

NONHOMOGENEOUS EUCLIDEAN FIRST-PASSAGE PERCOLATION AND DISTANCE LEARNING

P. GROISMAN, M. JONCKHEERE, AND F. SAPIENZA

ABSTRACT. Consider an i.i.d. sample from an unknown density function supported on an unknown manifold embedded in a high dimensional Euclidean space. We tackle the problem of learning a distance between points, able to capture both the geometry of the manifold and the underlying density. We define such a sample distance and prove the convergence, as the sample size goes to infinity, to a macroscopic one that we call *Fermat distance* as it minimizes a path functional, resembling Fermat principle in optics. The proof boils down to the study of geodesics in Euclidean first-passage percolation for nonhomogeneous Poisson point processes.

1. INTRODUCTION

The main motivation for this article is the following problem:

Let $Q_n = \{q_1, \dots, q_n\}$ be independent random points with common density supported on a Riemannian manifold. Define a distance in Q_n that captures both the intrinsic structure of the manifold and the density.

This problem arises naturally in tasks like clustering or dimensionality reduction of high-dimensional data, for which the notion of distance between points that is used is crucial. A typical example is the problem of clustering images according to their visual content (say, pictures of hand-writing digits). Even for black and white low-resolution pictures, as low as 30×30 pixels, the ambient space is already \mathbb{R}^{900} . Two important considerations in this kind of problems are:

- *Curse of dimensionality.* Euclidean or Minkowsky distance are not a good choice because in high dimensional spaces every two points of a typical large set are at similar distance [1].
- *Data support.* Real data usually lies in a manifold of much smaller dimension. They can be described with a few degrees of freedom, each of these representing one intrinsic variable that parametrize the manifold. In this context, the Euclidean distance can be very different from the geodesic one, which is more adequate.

Nevertheless, considering the geodesic distance might still not be good enough since it does not take into account the underlying density of the points given by f . For example, if f is given by a mixture (with equal weights) of two one-dimensional Gaussian distributions with means 0, 10 and variances 1 and 2, respectively, we would like the point 5 to be closer

2010 *Mathematics Subject Classification.* 60D05, 60K35, 60K37, 62G05, 60G55.

Key words and phrases. Distance learning, Euclidean First-Passage Percolation, nonhomogeneous point processes.

to 10 than to 0. Of course, for a real case scenario the manifold and the density function f are unknown, but for many learning tasks, it is certainly preferable to define a distance that takes both into account.

A fundamental step towards solving this problem was done by Tenenbaum, de Silva and Landford with Isomap, [18]. This estimator was shown to achieve better results for dimensionality reduction tasks by estimating the geodesic distance between points. However, it is independent of the density f from which points are sampled and, as a consequence, it is unable to give/use information about it. In [6, 7] the authors consider sample statistics that capture the intrinsic dimension of the manifold and the intrinsic entropy. They take into account both the manifold structure and the density function. More general results in the manifold setting are given in [15]. Although their estimators have a similar flavor to our proposal, they are different in the sense that they consider global properties of the manifold while we are concerned with the distance between pairs of points. The use of density based distance was explored in several papers, [2, 3, 5, 14].

In this article, we elaborate on a distance learning methodology that we introduce in [17], focusing here mainly on the mathematical aspects of the problem. We define the *Fermat distance*, a macroscopic quantity that measures the distance between two points in a manifold in this context, and the *Sample Fermat distance* as a distance inferred from the data that estimates the former one. Our contribution is then three-fold:

- *Consistency.* We show that a scaled version of the sample Fermat distance converges almost surely towards the macroscopic Fermat distance, on connected open sets of Euclidean space and as a consequence also on manifolds.
- *Convergence of geodesics.* We show that sample geodesics (minimizers of our microscopic action functional) do converge towards macroscopic minimizers of the macroscopic functional that defines the Fermat distance. The core of the proof is a bound from above for the arc length of geodesics in nonhomogeneous Euclidean first-passage percolation.
- *Complexity.* We show that with large probability the sample Fermat distance can be computed in $\mathcal{O}(n^2 \log^2 n)$ operations by restricting ourselves to “local” paths.

These fundamental mathematical properties shed light on the potential efficiency of the sample Fermat distance for unsupervised learning tasks. For a more detailed discussion on real applications to distance and manifold learning, clustering, dimensionality reduction and comparison with other methods, computational aspects, etc. we refer to [17] (see also [6, 18] on alternative proposals). A practical implementation of an algorithm computing our proposed distance can be downloaded from <http://www.aristas.com.ar/fermat/index.html>.

As a byproduct of the analysis we obtain a *Shape Theorem* for first-passage percolation in nonhomogeneous point processes as in [8] (see Corollary 2.5 below). In our setting, due to the inhomogeneous space, a huge family of shapes rather than Euclidean ball can be obtained depending on the density of points.

