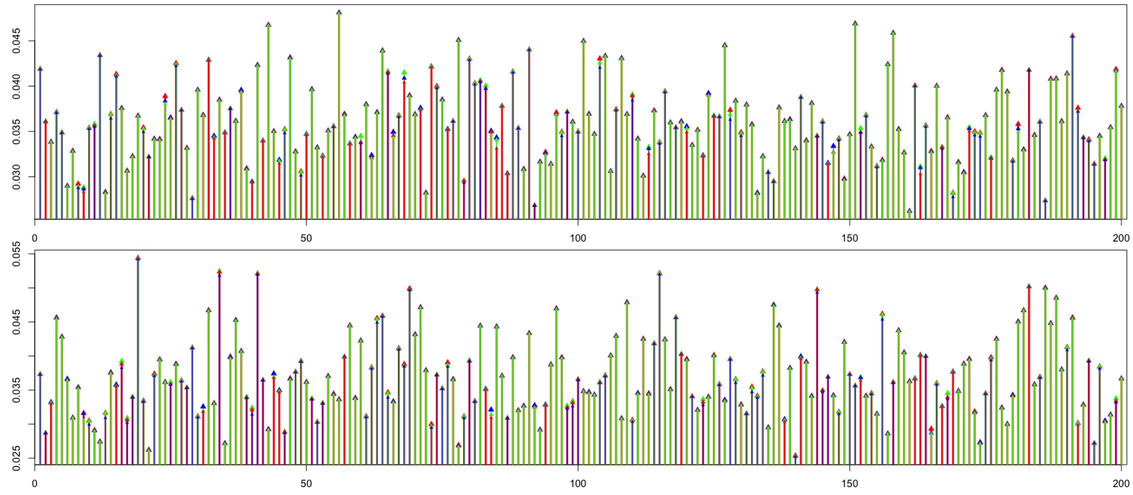


Figure 2.5: *Plots of errors for Model 1 (upper panel) and Model 2 (lower panel), each using 200 random samples; heights of red, green and blue triangles represent estimation errors for Algorithms 1,2 and SCMS, respectively.*

where $p \in (0, 1)$. 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 $\big(0, \frac{1}{\sqrt{2}}\big)$.

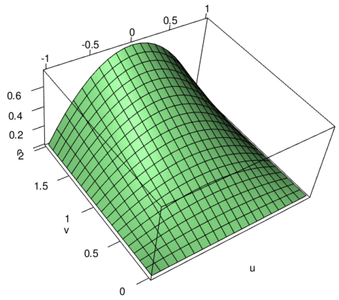


Figure 2.6: *Surface plot of the function f in (2.12).*

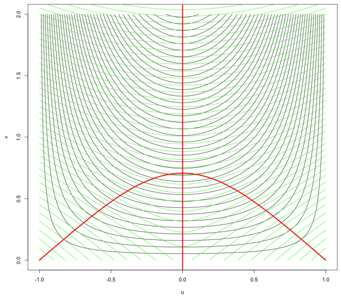


Figure 2.7: *contour lines (black), ridge lines (red), and integral curves driven by $V_\perp$ (green).*

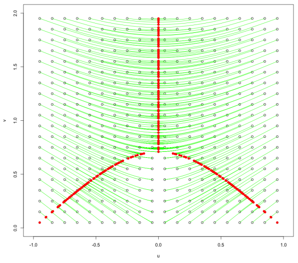


Figure 2.8: *green curves are the trajectories of ξ^f, and the red dots are the limit points.*

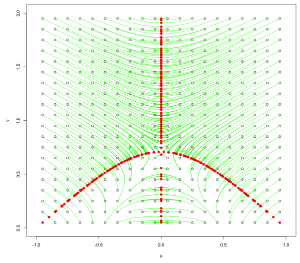


Figure 2.9: *green curves are the trajectories of ξ^η, and the red dots are the limit points.*

The plots of this function when $p = 0$ and the ridge are given in Figures 2.6 and 2.7. Note that here p presents the proportion of a uniform background noise and does not affect the location of the ridge set in the model. In Figures 2.8 and 2.9 we compare the ideas behind the SCMS algorithm and our new algorithms. It can be seen that a piece of the ridge $S_{\text{missing}} := \{0\} \times (0, 1/\sqrt{2})$ is failed to be detected using the idea of the SCMS algorithm in this example.

We also tested our new algorithms and SCMS on random samples generated from the density in (2.12) with $p = 0.3$. For all the three algorithms, we used the same 200 replicates of size 10000, the same bandwidth 0.3, and the same 50×50 grid points as the starting points. We set $\tau = 0.005$ for Algorithm 2. To handle the boundary effect of the kernel estimation especially near the top boundary $[-1, 1] \times \{2\}$, caused by the abrupt change in density beyond the boundary, for each sample we doubled the original sample size by adding a reflected data set across this boundary, and then only kept the ridge estimates within $[-1 + \delta, 1 - \delta] \times [\delta, 2 - \delta]$ for $\delta = 0.1$. The true ridge in this region is used to compute the errors of estimation in terms of the Hausdorff distance. The results are shown in Figure 2.10 and Table 2, with an illustrative case given in Figure 2.11. It is clear to see that the ridge is estimated well by our Algorithms 1 and 2. Notably, our algorithms can find a piece of ridge corresponding to S_{missing} but SCMS fails to find it. In addition, the estimated ridge pieces by the SCMS algorithm near near $(0, 1/\sqrt{2})$ are severely broken, while they are almost connected in the outputs of our algorithms.

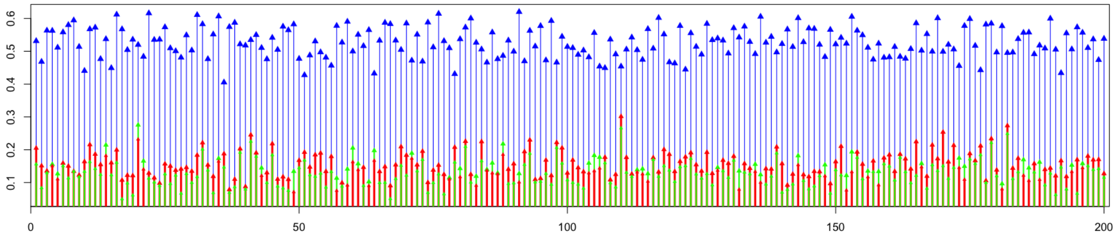


Figure 2.10: *Plots of errors for the model in* (2.12)*, using 200 random samples; heights of the red, green and blue triangles represent estimation errors for Algorithms 1,2 and SCMS, respectively.*

	Alg 1	Alg 2	SCMS
Errors	0.1514 (0.0402)	0.1321 (0.0397)	0.5273 (0.0455)

Table 2: *Means and standard deviations (in parentheses) of the errors shown in Figure 2.10.*

2.9 Real Data Application

We apply our algorithms to a data set of active and extinct volcanoes in Japan available at `https://en.wikipedia.org/wiki/List_of_volcanoes_in_Japan`. The locations of these volcanoes exhibit a clear filamentary structure with three major branches sharing an intersection. The results using SCMS and our algorithms are shown in Figure 2.12. We used the same bandwidth for all the three algorithms based on an optimal selection for the second derivatives of the kernel density estimation. Using all the sample points as starting

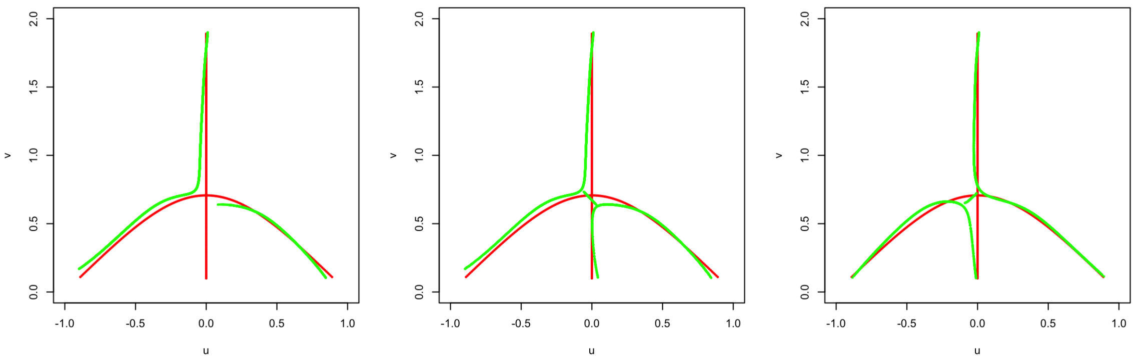


Figure 2.11: *Plots from left to right show the outputs of the SCMS, our Algorithms 1 and 2 (green dots). The red lines indicate the ridge model as given in Figure 2.9.*

points, the outputs of the algorithms are shown in the three left panels. It can be seen that all the three algorithms can capture the three major branches in the data, however, a careful examination reveals that the output of the SCMS algorithm seems to have big gap near the intersection of the three branches. To further investigate this issue, we ran each algorithm with a new set of starting values while keeping all the tuning parameters the same. The results are shown in the three right panels. The new starting points are constructed as follows: For each of the n outputs of an algorithm, connect each of the 20-nearest neighbors among the original data points to the output point by a line segment. On each of these 20 line segments choose 10 equidistant points. The resulting $200 \times n$ points are the new starting values of the respective algorithm. The idea underlying this construction is to find starting values that form a dense neighborhood of the true ridge lines. We observe that, although these start points fill the gap near the intersection well, the detected branches of SCMS algorithm are still clearly separated, while the branches are connected better in the outputs of our algorithms. This is consistent with our arguments that SCMS algorithm may miss some parts of the ridges.

3. Main results

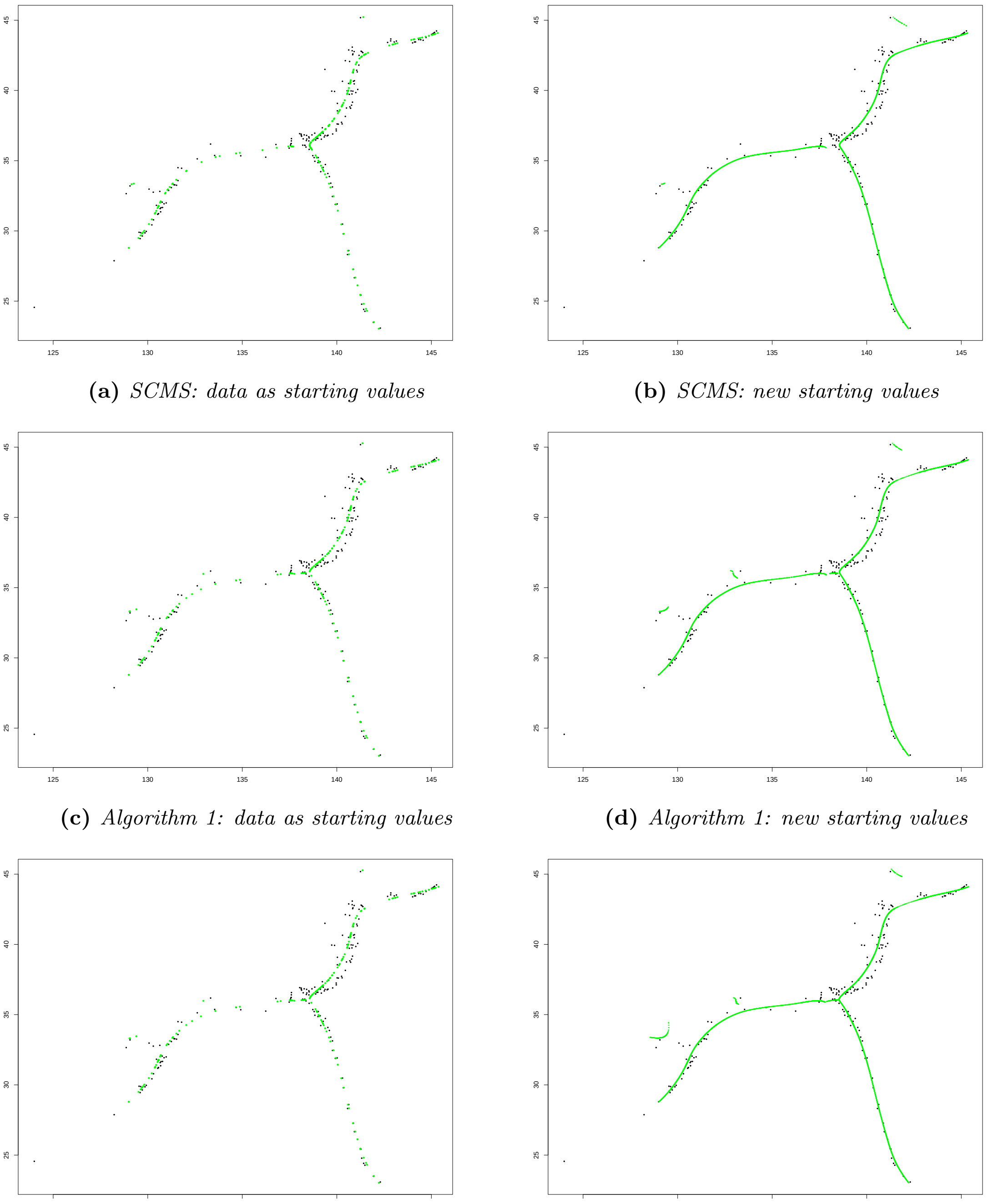
3.1 Assumptions and some technical implications

Before we state our assumptions, we first introduce some notation. For $\alpha = (\alpha_1, \cdots, \alpha_d)^T \in \mathbb{N}^d$, let $|\alpha| = \alpha_1 + \cdots + \alpha_d$. For a function $g : \mathbb{R}^d \to \mathbb{R}$ with partial derivatives of order $|\alpha|$, define

$$\partial^{(\alpha)} g(x) = \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} g(x), \ x \in \mathbb{R}^d.$$

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

$$d_H(\mathcal{A}, \mathcal{B}) = \max\{\sup_{x \in \mathcal{A}} d(x, \mathcal{B}), \ \sup_{x \in \mathcal{B}} d(x, \mathcal{A})\}. \tag{3.1}$$



(a) *SCMS: data as starting values*

(b) *SCMS: new starting values*

(c) *Algorithm 1: data as starting values*

(d) *Algorithm 1: new starting values*

(e) *Algorithm 2: data as starting values*

(f) *Algorithm 2: new starting values*

Figure 2.12: *Outputs of the three algorithms using the original sample as initialization (left panels) and new initilization (right panels). The black dots are the locations of the volcanoes in Japan, and the green dots are the estimated ridge points.*

Let $\partial\mathcal{A}$ be the boundary of $\mathcal{A}$. For $\delta > 0$ we define $\mathcal{A}^\delta = \bigcup_{x\in\mathcal{A}} \mathcal{B}(x,\delta)$, where $\mathcal{B}(x,\delta)$ denotes the (open) ball with midpoint x and radius δ. For any $c_1, c_2 \in \mathbb{R}$, let $c_1 \vee c_2 = \max(c_1, c_2)$ and $c_1 \wedge c_2 = \min(c_1, c_2)$. Finally, we will use $\|A\|_F$ to denote Frobenius-norm of a matrix A.

Recall that $\xi^{\widehat{\eta}}(x)$ and $\xi^{\widehat{\eta}_\tau}(x)$ denote the projections of the gradients of $\widehat{\eta}(x)$ and $\widehat{\eta}_\tau(x)$, respectively, onto the subspace spanned by the trailing $(d-k)$ eigenvectors of $\nabla^2\widehat{\eta}$ and $\nabla^2\widehat{\eta}_\tau$, respectively. Similarly, we define the corresponding population quantities $\xi^f(x)$ and $\xi^\eta(x)$ (see Section 2.5). With slight abuse of notation, we use 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 $m \geq 4$).

$\mathbf{(A1)}_{f,m}$ $f > 0$ is a density on $\mathbb{R}^d$ such that, for some $\epsilon_0 > 0$, $[-\varepsilon_0, 1+\varepsilon_0]^d$ is contained in the support of f; the partial derivatives of f up to order m exist and are bounded and continuous.

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

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

$\mathbf{(K)}_m$ 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

$$\Big\{x \mapsto \partial^{(\alpha)} K\Big(\frac{x-y}{h}\Big) : y \in \mathbb{R}^d, h > 0\Big\}$$

is a VC-class (see van der Vaart and Wellner, 1996). Also assume $\int_{\mathbb{R}^d} K(x)\|x\|^2 dx < \infty$.

