J. Appl. Prob. **55,** 1001–1013 (2018)
*doi:*10.1017/jpr.2018.66
© *Applied Probability Trust* 2018

CONVERGENCE RATES FOR ESTIMATORS OF GEODESIC DISTANCES AND FRÈCHET EXPECTATIONS

CATHERINE AARON,* *Université Clermont Auvergne*

OLIVIER BODART,** *Université Jean Monnet*

Abstract

Consider a sample $\mathcal{X}_n = \{X_1, \dots, X_n\}$ of independent and identically distributed variables drawn with a probability distribution $\mathbb{P}_X$ supported on a compact set $M \subset \mathbb{R}^d$. In this paper we mainly deal with the study of a natural estimator for the geodesic distance on M. Under rather general geometric assumptions on M, we prove a general convergence result. Assuming M to be a compact manifold of known dimension $d' \leq d$, and under regularity assumptions on $\mathbb{P}_X$, we give an explicit convergence rate. In the case when M has no boundary, knowledge of the dimension d' is not needed to obtain this convergence rate. The second part of the work consists in building an estimator for the Fréchet expectations on M, and proving its convergence under regularity conditions, applying the previous results.

Keywords: Geometric inference; geodesic distance; statistics on manifolds; Fréchet expectations

2010 Mathematics Subject Classification: Primary 62-07
Secondary 62G05; 62G20; 62H99

1. Introduction

Let $\mathbb{P}_X$ be a probability distribution supported on a compact set $M \subset \mathbb{R}^d$, $d \geq 2$; that is, M is the smallest closed set in $\mathbb{R}^d$ of probability 1. Let $\mathcal{X}_n = \{X_1, \dots, X_n\}$ be a sample of independent and identically distributed (i.i.d) variables drawn on M with the distribution $\mathbb{P}_X$. Our first aim is the study of a rather classical estimator of the geodesic distance on the unknown set M.

The way to build this estimator is quite intuitive (see, for example, [18]): given $r > 0$, build a graph interconnecting all the pairs (X_i, X_j) of the sample $\mathcal{X}_n$ such that $\|X_i - X_j\| \leq r$. The geodesic distance between any two points X_k and X_l of the sample is then estimated by the length of the shortest path connecting X_k and X_l in the graph (see Definition 2.1 for details). This path (and its length) can be computed with optimal complexity by using Dijkstra's algorithm (see, for example, [3] for a presentation of this algorithm). As usual in such problems, $r = r_n$ must be a conveniently chosen sequence. First, it must converge to 0 as $n \to \infty$. Moreover, this convergence has to be slow enough for the path realizing the estimator to be smooth enough.

Received 20 July 2017; revision received 9 November 2018.

* Postal address: Laboratoire de Mathématiques Blaise Pascal (UMR 6620 - CNRS), Université Clermont Auvergne (Clermont-Ferrand 2), 63177 Aubière cedex, France. Email address: catherine.aaron@uca.fr

** Postal address: Institut Camille Jordan Faculté des Sciences et Techniques, Université Jean Monnet, 42023 Saint-Étienne Cedex 2, France.

To the best of the authors' knowledge, the asymptotic behavior of such an estimator has not been studied yet. In a similar framework, the estimator

$$\widehat{L}_p(X_i, X_j) = \min\biggl\{\sum_{k=1}^{K} \|X_{i_{k+1}} - X_{i_k}\|^p,\ i_1 = i,\ i_K = j\biggr\}, \qquad p \geq 1,$$

was studied in [8], generalizing the results in [7] when M is a compact manifold without boundary. The minimum was computed on all the paths in the graph connecting X_i to X_j. The estimator studied in this paper roughly reads as follows (see Definition 2.1):

$$\min\biggl\{\sum_{k=1}^{K} \|X_{i_{k+1}} - X_{i_k}\|,\ i_1 = i,\ i_K = j \text{for all } k \text{ such that } \|X_{i_{k+1}} - X_{i_k}\| \leq r_n\biggr\},$$

where (r_n) is a conveniently chosen sequence converging to 0 as $n \to \infty$. Here the minimum is then computed over the paths whose vertices satisfy a proximity criterion.

As mentioned earlier, $\widehat{L}_p$ is computed on the whole graph. We have $\widehat{L}_1(X_i, X_j) = \|X_i - X_j\|$, whereas a power $p > 1$ tends to 'select' a path in the graph which is 'close to' the manifold. When M is a d'-manifold without boundary, and the probability distribution $\mathbb{P}_X$ has a density f_X, Hwang *et al.* [8] proved that, for $p > 1$,

$$\biggl|\frac{\widehat{L}_p(X_i, X_j)}{C_{d',p}\, n^{(1-p)/d'}} - L_p(X_i, X_j)\biggr| \overset{\text{a.s.}}{\to} 0, \qquad \text{for all } (i, j),\ n \to \infty,$$

where $C_{d',p}$ is a positive constant and $L_p(x, y)$ is the geodesic distance on M endowed with the metric $f_X^{2(1-p)/d'} I_{d'}$ (where $I_{d'}$ is the identity matrix of size d'). Thus, the estimator $\widehat{L}_p$ can only estimate the canonical geodesic distance (that is, M endowed with the identity) when observations are uniformly drawn, while our estimator does not requires such an hypothesis. Moreover, we obtain convergence rates while none are provided for $\widehat{L}$.