Related work. After a preprint version of this paper was uploaded to Arxiv, McKenzie and Damelin [11] proposed independently similar clustering applications as in [17], focusing more specifically on spectral clustering. They use the term *power weighted shortest path metric* (p-wspm) for what we call *Fermat distance* and consider a slightly different setting. They

assume that the data is sampled from a density supported in a disjoint union of manifolds and they use this distance in a different way. While we consider K -medoids algorithm for clustering, they perform spectral clustering (note that both algorithms depend strongly on the notion of distance between data points that is used). They also present a fast algorithm for computing these distances. They show in specific synthetic and real-data examples that spectral clustering with Fermat distance outperforms the classical spectral clustering algorithm with Euclidean distance, similar to what we show in [17] with K -medoids. On a different context, in [4] the use of Fermat distance in combination with Topological Data Analysis techniques is considered to understand the topology of gene expressions in healthy and cancer tissue. By means of persistent homology, the authors are able to distinguish the topology of these two datasets.

2. DEFINITIONS AND MAIN RESULTS

Following [8, 9], let Q be a non-empty, locally finite, subset of \mathbb{R}^d . We refer to the elements $q \in Q$ as *particles*. For any $x \in \mathbb{R}^d$ we denote $q(x)$ the center of the Voronoi cell of x with respect to Q . That is, $q(x)$ is the particle closest to x in Euclidean distance. Given $x, y \in \mathbb{R}^d$, a *path* from x to y is a finite sequence of particles (q_1, \dots, q_k) with $k \geq 2$, $q_1 = q(x)$ and $q_k = q(y)$. The line segment from x to y is denoted \overline{xy} and $\overline{(q_1, \dots, q_k)}$ denotes the polygonal path of line segments $\overline{q_1q_2}, \overline{q_2q_3}, \dots, \overline{q_{k-1}q_k}$. We also use $|\overline{(q_1, \dots, q_k)}|$ for its arc length, $|x|$ for the Euclidean norm of x and for $a > 0$, $C \subset \mathbb{R}^d$, $B(C, a)$ is the set given by

$$B(C, a) = \bigcup_{z \in C} B(z, a),$$

where $B(z, a)$ is the open ball centered at z with radius a with respect to the Euclidean norm. We can now define the sample Fermat distance.

Definition 2.1. For $\alpha \geq 1$ and $x, y \in \mathbb{R}^d$, we define the *sample Fermat distance* as

$$D_{Q,\alpha}(x, y) = \inf \left\{ \sum_{j=1}^{K-1} |q_{j+1} - q_j|^\alpha : (q_1, \dots, q_K) \text{ is a path from } x \text{ to } y, K \geq 1 \right\}. \quad (2.1)$$

Notice that $D_{Q,\alpha}$ satisfies the triangular inequality and defines a metric over Q and a pseudometric over \mathbb{R}^d . When not strictly necessary we will drop the dependence of all these quantities on α and Q .

This distance was considered previously in [8, 9] to construct continuous models of first-passage percolation and extended in [10] to the setting in which Q is a set of independent random points with common density f supported on a Riemannian manifold. The key difference between [10] and our setting is that in [10] the authors consider intrinsic geodesic distance raised to the power α as weights in (2.1) instead of the Euclidean distance, while we keep the last one. Their goal is to work in the (more general) intrinsic manifold setting while ours is to define a (computable) estimator of a suitable distance. Although in both cases we get the same limiting object, the microscopic objects are different and so do the proofs, with the exception of the case in which S is an open convex set of \mathbb{R}^d , where geodesic and Euclidean distance do coincide. In that case, our Corollary 2.4 below is indeed contained in [10]. The Shape Theorem, Corollary 2.5, the convergence of geodesics, Corollary 2.10 and

the local nature of geodesics, Proposition 2.12, crucial to compute geodesics, are new (even in the case of an open convex set). When S is not convex, or a manifold with dimension strictly smaller than the one of the ambient space, also Corollary 2.4 is new.

We will focus on (but not restrict ourselves to) the case in which Q is a Poisson Point Process (PPP) although the intensity function will be different in different instances. We assume that all the processes involved are constructed in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. All the “almost sure” statements are with respect to \mathbb{P} . We write $Q \sim \text{Poisson}(S, g)$ when Q is a PPP on S with intensity function g with respect to volume element on S . We include here the possibility that S is a manifold with dimension smaller than d .

Notice that for $\alpha = 1$ this distance coincides with the Euclidean distance but for $\alpha > 1$ large jumps are discouraged and this results in a different distance that penalizes paths in which consecutive points are far away to each other. We also call $r_{Q,\alpha}(x, y)$ the unique path along which $D_{Q,\alpha}(x, y)$ is achieved when it is defined (that is the case a.s. if, for example, x, y are deterministic and Q is a PPP, [8]).

Next we define a macroscopic version of the sample Fermat distance that we call the *macroscopic Fermat distance* or simply the Fermat distance or as follows.