$\mathbf{(L)}$ 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 $\mathbf{(A1)}_{f,m}$ 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 $\mathbf{(A2)}_f$ 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 $\mathbf{(A3)}_f$ implies that Ridge(f) is a manifold (without boundary). E.g., see Theorem 5.12 in Lee (2013). In particular, this precludes the existence of intersections of the ridge. For ridges with boundary or intersections, all our results except for Theorem 18(iii) and Corollary 20(iii) still apply to the part on the ridges strictly bounded away from the boundary and intersections. (v) The VC-class assumption in $\mathbf{(K)}_m$ holds if K is of the form $K(x) = \phi(p(x))$ with $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ 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 4 *Under* $\mathbf{(A1)}_{f,4}$*,* $\mathbf{(A2)}_f$*, and* $\mathbf{(A3)}_f$*, Ridge(f) is a compact set; we have $\lambda_1^\eta(x) = \cdots = \lambda_k^\eta(x) = 0$ for all $x \in$ Ridge(f), and there exist positive constants A, α and $\delta' \leq \delta$, such that for all $x \in$ Ridge$(f)^{\delta'}$,*

$$-\alpha \geq \lambda_{k+1}^\eta(x) \cdots \geq \lambda_d^\eta(x) \geq -A. \tag{3.2}$$

Moreover, the columns of $V_\perp^\eta(x)$ span the normal space of Ridge(f) at $x \in$ Ridge(f).

Recall the definition of $S_\epsilon^{\eta,f}$ given in (2.8) and notice that $S_0^{\eta,f} = \mathrm{Ridge}(f)$. The following lemma states that $S_\epsilon^{\eta,f}$ is in a neighborhood Ridge(f) of radius in the order of $\sqrt{\epsilon}$ when $\epsilon \to 0$.

Lemma 5 *Assuming* $\mathbf{(A1)}_{f,4}$*,* $\mathbf{(A2)}_f$*, and* $\mathbf{(A3)}_f$*, there exist constants $L_1, L_2, \epsilon_0 > 0$ such that for all $0 < \epsilon \leq \epsilon_0$,*

$$L_1\sqrt{\epsilon} \leq d_H(\partial S_\epsilon^{\eta,f}, \mathit{Ridge}(f)) \leq L_2\sqrt{\epsilon}, \tag{3.3}$$

where d_H is the Hausdorff distance.

Since $\nabla\eta(x) = \mathbf{0}$ for $x \in \mathrm{Ridge}(f)$, Lemma 4 implies that Ridge$(f) \subset$ Ridge(η). The following lemma states that in a neighborhood of the ridge of f, the ridge of the ridgeness function η equals the ridge of f. Recall that $\eta(x) = -\frac{1}{2}\|V_\perp^f(x)^\top \nabla f(x)\|^2 = -\frac{1}{2}\|\xi^f(x)\|^2$.

Lemma 6 *Assuming* $\mathbf{(A1)}_{f,4}$*,* $\mathbf{(A2)}_f$*, and* $\mathbf{(A3)}_f$*, there exists an $\epsilon_0 > 0$ such that*

$$\mathrm{Ridge}(f) = \mathrm{Ridge}(\eta) \cap S_\epsilon^{\eta,f} \qquad \forall\, 0 < \epsilon \leq \epsilon_0.$$

3.2 Convergence results

Brief outline of this section: As the algorithms are targeting Ridge$(\widehat{f}\,)$, while the theoretical target is Ridge(f), we first control the distance between these two sets (see Theorem 7). Then, in Theorem 9, we consider the continuous version of the algorithms, where the paths traced by the algorithm are replaced by the integral curves generated by our ridgeness functions. Finally, we consider the discrete version (see Theorem 11), i.e. the actual algorithms, and control the distance between the limit points of the algorithms and Ridge$(\widehat{f}\,)$.

Theorem 7 *Assume* $\mathbf{(A1)}_{f,4}$*,* $\mathbf{(A2)}_f$*,* $\mathbf{(A3)}_f$*,* $\mathbf{(K)}_4$*, $h = o(1)$, and $\frac{\log n}{nh^{d+8}} = o(1)$ as $n \to \infty$. Then, for any $B > 0$ and n large enough, there exists a constant $C > 0$ such that with probability at least $1 - n^{-B}$:*

$$d_H(\mathrm{Ridge}(f), \mathrm{Ridge}(\widehat{f}\,)) \leq C\Big(\Big(\frac{\log n}{nh^{d+4}}\Big)^{1/2} + h^2\Big). \tag{3.4}$$

Remark 8 *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, they require that $\|[I_d - \Pi^f(x)]\nabla f(x)\| f_{\max,3}(x) \leq \beta^2/(2d^3)$ for all $x \in$ Ridge$(f)^\delta$ in their assumption (A2), where $f_{\max,3}(x)$ is the maximum of the absolute values of all the third partial derivatives of f at x, and β is essentially the same one as given in our* $\mathbf{(A2)}_f$*. Instead of this assumption, we use* $\mathbf{(A3)}_f$*, which is weaker (see Chen et al., 2015, Lemma 2).*

3.2.1 Continuous versions of the algorithms

Further notation, Part 1. Recall the definition of the vector field $\xi^g : \mathbb{R}^d \to \mathbb{R}^d$ given in Section 2.5 for a twice differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, and let $x_0 \in \mathbb{R}^d$. Consider the system of ODEs

$$\frac{dv(t)}{dt} = \xi^g(v(t)),\ t \in \mathbb{R},$$

where $v : \mathbb{R} \to \mathbb{R}^d$ with $v(0) = x_0$. We denote the flow generated by this system of ODEs driven by ξ^g as

$$\gamma^g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d,$$

i.e. $\gamma^g(x, 0) = x$ for $x \in \mathbb{R}^d$, $\gamma^g(\gamma^g(x, t), s) = \gamma^g(x, t + s)$ and $\frac{\partial}{\partial t}\gamma^g(x, t) = \xi^g(\gamma^g(x, t))$, for $s, t \in \mathbb{R}$.

With this notation applied to $g = \widehat{\eta}$ and $g = \widehat{\eta}_\tau$, we now have the following result:

Theorem 9 *Assume* $\textbf{(A1)}_{f,4}$*,* $\textbf{(A2)}_f$*,* $\textbf{(A3)}_f$*,* $\textbf{(K)}_4$*,* $h = o(1)$ *and* $\frac{\log n}{nh^{d+8}} = o(1)$ *as* $n \to \infty$*. Then, for any* $B > 0$ *and* n *large enough, with probability at least* $1 - n^{-B}$*:*

(i) $\widehat{f}$ *satisfies assumptions* $\textbf{(A1)}_{\widehat{f},4}$ *,* $\textbf{(A2)}_{\widehat{f}}$*, and* $\textbf{(A3)}_{\widehat{f}}$*;*

(ii) Continuous version of Algorithm 1: there exists $\epsilon_0 > 0$ *such that, for all* $0 \le \epsilon \le \epsilon_0$*,*

$$\mathrm{Ridge}(\widehat{f}\,) = \{\lim_{t\to\infty} \gamma^{\widehat{\eta}}(x, t), x \in \partial S_\epsilon^{\widehat{\eta},\widehat{f}}\} = \mathrm{Ridge}(\widehat{\eta}\,) \cap S_\epsilon^{\widehat{\eta},\widehat{f}}.$$

(iii) Continuous version of Algorithm 2: under the additional assumptions $\textbf{(A1)}_{f,6}$*,* $\textbf{(K)}_6$*,* $\textbf{(L)}$ *and* $\frac{\log n}{nh^{d+12}} = o(1)$*, and for* $\epsilon > 0$ *small enough,*

a) $\mathrm{Ridge}(\widehat{\eta}_\tau) \cap S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}} = \{\lim_{t\to\infty} \gamma^{\widehat{\eta}_\tau}(x, t), x \in \partial S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}\}$ *for* $\tau > 0$ *small enough;*

b) $d_H(\mathrm{Ridge}(\widehat{\eta}_\tau) \cap S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}, \mathrm{Ridge}(\widehat{f}\,)) \le C_1\tau^2$ *for a constant* C_1 *and* τ *small enough;*

c) $d_H(\mathrm{Ridge}(\widehat{\eta}_\tau) \cap S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}, \mathrm{Ridge}(f)) \le C_2((\frac{\log n}{nh^{d+4}})^{1/2} + h^2 + \tau^2)$ *for a constant* C_2 *and* τ *small enough.*

Remark 10 *Without the additional assumptions in part (iii) of the above theorem, we can still show that* $\sup_{x\in[0,1]^d} |\partial^{(\alpha)}\widehat{\eta}(x) - \partial^{(\alpha)}\widehat{\eta}_\tau(x)| = o_p(1)$ *for all* $|\alpha| = 0, 1, 2$*, and further* $d_H(\mathrm{Ridge}(\widehat{\eta}_\tau) \cap S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}, \mathrm{Ridge}(f)) = o_p(1)$ *as* $\tau \to 0$*.*

3.2.2 Discrete approximation: the Euler method

Here we study discrete versions of the algorithms given above. To this end, we need a discrete version of the continuous flows defined above:

Further notation, Part 2. For a twice differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ and a constant $a > 0$ let the sequence $\gamma_a^g(x, \ell)$, $\ell = 0, 1, \cdots$, be defined as

$$\gamma_a^g(x, 0) = x \qquad \text{and} \qquad \gamma_a^g(x, \ell + 1) = \gamma_a^g(x, \ell) + a\,\xi^g(\gamma_a^g(x, \ell)), \quad \ell = 1, 2, \ldots$$

This is a discrete approximation to $\gamma^g(x, t)$, $t \ge 0$ using Euler's method.

The following result applies this notation with $g = \widehat{\eta}$ and $g = \widehat{\eta}_\tau$. It says that these discretized approximations can be used to recover the corresponding ridges.

Theorem 11 *Assume* $\mathbf{(A1)}_{f,6}$, $\mathbf{(A2)}_f$, $\mathbf{(A3)}_f$, $\mathbf{(K)}_6$, $\mathbf{(L)}$, $h = o(1)$ *and* $\frac{\log n}{nh^{d+12}} = o(1)$ *as* $n \to \infty$. *Then, for any* $B > 0$ *and* n *large enough, with probability at least* $1 - n^{-B}$:

(i) Algorithm 1: for $\epsilon \geq 0$ *small enough,* $\lim_{\ell\to\infty} \gamma_a^{\widehat{\eta}}(x,\ell)$ *exists for all* $x \in \partial S_\epsilon^{\widehat{\eta},\widehat{f}}$, *and*

$$d_H(\text{Ridge}(\widehat{f}), R_a(\widehat{f})) \leq C_1 a^{1-\sigma_0-\mu}, \tag{3.5}$$

for a constant C_1 *and* τ *small enough, where* $R_a(\widehat{f}) = \{\lim_{\ell\to\infty} \gamma_a^{\widehat{\eta}}(x,\ell), x \in \partial S_\epsilon^{\widehat{\eta},\widehat{f}}\}$.

(ii) Algorithm 2: for $\epsilon \geq 0$ *small enough,* $\lim_{\ell\to\infty} \gamma_a^{\widehat{\eta}_\tau}(x,\ell)$ *exists for all* $x \in \partial S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}$, *and*

$$d_H(\text{Ridge}(\widehat{f}), R_{\tau,a}(\widehat{f})) = C_2(a^{1-\sigma_0-\mu} + \tau^2), \tag{3.6}$$

for a constant C_2 *and* τ *and* a *small enough, where* $R_{\tau,a}(\widehat{f}) = \{\lim_{\ell\to\infty} \gamma_a^{\widehat{\eta}_\tau}(x,\ell), x \in \partial S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}\}$.

In the above rates, $0 < \sigma_0 < 1$ *is given in (B.69), and* $\mu > 0$ *is arbitrarily small.*

Remark 12 *By using Corollary 20 given below, one can also show that, for* $d = 1$, $\text{Ridge}(\widehat{f}) = R_a(\widehat{f})$ *and* $\text{Ridge}(\widehat{\eta}_\tau) \cap S_\epsilon^{\widehat{\tau},\widehat{f}} = R_{\tau,a}(\widehat{f})$ *with probability at least* $1 - n^{-B}$ *for any* $B > 0$ *and* n *large enough.*

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

This important section can be interpreted as providing population level versions of our main convergence results for the proposed algorithms presented above. Indeed, the algorithms can be interpreted as "perturbed versions" of corresponding population level versions. We will discuss the precise meaning of this in what follows, and we also indicate how this correspondence is being used to prove the convergence results for the algorithms. This section will also provide additional insights into why the algorithms proposed in this work do not suffer from the theoretical gaps of the SCMS algorithm - cf. Section 4.2.1.

The population level results are using theory for Ordinary Differential Equations (ODE). These ODEs provide the mathematical (population level) model for our algorithm. For the original mean shift algorithm, this analogy has been used in Arias-Castro et al. (2016) and Arias-Castro and Qiao (2025+). For the Subspace Constraint Mean Shift algorithm, see Genovese et al. (2014) and Qiao and Polonik (2016).

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.

4.1 Some useful background knowledge of ODEs

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

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

where $x : \mathbb{R} \to \mathbb{R}^d$ with $x(0) = x_0 \in \mathbb{R}^d$, and $U : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field. Let $\pi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ 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 $\mathbf{C}^1$ manifold. A point $\bar{x}$ is called an *equilibrium point*, or fixed point, if $U(\bar{x}) = 0$. Let $\mathcal{A} \subset S$ 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

$$E = \left\{ x \in S : \ \frac{dV(x(t))}{dt} = 0 \right\}, \tag{4.1}$$

$$M = \bigcup_{x_0 \in E} \big\{ \pi(x_0, t), t > 0 \, : \, \pi(x_0, t) \in E \text{ for all } t > 0 \big\}. \tag{4.2}$$

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

Theorem 13 *For all $x \in S$, $\pi(x, t) \to M$ as $t \to \infty$.*

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

4.2 ODE theory for ridge extraction

Recall that $\frac{\partial \gamma^\eta(x,t)}{\partial t} = \xi^\eta(\gamma^\eta(x, t)), t \in \mathbb{R}$, which is the mathematical model of Algorithms 1 and 2. There exists a positive $\epsilon > 0$ such that $S_\epsilon^{\eta,f}$ is a positively invariant set corresponding to this flow. This can be seen as follows. Consider the derivative of $\eta(\gamma^\eta(x, t))$ with respect to t:

$$\begin{aligned} \frac{\partial \eta(\gamma^\eta(x,t))}{\partial t} &= [\nabla \eta(\gamma^\eta(x,t))]^\top V_\perp^\eta(\gamma^\eta(x,t))[V_\perp^\eta(\gamma^\eta(x,t))]^\top \nabla \eta(\gamma^\eta(x,t)) \\ &= \|\xi^\eta(\gamma^\eta(x,t))\|^2 \geq 0. \end{aligned} \tag{4.3}$$

In other words, η is always non-decreasing along the trajectories of the flow. Indeed, notice that for $\epsilon > 0$ small enough, $\|\xi^\eta(\gamma^\eta(x, t))\| > 0$ for all $x \in S_\epsilon^{\eta,f} \backslash \text{Ridge}(f)$ and $t \in (0, \infty)$ based on Lemma 6. It follows that $S_\epsilon^{\eta,f}$ is a positively invariant set as we have claimed. Note that Ridge(f) is the set of all the equilibrium points in $S_\epsilon^{\eta,f}$. A natural choice for a Lyapunov function V is

$$V(x) = -\eta(x).$$

Since $\frac{\partial V(\gamma^\eta(x,t))}{\partial t} = -\langle \nabla \eta(\gamma^\eta(x, t)), \xi^\eta(\gamma^\eta(x, t)) \rangle = -\|\xi^\eta(\gamma^\eta(x, t))\|^2 \leq 0$, 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 (4.1) and (4.2), we have $E = M = \text{Ridge}(f)$.

Theorem 14 *Assume that* $\mathbf{(A1)}_{f,4}$, $\mathbf{(A2)}_f$, *and* $\mathbf{(A3)}_f$ *hold. There exists $\epsilon > 0$ such that:*

(i) For $x \in (0,1)^d$, let a path γ_x^η generated by $\gamma^\eta(x,t)$ be given by the set $\gamma_x^\eta = \{\gamma^\eta(x,t), t \in T_x\}$, where T_x is the largest open interval containing 0 such that $\gamma^\eta(x,t) \in (0,1)^d$ for all $t \in T_x$, and let $\Gamma = \{\gamma_x^\eta, x \in (0,1)^d\}$ be the collection of all these paths. For each $x \in S_\epsilon^{\eta,f} \backslash Ridge(f)$, the path γ_x^η is the unique path in Γ passing through x.

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

(iii) Ridge(f) $= \{\lim_{t\to\infty} \gamma^\eta(x,t) : x \in \partial S_\epsilon^{\eta,f}\}$.

Remark 15 *Part (iii) of Theorem 14 can be generalized as follows. Let $\mathcal{A}$ be a set such that for any $x \in \partial S_\epsilon^{\eta,f}$, there exists a point $y \in \mathcal{A}$ such that there is a finite $t_x \in (-\infty,\infty)$ such that $y = \gamma^\eta(x,t_x)$. Then by (iii) in Theorem 14, we have Ridge(f) $= \{\lim_{t\to\infty} \gamma^\eta(x,t) : x \in \mathcal{A}\}$.*

4.2.1 Further insights into differences between SCMS and our algorithms

Using our notation introduced above, $\gamma^f(x,t)$ denotes the flow corresponding to $\frac{dx(t)}{dt} = \xi^f(x(t)), t \in \mathbb{R}$, which is the model for the SCMS algorithm. We can now see the major difference between using $\gamma^f(x,t)$ and using the flow $\gamma^\eta(x,t)$ considered in our approach. For $\gamma^f(x,t)$, we have

$$\begin{aligned}\frac{\partial f(\gamma^f(x,t))}{\partial t} &= [\nabla f(\gamma^f(x,t))]^\top V_\perp^f(\gamma^f(x,t))[V_\perp(\gamma^f(x,t))]^\top \nabla f(\gamma^f(x,t)) \\ &= \|\xi^f(\gamma^f(x,t))\|^2 \geq 0. \qquad (4.4)\end{aligned}$$

In other words, it is the height of f that increases along the path of $\gamma^f(x,t)$ as t increases. In contrast to that, it is the ridgeness $\eta(x)$ that increases along the path of $\gamma^\eta(x,t)$. In general, while the height and ridgeness are closely related, they are two different quantities. This provides a different point of view for the SCMS algorithm, and also shows its difference to our approach.