We will show, under quite general assumptions on the support M, that choosing $r_n = d_h(\mathcal{X}_n, M)^{2/3}$ appears to be convenient (Theorem 2.1). Here and throughout the paper, $d_h(A, B)$ denotes the Hausdorff distance between the sets A and B; that is,

$$d_h(A, B) = \max\biggl\{\sup_{a\in A}\Bigl(\inf_{b\in B} \|a - b\|\Bigr), \sup_{b\in B}\Bigl(\inf_{a\in A} \|a - b\|\Bigr)\biggr\}.$$

Assuming that M is a d'-manifold, $d' \leq d$, and assuming some regularity for the distribution $\mathbb{P}_X$, we show that $d_h(\mathcal{X}_n, M) = \mathcal{O}(\ln n/n)^{1/d'}$ everywhere almost surely (e.a.s.), allowing us to find the convergence rate of our estimator when the dimension d' is known (Corollary 2.1). When d' is unknown, and M is supposed to have no boundary, Corollary 2.2 contains an estimator of r_n which allows us to obtain the same convergence rate.

Eventually, we will apply these results to the estimation of the Fréchet expectations, as defined in [13], of the distribution $\mathbb{P}_X$ on M (Theorem 2.2).

Using the estimated geodesic distance in place of the Euclidean distance has become frequent in different fields of application in order to take the nonlinearity of the data into account. In [18], the authors proposed to apply the multidimensional scaling (see, for example, [9]) to the array of geodesic distances between points. This idea opened the way to the use of the geodesic distance in dimension reduction (see [5], [10]–[12], and [16]). In [2] and [6], the question of intrinsic dimension estimation using graph-based statistics was studied. In particular, in [6] the authors proposed a generalization of the correlation dimension where the Euclidean distance is

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

replaced by the (estimated) geodesic distance. This approach has the advantage that it is less sensitive to the (difficult) question of the choice of parameter (see also [17]). In [13], the author raised the question of the generalization of classical statistical quantities (such as the mean and median) to the case of data supported on Riemannian manifolds.

The paper is organized as follows. In Section 2 we state the general framework, main definitions, and results. In Section 2.1 we present the results concerning the estimation of the geodesic distance on the support M (Theorem 2.1 and Corollaries 2.1 and 2.2), while in Section 2.2 we present the theorem for the Fréchet expectations estimator (Theorem 2.2). Section 3 is devoted to the proofs of the results.

2. General framework and main results

2.1. Estimating geodesic distances

Let us first start with the definition of our estimator.

Definition 2.1. Let $\mathcal{X}_n = \{X_1, \dots X_n\}$ be a set of n i.i.d. random variables with distribution $\mathbb{P}_X$ supported on a compact set $M \subset \mathbb{R}^d$, $d \geq 2$. With $r_n > 0$ being a given number, let $\mathcal{G}_{r_n}(\mathcal{X}_n)$ be the graph whose edges are the segments $[X_i, X_j]$ such that $\|X_i - X_j\| \leq r_n$.

For $(i, j) \in \{1, \dots, n\}^2$, let, if it exists, $\hat{\gamma}_{r_n}(X_i, X_j)$ be the shortest path (in the Euclidean norm) connecting X_i and X_j in $\mathcal{G}_{r_n}(\mathcal{X}_n)$, and $|\hat{\gamma}_{r_n}(X_i, X_j)|$ its length.

We aim to prove, for a class of convenient compact sets in $\mathbb{R}^d$, that $|\hat{\gamma}_{r_n}(X_i, X_j)|$ is an estimator of the geodesic distance $\gamma(X_i, X_j)$ on M, with good convergence properties.

Definition 2.2. Let $M \subset \mathbb{R}^d$ be a compact set, M is said to be K_M-geodesically smooth (GS) for some positive number K_M if:

(i) for all $(x, y) \in M^2$, there exists a geodesic path $\gamma_{x\to y}$ of class $\mathcal{C}^1$ that links x to y;

(ii) there exists a real function β such that $\lim_{t\to 0} \beta(t) = 0$ and, for all $(x, y) \in M^2$, $|\gamma_{x\to y}| \leq \beta(\|x - y\|)$;

(iii) let $\Gamma_{x\to y}\colon [0, |\gamma_{x\to y}|] \to \mathbb{R}^d$ be the parametrization of $\gamma_{x\to y}$ such that $\Gamma_{x\to y}(s)$ is the point of $\gamma_{x\to y}$ that is at a (curvilinear) distance s from x (along the geodesic curve). For all $(x, y) \in M^2$, the gradient of $\Gamma_{x\to y}$, denoted $\dot{\Gamma}_{x\to y}$, is K_M-Lipschitz continuous.

A compact manifold of class $\mathcal{C}^2$ with no boundary satisfies the assumptions of Definition 2.2, but we can build more general examples of such sets (that is, compact sets with $\mathcal{C}^1$ geodesic curves which have K_M-Lipschitz tangent maps). As an example, in Figure 1 we present two examples of GS-sets (sets 1 and 2), and one which is not. Note that the middle panel in Figure 1, while satisfying the GS property, is not a manifold.

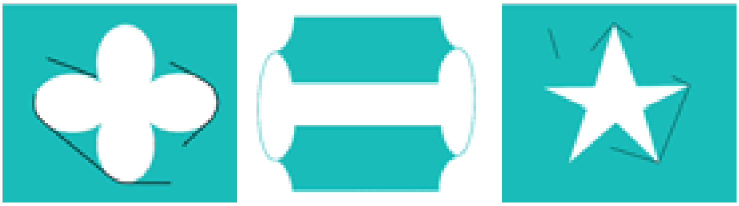

Figure 1: The sets are the shaded areas. *Left:* this is GS (with some geodesic curves depicted). *Middle:* this is also GS (but is not a manifold). *Right:* this is not GS: some geodesic curves are not smooth enough.

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

Theorem 2.1. *Let $\hat{\gamma}_{r_n}$ be the estimator introduced in Definition 2.1. Assume that there exists a sequence $\rho_n \overset{\text{a.s.}}{\to} 0$ such that $\rho_n \geq d_h(\mathcal{X}_n, M)$ (e.a.s), and let (r_n) be a sequence such that $r_n > 2\rho_n$ and $\rho_n/r_n \overset{\text{a.s.}}{\to} 0$. Then*

$$\max_{i,j} ||\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|| = \mathcal{O}\Bigl(\max\Bigl(r_n, \frac{\rho_n^2}{r_n^2}\Bigr)\Bigr) \quad \textit{e.a.s.} \tag{2.1}$$

Assuming that $r_n > 2\rho_n$ ensures the existence of $|\hat{\gamma}_{r_n}(X_i, X_j)|$ for all i and j. The first part of the proof will clearly illustrate this fact (see Section 3.1).