Definition 2.2. For a continuous and positive function $f, \beta \geq 0$ and $x, y \in S$ we define the *Fermat distance* $\mathcal{D}_{f,\beta}(x, y)$ as

$$\mathcal{T}_{f,\beta}(\gamma) = \int_{\gamma} f^{-\beta}, \quad \mathcal{D}_{f,\beta}(x, y) = \inf_{\gamma} \mathcal{T}_{f,\beta}(\gamma). \quad (2.2)$$

Here the infimum is taken over all continuous and rectifiable paths γ contained in \bar{S} , the closure of S , that start at x and end at y ; and the integral is understood with respect to arc length given by Euclidean distance.

We will omit the dependence on β and f when not strictly necessary. This definition coincides with Fermat Principle in optics for the path followed by light in a non-homogeneous media when the refractive index is given by $f^{-\beta}$. We will call the minimizer γ^* in (2.2) a *macroscopic f -geodesic* between x and y . Observe that f -geodesics are likely to lie in regions where f is large.

2.1. Consistency. Our main result consists in proving that the sample Fermat distance when appropriately scaled converges to the Fermat distance. In other words, the scaled sample Fermat distance is a strongly consistent estimator of the macroscopic one.

Theorem 2.3. *Let $S \subset \mathbb{R}^d$ be an open connected set with C^1 (or empty) boundary. Let $f: \bar{S} \rightarrow [m_f, M_f]$ be a continuous intensity function. Assume $m_f > 0$. For each $n \in \mathbb{N}$ let $Q_n \sim \text{Poisson}(S, nf)$. Given $x, y \in S$ and $\varepsilon > 0$, there exist constants μ, c_1, c_2 and n_0 such that*

$$\mathbb{P}(|n^\beta D_{Q_n,\alpha}(x, y) - \mu \mathcal{D}_{f,\beta}(x, y)| > \varepsilon) \leq e^{-c_1 n^{c_2}}, \quad (2.3)$$

for $\beta = (\alpha - 1)/d$ and every $n \geq n_0$. In particular

$$\lim_{n \rightarrow \infty} n^\beta D_{Q_n,\alpha}(x, y) = \mu \mathcal{D}_{f,\beta}(x, y) \quad \text{almost surely.}$$

The Poisson assumption can be replaced by assuming Q is an i.i.d. sample with density f using a simple large deviations estimate.

Corollary 2.4. *The same result holds if we replace Q_n by a set of n independent points with common density f .*

If we “invert” Theorem 2.3 as in [8, Theorem 1] we obtain the Shape Theorem. For $t > 0$ define

$$\mathbb{B}_n(x, t) = \{y \in \mathbb{R}^d : D_{Q_n, \alpha}(x, y) < t\} \text{ and } \mathcal{B}(x, t) = \{y \in \mathbb{R}^d : \mathcal{D}_{f, \beta}(x, y) < t\}.$$

Corollary 2.5 (Shape Theorem). *In the setting of Theorem 2.3, let $x \in S$, ε with $0 < \varepsilon < \mu^{-1}$ and t such that $\mathcal{B}(x, (\mu^{-1} + \varepsilon)t) \subset S$. Then, a.s. there is n_0 such that for $n \geq n_0$ we have*

$$\mathcal{B}(x, (\mu^{-1} - \varepsilon)t) \subset n^\beta \mathbb{B}_n(x, t) \subset \mathcal{B}(x, (\mu^{-1} + \varepsilon)t).$$

Remark 2.6. If S is not connected and x and y belong to different connected components of \bar{S} , we have $\mathcal{D}_{f, \beta}(x, y) = \liminf n^\beta D_{Q_n}(x, y) = \infty$ a.s. If x, y belong to the same connected component, we can restrict ourselves to this component. So, the connectedness assumption can be dropped and is assumed for simplicity.

If the Euclidean norm in (2.1) is replaced by another distance, similar results can be obtained with the line integrals with respect to arc length replaced by line integrals with respect to the distance involved. It could be interesting to explore other choices.

The C^1 smoothness assumption for the boundary of S is not really necessary either and can be relaxed up to some point. For instance, it is enough (but actually not necessary) to suppose S to be locally convex at points of the boundary where it is not C^1 . Also, if we allow $m_f = 0$ (which can be done with some extra work), no regularity assumptions are needed at boundary points where f vanishes. In fact, the only problem one needs to deal with is the case in which the macroscopic geodesic intersects the boundary. This case can be avoided in several ways, but it can certainly happen if S is not convex and f is not negligible at the boundary. We assume in the sequel the stronger C^1 assumption to simplify the exposition.

As a consequence of Theorem 2.3, we obtain a similar result for points supported on a lower dimensional manifold. We will say that \mathcal{M} is an isometric d -dimensional C^1 manifold embedded in \mathbb{R}^D if there exists $S \subset \mathbb{R}^d$ an open connected set and $\phi : \bar{S} \rightarrow \mathbb{R}^D$ an isometric transformation such that $\phi(\bar{S}) = \mathcal{M}$. As we mentioned before, in real applications, we expect $d \ll D$, but this is not required.