4.3 Stability of the flows

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

$$\delta_0 = \sup_{x\in[0,1]^d} |\eta(x) - \widetilde{\eta}(x)|, \qquad (4.5)$$

$$\delta_1 = \sup_{x\in[0,1]^d} \|\nabla\eta(x) - \nabla\widetilde{\eta}(x)\|, \qquad (4.6)$$

$$\delta_2 = \sup_{x\in[0,1]^d} \|\nabla^2\eta(x) - \nabla^2\widetilde{\eta}(x)\|_F. \qquad (4.7)$$

Recall that, by definition, Ridge($\widetilde{\eta}$) $= \{x \in [0,1]^d : \xi^{\widetilde{\eta}}(x) = 0,\ \lambda_{k+1}^{\widetilde{\eta}}(x) < 0\}$ with $\xi^{\widetilde{\eta}}(x)$ and $\lambda_{\widetilde{\eta}}^{k+1}(x)$ as defined in Section 2.5.

Theorem 16 *Suppose that* **(A1)**$_{f,4}$, **(A1)**$_{\widetilde{f},4}$, **(A2)**$_f$, *and* **(A3)**$_f$ *hold. There exists $\epsilon_0 > 0$ such that for any $\epsilon \in (0,\epsilon_0]$ and for* $\max(\delta_0,\delta_1,\delta_2)$ *small enough we have the following:*

(i) For $x \in (0,1)^d$, let a path $\gamma_x^{\widetilde{\eta}}$ generated by $\gamma^{\widetilde{\eta}}(x,t)$ be given by the set $\gamma_x^{\widetilde{\eta}} = \{\gamma^{\widetilde{\eta}}(x,t), t \in T_x\}$, where T_x is the largest open interval containing 0 such that $\gamma^{\widetilde{\eta}}(x,t) \in (0,1)^d$ for all $t \in T_x$, and let $\Gamma = \{\gamma_x^{\widetilde{\eta}}, x \in (0,1)^d\}$ be the collection of all these paths. For each $x \in S_\epsilon^{\widetilde{\eta},f} \backslash Ridge(\widetilde{\eta})$, the path $\gamma_x^{\widetilde{\eta}}$ is the unique path in Γ passing through x.

(ii) For each $x \in S_\epsilon^{\widetilde{\eta},f}$, the path $\gamma^{\widetilde{\eta}}(x,t)$ converges to a point in Ridge($\widetilde{\eta}$)$\cap S_\epsilon^{\widetilde{\eta}}$, as $t \to +\infty$.

(iii) Ridge($\widetilde{\eta}$) $\cap S_\epsilon^{\widetilde{\eta},f} = \{\lim_{t\to\infty} \gamma^{\widetilde{\eta}}(x,t) : x \in \partial S_\epsilon^{\widetilde{\eta},f}\}$.

(iv) There exists a constant $C > 0$ such that $d_H(Ridge(\widetilde{\eta}) \cap S_\epsilon^{\widetilde{\eta},f}, Ridge(f)) \leq C \max(\delta_1, \delta_2)$.

Remark 17 *This theorem is applied to $f, \widetilde{\eta}$ being $\widehat{f}$ and $\widehat{\eta}_\tau$, respectively, for which small enough values of δ_0, δ_1, and δ_2 can be found with high probability. See the proof of Theorem 9.*

4.4 Euler's method

The following result about this discretization of the flow γ^η is the main result of this section. Recall that $\xi^\eta(x) = V_\perp^\eta(x) V_\perp^\eta(x)^\top \nabla\eta(x)$, and recall the definition of the discretized path $\gamma_a^\eta(x,\ell)$, $\ell = 0, 1, \cdots$ as defined in Section 3.2.2.

Theorem 18 *Suppose that assumptions* **(A1)**$_{f,6}$, **(A2)**$_f$, *and* **(A3)**$_f$ *hold. Let $R_a(f) = \{\lim_{\ell\to\infty} \gamma_a^\eta(x,\ell),\ x \in \partial S_\epsilon^{\eta,f}\}$. There exist $\epsilon_0 > 0, \delta_a > 0$ such that when $0 < a \leq \delta_a$ and $0 < \epsilon < \epsilon_0$, we have*

(i) $R_a(f) \subset$ Ridge(f), and

(ii) there exists a constant $C > 0$ such that $d_H(Ridge(f), R_a(f)) \leq C a^{1-\sigma_0}$, where $0 < \sigma_0 < 1$ is given in (B.69).

Moreover,

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

$$Ridge(f) = R_a(f).$$

Remark 19 *The last assertion of this theorem says that in the case of one-dimensional ridges we can recover the entire ridge of f (as in the continuous case - see Theorem 14(iii)), even when using the discretized algorithm, provided the step-size is small enough. We conjecture that the result also holds true for multi-dimensional ridges.*

Recall that we have analyzed the perturbed flow $\gamma^{\widetilde{\eta}}$ and its convergence in Section 4.3. We now assume that $\widetilde{\eta}$ is generated by a density function $\widetilde{f}$, i.e.,

$$\widetilde{\eta}(x) = -\frac{1}{2} \left\| V_\perp^{\widetilde{f}}(x))^\top \nabla \widetilde{f}(x) \right\|^2. \tag{4.8}$$

Using the same notation as defined in Section 3.2.2, consider the sequence $\gamma_a^{\widetilde{\eta}}(x,\ell)$, $\ell = 0, 1, \cdots$. Using Theorem 16, and following a very similar proof of Theorem 18, we obtain the following discretization result of the perturbed flows and ridges.

Corollary 20 *Suppose that assumptions* $\mathbf{(A1)}_{f,6}$*,* $\mathbf{(A1)}_{\widetilde{f},6}$*,* $\mathbf{(A2)}_f$*, and* $\mathbf{(A3)}_f$ *hold. Let* $R_a(\widetilde{f}\,) = \{\lim_{\ell\to\infty} \gamma_a^{\widetilde{\eta}}(x,\ell),\ x \in \partial S_\epsilon^{\widetilde{\eta},f}\}$*. 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) $R_a(\widetilde{f}\,) \subset Ridge(\widetilde{\eta}\,) \cap S_\epsilon^{\widetilde{\eta},f}$*, and*

(ii) there exists a constant $C > 0$ *such that* $d_H(Ridge(\widetilde{\eta}\,) \cap S_\epsilon^{\widetilde{\eta},f}, R_a(\widetilde{f}\,)) \leq Ca^{1-\sigma_0-\mu}$*, for an arbitrarily small* $\mu > 0$*.*

Moreover,

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

$$Ridge(\widetilde{\eta}\,) \cap S_\epsilon^{\widetilde{\eta},f} = R_a(\widetilde{f}\,).$$

The proof of Corollary 20 is similar to that of Theorem 18, and no details are presented.

5. Discussion

5.1 Other variants

Here we briefly discuss two other variants of our algorithms.

The first variant is based on the logarithm transformation. The original mean shift algorithm proposed by Fukunaga and Hostetler (1975) is a gradient ascent algorithm implicitly using the logarithm transformation of kernel density estimates, which is analyzed in Arias-Castro et al. (2016) and Arias-Castro and Qiao (2025+). More specifically, for any $x_0 \in \mathbb{R}$, they consider the sequence $x_{j+1} = x_j + a\nabla\widehat{r}(x_j)$, $j = 0, 1, 2, \cdots$, where $\widehat{r}$ is an estimate of $\log f$. The limit of the sequence is a local mode of $\widehat{f}$, provided some regularity assumptions hold.

A similar idea can be applied to our algorithms as well. Let $g : [0,\infty) \to (0,\infty)$ be a (known) twice differentiable increasing positive function. In Algorithm 1, we can replace the derivatives (and the induced eigenvalues and eigenvectors) of $\widehat{\eta}$ by those of $\log(g(\widehat{\eta}))$. Note that $\widehat{\eta}$ by definition is non-positive and this explains the need for a positive transformation function g before applying the logarithm. More specifically, the update step (2.9) in Algorithm 1 can be replaced by

$$y_i^{j+1} = y_i^j + a\xi^{\log g(\widehat{\eta})}(y_i^j).$$

Algorithm 2 can be modified in a similar way.

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

$$\widehat{s}_q(x) = \frac{\widehat{\eta}(x)}{[\widehat{f}(x)]^q}. \tag{5.1}$$

We have that $\widehat{s}_q \leq 0$ and $\widehat{s}_q = 0$ on $\text{Ridge}(\widehat{f}\,)$, and finding minimizers of $\widehat{s}_q$ is equivalent to that of $\widehat{\eta}$. Thus we can replace the derivatives (and the induced eigenvalues and eigenvectors) of $\widehat{\eta}$ by those of $\widehat{s}_q$. Recall that our basic algorithms can possibly return non-global modes of the ridgeness function depending on the start points, and thus requires a post-process step (see Section 2.7). The numerator in (5.1) is a penalization for low density. The purpose 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 above.

5.2 Mathematical models of SCMS algorithm and its convergence

The analyses and simulations in this paper show that the set that SCMS converges to is not always exactly the ridge set under its original definition, although they are clearly related. It is an interesting open problem to find the mathematical definition of the set of the limit points of SCMS. Once this is discovered, we believe the convergence analyses established in this paper can be useful to show the convergence of SCMS in the sense of recovering the entire set.

Appendix A. More details of algorithms

A.1 Discussion of why the SCMS algorithm might miss parts of ridges

Here we provide some more details for the argument made in Section 2.3 that the SCMS algorithm might miss some parts of a ridge, as observed in the second example in Section 2.8.

We consider the case $d = 2$. First we show that a ridge point does not necessarily have to be a local maximum of the integral curve traced by the SCMS algorithm. Such a point will thus not be identified as a ridge point (except in the trivial case where the starting point happens to be this ridge point). For a given point x_0 near the ridge, consider an integral curve $x(t)$ defined as

$$\frac{dx(t)}{dt} = V_\perp(x(t)), \;\; x(0) = x_0, \tag{A.1}$$

where $V_\perp(x)$ is the second unit eigenvector of the Hessian of f at x. We assume that the direction (sign) of $V_\perp(x)$ is determined such that it varies continuously with x. Note that $V_\perp$ is parallel to $V_\perp V_\perp^\top \nabla f =$ and so $x(t)$ has the same trajectory as the integral curve driven by $V_\perp V_\perp^\top \nabla f$. Using $V_\perp$ allows tracking integral curves both forward and backward. Indeed, the vector field $V_\perp V_\perp^\top \nabla f$ 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

$$\frac{df(x(t))}{dt} = \big\langle \nabla f(x(t)), \; V_\perp(x(t)) \big\rangle \tag{A.2}$$

and

$$\begin{aligned}\frac{d^2 f(x(t))}{dt^2} &= \Big\langle \nabla \big\langle \nabla f(x(t)), \; V_\perp(x(t)) \big\rangle, \; V_\perp(x(t)) \Big\rangle \\ &= \nabla f(x(t)) \; \nabla V_\perp(x(t)) \; V_\perp(x(t)) + \Big\langle \nabla^2 f(x(t)) \; V_\perp(x(t)), \; V_\perp(x(t)) \Big\rangle\end{aligned}$$

$$= \nabla f(x(t))\, \nabla V_\perp(x(t))\, V_\perp(x(t)) + \lambda_2(x(t)). \tag{A.3}$$

If $x(t)$ is a ridge point, then $\frac{df(x(t))}{dt} = 0$ in (A.2) by Definition 1, and the second term $\lambda_2(x(t))$ in (A.3) is negative. In general, however, the right-hand side of (A.3) may be not negative. In other words, depending on the sign in (A.3), ridge points on the trajectory driven by $V_\perp$ (or equivalently, by $V_\perp V_\perp^\top \nabla f$) can be local maxima, local minima or even saddle points. As shown in the second example in Section 2.8, following the direction of $V_\perp V_\perp^\top \nabla f$, a part of the ridge can be missed if the starting points are not chosen exactly on that part.

A.2 Formulas needed for implementing Algorithm 1

In Algorithm 1 we need to compute $\nabla \xi^{\widehat{f}}(x)$ and also the Hessian of $\widehat{\eta}(x)$ (in order to find its eigenvectors). In the following we provide some formulas that are useful for the implementation.

Let $\textsc{vec}$ be the matrix vectorization operator such that $\textsc{vec}(A)$ stacks all the columns of a matrix A into a vector.

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

$$\begin{aligned} \nabla \xi^{\widehat{f}}(x) &= (\nabla \widehat{f}(x)^\top \otimes \mathbf{I}_d) \nabla \Pi^{\widehat{f}}(x) + \Pi^{\widehat{f}}(x) \nabla^2 \widehat{f}(x) \\ &= (\nabla \widehat{f}(x)^\top \otimes \mathbf{I}_d) \nabla \Pi^{\widehat{f}}(x) + V_\perp^{\widehat{f}}(x) \Lambda_\perp^{\widehat{f}}(x) V_\perp^{\widehat{f}}(x)^\top, \end{aligned} \tag{A.4}$$

where $\Lambda_\perp^{\widehat{f}}(x) = \mathrm{diag}(\lambda_{r+1}^{\widehat{f}}, \cdots, \lambda_d^{\widehat{f}})$, $\nabla \Pi^{\widehat{f}}(x) = \frac{d\ \textsc{vec}[\Pi^{\widehat{f}}(x)]}{dx^\top} \in \mathbb{R}^{d^2 \times d}$, and $\otimes$ denotes Kronecker product.

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

$$\nabla^2 \widehat{\eta}(x) = -(\mathbf{I}_d \otimes \xi^{\widehat{f}}(x)^\top) \nabla(\nabla \xi^{\widehat{f}}(x)) - \nabla \xi^{\widehat{f}}(x))^\top \nabla \xi^{\widehat{f}}(x), \tag{A.5}$$

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

$$\begin{aligned} \nabla(\nabla \xi^{\widehat{f}}(x)) =& [(\nabla \Pi^{\widehat{f}}(x))^\top \otimes \mathbf{I}_d] \nabla(\nabla \widehat{f}(x)^\top \otimes \mathbf{I}_d) \\ &+ (\mathbf{I}_d \otimes (\nabla \widehat{f}(x)^\top \otimes \mathbf{I}_d)) \nabla(\nabla \Pi^{\widehat{f}}(x)) \\ &+ (\nabla^2 \widehat{f}(x) \otimes \mathbf{I}_d) \nabla \Pi^{\widehat{f}}(x) \\ &+ (\mathbf{I}_d \otimes \Pi^{\widehat{f}}(x)) \nabla(\nabla^2 \widehat{f}(x)). \end{aligned}$$

Here $\nabla(\nabla \widehat{f}(x)^\top \otimes \mathbf{I}_d) = (\mathbf{I}_d \otimes \textsc{vec}(\mathbf{I}_d)) \nabla^2 \widehat{f}(x)$. Furthermore, we can explicitly find the expressions of $\nabla \Pi^{\widehat{f}}(x)$ and $\nabla(\nabla \Pi^{\widehat{f}}(x))$, which by the chain rule involves the third and fourth derivatives of $\widehat{f}$, respectively. For example, for $\ell = 1, \cdots, d$, the ℓ-th column of $\nabla \Pi^{\widehat{f}}(x)$ is given by the vectorization of

$$\sum_{i=1}^{k} \sum_{j=k+1}^{d} \frac{1}{\lambda_j^{\widehat{f}}(x) - \lambda_i^{\widehat{f}}(x)} [V_j^{\widehat{f}}(x) V_j^{\widehat{f}}(x)^\top \widehat{H}_\ell(x) V_i^{\widehat{f}}(x) V_i^{\widehat{f}}(x)^\top + V_i^{\widehat{f}}(x) V_i^{\widehat{f}}(x)^\top \widehat{H}_\ell(x) V_j^{\widehat{f}}(x) V_j^{\widehat{f}}(x)^\top], \tag{A.6}$$

where $\widehat{H}_\ell(x)$ is the partial derivative of $\nabla^2 \widehat{f}(x)$ with respective to the ℓ-th component of x. See Qiao (2025+) for some relevant calculations.