We can then assume that the sequence $r_n = d_h(\mathcal{X}_n, M)^{2/3}$ is an optimal choice. However, even though it is known that $d_h(\mathcal{X}_n, M) \overset{\text{a.s.}}{\to} 0$ (see [4]), the rate of this convergence is unknown in general. Thus, in order to obtain a convergence rate for our estimator, we need to make extra assumptions on the set M and the probability distribution $\mathbb{P}_X$.

Definition 2.3. Let $\delta > 0$. A probability measure $\mathbb{P}_X$ supported on $M \subset \mathbb{R}^d$ is said to be δ-standard with respect to a measure μ if there exists $\lambda > 0$ such that $\mathbb{P}_X(\mathcal{B}(x, \varepsilon)) \geq \delta\mu(\mathcal{B}(x, \varepsilon))$ for all $x \in M$ and $\varepsilon \in]0, \lambda]$.

We then have the following result.

Corollary 2.1. *Let $M \subset \mathbb{R}^d$, $d \geq 2$, be a d'-dimensional compact manifold of class $\mathcal{C}^1$ satisfying the GS property for some number $K_M > 0$. Let $\mathbb{P}_X$ be a probability distribution on M. Assume, for some number $\delta > 0$, that $\mathbb{P}_X$ is δ-standard with respect to the measure induced on M by the Lebesgue measure in $\mathbb{R}^d$.*

If the sequence (r_n) in Definition 2.1 is such that

$$\Bigl(A_0 \frac{\ln n}{n}\Bigr)^{2/3d'} \leq r_n \leq \Bigl(A_1 \frac{\ln n}{n}\Bigr)^{2/3d'},$$

with $A_0 > 0$ and $A_1 > 0$, then

$$\max_{i,j} ||\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|| = \mathcal{O}\Bigl(\Bigl(\frac{\ln n}{n}\Bigr)^{2/3d'}\Bigr) \quad \textit{e.a.s.}$$

As usual when dealing with estimation problems, the sequence of radii (r_n) in the previous theorem remains abstract. In particular, the dimension d' of the support is generally unknown. However, making extra assumptions on the support M and the density of the distribution, we can accurately estimate the sequence of radii, with no need for estimating d'. This last fact is indeed worth emphasizing, since knowledge of the geodesic distance is known to be useful for a good estimation of the dimension of a manifold (see, for example, [6]).

Let $L_n = \max_i(\min_{j\neq i} \|X_i - X_j\|)$ and let θ_n be the longest edge of the minimal spanning tree of the sample. Up to a rescaling of the data, we can suppose that $\max_i(\max_j ||X_i - X_j||) \leq 1$. Then we have $L_n = \max_i(\min_{j\neq i} \|X_i - X_j\|) \leq 1$ and $\theta_n \leq 1$; hence, $L_n^{2/3} \geq L_n$ and $\theta_n^{2/3} \geq \theta_n$.

Choosing a sequence of radii satisfying $r_n \geq \theta_n$ ensures the existence of the estimator $|\hat{\gamma}_{r_n}(X_i, X_j)|$. Conjecturing that the results of [14] can be generalized to the case of data drawn on a smooth manifold with a density close to the uniform one leads to the choice of $r_n = c.\theta_n^{2/3}$ with $c \geq 1$. If the conjecture is correct, this would guarantee the existence of the estimator and provide optimal convergence rates. More practically, in order to prove a theoretical result, we are led to choosing r_n in relation to L_n. This only ensures the existence of our estimator asymptotically.

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

Corollary 2.2. *Let $M \in \mathbb{R}^d$, $d \geq 2$, be a d'-dimensional compact manifold, $d' < d$ of class $\mathcal{C}^2$ with no boundary, and $\mathbb{P}_X$ be a probability distribution on M with continuous probability density f_X bounded from below on M by a positive constant f_0. Then, for any $c > 0$, setting $r_n = c(\max_i(\min_{j \neq i} \|X_i - X_j\|))^{2/3}$ in Definition 2.1, we have*

$$\max_{i,j} ||\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|| = \mathcal{O}\bigg(\bigg(\frac{\ln n}{n}\bigg)^{2/3d'}\bigg) \qquad \textit{e.a.s.}$$

The assumptions of this corollary imply those of Theorem 2.1; they allow us to explicitly build a convenient sequence of radii (r_n) only from the sample. To prove Theorem 2.1 we use a result due to Penrose (see [15]) which applies only in the case when M has no boundary. However, numerical simulations on $\mathcal{C}^2$ sets with a boundary satisfying the GS assumption lead us to think that the result is also true for such sets.

The question of the choice of the sequence (r_n) remains a difficult subject. In our framework, we propose the following decision rule: first, in the absence of *a priori* knowledge on the data, and when the support M can have several arcwise connected components (that is, data classes), we choose r_n of order $L_n^{2/3}$. This sequence of radii will converge to 0 and allow us to identify the different classes in the data with optimal convergence rate (although the existence of the estimator is only ensured asymptotically). If we know *a priori* that the support is arcwise connected, choosing $r_n = c.\theta_n^{2/3}$ with $c \geq 1$ may be a convenient choice, even if the asymptotic properties of the estimator are conjecture-based.