Theorem 2.7. *Assume \mathcal{M} is an isometric C^1 d -dimensional manifold embedded in \mathbb{R}^D and $f : \mathcal{M} \rightarrow \mathbb{R}_+$ is a continuous probability density function. Let $Q_n = \{q_1, \dots, q_n\}$ be independent random points with common density f . Then, for $\alpha > 1$ and $x, y \in \mathcal{M}$ we have*

$$\lim_{n \rightarrow \infty} n^\beta D_{Q_n, \alpha}(x, y) = \mu \mathcal{D}_{f, \beta}(x, y) \quad \text{almost surely.}$$

Here $\beta = (\alpha - 1)/d$ and μ is a constant depending only on α and d ; the minimization is carried over all rectifiable curves $\gamma \subset \mathcal{M}$ that start at x and end at y .

Remark 2.8. Notice that the scaling factor $\beta = (\alpha - 1)/d$ depends on the intrinsic dimension of the manifold, instead of the dimension D of the ambient space.

2.2. Geodesics. Once, we have the convergence of the distances, it is natural to ask for the convergence of the geodesics. An important step towards this result is to prove that sample geodesic arc length is bounded. This result is not straightforward and follows from geometric arguments combined with large deviations.

Proposition 2.9. *Let $S \subset \mathbb{R}^d$ be a bounded connected open set with C^1 boundary, $Q_n \sim \text{Poisson}(S, nf)$ and $\delta > 0$. Then, there exists positive constants ℓ, c_2, c_3 and n_0 , with $c_3(\delta)$ depending on δ , such that if $x, y \in S$, $|x - y| > \delta$, then for all $n > n_0$*

$$\mathbb{P} \left(\overline{|r_{Q_n, \alpha}(x, y)|} > \ell \right) \leq \exp(-c_3 n^{c_2}). \quad (2.4)$$

As a consequence, we have

$$\limsup_{n \rightarrow \infty} \overline{|r_{Q_n, \alpha}(x, y)|} \leq \ell \quad \text{almost surely.}$$

The constant c_2 is the same as in Theorem 2.3. Having obtained this upper bound on the arc length of geodesics, we can show that under suitable conditions the microscopic geodesics converge to the macroscopic one.

Corollary 2.10. *Let $S \subset \mathbb{R}^d$ be a bounded connected open set and $Q_n \sim \text{Poisson}(S, nf)$. If there is a unique macroscopic f -geodesic γ^* , then $\overline{r_{Q_n, \alpha}(x, y)}$ converges uniformly to γ^* almost surely.*

2.3. Complexity. Finally, we turn our attention to the computability of the sample Fermat distance. Computing the minimum in (2.1) for every two points in Q_n requires a search in a discrete set of size larger than $n!$. By means of Floyd-Warshall algorithm, this task can be done in $\mathcal{O}(n^3)$ operations. We prove that we can restrict the search to paths in which each particle q_i of the path is a k -th nearest neighbor of q_{i-1} for $k \approx \log n$. Based on this fact, Dijkstra algorithm requires $\mathcal{O}(n^2 \log^2 n)$ operations to compute the distances between every two points in the sample.

Given $k \geq 1$ and $q \in Q_n$, the k -th nearest neighbor of q , denoted by $q^{(k)}$, is defined by

$$q^{(1)} = \arg \min_{q' \in Q_n \setminus \{q\}} |q' - q|, \quad q^{(k)} = \arg \min_{q' \in Q_n \setminus \{q, q^{(1)}, \dots, q^{(k-1)}\}} |q' - q| \quad \text{for } k > 1.$$

We use the lexicographic order to break ties. Also denote $\mathcal{N}_k(z) = \{q^{(1)}, q^{(2)}, \dots, q^{(k)}\}$ the set of k -nearest neighbors of q . We can now define the *restricted* sample Fermat distance as follows:

Definition 2.11. For $x, y \in Q_n$, $\alpha \geq 1$ and $k \in \mathbb{N}$, we define

$$D_{Q_n}^k(x, y) = \min \left\{ \sum_{i=1}^{K-1} |q_{i+1} - q_i|^\alpha : q_1 = x, q_K = y, q_{i+1} \in \mathcal{N}_k(q_i), 1 \leq i \leq K-1 \right\}.$$

We have the following quantitative approximation result:

Proposition 2.12. *In the setting of Theorem 2.3, given $\varepsilon > 0$, there exist positive constants c_4, c_5 such that if $k > c_4 \log(n/\varepsilon) + c_5$ we have*

$$\mathbb{P} \left(D_{Q_n}^k(x, y) = D_{Q_n}(x, y) \right) > 1 - \varepsilon.$$

In other words, with probability at least $1 - \varepsilon$, the minimizing path (q_1, \dots, q_{k_n}) satisfies $q_{i+1} \in \mathcal{N}_k(q_i)$ for every $i = 1, \dots, K - 1$.