Appendix B. Proofs

Proof of Lemma 4. First we show that Ridge(f) is a compact set. Notice that by assumption $\mathbf{(A2)}_f$, we can write

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

because the set $\{x \in [0,1]^d : \xi^f(x) = 0,\ \lambda_{k+1}^f(x) = 0\}$ is empty. If we treat $\Pi^f(x)$ as a function of $\nabla^2 f(x)$, due to the assumed positive gap between $\lambda_k^f(x)$ and $\lambda_{k+1}^f(x)$ in assumption $\mathbf{(A2)}_f$, $\Pi^f(x)$ is an analytical matrix-valued function on the space of real symmetric matrices by the classical matrix perturbation theory (see Kato, 2013), which further implies that $\Pi^f(x)$ is a Lipschitz function of x by assumption $\mathbf{(A1)}_{f,4}$. Since both $\xi^f(x)$ and λ_{k+1}^f are continuous functions, Ridge(f) is closed, and hence compact because it is defined on the compact set $[0,1]^d$.

We have $\nabla^2\eta(x) = -(\mathbf{I}_d \otimes \xi^f(x)^\top)\nabla(\nabla\xi^f(x)) - \nabla\xi^f(x)^\top\nabla\xi^f(x)$. Thus, for $x \in \text{Ridge}(f)$, $\xi^f(x) = 0$ and $\nabla^2\eta(x) = -\nabla\xi^f(x)^\top\nabla\xi^f(x)$, which, by using assumption $\mathbf{(A3)}_f$, implies that the rank of $\nabla^2\eta(x)$ is $d-k$. Hence we have for $x \in \text{Ridge}(f)$,

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

Because of the compactness of Ridge(f) and the continuity of λ_j^η on Ridge(f), there exist positive constants α' and A' such that

$$-\alpha' \geq \lambda_{k+1}^\eta(x) \cdots \geq \lambda_d^\eta(x) \geq -A' \tag{B.1}$$

for all $x \in \text{Ridge}(f)$.

Since $\xi^f(x) = 0$ for all $x \in \text{Ridge}(f)$, under assumption $\mathbf{(A1)}_{f,4}$, when δ' is small enough, for all $x \in \text{Ridge}(f)^{\delta'}$,

$$\begin{aligned}
\iota(x) &:= \|(\mathbf{I}_d \otimes \xi^f(x)^\top)\nabla(\nabla\xi^f(x))\|_F \\
&\leq \sqrt{d}\|\xi^f(x)\|_F\|\nabla(\nabla\xi^f(x))\|_F \\
&\leq 2\sqrt{d} \sup_{x \in \text{Ridge}(f)^{\delta'}} [\|\nabla\xi^f(x)\|_F\|\nabla(\nabla\xi^f(x))\|_F]\, d(x, \text{Ridge}(f)) \\
&=: c_0 d(x, \text{Ridge}(f)).
\end{aligned}$$

Also for all $x, y \in \text{Ridge}(f)^{\delta'}$,

$$\begin{aligned}
&\|\nabla\xi^f(x)^\top\nabla\xi^f(x) - \nabla\xi^f(y)^\top\nabla\xi^f(y)\|_F \\
&\leq \|[\nabla\xi^f(x) - \nabla\xi^f(y)]^\top\nabla\xi^f(x)\|_F + \|[\nabla\xi^f(x) - \nabla\xi^f(y)]^\top\nabla\xi^f(y)\|_F \\
&\leq 2 \sup_{x \in \text{Ridge}(f)^{\delta'}} [\|\nabla\xi^f(x)\|_F\|\nabla(\nabla\xi^f(x))\|_F]\, \|x - y\| \\
&= c_0\|x - y\|.
\end{aligned}$$

For any $x \in \text{Ridge}(f)^{\delta'}$, let $y_x \in \text{Ridge}(f)$ be such that $\|x - y\| = d(x, \text{Ridge}(f))$. By using Weyl's inequality (see Serre, 2002, page 15), we have for all $x \in \text{Ridge}(f)^{\delta'}$, and $j = k+1, \cdots, d$,

$$|\lambda_j^\eta(x) - \lambda_j^\eta(y_x)| \leq \|\nabla^2\eta(x) - \nabla^2\eta(y_x)\|_F$$

$$\begin{aligned}&\leq \iota(x) + \|\nabla\xi^f(x)^\top\nabla\xi^f(x) - \nabla\xi^f(y)^\top\nabla\xi^f(y)\|_F\\ &\leq c_0[d(x,\mathrm{Ridge}(f)) + \|x-y\|]\\ &\leq 2c_0\delta'.\end{aligned}$$

Therefore (3.2) holds for $\delta' > 0$ small enough by noticing (B.1).

Since $\xi^f(x) = 0$ for all $x \in \mathrm{Ridge}(f)$, the row vectors of $\nabla\xi^f(x)$ span the normal space of Ridge(f) at x, denoted by $\mathcal{N}(x)$. Note that $\mathcal{N}(x) = \{\nabla\xi^f(x)^\top a :\ a \in \mathbb{R}^d\} = \{\nabla\xi^f(x)^\top\nabla\xi^f(x)a :\ a \in \mathbb{R}^d\}$, 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 $V_\perp^\eta(x)$.

■

Proof of Lemma 5. For any $b \in \mathbb{R}^d$, consider

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

and note that $S_\epsilon^{\eta,f} = \bigcup_{\{b:\frac{1}{2}\|b\|^2\leq\epsilon\}} M_b^f$ for all $\epsilon \geq 0$. Recall δ' defined in Lemma 4 and let $\mathcal{A} = [0,1]^d\backslash\{t \in [0,1]^d : \eta(t) = 0\}^{\delta'}$. Since $\mathcal{A}$ is a compact set, we have $r := -\sup_{x\in\mathcal{A}}\eta(x) > 0$. Then assumption $\mathbf{(A2)}_f$ implies that $M_b^f \subset \mathrm{Ridge}(f)^{\delta'}$ when $\frac{1}{2}\|b\|^2 < r$, which is also what we assume below. We will first show that for all $\|b\|$ small enough,

$$C_1\|b\| \leq \sup_{x\in M_b^f} d(x,\mathrm{Ridge}(f)) \leq C_2\|b\|,$$

for some positive constants C_1 and C_2.

For any $x \in M_b^f$, let x_0 be its projection point onto Ridge(f). There exists a unit vector $u = u(x_0) \in \mathcal{N}(x_0)$ where $\mathcal{N}(x_0)$ the normal space to the ridge at x_0, and $\delta_x > 0$ such that $x = x_0 + \delta_x u$. Using a Taylor expansion, we get

$$\xi^f(x) = \xi^f(x_0) + \nabla\xi^f(x_0)(x-x_0) + R_0(x) = \delta_x\nabla\xi^f(x_0)u + R_0(x),$$

where $\|R_0(x)\| \leq \kappa_0\delta_x^2$, for a constant $\kappa_0 > 0$ for all $x \in M_b^f$. Therefore

$$\|b\|^2 = \|\xi^f(x)\|^2 = \delta_x^2 u^\top\nabla\xi^f(x_0)^\top\nabla\xi^f(x_0)u + R_0'(x), \tag{B.2}$$

where $|R_0'(x)| \leq \kappa_0'\delta_x^3$, for a constant $\kappa_0' > 0$ for all $x \in M_b^f$. Let $\omega_{\max}(x)$ and $\omega_{\min}(x)$ be the largest and the $(d-k)$th largest eigenvalues of $\nabla\xi^f(x)^\top\nabla\xi^f(x)$, respectively. For $x_0 \in \mathrm{Ridge}(f)$, we have $\omega_{\min}(x_0) = -\lambda_{k+1}^\eta(x_0)$ and $\omega_{\max}(x_0) = -\lambda_d^\eta(x_0)$ because $\nabla^2\eta(x_0) = -\nabla\xi^f(x_0)^\top\nabla\xi^f(x_0)$. From Lemma 4 we know that for all $x_0 \in \mathrm{Ridge}(f)$, $0 < \alpha \leq \omega_{\min}(x_0) \leq \omega_{\max}(x_0) \leq A < 0$ for some positive constants A and α. Notice that u is in the space spanned by $V_{k+1}^\eta(x_0), \cdots, V_d^\eta(x_0)$. So when $\|b\|$ is small enough (and hence δ_x is small enough), it follows from (B.2) that

$$2A\delta_x^2 \geq 2|\lambda_d^\eta(x_0))|\delta_x^2 = 2\omega_{\max}(x_0)\delta_x^2 \geq \|b\|^2 \geq \frac{\omega_{\min}(x_0)}{2}\delta_x^2 = \frac{|\lambda_{k+1}^\eta(x_0)|}{2}\delta_x^2 \geq 2\alpha\delta_x^2. \tag{B.3}$$

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 $\epsilon = \frac{1}{2}\|b\|^2$ is small enough,

$$\sqrt{\frac{1}{A}}\sqrt{\epsilon} = \sqrt{\frac{1}{2A}}\|b\| \leq \sup_{x\in M_b^f} d(x, \text{Ridge}(f)) \leq \sqrt{\frac{2}{\alpha}}\|b\| = \sqrt{\frac{4}{\alpha}}\sqrt{\epsilon}. \tag{B.4}$$

The other direction can be proved in a similar way, as given as follows. Now let x be a point on Ridge(f), and x_0 be its projection onto M_b, such that $x = x_0 + \delta_x u$, where $\delta_x = \|x - x_0\| > 0$, and $u \in \mathcal{N}(x_0)$ is a unit normal vector of M_b^f at x_0. Following the above analysis (in particular (B.3)), 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 both $\omega_{\max}$ and $\omega_{\min}$ are continuous functions in a neighborhood of Ridge(f), we have that $4A\delta_x^2 \geq \|b\|^2 \geq \frac{1}{4}\alpha\delta_x^2$ for all $\|b\|$ small enough. Therefore for all $b \in \mathbb{R}^d$ such that $\epsilon = \frac{1}{2}\|b\|^2$ is small enough,

$$\sqrt{\frac{1}{2A}}\sqrt{\epsilon} \leq \sup_{x\in\text{Ridge}(f)} d(x, M_b) \leq \sqrt{\frac{8}{\alpha}}\sqrt{\epsilon}. \tag{B.5}$$

Combining (B.4) and (B.5), we obtain

$$\sqrt{\frac{1}{2A}}\sqrt{\epsilon} \leq d_H(\partial S_\epsilon^{\eta,f}, \text{Ridge}(f)) \leq \sqrt{\frac{8}{\alpha}}\sqrt{\epsilon}. \tag{B.6}$$

■

Proof of Lemma 6. First we show that Ridge$(f) \subset$ Ridge$(\eta)\cap S_\epsilon^{\eta,f}$. We have $\nabla\eta(x) = -\nabla\xi^f(x)^\top\xi^f(x)$ for $x \in [0,1]^d$. Thus, for $x \in$ Ridge(f) (implying that $\xi^f(x) = 0$), we have $\nabla\eta(x) = 0$, and hence $\xi^\eta(x) = 0$. Then Ridge$(f) \subset$ Ridge$(\eta) \cap S_\epsilon^{\eta,f}$ for ϵ small enough is a direct consequence of Lemma 4.

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

$$\|\xi^\eta(x)\| > 0 \quad \text{for all } x \in S_\epsilon^{\eta,f}\backslash\text{Ridge}(f), \tag{B.7}$$

which implies the result. It follows from Lemma 5 that for every $\epsilon > 0$ is small enough, there exists $\delta(\epsilon) > 0$ such that $S_\epsilon^{\eta,f} \subset \text{Ridge}(f)^{\delta(\epsilon)}$ where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. For any $x \in S_\epsilon^{\eta,f}\backslash$Ridge$(f)$, let $x_0 \in$ Ridge(f) be the projection of x onto Ridge(f), that is, there exists a $\delta_x \in (0, \delta(\epsilon))$, such that we can write $x = x_0 + \delta_x u$, where $u = u(x_0) \in \mathcal{N}(x_0)$ is a unit vector with $\mathcal{N}(x_0)$ the normal space to the ridge at x_0. Notice that by using Lemma 4, $\mathcal{N}(x_0) = \{V_\perp^\eta(x_0)a : a \in \mathbb{R}^{d-k}\}$. We will show that there exists an $\epsilon > 0$, such that $\|\nabla\eta(x)\| > 0$ for all $x \in S_\epsilon^{\eta,f} \setminus$ 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 $x_1 \in \mathcal{B}(x_0, \delta(\epsilon))$. Then we can write

$$\begin{aligned}\nabla\eta(x) &= \nabla\eta(x_0) + \nabla^2\eta(x_0)(x - x_0) + R_1(x)\\ &= \delta_x\nabla^2\eta(x_0)u + R_1(x),\end{aligned} \tag{B.8}$$

where $\|R_1(x)\| \le \kappa_1\delta_x^2$. Note that $\nabla^2\eta(x_0) = \sum_{j=k+1}^d \lambda_j^\eta(x_0)V_j^\eta(x_0)V_j^\eta(x_0)^\top$ by Lemma 4 and u is a unit vector in the space spanned by $V_{k+1}^\eta(x_0), \cdots, V_d^\eta(x_0)$. Hence $A \ge |\lambda_d^\eta(x_0)| \ge \|\nabla^2\eta(x_0)u\| \ge |\lambda_{k+1}^\eta(x_0)| \ge \alpha$, where α and A are positive constants given in Lemma 4. Thus, for $\epsilon > 0$ small enough (and hence δ_x is small), we have $\|\nabla\eta(x)\| > 0$ for $x \in S_\epsilon^{\eta,f} \setminus \mathrm{Ridge}(f)$.

Next we show that $\frac{|\langle\nabla\eta(x),u(x)\rangle|}{\|\nabla\eta(x)\|}$ is bounded from below. Using (B.8) we have that $|\langle\nabla\eta(x), u(x)\rangle| \ge \delta_x|\lambda_{k+1}^\eta(x_0)| - \kappa_1\delta_x^2$, and

$$\frac{|\langle\nabla\eta(x), u(x)\rangle|}{\|\nabla\eta(x)\|} \ge \frac{\delta_x|\lambda_{k+1}^\eta(x_0)| - \kappa_1\delta_x^2}{\delta_x|\lambda_d^\eta(x_0)| + \kappa_1\delta_x^2}.$$

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

$$\frac{|\langle\nabla\eta(x), u(x)\rangle|}{\|\nabla\eta(x)\|} \ge \frac{1}{2}\frac{|\lambda_{k+1}^\eta(x_0)|}{|\lambda_d^\eta(x_0)|}.$$

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

$$\frac{\|V_\perp^\eta(x_0)V_\perp^\eta(x_0)^\top\nabla\eta(x)\|}{\|\nabla\eta(x)\|} \ge \frac{1}{2}\frac{|\lambda_{k+1}^\eta(x_0)|}{|\lambda_d^\eta(x_0)|}. \tag{B.9}$$

Furthermore, the angle between the eigenspace spanned by $V_\perp^\eta(x_0)$ and $V_\perp^\eta(x)$ should be small if δ_x is small because $\nabla^2\eta$ is a continuous function of x. This can be seen by using the Davis-Kahan inequality:

$$\|V_\perp^\eta(x_0)V_\perp^\eta(x_0)^\top - V_\perp^\eta(x)V_\perp^\eta(x)^\top\|_F \le \frac{2\sqrt{2}\|\nabla^2\eta(x_0) - \nabla^2\eta(x)\|_F}{|\lambda_{k+1}^\eta(x_0)|} \le \frac{2\sqrt{2}\kappa_1\delta_x}{|\lambda_{k+1}^\eta(x_0)|}. \tag{B.10}$$

Hence using (B.9) and (B.10),

$$\frac{\|\xi^\eta(x)\|}{\|\nabla\eta(x)\|} = \frac{\|V_\perp^\eta(x)V_\perp^\eta(x)^\top\nabla\eta(x)\|}{\|\nabla\eta(x)\|} \ge \frac{1}{2}\frac{|\lambda_{k+1}^\eta(x_0)|}{|\lambda_d^\eta(x_0)|} - \frac{2\sqrt{2}\kappa_1\delta_x}{|\lambda_{k+1}^\eta(x_0)|},$$

which can be bounded from below by a positive constant when ϵ (and hence δ_x) is small enough. This is (B.7).