2.2. Estimating Fréchet expectations

In this section we assume the set M to be a compact d'-manifold of class $\mathcal{C}^2$. Following the ideas of Pennec (see [13]), we consider the Fréchet expectations of the random variable X (which distribution is supported on M), that is,

$$\mathbb{E}_k^{\mathrm{Fr}}(X) = \arg\min_{x \in M} \mathbb{E}(|\gamma_{x \to X}|^k), \qquad k \in \mathbb{N}^*, \tag{2.2}$$

which are generalizations of the expected value for $k = 2$ and of the median (or depth) for $k = 1$. As pointed out in [13], these expectations are not necessarily unique. For example, if M is a sphere and $\mathbb{P}_X$ the uniform distribution, then obviously all the points of M realize the minimum in (2.2) (for any $k \geq 1$).

To avoid dealing with such situations, we are going to make the following assumption, considering that k is fixed:

(A) $\Phi(x) = \mathbb{E}(|\gamma_{x \to X}|^k)$ admits a unique minimum $x^* \in M$, Φ is of class $\mathcal{C}^2$ in a neighbourhood of x^*, and $H_\Phi(x^*)$ is positive definite,

where H_Φ denotes the hessian matrix of Φ (that is, $(H_\Phi)_{i,j} = \partial^2\Phi/(\partial x_i, \partial x_j)$).

Remark 2.1. We must note that Φ is a continuous function on M. Indeed, the triangle and Minkowski inequalities yield $|\Phi(x)^{1/k} - \Phi(y)^{1/k}| \leq |\gamma_{x \to y}|$ for any $(x, y) \in M^2$. The extra (local) regularity in assumption(A) is required for the sake of simplicity, allowing us to apply basic differential calculus results at the optimal point x^*.

The first part of this assumption is very strong, but the second part is not. For example, when $d' = 1$ and M is homeomorphic to a segment, explicit computations show that (A) holds for $k = 1$ if and only if $f_X(x^*) \neq 0$. For $k = 2$, when M is a bounded closed convex set of dimension d, the geodesic distance on M coincides with the Euclidean distance, the expectation $\mathbb{E}(X)$ lies in M, it minimizes the function $\Phi(x)$, and assumption (A) is satisfied

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

(with $H_\Phi \equiv 2I_d$). This leads us to think that, for $k = 2$, this condition is general enough and may hold for a wide class of regular submanifolds of $\mathbb{R}^d$.

In this section we study the behavior of the natural estimator of $\mathbb{E}_k^{\mathrm{Fr}}(X)$, that is,

$$\hat{\mathbb{E}}_{k,r_n}^{\mathrm{Fr}}(\mathcal{X}_n) = \arg\min_{X_i \in M} \frac{1}{n} \sum_j |\hat{\gamma}_{r_n}(X_i, X_j)|^k. \tag{2.3}$$

Theorem 2.2. *Assume that $M \subset \mathbb{R}^d$, $d \geq 2$, is a d'-dimensional manifold, $d' < d$ of class $\mathcal{C}^2$ with no boundary, and that $\mathbb{P}_X$ is a probability distribution on M with a continuous and bounded from below probability density f_X. Moreover, suppose that assumption (A) holds. Then, choosing $r_n = c(\max_i(\min_j \|X_i - X_j\|))^{2/3}$ in the definition of $\hat{\gamma}_{r_n}$, we have*

$$|\mathbb{E}_k^{Fr}(X) - \hat{\mathbb{E}}_{k,r_n}^{Fr}(\mathcal{X}_n)| = \mathcal{O}\left(\left(\frac{\ln n}{n}\right)^{\min(1/4, 1/3d')}\right) \quad \textit{e.a.s.}$$

3. Proofs of the results

Let us start with a result which is a direct consequence of the regularity of the set considered here.

Proposition 3.1. *If $M \subset \mathbb{R}^d$ is K_M-geodesically smooth then there exist $r_M > 0$ and $A_M > 0$, depending only on M, such that*

$$\|x - y\| \leq r_M \quad \Longrightarrow \quad |\gamma_{x\to y}| \leq \|x - y\| + A_M \|x - y\|^2 \quad \textit{for all } (x, y) \in M^2.$$

Proof. Let $(x, y) \in M^2$. Consider the parametrization $\Gamma_{x\to y}$ of the geodesic curve $\gamma_{x\to y}$ as in Definition 2.2. The map $\overset{\bullet}{\Gamma}$ being K_M-Lipschitz continuous, for all $t_0 \in [0, |\gamma_{x\to y}|]$, there exists $\varepsilon_{t_0} \colon [0, |\gamma_{x\to y}|] \to \mathbb{R}^d$ such that

$$\overset{\bullet}{\Gamma}(t) = \overset{\bullet}{\Gamma}(t_0) + K_M |t - t_0| \varepsilon_{t_0}(t), \qquad \|\varepsilon_{t_0}(t)\| \leq 1 \quad \text{for all } t \in [0, |\gamma_{x\to y}|].$$

Thus,

$$\int_0^{|\gamma_{x\to y}|} \overset{\bullet}{\Gamma}(t)\,\mathrm{d}t = \int_0^{|\gamma_{x\to y}|} (\overset{\bullet}{\Gamma}(t_0) + K_M |t - t_0| \varepsilon_{t_0}(t))\,\mathrm{d}t;$$

that is,

$$y - x = \overset{\bullet}{\Gamma}(t_0)|\gamma_{x\to y}| + K_M \int_0^{|\gamma_{x\to y}|} |t - t_0| \varepsilon_{t_0}(t)\,\mathrm{d}t.$$

Choosing $t_0 = \frac{1}{2}|\gamma_{x\to y}|$, and noting that with the chosen parametrization we have $\|\overset{\bullet}{\Gamma}(t_0)\| = 1$, we obtain

$$\|x - y\| \geq |\gamma_{x\to y}| - \tfrac{1}{4} K_M |\gamma_{x\to y}|^2 \quad \text{for all } (x, y) \in M^2.$$

Now assuming that $\|x - y\| \leq K_M^{-1}$, the following alternatives hold:

(i) either $|\gamma_{x\to y}| \geq (2 + 2\sqrt{1 - K_M \|x - y\|})/K_M$,

(ii) or $|\gamma_{x\to y}| \leq (2 - 2\sqrt{1 - K_M \|x - y\|})/K_M$.