While the previous result is certainly an improvement, it might still be unsatisfactory for large data sets. However, if n is very large, it is possible to appeal to greedy implementations. Given Q_n , let us consider a subset of *landmarks* $\tilde{Q} \subset Q_n$ with $|\tilde{Q}| = m$ and $m \ll n$. Then, we compute the minimum path between each of the m landmarks and the rest of the particles in Q_n using Dijkstra's algorithm on the k -nearest neighbor graph. This can be done in $\mathcal{O}(mkn \log n)$ operations. Then, we can bound the exact sample Fermat distance between any two points $q, q' \in Q_n$ by

$$\max_{\tilde{q} \in \tilde{Q}} |D_{Q_n}(q, \tilde{q}) - D_{Q_n}(q', \tilde{q})| \leq D_{Q_n}(q, q') \leq \min_{\tilde{q} \in \tilde{Q}} (D_{Q_n}(q, \tilde{q}) + D_{Q_n}(q', \tilde{q})),$$

see [16]. Notice that the bound from above holds with equality if there is a landmark $\tilde{q} \in \tilde{Q}$ in the shortest path between q and q' . Due to this fact, an interesting and important problem is to choose a good set of landmarks, [16].

2.4. Organization of the paper. The rest of the article is organized as follows. In Section 3, we prove several lemmas that lead to the proof of consistency, Theorem 2.3. Corollary 2.4 can be easily obtained by means of a large deviations principle for Poisson random variables and is left to the reader. In Section 4 we consider the original problem, i.e., the case in which Q_n is a random set of independent points with common density f supported on a manifold and we prove Theorem 2.7. We then obtain Corollary 2.10 as a consequence of Theorem 2.3 after proving that the arc length of microscopic geodesics is bounded, which is done in Section 5. Section 6 deals with computational considerations. We show that with large probability $(D_{Q_n}(q, q'))_{q, q' \in Q_n}$ can be computed in $\mathcal{O}(n^2 \log^2 n)$ operations by restricting ourselves to “local” paths.

3. NONHOMOGENEOUS PPP

We begin by proving the almost sure convergence of $n^\beta D_{Q_n}(x, y)$ to Fermat distance between x and y for nonhomogeneous PPP stated in Theorem 2.3. The proof will be split in several lemmas. The first step consists in considering homogeneous PPP in a convex set $S \subset \mathbb{R}^d$. This case has actually been treated in [8, 9] where the following is proved.

Proposition 3.1. [8, Lemma 3 and Lemma 4], [9, Theorem 2.2] *Assume $S \subset \mathbb{R}^d$ is an open convex set and let $Q_n \sim \text{Poisson}(S, n)$. There exists $0 < \mu < \infty$ such that for any $x, y \in S$ we have*

$$\lim_{n \rightarrow \infty} n^\beta D_{Q_n}(x, y) = \mu |x - y|, \quad \text{almost surely.}$$

Moreover, given $\delta > 0$ there exist positive constants λ, c_2, c_6, c_7 , with c_7 depending on δ , such that if $|x - y| > \delta$ then

$$\mathbb{P}(|n^\beta D_{Q_n}(x, y) - \mu |x - y|| \geq \lambda n^{-1/3d}) \leq c_6 \exp(-c_7 n^{c_2}).$$

for every $n \geq 1$.

The results of [8, 9] are proved in fact for the case in which Q_n is replaced by an intensity one PPP and instead of taking $n \rightarrow \infty$, the authors consider the limit as $|y| \rightarrow \infty$. The adaptation of those results to our setting to get 3.1 is straightforward by considering the rescaled process $n^{1/d}Q_n$ and using [9, Theorem 2.4] to show that if we have $\tilde{Q}_n \sim \text{Poisson}(S, n)$, $\tilde{Q}_n \sim \text{Poisson}(\mathbb{R}^d, n)$ and $x, y \in S$, then

$$\mathbb{P}(D_{Q_n}(x, y) \neq D_{\tilde{Q}_n}(x, y)) \leq c_6 \exp(-c_7 n^{c_2}).$$

By means of Proposition 3.1 we can obtain rough bounds for the nonhomogeneous case.