■

Proof of Theorem 7. Using Talagrand's inequality (see Sriperumbudur and Steinwart, 2012, Proposition A.5), there exists a constant $C > 0$ such that, for all $n \ge 1$, $h \in (0,1)$, $b > 1$ and $|\alpha| \le 4$ with $nh^{d+2|\alpha|} \ge (b \vee |\log h|)$, we have

$$\mathbb{P}\Big(\sup_{x\in[0,1]^d} |\partial^{(\alpha)}\widehat{f}(x) - \mathbb{E}\partial^{(\alpha)}\widehat{f}(x)| < C\sqrt{\frac{b \vee |\log h|}{nh^{d+2|\alpha|}}}\Big) \ge 1 - e^{-b}. \tag{B.11}$$

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

$$\sup_{x\in[0,1]^d} |\partial^{(\alpha)} f(x) - \mathbb{E}\partial^{(\alpha)} \widehat{f}(x)| = O(h^{(4-|\alpha|)\wedge 2}). \tag{B.12}$$

Then (B.11) and (B.12) imply that for any $B > 0$, on a set A_n with probability at least $1 - n^{-B}$, there exists a constant $C > 0$ such that

$$\sup_{x\in[0,1]^d} |\partial^{(\alpha)} f(x) - \partial^{(\alpha)} \widehat{f}(x)| \leq C \left(\left(\frac{\log n}{nh^{d+4}}\right)^{1/2} + h^{(4-|\alpha|)\wedge 2} \right),$$

for all $|\alpha| \leq 4$. The fact that the eigenvalues of a symmetric matrix M are Lipschitz continuous functions of M implies that for every $\delta > 0$ there exists n_0 such that for $n \geq n_0$, $\sup_{x\in[0,1]^d} |\lambda_{k+1}^{\widehat{f}}(x) - \lambda_{k+1}^{f}(x)| \leq \delta$ on A_n. By assumption $\mathbf{(A2)}_f$, this implies that on A_n we have $S_\epsilon^{\widehat{\eta},\widehat{f}} = S_\epsilon^{\widehat{\eta},f}$ for $\epsilon \geq 0$ small enough and n large enough. Denote $\rho_n = \sup_{x\in[0,1]^d} |\sqrt{-\widehat{\eta}(x)} - \sqrt{-\eta(x)}|$. It follows that for n large enough we have on A_n,

$$\text{Ridge}(\widehat{f}) = S_0^{\widehat{\eta},\widehat{f}} = S_0^{\widehat{\eta},f} \subset \left\{x \in [0,1]^d : \; \eta(x) \geq -\rho_n^2, \, \lambda_{k+1}^f(x) < 0\right\} = S_{\rho_n^2}^{\eta,f}.$$

We then have on A_n,

$$\sup_{x\in\text{Ridge}(\widehat{f})} d(x, \text{Ridge}(f)) \leq \sup_{x\in\text{Ridge}(\widehat{f})} d(x, S_{\rho_n^2}^{\eta,f}) \leq C_1\rho_n \leq C_2\Big(\left(\frac{\log n}{nh^{d+4}}\right)^{1/2} + h^2\Big), \tag{B.13}$$

for some constants $C_1, C_2 > 0$, which follows from Lemma 5 and the rate of convergence of ρ_n. Indeed, note that

$$\begin{aligned}
\sqrt{2}\rho_n &\leq \sup_{x\in[0,1]^d} \|\Pi^{\widehat{f}}(x)\nabla \widehat{f}(x) - \Pi^f(x)\nabla f(x)\| \\
&\leq \sup_{x\in[0,1]^d} \|[\Pi^{\widehat{f}}(x) - \Pi^f(x)]\nabla f(x)\| + \sup_{x\in[0,1]^d} \|\Pi^{\widehat{f}}(x)[\nabla \widehat{f}(x) - \nabla f(x)]\| \\
&\leq \sup_{x\in[0,1]^d} \|\Pi^{\widehat{f}}(x) - \Pi^f(x)\|_F \, \|\nabla f(x)\| + \sup_{x\in[0,1]^d} \|\nabla \widehat{f}(x) - \nabla f(x)\| \\
&\leq \frac{2\sqrt{2}}{\beta} \sup_{x\in[0,1]^d} \|\nabla^2 \widehat{f}(x) - \nabla^2 f(x)\|_F \sup_{x\in[0,1]^d} \|\nabla f(x)\| + \sup_{x\in[0,1]^d} \|\nabla \widehat{f}(x) - \nabla f(x)\|,
\end{aligned}$$

where the last step follows from the Davis-Kahan theorem (see Yu et al., 2015). Using (B.11) and (B.12) gives the asserted rate in (B.13). The result in Theorem 9(i) allows us to swap the roles of f and $\widehat{f}$ in (B.13) and get on a set with probability at least $1 - n^{-B}$,

$$\sup_{x\in\text{Ridge}(f)} d(x, \text{Ridge}(\widehat{f})) = C_3\Big(\left(\frac{\log n}{nh^{d+4}}\right)^{1/2} + h^2\Big), \tag{B.14}$$

for some positive constant C_3. We conclude the proof by combining (B.13) and (B.14).

■

Proof of Theorem 9. (i). Using (B.11) and (B.12), properties **(A1)**$_{\widehat{f},4}$ and **(A2)**$_{\widehat{f}}$ are consequences of the Lipschitz continuity of the eigenvalues as functions of symmetric matrices. For a symmetric matrix A, let $\lambda_{d-k}(A)$ be the $(d-k)$th largest eigenvalue of A. To show **(A3)**$_{\widehat{f}}$, notice that assumption **(A3)**$_f$ implies that for $\delta > 0$ small enough, there exists a constant $a_0 > 0$ such that

$$\inf_{x\in \text{Ridge}(f)^\delta} \lambda_{d-k}(\nabla\xi(x)^\top \nabla\xi(x)) \geq a_0.$$

Due to the perturbation stability of λ_{d-k}, we have $\inf_{x\in \text{Ridge}(f)^\delta} \lambda_{d-k}(\nabla\xi^{\widehat{f}}(x)^\top \nabla\xi^{\widehat{f}}(x)) \geq \frac{1}{2}a_0$ with probability at least $1-n^{-B}$ when n is large. This then implies the rank of $\nabla\xi^{\widehat{f}}(x)$ is at least $d-k$ for $x \in \text{Ridge}(\widehat{f}\,)$, if $\text{Ridge}(\widehat{f}\,) \subset \text{Ridge}(f)^\delta$, which occurs with probability $1-n^{-B}$ according to Theorem 7. Furthermore, using the expression of $\nabla\xi^{\widehat{f}}(x)$ in (A.4) and the calculation in (A.6), it can be seen that $V_i^{\widehat{f}}(x)^\top \nabla\xi^{\widehat{f}}(x) = 0$, $x \in \text{Ridge}(\widehat{f}\,)$, for all $i = 1, \cdots, k$, which implies that the rank of $\nabla\xi^{\widehat{f}}(x)$ is at most $d-k$ for $x \in \text{Ridge}(\widehat{f}\,)$. Hence **(A3)**$_{\widehat{f}}$ is satisfied.

(ii). Given part (i), this is a consequence of Lemma 6 and Theorem 14.

(iii). Given part (i), we have the following: For some $\delta > 0$, we have for $|\alpha| = 0, 1, 2$,

$$\sup_{x\in[0,1]^d} |\partial^{(\alpha)}\widehat{\eta}(x) - \partial^{(\alpha)}\widehat{\eta}_\tau(x)| = O(\tau^2). \tag{B.15}$$

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 τ small, we have $\text{Ridge}(\widehat{\eta}_\tau) = \{\lim_{t\to\infty} \gamma^{\widehat{\eta}_\tau}(x,t), x \in \partial S_\epsilon^{\widehat{\eta}_\tau,\widehat{f}}\}$, as a result of Theorem 16(iii). This is a). Part b) is a consequence of Theorem 16 (iv). Part c) immediately follows from part b) and Theorem 7.

■

Proof of Theorem 11 For part (i), the convergence of the sequence follows from Theorem 9(i) and Facts 1 and 2 in the proof of Theorem 18. Then the rates of convergence is a result after applying Theorem 9(i) and Corollary 20. The result in part (ii) is a consequence of Theorem 9 and Corollary 20.

■

Proof of Theorem 14 The assertion of part (i) is very similar to Lemma 2 in Genovese et al. (2014) and the proof is similar, too. Details are omitted. Part (ii) follows from LaSalle's Invariance Principle given in Theorem 13. See the beginning of Section 4.2 for how it is applied to this setting. Next we prove part (iii), for which we will use the technique in the proof of Theorem 2.6 in Nicolaescu (2011). Let

$$\xi_r^\eta := \frac{\xi^\eta}{\|\xi^\eta\|^2}, \tag{B.16}$$

and let γ_r^η be the flow generated by this rescaled vector field as defined in Section 3.2.1, that is, $\frac{\partial \gamma_r^\eta(x,t)}{\partial t} = \xi_r^\eta(\gamma_r^\eta(x,t)), \quad \gamma_r^\eta(x,0) = x$ for all $x \in S_\epsilon^{\eta,f}$. Here we require ϵ to be small enough that $S_\epsilon^{\eta,f} \subset (0,1)^d$, which is possible following Lemma 5. We first show part (iii) with γ^η replaced by γ_r^η. Observe that

$$\frac{\partial \eta(\gamma_r^\eta(x,t))}{\partial t} = [\nabla \eta(\gamma^\eta(x,t))]^\top \xi_r^\eta(\gamma_r^\eta(x,t)) = 1. \tag{B.17}$$

In other words, the level of η can be used to parametrize the integral curves $\gamma_r^\eta(x,t)$, so that for any $\epsilon_1, \epsilon_2 \in (0,\epsilon)$ with $\epsilon_1 > \epsilon_2$, we have

$$\partial S_{\epsilon_1}^{\eta,f} = \{\gamma_r^\eta(x, \epsilon_2 - \epsilon_1) : x \in \partial S_{\epsilon_2}^{\eta,f}\}, \tag{B.18}$$

$$\partial S_{\epsilon_2}^{\eta,f} = \{\gamma_r^\eta(x, \epsilon_1 - \epsilon_2) : x \in \partial S_{\epsilon_1}^{\eta,f}\}. \tag{B.19}$$

Both (B.18) and (B.19) are similar to Theorem 2.6 in Nicolaescu (2011). We only show (B.18). Using (B.17), it is clear that for any $x \in \partial S_{\epsilon_2}^{\eta,f}$, $\gamma_r^\eta(x, \epsilon_2 - \epsilon_1) \in \partial S_{\epsilon_1}^{\eta,f}$. This means $\{\gamma_r^\eta(x, \epsilon_2 - \epsilon_1) : x \in \partial S_{\epsilon_2}^{\eta,f}\} \subset \partial S_{\epsilon_1}^{\eta,f}$. Note that for any $x \in \partial S_{\epsilon_1}^{\eta,f}$, there exists $\tilde{x} = \gamma_r^\eta(x, \epsilon_1 - \epsilon_2)$ so that $x = \gamma_r^\eta(\tilde{x}, \epsilon_2 - \epsilon_1)$. This means that $\partial S_{\epsilon_1}^{\eta,f} \subset \{\gamma_r^\eta(x, \epsilon_2 - \epsilon_1) : x \in \partial S_{\epsilon_2}^{\eta,f}\}$. Hence (B.18) is verified.

By using (B.18) and (B.19), Lemma 5 implies that, for all $\epsilon' > 0$ small enough

$$L_1\sqrt{\epsilon'} \leq d_H(\{\gamma_r^\eta(x, \epsilon - \epsilon') : x \in \partial S_\epsilon^{\eta,f}\}, \mathrm{Ridge}(f)) \leq L_2\sqrt{\epsilon'}. \tag{B.20}$$

Next we show that

$$\mathrm{Ridge}(f) = \{\lim_{t \to \epsilon^-} \gamma_r^\eta(x,t) : x \in \partial S_\epsilon^{\eta,f}\}. \tag{B.21}$$

As shown below, $\gamma_r^\eta(x,\cdot)$ and $\gamma^\eta(x,\cdot)$ have the same trajectories, and in particular,

$$\lim_{t \to \infty} \gamma^\eta(x,t) = \lim_{s \to \epsilon^-} \gamma_r^\eta(x,s) \text{ for all } x \in S_\epsilon^{\eta,f}. \tag{B.22}$$

Recall that Ridge(f) is a compact set by Lemma 4. The set $\{\lim_{t \to \epsilon^-} \gamma_r^\eta(x,t) : x \in \partial S_\epsilon^{\eta,f}\}$ is also compact. This is because $\lim_{t \to \epsilon^-} \gamma_r^\eta(x,t)$ is a continuous function of x (see Fact 6 below) and $\partial S_\epsilon^{\eta,f}$ is a compact set. Suppose that (B.21) is not true. Then there must exist a constant $\delta_0 > 0$ such that

$$\delta_0 \leq d_H(\{\lim_{t \to \epsilon^-} \gamma_r^\eta(x,t) : x \in \partial S_\epsilon^{\eta,f}\}, \mathrm{Ridge}(f)).$$

It follows from follows from (B.22) and (ii) that $\{\lim_{t \to \epsilon^-} \gamma_r^\eta(x,t) : x \in \partial S_\epsilon^{\eta,f}\} \subset \mathrm{Ridge}(f)$, and hence there exists $x_0 \in \mathrm{Ridge}(f)$ such that

$$\inf_{x \in \partial S_\epsilon^{\eta,f}} \|x_0 - \lim_{t \to \epsilon^-} \gamma_r^\eta(x,t)\| \geq \delta_0. \tag{B.23}$$

By (B.20), there exists $x_1 \in \partial S_\epsilon^{\eta,f}$ and $\epsilon' = (\delta_0/(3(L_2 \vee C')))^2$, where C' is given in Fact 4 below, such that with $x_2 = \gamma_r^\eta(x_1, \epsilon - \epsilon')$, $\|x_2 - x_0\| \leq L_2\sqrt{\epsilon'} \leq \delta_0/3$. Let $x_3 = \lim_{t \to \epsilon^-} \gamma_r^\eta(x_0, t)$. Using Fact 4, we then have $\|x_2 - x_3\| \leq \int_0^{\epsilon'} \|\gamma_r^\eta(x,t)\| dt \leq C'\sqrt{\epsilon'} \leq \delta_0/3$.

The triangle inequality gives $\|x_3 - x_0\| \leq 2\delta_0/3 < \delta_0$, which contradicts (B.23) and therefore (B.21) has to be true, which by (B.22) implies (iii).

Next we show that $\gamma_r^\eta(x,\cdot)$ and $\gamma^\eta(x,\cdot)$ have the same trajectories, which makes the proof of (iii) complete. To this end we show a reparameterization relation between the two flows. For each $x \in \partial S_\epsilon^{\eta,f}$, let

$$\begin{aligned} s(t) &:= \int_0^t \|\xi^\eta(x,u)\|^2 du \\ &= \int_0^t \left[\frac{\partial \eta(\gamma^\eta(x,u))}{\partial u}\right] du \\ &= \eta(\gamma^\eta(x,t)) - \eta(\gamma^\eta(x,0)) \\ &= \eta(\gamma^\eta(x,t)) - \eta(x), \end{aligned}$$

where the first equality is using (4.3). 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

$$\frac{dt(s)}{ds} = \frac{1}{\|\xi^\eta(x,t(s))\|^2}. \tag{B.24}$$

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

$$\frac{\partial \gamma^\eta(x,t(s))}{\partial s} = \xi_r^\eta(\gamma^\eta(x,t(s))), \quad \gamma^\eta(x,t(0)) = x. \tag{B.25}$$

Note that as $t \to \infty$, we have $s(t) \to -\eta(x) = \epsilon$ for all $x \in S_\epsilon^{\eta,f}$, because $\lim_{t\to\infty} \gamma^\eta(x,t) \in \text{Ridge}(f)$. Hence we get (B.22).

■

Proof of Theorem 16. First we show that

$$\text{Ridge}(\widetilde{\eta}) \cap S_\epsilon^{\widetilde{\eta},f} = R(\widetilde{\eta}) \cap S_\epsilon^{\widetilde{\eta},f}, \tag{B.26}$$

where $R(\widetilde{\eta}) = \{x \in [0,1]^d : \xi^{\widetilde{\eta}}(x) = 0\}$. Using Lemma 4 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 $x \in \text{Ridge}(f)^{\delta'}$,

$$-\frac{1}{2}\alpha \geq \lambda_{k+1}^{\widetilde{\eta}}(x) \geq \cdots \geq \lambda_d^{\widetilde{\eta}}(x) \geq -2A. \tag{B.27}$$

Then for any ϵ small enough we can choose $\delta_0 \in [0,\epsilon)$ small enough such that

$$S_{\epsilon-\delta_0}^{\eta,f} \subset S_\epsilon^{\widetilde{\eta},f} \subset S_{\epsilon+\delta_0}^{\eta,f} \subset \text{Ridge}(f)^{\delta'}, \tag{B.28}$$

and hence we get (B.26). Then (i) and (ii) follows from similar arguments for Theorem 14 (i) and (ii).

Next we prove (iii).

Following a similar argument as in the proof of Lemma 5, we can show that for all $\epsilon' > 0$ small enough

$$\widetilde{L}_1\sqrt{\epsilon'} \leq d_H(\partial S_{\epsilon'}^{\widetilde{\eta},f}, \text{Ridge}(\widetilde{\eta}) \cap S_\epsilon^{\widetilde{\eta},f}) \leq \widetilde{L}_2\sqrt{\epsilon}, \tag{B.29}$$

for some constants $\widetilde{L}_1, \widetilde{L}_2 > 0$. Then (iii) can be proved following the similar arguments for Theorem 14 (iii). Using similar arguments given in the proof of Theorem 7, we can show that there exists a constant $C > 0$ such that when $\max(\delta_0, \delta_1, \delta_2)$ and $\epsilon > 0$ are small enough,

$$d_H(\text{Ridge}(\widetilde{\eta}) \cap S_\epsilon^{\widetilde{\eta},f}, \text{Ridge}(\eta) \cap S_\epsilon^{\widetilde{\eta},f}\big) \leq C \max(\delta_1, \delta_2). \tag{B.30}$$

The result in (iv) follows from (B.28), (B.30) and Lemma 6.