For $\|x - y\|$ small enough, the case (i) is impossible due to Definition 2.2(ii). Therefore, there exists $r_M \leq K_M^{-1}$ such that, for $\|x - y\| \leq r_M$, the alternative, case (ii), holds. Making a Taylor expansion of $\|x - y\|$ completes the proof. □

3.1. Proof of Theorem 2.1

Let $(i, j) \in \{1, \dots n\}^2$, $i \neq j$, and let γ_{ij} be the geodesic curve between X_i and X_j. Consider a partition $\{x_0, \dots, x_K\}$ of γ_{ij} such that

$$x_0 = X_i, \qquad x_K = X_j, \tag{3.1}$$

$$K = \left\lceil \frac{|\gamma_{X_i \to X_j}|}{r_n - 2\rho_n} \right\rceil, \tag{3.2}$$

$$|\gamma_{x_k \to x_{k+1}}| = \frac{|\gamma_{X_i \to X_j}|}{K}, \tag{3.3}$$

so that

$$|\gamma_{x_k \to x_{k+1}}| = r_n - 2\rho_n, \quad k = 0, \dots, K-2, \qquad |\gamma_{x_{K-1} \to x_K}| < r_n - 2\rho_n. \tag{3.4}$$

We have

$$|\gamma_{X_i \to X_j}| = \sum_{k=0}^{K-1} |\gamma_{x_k \to x_{k+1}}| \geq \sum_{k=0}^{K-1} \|x_k - x_{k+1}\|. \tag{3.5}$$

From the definition of ρ_n, for any $k \in \{0, \dots, K\}$, there exists $i_k \in \{1, \dots, n\}$ such that $\|X_{i_k} - x_k\| \leq \rho_n$. For the sake of simplicity, denote

$$Y_k = X_{i_k}, \qquad \varepsilon_k = Y_k - x_k, \qquad U_k = \frac{x_k - x_{k+1}}{\|x_k - x_{k+1}\|}.$$

Recall that

$$\|\varepsilon_k\| \leq \rho_n, \qquad k = 0, \dots, K-1. \tag{3.6}$$

For $k \in \{0, \dots, K-1\}$,

$$\begin{aligned}\|Y_k - Y_{k+1}\|^2 &= \|\varepsilon_k + (x_k - x_{k+1}) - \varepsilon_{k+1}\|^2 \\ &= \|x_k - x_{k+1}\|^2 + 2\langle x_k - x_{k+1} \mid \varepsilon_k - \varepsilon_{k+1}\rangle + \|\varepsilon_k - \varepsilon_{k+1}\|^2 \\ &= \|x_k - x_{k+1}\|^2 \times \left(1 + 2\frac{\langle U_k \mid \varepsilon_k - \varepsilon_{k+1}\rangle}{\|x_k - x_{k+1}\|} + \frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|^2}\right);\end{aligned}$$

that is, taking the square root of this equality, and noting that $\sqrt{1+t} \leq 1 + \frac{1}{2}t$, $t \geq -1$,

$$\begin{aligned}\|Y_k - Y_{k+1}\| &\leq \|x_k - x_{k+1}\| \times \left(1 + \frac{\langle U_k \mid \varepsilon_k - \varepsilon_{k+1}\rangle}{\|x_k - x_{k+1}\|} + \frac{1}{2}\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|^2}\right) \\ &\leq \|x_k - x_{k+1}\| + \langle U_k \mid \varepsilon_k - \varepsilon_{k+1}\rangle + \frac{1}{2}\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|}.\end{aligned}$$

In view of (3.1)–(3.4), the length of the last segment, that is, $\|x_{K-1} - x_K\|$, is not bounded from below, hence, we shall treat the cases $k < K-1$ and $k = K-1$ separately. From (3.5), we have

$$|\gamma_{X_i \to X_j}| \geq \|x_{K-1} - x_K\| + \sum_{k=0}^{K-2} \|Y_k - Y_{k+1}\| - \frac{1}{2}S_1 - S_2, \tag{3.7}$$

with

$$S_1 = \sum_{k=0}^{K-2} \frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|}, \qquad S_2 = \sum_{k=0}^{K-2} \langle U_k \mid \varepsilon_k - \varepsilon_{k+1}\rangle. \tag{3.8}$$

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

We first study S_1. From (3.4) and Proposition 3.1, we have for $k \in \{0, \dots, K-2\}$,

$$r_n - 2\rho_n - A_M(r_n - 2\rho_n)^2 \le \|x_k - x_{k+1}\| \le r_n - 2\rho_n \le r_n, \tag{3.9}$$

with $A_M > 0$ depending only on M.

Then, for n large enough to have $u_n = 2\rho_n/r_n + A_M((r_n - 2\rho_n)^2/r_n) < 1$ and applying the fact that $1/(1-u) \le 1+u$ when $u \in [0, 1[$, we have, for all $k \in \{0, \dots, K-2\}$,

$$\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|} \le \frac{4\rho_n^2}{r_n - 2\rho_n - A_M(r_n - 2\rho_n)^2} \le \frac{4\rho_n^2}{r_n}\left(1 + \frac{2\rho_n}{r_n} + A_M\frac{(r_n - 2\rho_n)^2}{r_n}\right). \tag{3.10}$$

Thus,

$$\frac{\|\varepsilon_k - \varepsilon_{k+1}\|^2}{\|x_k - x_{k+1}\|} \le \frac{4\rho_n^2}{r_n}(1 + o(1)).$$

The definition of ρ_n implies that $r_n - 2\rho_n \sim r_n$. Moreover, since the set M is compact and satisfies the GS assumption, $\gamma_{X_i \to X_j}$ is uniformly bounded for all $(i, j) \in \{1, \dots, n\}^2$. Hence, there exists a constant $L_M > 0$ such that

$$0 < K \le \frac{L_M}{r_n}, \tag{3.11}$$

where K is defined by (3.2), and we have

$$S_1 \le L_M\left(\frac{4\rho_n^2}{r_n^2} + o\left(\frac{\rho_n^2}{r_n^2}\right)\right). \tag{3.12}$$

Note that the bound in (3.10) is uniform in (i, j); therefore, the same holds for (3.12).