Lemma 3.2. *Let $S \subset \mathbb{R}^d$ be an open bounded connected set with C^1 (or empty) boundary and $f: S \rightarrow [m_f, M_f]$ measurable. Assume $m_f > 0$. Let $\delta > 0$ and $x, y \in S$ with $|x - y| > \delta$ and $Q_n \sim \text{Poisson}(S, nf)$. Then, for all $\varepsilon > 0$ there exist $n_0 = n_0(\varepsilon)$ and a positive constant $c_8 = c_8(\delta)$ such that for all $n > n_0$,*

$$\mathbb{P}\left(n^\beta D_{Q_n}(x, y) \leq \mu M_f^{-\beta} |x - y| - \varepsilon\right) \leq \exp(-c_8 (m_f n)^{c_2}), \quad (3.1)$$

$$\mathbb{P}\left(n^\beta D_{Q_n}(x, y) \geq \mu m_f^{-\beta} \mathcal{D}_0(x, y) + \varepsilon\right) \leq \exp(-c_8 (M_f n)^{c_2}). \quad (3.2)$$

Here $\mathcal{D}_0(x, y) := \mathcal{D}_{f,0}(x, y)$ is the geodesic distance between x and y defined in accordance with (2.2).

Proof. Denote $\text{co}(S)$ the convex hull of S . Given two locally finite configurations $Q \subset \tilde{Q}$, we have $D_{\tilde{Q}}(x, y) \leq D_Q(x, y)$. Enlarge the probability space to consider two homogeneous PPP $Q_n^- \sim \text{Poisson}(S, nm_f)$ and $Q_n^+ \sim \text{Poisson}(\text{co}(S), nM_f)$, coupled with Q_n (see for instance [12, Section 3.2.2]) to guarantee that $Q_n^- \subset Q_n \subset Q_n^+$. Then

$$\mathbb{P}\left(n^\beta D_{Q_n}(x, y) \leq \mu M_f^{-\beta} |x - y| - \varepsilon\right) \leq \mathbb{P}\left(n^\beta D_{Q_n^+}(x, y) \leq \mu M_f^{-\beta} |x - y| - \varepsilon\right)$$

Choosing n_0 such that $\varepsilon > \lambda(n_0 m_f)^{-1/3d}$, by means of Proposition 3.1 we get (3.1). To prove (3.2) we proceed similarly, but we need to be more careful. Since S is open and connected, we can consider a polygonal $\gamma = \overline{(x_0, \dots, x_k)} \subset S$ from x to y with

$$|\overline{(x_0, \dots, x_k)}| < \mathcal{D}_0(x, y) + \frac{m_f^\beta \varepsilon}{2\mu} \quad \text{and} \quad B\left(\frac{x_{i+1} + x_i}{2}, |x_{i+1} - x_i|\right) \subset S,$$

for every $0 \leq i \leq k - 1$. We claim that k can be taken uniformly bounded for every two points $x, y \in S$. To see that, we proceed by contradiction. Fix one point $z \in S$ and assume there is a sequence of points $z_n \in S$ with the property that any polygonal from z to z_n contained in S is composed by at least n line segments. Since \bar{S} is compact we can extract a convergent subsequence $z_{n_j} \rightarrow z^* \in \bar{S}$. If $z^* \in S$, there is ball centered at z^* contained in S with points z_n for large n . Then we can easily construct polygonals contained in S from z to z_n with a bounded number of line segments, a contradiction. Then it should hold that $z^* \in \partial S$ but since ∂S is C^1 we can proceed in the same way to obtain again a contradiction.

Denote

$$Q_n^i = Q_n^- \cap B\left(\frac{x_{i+1} + x_i}{2}, |x_{i+1} - x_i|\right).$$

Proceeding as before, we get for every i ,

$$\mathbb{P}\left(n^\beta D_{Q_n}(x_i, x_{i+1}) \geq \mu m_f^{-\beta} |x_{i+1} - x_i| + \varepsilon\right) \leq \mathbb{P}\left(n^\beta D_{Q_n^i}(x_i, x_{i+1}) \geq \mu m_f^{-\beta} |x_{i+1} - x_i| + \varepsilon\right).$$

Then,

$$\begin{aligned} \mathbb{P}\left(n^\beta D_{Q_n}(x, y) \geq \mu m_f^{-\beta} \mathcal{D}_0(x, y) + \varepsilon\right) &\leq \mathbb{P}\left(n^\beta \sum_{i=0}^k D_{Q_n}(x_i, x_{i+1}) \geq \frac{\mu}{m_f^\beta} |(x_0, \dots, x_k)| + \frac{\varepsilon}{2}\right) \\ &\leq \sum_{i=0}^k \mathbb{P}\left(n^\beta D_{Q_n^i}(x_i, x_{i+1}) \geq \mu m_f^{-\beta} |x_{i+1} - x_i| + \frac{\varepsilon}{2k}\right). \end{aligned}$$

Using again Proposition 3.1 we get (3.2). \square

The second step is to show that the distance between consecutive particles in the optimal path vanishes as $n \rightarrow \infty$.