■

Proof of Theorem 18. We will first prove several facts that then lead to Theorem 18. Throughout the proofs below we assume without further mention that assumptions **(A1)**$_{f,6}$, **(A2)**$_f$, and **(A3)**$_f$ hold.

Notation: In order to ease the notation in this long proof, we will drop the superscript η and write, for instance, $\xi, \gamma, \gamma_a, \lambda_{k+1}(x), \Pi(x), V_{k+1}(x), V_\perp(x)$ for $\xi^\eta, \gamma^\eta, \gamma_a^\eta, \lambda_{k+1}^\eta(x), \Pi^\eta(x)$, $V_{k+1}^\eta(x)$ and $V_\perp^\eta(x)$, respectively. We also write S_ϵ for $S_\epsilon^{\eta,f}$. Recall that the superscript notation is defined in Section 2.5 and in Sections 3.2.1 and 3.2.2.

Fact 1 *If $\epsilon, a > 0$ are small enough, then, for any starting point $x \in \partial S_\epsilon$, the sequence $\gamma_a(x,\ell), \ell = 0, 1, 2, \cdots$ stays in S_ϵ and converges to a ridge point in S_ϵ.*

Proof: By (B.6), we can choose ϵ small enough that $S_\epsilon \subset \text{Ridge}(f)^{(\frac{1}{2}\delta')}$, where δ' is given in Lemma 4. Denote $S_{\epsilon,\delta'} = \big(S_\epsilon\big)^{(\frac{1}{2}\delta')}$. Under assumption **(A2)**$_f$ we have

$$\sup_{x \in S_{\epsilon,\delta'}} \lambda_{k+1}^f(x) \leq -\beta < 0. \tag{B.31}$$

Since S_ϵ is a compact set when ϵ is small enough and $\|\xi\|$ is a continuous function on S_ϵ with $\|\xi\| = 0$ on Ridge(f), we can choose ϵ small enough that $\sup_{x \in S_\epsilon} \|\xi(x)\| < \frac{1}{2}\delta'$.

Let $\kappa(y) = \sup\{\|\nabla^2 \eta(z)\|_{\text{op}} : z \in \mathcal{B}(y, \|\xi(y)\|)\}$, where $\|\cdot\|_{\text{op}}$ is the operator norm of a matrix. From Lemma 4 it is known that $0 < \alpha < \kappa(y) < A < \infty$, for all $y \in S_\epsilon$. Choose $0 < a < A^{-1} \wedge 1$. Using a Taylor expansion, we have for any $y \in S_\epsilon$,

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

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

$$\eta(y + a\xi(y)) \geq \eta(y) + \frac{1}{2}a\|\xi(y)\|^2 \geq \eta(y).$$

Thus, the sequence $\eta(\gamma_a(x,\ell)), \ell = 0, 1, 2, \cdots$ is upper bounded by 0 and increasing, and therefore convergent. We can see that if $\gamma_a(x,\ell) \in S_\epsilon$ then $\gamma_a(x,\ell+1) \in S_{\epsilon,\delta'}$ and hence $\gamma_a(x,\ell+1) \in S_\epsilon$ using (B.31). In other words, $\gamma_a(x,\ell)$ stays in S_ϵ for all $\ell \geq 0$. Moreover, using the above inequality, we have

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

and therefore

$$\lim_{\ell \to \infty} \|\xi(\gamma_a(x,\ell))\| \to 0.$$

Observe further that by using Fact 2 below, $\{\gamma_a(x,\ell)\}_\ell$ 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_ϵ).

■

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

$$\|\gamma_a(x,\ell+2) - \gamma_a(x,\ell+1)\| \leq (1 - c_1 a)\|\gamma_a(x,\ell+2) - \gamma_a(x,\ell)\|. \tag{B.32}$$

As a consequence, the maximal length of the discretized paths with starting points in ∂S_ϵ is bounded by $C\sqrt{\epsilon}$ for a constant $C > 0$ not depending on a , i.e.,

$$\sup_{x \in \partial S_\epsilon} \sum_{\ell=0}^{\infty} \|\gamma_a(x,\ell+1) - \gamma_a(x,\ell)\| \leq C\sqrt{\epsilon}.$$

Proof: We assume that the a and ϵ are small enough that the sequence $\gamma_a(x,\ell), \ell = 0, 1, 2, \cdots$ stays in S_ϵ, as given in Fact 1 throughout the proof. First notice that if there exists $\ell_0 \geq 0$ such that $\xi(\gamma_a(x,\ell_0)) = 0$, then $\gamma_a(x,\ell_0) = \gamma_a(x,\ell_0+1) = \gamma_a(x,\ell_0+2) = \cdots$, and the conclusion of this fact is valid. We thus can assume that $\xi(\gamma_a(x,\ell)) \neq 0$ for all $\ell \geq 0$. We have the following Taylor expansion

$$\begin{aligned}
\|\xi(\gamma_a(x,\ell+1))\|^2 &= \|\xi(\gamma_a(x,\ell)\|^2 + \big\langle \nabla\|\xi(y)\|^2|_{y=\gamma_a(x,\ell)}, \gamma_a(x,\ell+1) - \gamma_a(x,\ell)\big\rangle \\
&\quad + \frac{1}{2}[\gamma_a(x,\ell+1) - \gamma_a(x,\ell)]^\top \{\nabla^2\|\xi(y)\|^2|_{y=\widetilde{y}_\delta}\}[\gamma_a(x,\ell+1) - \gamma_a(x,\ell)] \\
&= \|\xi(\gamma_a(x,\ell)\|^2 + a\big\langle \nabla\|\xi(y)\|^2|_{y=\gamma_a(x,\ell)}, \xi(\gamma_a(x,\ell)\big\rangle \\
&\quad + a^2\frac{1}{2}[\xi(\gamma_a(x,\ell)]^\top \{\nabla^2\|\xi(y)\|^2|_{y=\widetilde{y}_{t,\ell}}\}[\xi(\gamma_a(x,\ell)]
\end{aligned}$$

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

Let $\nu(y) = \lambda_{\max}[\nabla^2\|\xi(y)\|^2]$, where $\lambda_{\max}(B)$ denotes the maximum eigenvalue of a symmetric matrix B. For each $y \in \text{Ridge}(f)$, because $\nabla\eta(y) = 0$ and $\xi(y) = 0$, we can write

$$\nabla^2\|\xi(y)\|^2 = 2[\nabla\xi(y)]^\top\nabla\xi(y) = 2\nabla^2\eta(y)V_\perp(y)[V_\perp(y)]^\top\nabla^2\eta(y).$$

Recalling that $V_\perp(y)$ is the matrix built by the trailing $k-d$ (unit) eigenvectors of $\nabla^2\eta(y)$, we see that $\nu(y) = 2[\lambda_d(y)]^2$ for $y \in \text{Ridge}(f)$. Since ν is a continuous function on $\text{Ridge}(f)^{\delta'}$, where δ' is given in Lemma 4, we can find an $\epsilon > 0$ small enough that $0 < \lambda^*_{\max} := \sup_{y \in S_\epsilon} \nu(y) < \infty$. We thus obtain

$$\|\xi(\gamma_a(x,\ell+1)\|^2$$

$$\leq \|\xi(\gamma_a(x,\ell)\|^2 + 2a[\xi(\gamma_a(x,\ell)]^\top[\nabla\xi(\gamma_a(x,\ell)][\xi(\gamma_a(x,\ell)] + \frac{1}{2}a^2\|\xi(\gamma_a(x,\ell))\|^2\lambda^*_{\max}. \tag{B.33}$$

Hence

$$\begin{aligned}\frac{\|\gamma_a(x,\ell+2)-\gamma_a(x,\ell+1)\|^2}{\|\gamma_a(x,\ell+1)-\gamma_a(x,\ell)\|^2} &= \frac{\|\xi(\gamma_a(x,\ell+1))\|^2}{\|\xi(\gamma_a(x,\ell))\|^2}\\ &\leq 1 + 2a\frac{[\xi(\gamma_a(x,\ell))]^\top[\nabla\xi(\gamma_a(x,\ell))][\xi(\gamma_a(x,\ell))]}{\|\xi(\gamma_a(x,\ell))\|^2} + \frac{1}{2}a^2\lambda^*_{\max}. \end{aligned}\tag{B.34}$$

For any $y \in S_\epsilon$, let $\mathcal{N}(y) = \{\Pi(y)u : \; u \in \mathbb{R}^d\}$, which is the $(d-k)$-dimensional subspace spanned by the orthonormal eigenvectors $V_{k+1}(y), \cdots, V_d(y)$. Also let $\mathcal{S}_\circ(y) = \{v/\|v\|, v \in \mathcal{N}(y) \setminus \{\mathbf{0}\}\}$, which is a $(d-k)$-dimensional unit sphere in $\mathbb{R}^d$. Define

$$\bar\beta(y) = \sup_{v\in\mathcal{N}(y)\setminus\{\mathbf{0}\}} \frac{v^\top\nabla\xi(y)v}{\|v\|^2} = \sup_{w\in\mathcal{S}_\circ(y)} w^\top\nabla\xi(y)w. \tag{B.35}$$

Then from (B.34) we can write

$$\frac{\|\gamma_a(x,\ell+2)-\gamma_a(x,\ell+1)\|}{\|\gamma_a(x,\ell+1)-\gamma_a(x,\ell)\|} \leq 1 + a\bar\beta(\gamma_a(x,\ell))) + \frac{1}{4}a^2\lambda^*_{\max}. \tag{B.36}$$

For any $y \in \text{Ridge}(f)$, we can write $\nabla\eta(y) = 0$ and $\nabla\xi(y) = V_\perp(y)[V_\perp(y)]^\top\nabla^2\eta(y) = \sum_{j=k+1}^d \lambda_j(y)V_j(y)V_j(y)^\top$, and therefore

$$\bar\beta(y) = \lambda_{k+1}(y) < 0. \tag{B.37}$$

Next we will show that $\bar\beta$ is continuous in $\text{Ridge}(f)^\delta$. To this end, we will, for any $y, y_0 \in \text{Ridge}(f)^\delta$, find a bijective map $q_{y,y_0} : \mathcal{S}_\circ(y) \to \mathcal{S}_\circ(y_0)$. Let $\pi/2 \geq \theta_1 \geq \cdots \geq \theta_{d-k} \geq 0$ be the principal angles between $\mathcal{N}(y)$ and $\mathcal{N}(x_0)$, and suppose that $u_1(\tilde y), \cdots, u_k(\tilde y)$ are the associated principal vectors for $\mathcal{N}(\tilde y)$, where $\tilde y \in \{y, y_0\}$, respectively. In other words, if the singular value decomposition of $V_\perp(y)^\top V_\perp(y_0)$ is given by $P\Sigma R^\top$, where P and R are $(d-k)\times(d-k)$ orthogonal matrices and Σ is a $(d-k)\times(d-k)$ diagonal matrix, then $[u_1(y), \cdots, u_k(y)] = V_\perp(y)P$, $\Sigma = \cos\Theta$, and $[u_1(y_0), \cdots, u_k(y_0)] = V_\perp(y_0)R$, where $\Theta = \text{diag}(\theta_1, \cdots, \theta_{d-k})$. Using the Davis-Kahan theorem and Lemma 4, we have

$$\|\sin\Theta\|_F \leq \frac{2}{\beta}\|\nabla^2\eta(y) - \nabla^2\eta(y_0)\|_F. \tag{B.38}$$

We choose $\|y-y_0\|$ small enough that $\theta_1 \leq 2\pi/3$, so that $\sin\theta_i \geq \sin(\theta_i/2)$ for $i = 1, \cdots, d-k$. Let $U(y,y_0) = [\bar u_1(y,y_0), \cdots, \bar u_{d-k}(y,y_0)]$, where $\bar u_i(y,y_0) = \frac{1}{\sqrt{2(1+\cos\theta_i)}}[u_i(y) + u_i(y_0)]$, $i = 1, \cdots, d-k$, and $\bar{\mathcal{N}}(y,y_0)$ be the column space of $U(y,y_0)$. Note that the columns of $U(y,y_0)$ are orthonormal and $\bar{\mathcal{N}}(y,y_0)$ is a subspace about which $\mathcal{N}(y)$ and $\mathcal{N}(y_0)$ are symmetric. Therefore the images of $\mathcal{S}_\circ(x)$ and $\mathcal{S}_\circ(x_0)$ under the projection map $\Omega_{y,y_0} := U(y,y_0)[U(y,y_0)]^\top$ are the same. For $w \in \mathcal{S}_\circ(y)$, define $q_{y,y_0}(w)$ in such a way that $\Omega_{y,y_0}q_{y,y_0}(w) = w$. Here $q_{y,y_0}(w)$ is uniquely defined because Ω_{y,y_0} is bijective from either $\mathcal{N}(y)$ or $\mathcal{N}(y_0)$ to $\bar{\mathcal{N}}(y,y_0)$. For $w \in \mathcal{S}_\circ(y)$, using (B.38), we have

$$\|q_{y,y_0}(w) - w\| = 2\sqrt{\sin^2\left(\frac{\theta_1}{2}\right) + \cdots + \sin^2\left(\frac{\theta_{d-k}}{2}\right)} \leq 2\|\sin\Theta\|_F \leq \frac{4}{\beta}\|\nabla^2\eta(y) - \nabla^2\eta(y_0)\|_F. \tag{B.39}$$

Then we can write

$$\begin{aligned}
|\bar{\beta}(y) - \bar{\beta}(y_0)| &\leq \Big| \sup_{w \in \mathcal{S}_\circ(y)} w^\top \nabla\xi(y) w - \sup_{w \in \mathcal{S}_\circ(y)} w^\top \nabla\xi(y_0) w \Big| \\
&\quad + \Big| \sup_{w \in \mathcal{S}_\circ(y)} w^\top \nabla\xi(y_0) w - \sup_{w \in \mathcal{S}_\circ(y_0)} w^\top \nabla\xi(y_0) w \Big| \\
&\leq \sup_{w \in \mathcal{S}_\circ(y)} \Big| w^\top \nabla\xi(y) w - w^\top \nabla\xi(y_0) w \Big| \\
&\quad + \sup_{w \in \mathcal{S}_\circ(y)} \Big| w^\top \nabla\xi(y_0) w - q_{y,y_0}(w)^\top \nabla\xi(y_0) q_{x,x_0}(w) \Big| \\
&\leq \|\nabla\xi(y) - \nabla\xi(y_0)\|_F + 2\|\nabla\xi(y_0))\|_F \sup_{w \in \mathcal{S}_\circ(y)} \|w - q_{y,y_0}(w)\| \\
&\leq \|\nabla\xi(y) - \nabla\xi(y_0)\|_F + \frac{8\|\nabla\xi(y_0))\|_F}{\beta} \|\nabla^2\eta(y) - \nabla^2\eta(y_0)\|_F. \qquad \text{(B.40)}
\end{aligned}$$

Using the boundedness of $\|\nabla\xi\|_F$, we obtain that $\bar{\beta}(y)$ is a uniformly continuous function on $\text{Ridge}(f)^\delta$. Because the ridge is a compact set, using (B.37), we are able to find $\epsilon > 0$ such that

$$\beta_0 := \sup_{y \in S_\epsilon} \bar{\beta}(y) < 0, \qquad \text{(B.41)}$$

and from (B.36) we have that for all $\ell \geq 0$,

$$\frac{\|\gamma_a(x, \ell+2) - \gamma_a(x, \ell+1)\|}{\|\gamma_a(x, \ell+1) - \gamma_a(x, \ell)\|} \leq 1 + a\bar{\beta}(\gamma_a(x, \ell))) + \frac{1}{4}a^2\lambda^*_{\max} \leq 1 + a\beta_0 + \frac{1}{4}a^2\lambda^*_{\max}.$$

We require that $a \in (0, -\frac{2\beta_0}{\lambda^*_{\max}})$. 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 (B.32) with $c_1 = -\beta_0/2$. Let

$$\kappa^\dagger(\epsilon) = \sup_{x \in \partial S_\epsilon} \|\xi(x)\|. \qquad \text{(B.42)}$$

Notice that $0 < 1 + a\beta_0 < 1$. We can then write for any $x \in \partial S_\epsilon$,

$$T_a(x) := \sum_{\ell=0}^{\infty} \|\gamma_a(x, \ell+1) - \gamma_a(x, \ell)\| \leq \sum_{\ell=0}^{\infty} a\|\xi(x)\| \Big(1 + \frac{1}{2}a\beta_0\Big)^\ell \leq \frac{2\kappa^\dagger(\epsilon)}{-\beta_0}. \qquad \text{(B.43)}$$

For any $x \in \partial S_\epsilon$,

$$\|\xi(x)\| \leq \|V_\perp(x)V_\perp(x)^\top\|_F \|\nabla\eta(x)\| = \sqrt{d-k}\|\nabla\eta(x)\|. \qquad \text{(B.44)}$$

Recall that the notation used in the proof of Lemma 6, in particular, x_0 is the projection of x onto $\text{Ridge}(f)$. It follows from (B.8) that

$$\|\nabla\eta(x)\| \leq A\|x - x_0\| + \kappa_1\|x - x_0\|^2,$$

where A is given in Lemma 4. Then using (B.6), we get

$$\|\nabla\eta(x)\| \leq A\sqrt{\frac{8}{\alpha}}\sqrt{\epsilon} + \kappa_1\frac{8}{\alpha}\epsilon \leq (A + \kappa_1)\sqrt{\frac{8}{\alpha}}\sqrt{\epsilon},$$

where the last inequality is true when $8\epsilon \leq \alpha$. Combined with (B.44), this leads to

$$\kappa^\dagger(\epsilon) \leq (A + \kappa_1)\sqrt{\frac{8(d-k)}{\alpha}}\sqrt{\epsilon}. \tag{B.45}$$

It then follows from (B.43) that $T_a(x) \leq C\sqrt{\epsilon}$, where $C = \frac{2}{-\beta_0}(A + \kappa_1)\sqrt{\frac{8(d-k)}{\alpha}}$.