Now, since the set M is K_M-geodesically smooth we obtain, reasoning as in the beginning of the proof of Proposition 3.1,

$$\|\,\|x_{k+1} - x_k\|(U_k - (r_n - 2\rho_n))\dot{\Gamma}_{x_k \to x_{k+1}}(0)\| \le \tfrac{1}{2}K_M(r_n - 2\rho_n)^2.$$

Thus, applying Proposition 3.1 for $k \in \{0, \dots, K-2\}$, we have $U_k = \dot{\Gamma}_{x_k \to x_{k+1}}(0) + Z_k$ with $\|Z_k\| \le (A_M + \frac{1}{2}K_M)(r_n - 2\rho_n)$ uniformly in (i, j). Now, noting that $\dot{\Gamma}_{x_k \to x_{k+1}}(0) = \dot{\Gamma}_{x_0 \to x_K}(k(r_n - 2\rho_n))$ and due to the Lipshitz continuity of $\dot{\Gamma}$, we have

$$\|U_k - U_{k+1}\| \le \big(A_M + \tfrac{3}{2}K_M\big)(r_n - 2\rho_n). \tag{3.13}$$

We can now write S_2 as

$$S_2 = \sum_{k=1}^{K-2}\langle U_k - U_{k-1} \mid \varepsilon_k\rangle + \langle U_0 \mid \varepsilon_0\rangle - \langle U_{K-2} \mid \varepsilon_{K-1}\rangle;$$

hence, in view of (3.6), (3.9), (3.11), and (3.13) we have

$$|S_2| \le \big(L_M\big(A_M + \tfrac{3}{2}K_M\big) + 2\big)\rho_n.$$

Combining this last inequality with (3.7), (3.8), and (3.12), we obtain the existence of explicit constants $B_M > 0$ and $C_M > 0$ depending only on M such that, for large enough n and for all (i, j),

$$|\gamma_{X_i \to X_j}| \ge \sum_{k=0}^{K-2}\|Y_k - Y_{k+1}\| - B_M\rho_n - C_M\frac{\rho_n^2}{r_n^2}.$$

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

Thus,

$$|\gamma_{X_i\to X_j}| \geq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\| - \|Y_{K-1} - Y_K\| - B_M\rho_n - C_M\frac{\rho_n^2}{r_n^2}.$$

Recall that, for all $k \in \{0, \dots, K-1\}$, we have $\|Y_k - Y_{k+1}\| = \|(x_k - x_{k+1}) - (\varepsilon_k - \varepsilon_{k+1})\|$; hence, the triangle inequality, (3.4), and (3.6) yield

$$\|Y_k - Y_{k+1}\| \leq r_n,\ k \in \{0, \dots, K-1\}. \tag{3.14}$$

Applying (3.14) for $k = K - 1$, we first obtain

$$|\gamma_{X_i\to X_j}| \geq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\| - r_n - B_M\rho_n - C_M\frac{\rho_n^2}{r_n^2}. \tag{3.15}$$

From (3.14), the path $Y_0, \dots, Y_K$ belongs to the graph $\mathcal{G}_{r_n}(\mathcal{X}_n)$, so we clearly have

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \leq \sum_{k=0}^{K-1} \|Y_k - Y_{k+1}\|;$$

therefore, since $\rho_n = o(r_n)$ and in view of (3.15), we have, for large enough n,

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \leq |\gamma_{X_i\to X_j}| + r_n + B_M\rho_n + C_M\frac{\rho_n^2}{r_n^2}. \tag{3.16}$$

We now prove the following inequality:

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \geq |\gamma_{X_i\to X_j}| - 2A_M L_M r_n. \tag{3.17}$$

For the sake of clarity, we omit the superscripts in Definition 2.1 and denote $Z_0 = X_i, Z_1, \dots, Z_{L_1}, Z_L = X_j$ the nodes of the graph $\mathcal{G}_n^{i,j}$ realizing the path $\widehat{\gamma}_{r_n}(X_i, X_j)$. Proposition 3.1 yields

$$|\gamma_{X_i\to X_j}| \leq \sum_{k=0}^{L-1} |\gamma_{Z_k\to Z_{k+1}}| \leq \sum_{k=0}^{L-1}(\|Z_k - Z_{k+1}\| + A_M\|Z_k - Z_{k+1}\|^2).$$

Noting, from Definition 2.1, that $\|Z_k - Z_{k+1}\| \leq r_n$, we obtain

$$|\widehat{\gamma}_{r_n}(X_i, X_j)| \geq |\gamma_{X_i\to X_j}| - A_M L r_n^2. \tag{3.18}$$

We now obtain a bound for the number of nodes L in the path $\widehat{\gamma}_{r_n}(X_i, X_j)$. Necessarily, by construction, we have

$$\|Z_k - Z_{k+1}\| + \|Z_{k+1} - Z_{k+2}\| > r_n, \qquad k = 0, \dots, L-2. \tag{3.19}$$

Indeed, if this was not the case, we would have $\|Z_k - Z_{k+2}\| \leq r_n$; hence, the path $\{Z_k, Z_{k+2}\}$ would be shorter in the graph $\mathcal{G}_n^{i,j}$ than the path $\{Z_k, Z_{k+1}, Z_{k+2}\}$ which is impossible. Therefore, summing up (3.19) for $k \in \{0, \dots, L-2\}$, we obtain

$$Lr_n^2 \leq 2\,|\widehat{\gamma}_{r_n}(X_i, X_j)| + r_n^2;$$

hence, in view of (3.16), (3.18), and recalling that $|\gamma_{X_i\to X_j}|$ is uniformly bounded, we obtain (3.17).