Lemma 3.3. *In the setting of Theorem 2.3, let (q_1, \dots, q_{k_n}) be the minimizing path. Given $\delta > 0$, there exists a positive constant c_9 such that*

$$\mathbb{P}\left(\max_{i < k_n} |q_i - q_{i+1}| > \delta\right) \leq \exp(-c_9 n).$$

Proof. For any two consecutive points q_i, q_{i+1} in the optimal path we have

$$Q_n \cap \{z \in S : |z - q_{i+1}|^\alpha + |z - q_i|^\alpha < |q_{i+1} - q_i|^\alpha\} = \emptyset.$$

Observe that we can choose κ_1 depending only on α and d such that the region $\{z \in S : |z - q_{i+1}|^\alpha + |z - q_i|^\alpha < |q_{i+1} - q_i|^\alpha\}$ contains a cube of edge size $\kappa_1 |q_{i+1} - q_i|$. Consider a family \mathcal{C} of cubes of edge size $\kappa_1 \delta / 2$ with vertices in $(\kappa_1 \delta / 2) \mathbb{Z}^d$.

Notice that the number of cubes in this family that intersect S is finite. Each of these cubes has no particles with probability bounded by $\exp(-c_{10} n)$. If $\max_{i < k_n} |q_i - q_{i+1}| > \delta$, then there is a cube in S with side $\kappa_1 \delta$. Such a cube must contain a cube in \mathcal{C} . \square

Next, we prove that in order to find the optimal path between x and y we can restrict ourselves to certain neighborhoods of any path $\gamma_{xy} \subset S$ that starts at x and ends at y . This fact will be used both for points that are close to each other as well as for points that are at a large distance. Denote, $m_f^\gamma = \inf\{f(z) : z \in B(x, 2|\gamma|)\}$.

Lemma 3.4. *In the setting of Theorem 2.3, given $\delta > 0$, there exist positive constants c_{11} and n_0 such that for every $x, y \in S$ with $|x - y| > \delta$ and a path $\gamma \subset S$ from x to y we have,*

$$\mathbb{P}\left(D_{Q_n}(x, y) \neq D_{Q_n \cap B(x, \tilde{a}|\gamma|)}(x, y)\right) \leq \exp(-c_{11} n^{c_2}), \quad (3.3)$$

for every $n > n_0$ and $\tilde{a} = 3(M_f/m_f^\gamma)^\beta$. In particular,

(i) if S is bounded

$$\mathbb{P}\left(D_{Q_n}(x, y) \neq D_{Q_n \cap B(x, a|\overline{xy}|)}(x, y)\right) \leq \exp(-c_{11} n^{c_2}).$$

with

$$a = \tilde{a} \sup_{|z-w| \geq \delta} \frac{D_0(z, w)}{|z-w|} < \infty.$$

(ii) if $\overline{xy} \subset S$, we have

$$\mathbb{P}\left(D_{Q_n}(x, y) \neq D_{Q_n \cap B(x, \tilde{a}|\overline{xy}|)}(x, y)\right) \leq \exp(-c_{11} n^{c_2}).$$

Proof. Let $z \notin B(x, a|\gamma|) \cap S$. Given $\delta_1 < \mu(m_f^\gamma)^{-\beta}|\gamma|/3$, consider the events

$$\begin{aligned} A_n^z &= \left\{ n^\beta D_{Q_n}(x, z) \leq n^\beta D_{Q_n}(x, y) + \delta_1 \right\} \\ E_n^z &= \left\{ n^\beta D_{Q_n}(x, z) \geq \mu M_f^{-\beta} |x-z| - \delta_1 \right\} \\ F_n &= \left\{ n^\beta D_{Q_n \cap B(x, 2|\gamma|)}(x, y) \leq \mu(m_f^\gamma)^{-\beta}|\gamma| + \delta_1 \right\}. \end{aligned}$$

In $A_n^z \cap E_n^z \cap F_n$ we get

$$\begin{aligned} \mu M_f^{-\beta} |x-z| &\leq n^\beta D_{Q_n}(x, z) + \delta_1 \leq n^\beta D_{Q_n}(x, y) + 2\delta_1 \\ &\leq n^\beta D_{Q_n \cap B(x, 2|\gamma|)}(x, y) + 2\delta_1 \leq \mu(m_f^\gamma)^{-\beta}|\gamma| + 3\delta_1 \\ &< 2\mu(m_f^\gamma)^{-\beta}|\gamma|. \end{aligned}$$

Since $z \notin B(x, a|\gamma|)$ implies $|x-z| > a|\gamma|$ and $a = 3(M_f/m_f^\gamma)^\beta$, we have $A_n^z \cap E_n^z \cap F_n = \emptyset$. By Lemma 3.2 there exist $c_8(\delta)$, $n_0(\delta)$ independent of z and a positive constant c_2 such that

$$\mathbb{P}(A_n^z) \leq \mathbb{P}((E_n^z)^c) + \mathbb{P}(F_n^c) \leq 2 \exp(-c_8(m_f n)^{c_2}) \quad \text{for all } n > n_0.$$