■

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 ϵ is small enough: Let $x, x_0 \in \partial S_\epsilon$ be two starting points. For any $\omega > 0$, there exists $\delta_\omega > 0$ such that when $\|x - x_0\| \leq \delta_\omega$, we have*

$$\|\lim_{\ell\to\infty}\gamma_a(x,\ell) - \lim_{\ell\to\infty}\gamma_a(x_0,\ell)\| \leq \omega.$$

Proof: Note that for $\tilde{x} \in \{x, x_0\}$,

$$\gamma_a(\tilde{x}, \ell+1) = \tilde{x} + \sum_{i=0}^{\ell}[\gamma_a(\tilde{x}, i+1) - \gamma_a(\tilde{x}, i)] = \tilde{x} + a\sum_{i=0}^{\ell}\xi(\gamma_a(\tilde{x}, i)).$$

Then using a Taylor expansion we have

$$\begin{aligned}\gamma_a(x,\ell+1) - \gamma_a(x_0,\ell+1) &= (x - x_0) + a\sum_{i=0}^{\ell}\{\xi(\gamma_a(x,i)) - \xi(\gamma_a(x_0,i))\}\\ &= (x - x_0) + a\sum_{i=0}^{\ell}\int_0^1 \nabla\xi(\gamma_{a,i}(t,x,x_0))dt\{\gamma_a(x,i) - \gamma_a(x_0,i)\},\end{aligned}$$

where $\gamma_{a,i}(t,x,x_0) = t\gamma_a(x,i) + (1-t)\gamma_a(x_0,i)$. As in the proof of Fact 1, we can choose ϵ small enough such that $S_\epsilon \subset \text{Ridge}(f)^{(\frac{1}{2}\delta')}$. Recalling that $\|\xi(x)\| = 0$ for $x \in \text{Ridge}(f)$, we obtain from Fact 2 that ϵ and a can be chosen small enough that $\max\{T_a(x), T_a(x_0)\} \leq \frac{1}{6}\delta'$ for all $x, x_0 \in \partial S_\epsilon$, where T_a is defined in (B.43). Suppose $\|x - x_0\| \leq \frac{1}{6}\delta'$ so that by elementary geometric facts, $\gamma_{a,i}(t,x,x_0) \in \text{Ridge}(f)^{\delta'}$ for all $i \geq 0$, and $t \in [0,1]$. Hence with

$$\kappa^* := \sup_{x\in\text{Ridge}(f)^{\delta'}}\|\nabla\xi(x))\|_F < \infty, \tag{B.46}$$

we have

$$\|\gamma_a(x,\ell+1)-\gamma_a(x_0,\ell+1)\| \leq \|x-x_0\| + a\kappa^* \sum_{i=0}^{\ell} \|\gamma_a(x,i)-\gamma_a(x_0,i)\|. \tag{B.47}$$

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*

$$y_n \leq c + \sum_{0\leq k<n} g_k y_k,\ n\geq 0,$$

then

$$y_n \leq c\exp\left(\sum_{0\leq j<n} g_j\right),\ n\geq 0.$$

Applying this inequality to (B.47), we get

$$\|\gamma_a(x,\ell+1)-\gamma_a(x_0,\ell+1)\| \leq \|x-x_0\|\exp(a\kappa^*\ell). \tag{B.48}$$

Recall $\kappa^\dagger = \kappa^\dagger(\epsilon)$ given in (B.42). Using the argument in the proof of Fact 2 (in particular, see (B.43)), we have that for any positive integer N,

$$\begin{aligned}\|\gamma_a(x,N)-\lim_{\ell\to\infty}\gamma_a(x,\ell)\| \leq & \sum_{\ell=N}^{\infty}\|\gamma_a(x,\ell+1)-\gamma_a(x,\ell)\|\\ \leq & a\kappa^\dagger\sum_{\ell=N}^{\infty}[1-ac_1]^\ell\\ = & \kappa^\dagger\frac{[1-ac_1]^N}{c_1},\end{aligned} \tag{B.49}$$

where c_1 is given in (B.32) and a is small enough that $0<1-ac_1<1$. We then obtain for $N_\omega := \log(\frac{1}{3\kappa^\dagger}c_1\omega)/\log(1-ac_1)$, and any $x, x_0\in\partial S_\epsilon$,

$$\|\gamma_a(x,N_\omega+1)-\lim_{\ell\to\infty}\gamma_a(x,\ell)\| \leq \frac{\omega}{3}, \tag{B.50}$$

$$\|\gamma_a(x_0,N_\omega+1)-\lim_{\ell\to\infty}\gamma_a(x_0,\ell)\| \leq \frac{\omega}{3}. \tag{B.51}$$

Using the fact $\frac{t}{1+t}\leq\log(1+t)\leq t$, for all $t>-1$, we get

$$1-ac_1 \leq \frac{-ac_1}{\log(1-ac_1)} \leq 1,$$

and consequently

$$\exp(a\kappa^* N_\omega) = \exp\left(\frac{-ac_1}{\log(1-ac_1)}\Big[-\frac{\kappa^*}{c_1}\log(\frac{c_1\omega}{3\kappa^\dagger})\Big]\right) \leq \exp\left(-\frac{\kappa^*}{c_1}\log\left(\frac{c_1\omega}{3\kappa^\dagger}\right)\right).$$

Using (B.48), we choose $\delta_\omega > 0$ small enough (independent of a once it satisfied the above requirements) such that when $\|x - x_0\| \leq \delta_\omega$,

$$\begin{aligned}\|\gamma_a(x, N_\omega + 1) - \gamma_a^*(x_0, N_\omega + 1)\| \leq \|x - x_0\| \exp(a\kappa^* N_\omega)& \\ \leq \delta_\omega \exp\left(-\frac{\kappa^*}{c_1} \log\left(\frac{c_1\omega}{3\kappa^\dagger}\right)\right) \leq \frac{\omega}{3}&.\end{aligned}$$

Combining this with (B.50) and (B.51), we have

$$\begin{aligned}&\|\lim_{\ell\to\infty} \gamma_a(x,\ell) - \lim_{\ell\to\infty} \gamma_a(x_0,\ell)\| \\ \leq &\|\gamma_a(x, N_\omega + 1) - \lim_{\ell\to\infty} \gamma_a(x,\ell)\| + \|\gamma_a(x, N_\omega + 1) - \gamma_a(x_0, N_\omega + 1)\| \\ &\qquad + \|\gamma_a(x_0, N_\omega + 1) - \lim_{\ell\to\infty} \gamma_a(x_0,\ell)\| \\ \leq &\,\omega.\end{aligned}$$

This proves the assertion of this lemma. Note that the set ∂S_ϵ is a compact set, so the continuity of the function $x \mapsto \lim_{\ell\to\infty} \gamma_a(x,\ell)$ on this set is equivalent to its uniform continuity.

■

The following fact is a continuous version of Fact 2.

Fact 4 *When $\epsilon > 0$ is small enough, the maximal length of the paths $\gamma(x,t)$ with starting points in ∂S_ϵ is bounded by $C'\sqrt{\epsilon}$ for a constant $C' > 0$ not depending on a , i.e.,*

$$\sup_{x\in S_\epsilon} \int_0^\infty \|\xi(\gamma(x,t))\|dt \leq C'\sqrt{\epsilon}.$$

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

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

For each $\ell \geq 0$, using a Taylor expansion of the function $s \mapsto \int_{t_\ell}^s \|\xi(\gamma(x,t))\|dt$, we have

$$\int_{t_\ell}^{t_{\ell+1}} \|\xi(\gamma(x,t))\|dt = a\|\xi(\gamma(x,t_\ell))\| + \frac{1}{2}a^2 \frac{\partial}{\partial t}\|\xi(\gamma(x,t))\|\Big|_{t=\tilde{t}_\ell}, \tag{B.52}$$

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

$$\begin{aligned}\frac{\partial}{\partial t}\|\xi(\gamma(x,t))\| &= \left\{[\nabla\|\xi(y)\|]^\top \xi(y)\right\}\Big|_{y=\gamma(x,t)} \\ &= \{\|\xi(y)\|\theta(y)\}\,|_{y=\gamma(x,t)},\end{aligned} \tag{B.53}$$

where $\theta(y) = \xi_\diamond(y)^\top \nabla\xi(y)\xi_\diamond(y)$ with $\xi_\diamond(y) = \xi(y)/\|\xi(y)\|$. So from (B.52) we can write

$$\int_{t_\ell}^{t_{\ell+1}} \|\xi(\gamma(x,t))\|dt = a\|\xi(\gamma(x,t_\ell))\| + \frac{1}{2}a^2\|\xi(\gamma(x,\tilde{t}_\ell))\|\theta(\gamma(x,\tilde{t}_\ell)). \tag{B.54}$$

Recall that in the proof of Fact 2 we have shown that when $\epsilon > 0$ is small enough,

$$\sup_{y \in S_\epsilon \backslash \mathrm{Ridge}(f)} \theta(y) \le \sup_{y \in S_\epsilon} \bar{\beta}(y) = \beta_0 < 0, \tag{B.55}$$

which, by (B.53), implies that

$$\|\xi(\gamma(x,t))\| \text{ is a decreasing function of } t. \tag{B.56}$$

Corresponding to the definition of $\bar{\beta}$ in (B.35), define

$$\underline{\beta}(y) = \inf_{v \in \mathcal{N}(y) \backslash \{\mathbf{0}\}} \frac{v^\top \nabla\xi(y) v}{\|v\|^2} = \inf_{w \in \mathcal{S}_\circ(y)} w^\top \nabla\xi(y) w. \tag{B.57}$$

Using an argument similar to (B.39) and (B.40), we can show that $\underline{\beta}$ is a continuous function on $\mathrm{Ridge}(f)^{\delta'}$. Similar to (B.37), we see that for any $y \in \mathrm{Ridge}(f)$, $\underline{\beta}(y) = \lambda_d(y) < 0$. Therefore we can find $\epsilon > 0$ small enough that $\beta_1 := \inf_{y \in S_\epsilon} \underline{\beta}(y) > -\infty$. Note that $\beta_1 \le \beta_0 < 0$. Then

$$\inf_{y \in S_\epsilon \backslash \mathrm{Ridge}(f)} \theta(y) \ge \beta_1. \tag{B.58}$$

Let $\rho_\ell(x) = \gamma(x, t_{\ell+1}) - \gamma(x, t_\ell)$. Similar to (B.33) we have

$$\begin{aligned} &\|\xi(\gamma(x, t_{\ell+1}))\|^2 \\ &\le \|\xi(\gamma(x,t_\ell))\|^2 + 2[\rho_\ell(x)]^\top [\nabla\xi(\gamma(x,t_\ell))]\xi(\gamma(x,t_\ell)) + \frac{1}{2}\|\rho_\ell(x)\|^2 \lambda^*_{\max}. \end{aligned} \tag{B.59}$$

We have the following Taylor expansion

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

where $\bar{t}_\ell = t_\ell + \bar{\delta}_\ell \times a$ for some $\bar{\delta}_\ell \in [0,1]$. Note that $\bar{t}_\ell$ in the Taylor expansion may be different for the entries of the vector ρ_ℓ. Recall κ^* given in (B.46). Then we have

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

where we have used (B.56), and hence

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

Then from (B.59) we have

$$\begin{aligned} &\|\xi(\gamma(x,t_{\ell+1}))\|^2 \\ &\quad \le \|\xi(\gamma(x,t_\ell))\|^2 + 2a[\xi(\gamma(x,t_\ell))]^\top[\nabla\xi(\gamma(x,t_\ell))]\xi(\gamma(x,t_\ell)) \\ &\qquad\quad + 2a^2[[\nabla\xi(y)\xi(y)]\big|_{y=\gamma(x,\bar{t}_\ell)}]^\top[\nabla\xi(\gamma(x,t_\ell))]\xi(\gamma(x,t_\ell)) + \frac{1}{2}\|\rho_\ell(x)\|^2\lambda^*_{\max} \\ &\quad \le \|\xi(\gamma(x,t_\ell))\|^2 + 2a[\xi(\gamma(x,t_\ell))]^\top[\nabla\xi(\gamma(x,t_\ell))]\xi(\gamma(x,t_\ell)) \end{aligned}$$

$$+ 2a^2(\kappa^*)^2\|\xi(\gamma(x,t_\ell))\|^2 + \frac{1}{2}(a+a^2\kappa^*)^2\lambda^*_{\max}\,\|\xi(\gamma(x,t_\ell))\|^2.$$

This leads to

$$\begin{aligned}&\frac{\|\xi(\gamma(x,t_{\ell+1}))\|^2}{\|\xi(\gamma(x,t_\ell))\|^2}\\ &\leq 1+2a\frac{[\xi(\gamma(x,t_\ell))]^\top[\nabla\xi(\gamma(x,t_\ell))]\xi(\gamma(x,t_\ell))}{\|\xi(\gamma(x,t_\ell))\|^2}+2a^2(\kappa^*)^2+\frac{1}{2}(a+a^2\kappa^*)^2\lambda^*_{\max}\\ &=1+2a\theta(\gamma(x,t_\ell))+2a^2(\kappa^*)^2+\frac{1}{2}(a+a^2\kappa^*)^2\lambda^*_{\max}.\end{aligned}$$

Therefore

$$\frac{\|\xi(\gamma(x,t_{\ell+1}))\|}{\|\xi(\gamma(x,t_\ell))\|}\leq 1+a\theta(\gamma(x,t_\ell))+a^2(\kappa^*)^2+\frac{1}{4}(a+a^2\kappa^*)^2\lambda^*_{\max}. \tag{B.60}$$

From (B.54) we have

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

As we have shown in (B.55), $\theta(\gamma(x,\tilde{t}_{\ell+1}))<0$, for all $\ell\geq 0$. Hence by (B.60),

$$\begin{aligned}\int_{t_{\ell+1}}^{t_{\ell+2}}\|\xi(\gamma(x,t))\|dt &\leq a\;\|\xi(\gamma(x,t_{\ell+1}))\|\\ &\leq a\;\|\xi(\gamma(x,t_\ell))\|\left[1+a\theta(\gamma(x,t_\ell))+a^2(\kappa^*)^2+\frac{1}{4}(a+a^2\kappa^*)^2\lambda^*_{\max}\right].\end{aligned} \tag{B.61}$$

Now we turn back to $\int_{t_\ell}^{t_{\ell+1}}\|\xi(\gamma(x,t))\|dt$. Using (B.54) – (B.56), we have

$$\int_{t_\ell}^{t_{\ell+1}}\|\xi(\gamma(x,t))\|dt\geq a\|\xi(\gamma(x,t_\ell))\|+\frac{1}{2}a^2\|\xi(\gamma(x,t_\ell))\|\theta(\gamma(x,\tilde{t}_\ell)). \tag{B.62}$$

A Taylor expansion for $\theta(\gamma(x,\tilde{t}_\ell))$ gives

$$\theta(\gamma(x,\tilde{t}_\ell)=\theta(\gamma(x,t_\ell)+(\delta_\ell a)\frac{\partial}{\partial t}\theta(\gamma(x,t))\big|_{t=\tilde{t}^*_\ell}, \tag{B.63}$$

where $\tilde{t}^*_\ell=(1-\delta^*_\ell)t_\ell+\delta^*_\ell t_{\ell+1}=t_\ell+\delta^*_\ell\times a$ for some $\delta^*_\ell\in(0,\delta_\ell)$. Here it can be shown that

$$\frac{\partial}{\partial t}\theta(\gamma(x,t))=\pi(\gamma(x,t)),$$

where

$$\begin{aligned}\pi(y)=\xi_\diamond(y)^\top\{[\nabla\xi(y)]^\top\nabla\xi(y)+\nabla\xi(y)\nabla\xi(y)\}\xi_\diamond(y)+\xi_\diamond(y)^\top[\xi(y)^\top\otimes\mathbf{I}_d]\nabla(\nabla\xi(y))\xi_\diamond(y)\\ -2[\xi_\diamond(y)^\top\nabla\xi(y)\xi_\diamond(y)]^2.\end{aligned}$$

When $\epsilon > 0$ is small enough, we have that

$$\kappa^{\ddagger} := \sup_{y \in S_\epsilon \setminus \mathrm{Ridge}(f)} |\pi(y)| < \infty.$$

Then from (B.63) we have

$$\theta(\gamma(x, \tilde{t}_\ell)) \geq \theta(\gamma(x, t_\ell)) - a\kappa^{\ddagger}.$$

Plugging this result into (B.62) we get

$$\int_{t_\ell}^{t_{\ell+1}} \|\xi(\gamma(x,t))\| dt \geq a \|\xi(\gamma(x, t_\ell))\| \Big\{ 1 + \frac{1}{2} a [\theta(\gamma(x, t_\ell)) - a\kappa^{\ddagger}] \Big\}. \tag{B.64}$$

Here we require $a < \min\{1, (\kappa^{\ddagger} - \beta_1)^{-1}\}$ so that the right-hand side of (B.64) is positive, by (B.58). Now combining (B.61) and (B.64) we have

$$\frac{\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi(\gamma(x,t))\| dt}{\int_{t_\ell}^{t_{\ell+1}} \|\xi(\gamma(x,t))\| dt} \leq \frac{1 + a\theta(\gamma(x,t_\ell)) + a^2(\kappa^*)^2 + \frac{1}{4}(a + a^2\kappa^*)^2 \lambda^*_{\max}}{1 + \frac{1}{2} a[\theta(\gamma(x,t_\ell)) - a\kappa^{\ddagger}]}. \tag{B.65}$$

Using (B.55), we can show that if we further require $a \leq \frac{-\beta_0}{4(\kappa^*)^2 + (1+\kappa^*)^2 \lambda^*_{\max} + 2\kappa^{\ddagger}}$, then

$$\frac{\int_{t_{\ell+1}}^{t_{\ell+2}} \|\xi(\gamma(x,t))\| dt}{\int_{t_\ell}^{t_{\ell+1}} \|\xi(\gamma(x,t))\| dt} \leq 1 + \frac{1}{4} a\beta_0. \tag{B.66}$$

Then

$$\begin{aligned} \int_0^\infty \|\xi(\gamma(x,t))\| dt &\leq \int_0^{t_1} \|\xi(\gamma(x,t))\| dt \sum_{i=0}^{\infty} \Big(1 + \frac{1}{4} a\beta_0\Big)^i \\ &\leq \frac{\int_0^{t_1} \|\xi(\gamma(x,t))\| dt}{1 - (1 + \frac{1}{4} a\beta_0)} \leq \frac{4\kappa^{\dagger}(\epsilon)}{-\beta_0} \leq C'\sqrt{\epsilon}, \end{aligned} \tag{B.67}$$

where $\kappa^{\dagger}(\epsilon)$ is given in (B.42) and $C' = 2C$ with C given in Fact 2, by using (B.45).