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

This inequality and (3.16) finally imply that

$$||\widehat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i \to X_j}|| \leq D_M \max\left(\frac{\rho_n^2}{r_n^2}, r_n\right),$$

where the constant $D_M > 0$ only depends on the manifold M. This yields the estimate (2.1) and concludes the proof.

3.2. Proof of Corollary 2.1

Reasoning as in [1], since M is of class $\mathcal{C}^1$, we can cover M with $\nu_n \leq C\, n$ (with $C > 0$) deterministic balls of radius $\varepsilon_n = (1/n)^{1/d'}$ with centers $x_i \in M$, $i \in \{1, \dots, \nu_n\}$. Let $\omega_{d'}$ be the volume of the d'-dimensional unit ball. Recall that $\mathbb{P}_X$ is δ-standard with respect to the d'-dimensional measure. We then classically have, for all $a > 0$,

$$\begin{aligned}\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'}\right) \\ = \mathbb{P}_X\left(\text{there exists } x \in M;\ \mathcal{B}\left(x, \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'}\right) \cap \mathcal{X}_n = \varnothing\right).\end{aligned}$$

The triangle inequality thus implies

$$\begin{aligned}\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'}\right) \\ \leq \mathbb{P}_X\left(\text{there exists } i;\ \mathcal{B}\left(x_i, \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'} - \varepsilon_n\right) \cap \mathcal{X}_n = \varnothing\right).\end{aligned}$$

Since $\mathbb{P}_X$ is standard, we have, for large enough n, $((2a/\delta\omega_{d'})(\ln n/n))^{1/d'} < \lambda$. Thus,

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'}\right) \leq \nu_n\left(1 - \delta\omega_{d'}\left(\left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'} - \varepsilon_n\right)^{d'}\right)^n.$$

A Taylor expansion of the right-hand side of the above inequality yields

$$\mathbb{P}_X\left(d_h(\mathcal{X}_n, M) \geq \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'}\right) \leq Cn^{1-2a+o(1)} \quad \text{for any } a > 0.$$

Applying the Borel–Cantelli lemma, we deduce that, for any $a > 1$,

$$d_h(\mathcal{X}_n, M) \leq \left(\frac{2a}{\delta\omega_{d'}} \frac{\ln n}{n}\right)^{1/d'} \quad \text{e.a.s.}$$

Eventually, applying Theorem 2.1 completes the proof.

3.3. Proof of Corollary 2.2

Let

$$t_n = \max_i\left(\min_j \|X_i - X_j\|\right).$$

Applying [15, Theorem 5.1, p. 958], we have

$$\frac{n\omega_{d'} t_n^{d'}}{\ln n} \overset{\text{a.s.}}{\to} f_0^{-1}.$$

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

Therefore, we easily deduce that

$$\left(\frac{1}{2f_0\,\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} \leq t_n \leq \left(\frac{2}{f_0\,\omega_{d'}}\,\frac{\ln n}{n}\right)^{1/d'} \quad \text{e.a.s.}$$

Since we have $r_n = (c\,t_n)^{2/3}$, the assumptions of Corollary 2.1 are fulfilled, which allows to conclude the proof.

3.4. Proof of Theorem 2.2

In view of assumption (A) and (2.3), we introduce the estimators

$$\overline{\Phi}(x) = \frac{1}{n}\sum_i (|\gamma_{x\to X_i}|^k),$$

$$\hat{\Phi}(x) = \frac{1}{n}\sum_i (|\hat{\gamma}_{r_n}(x, X_i)|^k). \tag{3.20}$$

First, we prove that there exists a deterministic constant $D > 0$ such that

$$\max_i |\hat{\Phi}(X_i) - \Phi(X_i)| = D\left(\frac{\ln n}{n}\right)^{\min\{2/3d', 1/2\}} \quad \text{e.a.s} \tag{3.21}$$

Indeed, the manifold M being compact, one can apply the Hoeffding inequality and obtain

$$\mathbb{P}_X(|\overline{\Phi}(x) - \Phi(x)| \geq \varepsilon_n) \leq 2\exp\left(-\frac{2n\varepsilon_n^2}{L^{2k}}\right) \quad \text{for all } x \in M,$$

where $L > 0$ is the constant introduced in the proof of Theorem 2.1. Hence,

$$\mathbb{P}_X(\text{there exists } i \in \{1, \dots, n\};\ |\overline{\Phi}(X_i) - \Phi(X_i)| \geq \varepsilon_n) \leq 2n\exp\left(-\frac{2n\varepsilon_n^2}{L^{2k}}\right).$$

Setting $\varepsilon_n = \sqrt{2}L^k\sqrt{\ln n/n}$ in this last inequality yields

$$\mathbb{P}_X\left(\max_i |\overline{\Phi}(X_i) - \Phi(X_i)| \geq \varepsilon_n\right) \leq 2n^{-3}$$

so that the Borel–Cantelli lemma allows to conclude that

$$\max_i |\overline{\Phi}(X_i) - \Phi(X_i)| = \mathcal{O}\left(\frac{\ln n}{n}\right)^{1/2} \quad \text{e.a.s.} \tag{3.22}$$

Now, noting that the assumptions of Corollary 2.2 are fulfilled, we have

$$\max_i |\hat{\Phi}(X_i) - \overline{\Phi}(X_i)| = \mathcal{O}\left(\frac{\ln n}{n}\right)^{2/3d'} \quad \text{e.a.s.}$$

Combining this with (3.22), we obtain (3.21).

Next, since Φ is continuous on the compact set M, and in view of assumption (A), the gradient of Φ vanishes at the (unique) minimum point x^*; hence, there exist $r_0 > 0$, $c_0 > 0$, $c_1 > 0$, and ε_0 such that

$$\Phi(x) \geq \Phi(x^*) + \varepsilon_0 \quad \text{for all } x \in M \cap B^c, \tag{3.23}$$

$$c_0\|x - x^*\|^2 \leq \Phi(x) - \Phi(x^*) \leq c_1\|x - x^*\|^2 \quad \text{for all } x \in M \cap \overline{B}, \tag{3.24}$$

where $B = \mathcal{B}(x^*, r_0)$ is the open ball in $\mathbb{R}^d$ of center x^* and radius r_0. The second inequality holds due to the positiveness of the Hessian matrix $H_\Phi(x^*)$.

Now, since the assumptions of Corollary 2.2 are satisfied, there exists $C > 0$ such that $d_h(\mathcal{X}_n, M) \le C(\ln n/n)^{1/d'}$. Thus, e.a.s., there exists $i_0 \in \{1, \dots, n\}$ such that

$$\|X_{i_0} - x^*\| \le C\biggl(\frac{\ln n}{n}\biggr)^{1/d'}.$$

For large enough n, we have $r_0 > C(\ln n/n)^{1/d'}$; hence, in view of (3.24), X_{i_0} satisfies

$$\Phi(X_{i_0}) \le \Phi(x^*) + c_1 C^2\biggl(\frac{\ln n}{n}\biggr)^{2/d'},$$

and from (3.21),

$$\hat{\Phi}(X_{i_0}) \le \Phi(x^*) + c_1 C^2\biggl(\frac{\ln n}{n}\biggr)^{2/d'} + D\biggl(\frac{\ln n}{n}\biggr)^{2\alpha},$$

with

$$\alpha = \min\biggl\{\frac{1}{3d'}, \frac{1}{4}\biggr\};$$

that is, for large enough n, there exists $i_0 \in \{1, \dots, n\}$ such that

$$\hat{\Phi}(X_{i_0}) \le \Phi(x^*) + 2D\biggl(\frac{\ln n}{n}\biggr)^{2\alpha}. \tag{3.25}$$

Assume now that n is large enough so that $\varepsilon_0 > 4D(\ln n/n)^{2\alpha}$. For any $i \in \{1, \dots, n\}$ such that

$$\|X_i - x^*\| \ge \frac{2D}{\sqrt{c_0}}\biggl(\frac{\ln n}{n}\biggr)^{\alpha},$$

in view of (3.23) and (3.24), this point satisfies

$$\Phi(X_i) \ge \Phi(x^*) + 4D\biggl(\frac{\ln n}{n}\biggr)^{2\alpha}.$$

Thus, in view of (3.21), we have

$$\|X_i - x^*\| \ge \frac{2D}{\sqrt{c_0}}\biggl(\frac{\ln n}{n}\biggr)^{\alpha} \quad \Longrightarrow \quad \hat{\Phi}(X_i) \ge \Phi(x^*) + 3D\biggl(\frac{\ln n}{n}\biggr)^{2\alpha}. \tag{3.26}$$

Finally, let $i^* \in \{1, \dots, n\}$ such that X_{i^*} realizes the minimum (2.3). From (3.20) and (3.25), it is clear that

$$\hat{\Phi}(X_{i^*}) \le \Phi(x^*) + 2D\biggl(\frac{\ln n}{n}\biggr)^{2\alpha},$$

and (3.26) allows us to conclude the proof.

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press

References

[1] Baillo, A., Cuevas, A. and Justel, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.* **28,** 765–782.

[2] Brito, M. R., Quiroz, A. J. and Yukich, J. E. (2013). Intrinsic dimension identification via graph-theoretic methods. *J. Multivariate Anal.* **116,** 263–277.

[3] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2001). *Introduction to Algorithms*, 3rd edn. MIT Press, Cambridge, MA.

[4] Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. Appl. Prob.* **36,** 340–354.

[5] Demartines, P. and Herault, J. (1997). Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Networks* **8,** 148–154.

[6] Granata, D. and Carnevale, V. (2016). Accurate estimation of the intrinsic dimension using graph distances: unraveling the geometric complexity of datasets. *Sci. Rep.* **6,** 31377.

[7] Howard, C. D. and Newman, C. M. (2001). Geodesics and spanning trees for Euclidean first-passage percolation. *Ann. Prob.* **29,** 577–623.

[8] Hwang, S. J., Damelin, S. B. and Hero III, A. O. (2016). Shortest path through random points. *Ann. Appl. Prob.* **26,** 2791–2823.

[9] Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29,** 1–27.

[10] Lee, J. A., Lendasse, A. and Verleyzen, M. (2004). Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis. *Neurocomputing* **57,** 49–76.

[11] Lennon, M., Mercier, G., Mouchot, M. C. and Hubert-Moy, L. (2001). Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images. *Proc. SPIE Image and Signal Process. for Remote Sens. VII*, **4541,** 157–168.

[12] Nilsson, J., Fioretos, T., Höglund, M. and Fontes, M. (2004). Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics* **20,** 874–880.

[13] Pennec, X. (2006). Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements. *J. Math. Imaging Vision* **25,** 127–154.

[14] Penrose, M. D. (1997). The longest edge of the random minimal spanning tree. *Ann. Appl. Prob.* **7,** 340–361.

[15] Penrose, M. D. (1999). A strong law for the largest nearest-neighbour link between random points. *J. London Math. Soc.* **60,** 951–960.

[16] Saul, L. K. and Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.* **4,** 119–155.

[17] Takens, F. (1985). On the numerical determination of the dimension of an attractor. In *Dynamical Systems and Bifurcations* (Lectures Notes Math. **1125**), Springer, Berlin, pp. 99–106.

[18] Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality a global geometric framework for nonlinear dimensionality reduction. *Science* **290,** 2319–2323.

https://doi.org/10.1017/jpr.2018.66 Published online by Cambridge University Press