■

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, $t_\ell = a\ell$.

Fact 5 *When $\epsilon > 0$ is small enough, there exists $a_0 > 0$ such that when $a \leq a_0$, we have*

$$\sup_{x \in \partial S_\epsilon} \| \lim_{\ell \to \infty} \gamma(x, t_\ell) - \lim_{\ell \to \infty} \gamma_a(x, \ell) \| \leq C a^{1 - \sigma_0}.$$

for some constant $C > 0$, where $0 < \sigma_0 < 1$ is given in (B.69).

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

$$\begin{aligned} e_{\ell+1} &= \gamma(x, t_{\ell+1}) - \gamma_a(x, \ell + 1) \\ &= e_\ell + [\gamma(x, t_{\ell+1}) - \gamma(x, t_\ell)] - [\gamma_a(x, \ell + 1) - \gamma_a(x, \ell)] \\ &= e_\ell + [\gamma(x, t_{\ell+1}) - \gamma(x, t_\ell) - a\xi(\gamma(x, t_\ell))] + a[\xi(\gamma(x, t_\ell)) - \xi(\gamma_a(x, \ell))]. \qquad \text{(B.68)} \end{aligned}$$

As in the proof of Fact 1, we choose ϵ small enough that $S_\epsilon \subset \text{Ridge}(f)^{(\frac{1}{2}\delta')}$. By using Facts 2 and 4, we can choose ϵ small enough that $\alpha\gamma(x, t_\ell) + (1 - \alpha)\gamma_a(x, \ell) \in \text{Ridge}(f)^{\delta'}$. for all $\alpha \in [0, 1]$ and $\ell \geq 0$. We then have

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

where κ^* is defined in (B.46). Moreover, we also have

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

Hence,

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

where $\kappa^\dagger = \kappa^\dagger(\epsilon)$ is given in (B.42). It then follows from (B.68) that

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

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

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

From (B.49) in the proof of Fact 3, we know

$$\|\gamma_a(x, \ell) - \lim_{\ell' \to \infty} \gamma_a(x, \ell')\| \leq \kappa^\dagger \frac{[1 - ac_1]^\ell}{c_1} \leq \frac{\kappa^\dagger}{c_1} e^{-\ell a c_1}.$$

Using (B.66) in the proof of Fact 4 and noticing that $c_1 = -\beta_0/2$, the same arguments as in the derivation of (B.49) give that

$$\|\gamma(x, t_\ell) - \lim_{\ell' \to \infty} \gamma(x, t_{\ell'})\| \leq 2\kappa^\dagger \frac{[1 - \frac{1}{2}ac_1]^\ell}{c_1} \leq \frac{2\kappa^\dagger}{c_1} e^{-\frac{1}{2}\ell a c_1}.$$

Hence combining the above three inequalities, we get

$$\|\lim_{\ell\to\infty}\gamma(x,t_\ell)-\lim_{\ell'\to\infty}\gamma_a(x,\ell')\|\le\min_{\ell\ge 0}\psi(\ell),\qquad\text{where}\quad\psi(\ell)=\frac{1}{2}\kappa^\dagger a e^{a\ell\kappa^*}+\frac{3\kappa^\dagger}{c_1}e^{-\frac{1}{2}\ell a c_1}.$$

Denote the upper bound we have imposed on a by a_0. By choosing

$$\ell=\ell_a(\sigma_0):=\left\lceil\frac{\sigma_0}{a\kappa^*}\log\frac{1}{a}\right\rceil,\ \text{where}\ \sigma_0=\frac{1}{1+c_1/(2\kappa^*)},\tag{B.69}$$

we have

$$\|\lim_{\ell\to\infty}\gamma(x,t_\ell)-\lim_{\ell\to\infty}\gamma_a(x,\ell)\|\le\psi(\ell_a(\sigma_0))\le\kappa^\dagger\Big(\frac{1}{2}e^{a_0\kappa^*}+\frac{3}{c_1}\Big)a^{1-\sigma_0}.\tag{B.70}$$

■

We have the following continuous version of Fact 3.

Fact 6 *The following holds when ϵ is small enough: Let $x,x_0\in\partial S_\epsilon$ be two starting points. For any $\omega>0$, there exists $\delta_\omega>0$ such that when $\|x-x_0\|\le\delta_\omega$, we have*

$$\Big\|\lim_{t\to\infty}\gamma(x,t)-\lim_{t\to\infty}\gamma(x_0,t)\Big\|\le\eta.$$

Proof: Note that for $\tilde{x}\in\{x,x_0\}\subset\partial S_\epsilon$,

$$\gamma(\tilde{x},t)=\tilde{x}+\int_0^t\xi(\gamma(\tilde{x},s))ds.$$

When ϵ is small enough, we obtain by a Taylor expansion that for any $t\ge 0$,

$$\begin{aligned}\|\gamma(x,t)-\gamma(x_0,t)\|&\le\|x-x_0\|+\int_0^t\|\xi(\gamma(x,s))-\xi(\gamma(x_0,s))\|ds\\&\le\|x-x_0\|+\kappa^*\int_0^t\|\gamma(x,s)-\gamma(x_0,s)\|ds,\end{aligned}\tag{B.71}$$

where κ^* is given in (B.46). Using Gronwall's inequality, we have

$$\|\gamma(x,t)-\gamma(x_0,t)\|\le\|x-x_0\|e^{\kappa^*t}.$$

This is similar to (B.48). The rest of the proof is similar to the part below (B.48) in the proof of Fact 3, where we replace the application of Fact 2 by that of Fact 4. Details are omitted.

■

With all the above facts, now we are ready to complete the proof of Theorem 18.

Proof: Part (i) follows from Fact 1. Part (ii) is a consequence of Theorem 14 and Fact 5.

To show (iii), we only need to show that $\text{Ridge}(f) \subset R_a(f)$ when $k = 1$, because of (i). 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 $\text{Ridge}(f)$ is a union of finitely many 1-dimensional closed curves when $k = 1$ under our assumptions (see the discussion of the assumptions in Sec 3.1). We focus on one of the closed curves (also call it $\text{Ridge}(f)$), and parametrize it by $\zeta : [0, 1] \to \text{Ridge}(f)$ with $\zeta(0) = \zeta(1)$. Without loss of generality, we set $\bar{x} = \zeta(0)$, and assume that the total length of the ridge is 1 and the parametrization is by the arclength. Under our assumptions, $\text{Ridge}(f)$ is C^2-smooth, and has positive reach (Scholtes, 2013). Due to Fact 5, for a fixed (small) ϵ_0, when a and ϵ are small enough, there exist $x_1, x_2 \in \partial S_\epsilon$ such that

$$\begin{aligned}
&\lim_{\ell\to\infty} \gamma_a(x_1, \ell) \in \{\zeta(\alpha) :\ \alpha \in (0, \epsilon_0]\}, \\
&\lim_{\ell\to\infty} \gamma_a(x_2, \ell) \in \{\zeta(\alpha) :\ \alpha \in [1 - \epsilon_0, 1)\}.
\end{aligned}$$

In other words, the limit points of $\gamma_a(x_1, \ell)$ and $\gamma_a(x_2, \ell)$ are not 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 $\overline{x_1x_2}$ be the shortest curve (a connected set) in ∂S_ϵ connecting x_1 and x_2. When a is small enough, we have

$$\Big\{ \lim_{t\to\infty} \gamma(x, t) :\ x \in \overline{x_1x_2} \Big\} \subset \Big\{ \zeta(\alpha) :\ \alpha \in [0, 2\epsilon_0] \cup [1 - 2\epsilon_0, 1] \Big\}. \tag{B.72}$$

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

$$\gamma_a(\overline{x_1x_2}, \infty) := \Big\{ \lim_{\ell\to\infty} \gamma_a(x, \ell) :\ x \in \overline{x_1x_2} \Big\}$$

is also connected. Since $\gamma_a(\overline{x_1x_2}, \infty) \subset \text{Ridge}(f)$, we must have $\bar{x} \in \gamma_a(\overline{x_1x_2}, \infty)$, because otherwise we have

$$\{\zeta(\alpha) :\ \alpha \in [\alpha_1, \alpha_2]\} \subset \gamma_a(\overline{x_1x_2}, \infty),$$

which, however, contradicts (B.72) and Fact 5. Thus there exists $\bar{x}_0 \in \overline{x_1x_2}$ such that $\lim_{\ell\to\infty} \gamma_a(\bar{x}_0, \ell) = \bar{x}$, and we complete the proof.

■

Acknowledgments

We thank the anonymous referees and the Action Editor for insightful comments that helped significantly improve the quality of this paper. This work was partially supported by the US National Science Foundation (DMS 1821154) and resources provided by the Office of Research Computing at George Mason University (URL: `https://orc.gmu.edu`).

References

Ery Arias-Castro and Wanli Qiao. Clustering by hill-climbing: Consistency results. *The Annals of Statistics*, 2025+. To appear. Available as arXiv preprint arXiv:2202.09023.

Ery Arias-Castro, David L Donoho, and Xiaoming Huo. Adaptive multiscale detection of filamentary structures in a background of uniform random points. *The Annals of Statistics*, pages 326–349, 2006.

Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

Adam S Backer, Maurice Y Lee, and William E Moerner. Enhanced dna imaging using super-resolution microscopy and simultaneous single-molecule orientation measurements. *Optica*, 3(6):659–666, 2016.

Jeffrey D Banfield and Adrian E Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American statistical Association*, 87(417):7–16, 1992.

Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 437–478. Springer, 2012.

Gérard Biau and Aurélie Fischer. Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58(3):1924–1939, 2011.

José E Chacón and Tarn Duong. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, pages 499–532, 2013.

Yen-Chi Chen, Christopher R Genovese, and L Wasserman. Asymptotic theory for density ridges. *Annals of Statistics*, 43(5):1896–1928, 2015.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1197–1203. IEEE, 1999.

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

Sylvain Delattre and Aurélie Fischer. On principal curves with a length constraint. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 56, 2020.

David Eberly. *Ridges in Image and Data Analysis*, volume 7. Springer Science & Business Media, 1996.

Jochen Einbeck, Gerhard Tutz, and Ludger Evers. Local principal curves. *Statistics and Computing*, 15:301–313, 2005.

Jochen Einbeck, Ludger Evers, and Coryn Bailer-Jones. Representing complex data using localized principal components with application to astronomical data. In *Principal manifolds for data visualization and dimension reduction*, pages 178–201. Springer, 2008.

Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Reconstruction and interpolation of manifolds. i: The geometric whitney problem. *Foundations of Computational Mathematics*, 20(5):1035–1133, 2020.

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

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. On the path density of a gradient field. *The Annals of Statistics*, 37(6A):3236–3271, 2009.

Christopher R Genovese, Marco Perone Pacifico, Verdinelli Isabella, Larry Wasserman, et al. Minimax manifold estimation. *Journal of machine learning research*, 13:1263–1291, 2012a.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. The geometry of nonparametric filament estimation. *Journal of the American Statistical Association*, 107(498):788–799, 2012b.

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

Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

John M Holte. Discrete gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009.

Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

Balázs Kégl, Adam Krzyzak, Tamás Linder, and Kenneth Zeger. Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence*, 22(3): 281–297, 2000.

John M Lee. *Introduction To Smooth Manifolds*. Springer New York, 2013.

Liviu Nicolaescu. *An invitation to Morse theory*. Springer New York, 2011.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39:419–441, 2008.

Deborah Nolan and David Pollard. U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799, 1987.

Dmitri Novikov, Stéphane Colombi, and Olivier Doré. Skeleton as a probe of the cosmic web: the two-dimensional case. *Monthly Notices of the Royal Astronomical Society*, 366 (4):1201–1216, 2006.

Umut Ozertem and Deniz Erdogmus. Locally defined principal curves and surfaces. *The Journal of Machine Learning Research*, 12:1249–1286, 2011.

Seppo Pulkkinen. Ridge-based method for finding curvilinear structures from noisy data. *Computational Statistics & Data Analysis*, 82:89–109, 2015.

Wanli Qiao. Asymptotic confidence regions for density ridges. *Bernoulli*, 27(2):946–975, 2021.

Wanli Qiao. Confidence regions for filamentary structures. *Information and Inference: A Journal of the IMA*, 2025+. To appear. Available as arXiv preprint arXiv:2311.17831.

Wanli Qiao and Wolfgang Polonik. Theoretical analysis of nonparametric filament estimation. *The Annals of Statistics*, pages 1269–1297, 2016.

Wanli Qiao and Wolfgang Polonik. Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electron. J. Statist.*, 13(1):985–1030, 2019.

Alberto Rodríguez-Casal. Set estimation under convexity type assumptions. In *Annales de l'IHP Probabilités et statistiques*, volume 43, pages 763–774, 2007.

Sebastian Scholtes. On hypersurfaces of positive reach, alternating steiner formulae and hadwiger's problem. *arXiv preprint arXiv:1304.4179*, 2013.

Chelsea Scott, Rachel Adam, Ramon Arrowsmith, Christopher Madugo, Joseph Powell, John Ford, Brian Gray, Rich Koehler, Stephen Thompson, Alexandra Sarmiento, et al. Evaluating how well active fault mapping predicts earthquake surface-rupture locations. *Geosphere*, 19(4):1128–1156, 2023.

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

Bharath Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In *Artificial Intelligence and Statistics*, pages 1090–1098. PMLR, 2012.

Derek C Stanford and Adrian E Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):601–609, 2000.

Radu S Stoica, Vicent J Martínez, and Enn Saar. A three-dimensional object point process for detection of cosmic filaments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 56(4):459–477, 2007.

Andrzej Szymczak, Arthur Stillman, Allen Tannenbaum, and Konstantin Mischaikow. Coronary vessel trees from 3d imagery: a topological approach. *Medical image analysis*, 10(4):548–559, 2006.

Robert Tibshirani. Principal curves revisited. *Statistics and computing*, 2:183–190, 1992.

Dag Tjøstheim, Martin Jullum, and Anders Løland. Statistical embedding: Beyond principal components. *Statistical Science*, 1(1):1–29, 2023.

Aad van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, 1996.

Jakob J Verbeek, Nikos Vlassis, and B Kröse. A k-segments algorithm for finding principal curves. *Pattern Recognition Letters*, 23(8):1009–1017, 2002.

Suyi Wang, Yusu Wang, and Yanjie Li. Efficient map reconstruction and augmentation via topological methods. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2015.

Stephen Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer New York, 2003.

Zhigang Yao, Jiaji Su, and Bingjie Li. Manifold fitting: An invitation to statistics. *arXiv preprint arXiv:2304.07680*, 2023.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yikun Zhang and Yen-Chi Chen. Linear convergence of the subspace constrained mean shift algorithm: from euclidean to directional data. *Information and Inference: A Journal of the IMA*, 12(1):210–311, 2023.