Assume $D_{Q_n}(x, y) < D_{Q_n \cap B(x, a|\gamma|)}(x, y)$ and $\{\max_{i < k_n} |q_i - q_{i+1}| < a|\gamma|\}$. Then there is a particle $q \in Q_n \cap B(x, a|\gamma|)^c \cap B(x, 2a|\gamma|)$ with

$$D_{Q_n}(x, y) = D_{Q_n}(x, q) + D_{Q_n}(q, y) \geq D_{Q_n}(x, q).$$

Consider the following covering

$$S \cap (B(x, 2a|\gamma|) \setminus B(x, a|\gamma|)) \subset \bigcup_{v \in \mathcal{V}} B(v, \delta_0 n^{-1/d}).$$

Here $\mathcal{V} \subset S \setminus B(x, a|\gamma|)$ is a finite set of points that can be chosen in such a way that $\#\mathcal{V} \leq \kappa_2 n$ for some constant κ_2 and $(2\delta_0)^\alpha < \delta_1$. Let $v_q \in \mathcal{V}$ be such that $q \in B(v_q, \delta_0 n^{-1/d})$. If q is the closest particle in Q_n to v_q , then $D_{Q_n}(q, v_q) = 0$. If that is not the case, there is another particle in $B(v_q, \delta_0 n^{-1/d})$ and consequently we have $D_{Q_n}(q, v_q) < (2\delta_0 n^{-1/d})^\alpha$. From triangular inequality we get

$$n^\beta D_{Q_n}(x, q) \geq n^\beta D_{Q_n}(x, v_q) - n^\beta D_{Q_n}(q, v_q) \geq n^\beta D_{Q_n}(x, v_q) - \delta_1 n^{-\alpha/d} \geq n^\beta D_{Q_n}(x, v_q) - \delta_1.$$

Hence

$$\begin{aligned}
& \mathbb{P}\left(D_{Q_n}(x, y) \neq D_{Q_n \cap B(x, a|\gamma|)}(x, y), \max_{i < k_n} |q_i - q_{i+1}| < a|\gamma|\right) \\
& \leq \mathbb{P}\left(\exists v \in \mathcal{V} : n^\beta D_{Q_n}(x, y) \geq n^\beta D_{Q_n}(x, v) - \delta_1\right) \\
& \leq \sum_{v \in \mathcal{V}} \mathbb{P}((A_n^v)^c) \\
& \leq 2\kappa_2 n \exp(-c_8(m_f n)^{c_2}) \quad \forall n > n_0.
\end{aligned}$$

From Lemma 3.3 and the fact that $c_2 < 1/d \leq 1$ ([9]), we get (3.3). \square

We are ready to prove the upper bound in (2.3).

Lemma 3.5 (Upper bound). *In the setting of Theorem 2.3, there are positive constants c_{12} and n_0 such that*

$$\mathbb{P}(n^\beta D_{Q_n}(x, y) > \mu \mathcal{D}_{f, \beta}(x, y) + \varepsilon) \leq \exp(-c_{12} n^{c_2}).$$

for all $n > n_0$.

Proof. Let $\gamma^* \subset S$ be a continuous and rectifiable curve that starts at x and ends at y and such that $\int_{\gamma^*} \frac{1}{f^\beta} < \mathcal{D}_{f, \beta}(x, y) + \varepsilon/(4\mu)$. If $\varepsilon < 1$, the arc length $|\gamma^*|$ is bounded above by

$$|\gamma^*| < \ell^* := M_f^\beta \left(\mathcal{D}_{f, \beta}(x, y) + \frac{1}{4\mu} \right).$$

Let us consider a finite set of points $z_1, z_2, \dots, z_M \in \gamma^*$ sorted according to a parametrization of γ^* that starts at x and ends at y , such that $z_1 = x$, $z_M = y$ and $\delta < |z_{i+1} - z_i| < 2\delta$. Notice that $M = M(\delta) < \ell^*/\delta$. Let γ_i^* be the part of γ^* that connects z_i and z_{i+1} . Then

$$\int_{\gamma^*} \frac{1}{f^\beta} = \sum_{i=1}^{M-1} \int_{\gamma_i^*} \frac{1}{f^\beta}.$$

Since $f^{-\beta}$ is integrable and uniformly continuous in $\overline{S \cap B(x, a|\gamma|)}$, we can choose $\delta > 0$ such that

- (i) $\sum_{i=1}^{M-1} \left(\min_{\gamma_i^*} f \right)^{-\beta} |z_i - z_{i+1}| < \int_{\gamma^*} \frac{1}{f^\beta} + \frac{\varepsilon}{4}$,
- (ii) $|z - z'| < \delta \implies |f^{-\beta}(z) - f^{-\beta}(z')| < \varepsilon_2 := \varepsilon m_f^\beta / (4\mu \ell^*)$.
- (iii) $\text{co}(B(\gamma_i^*, \delta)) \subset S$.

Recall here that $\text{co}(B)$ denotes the convex hull of B . For $i = 1, 2, \dots, M-1$ consider the set $C_i = \text{co}(B(\gamma_i^*, \delta))$. On the one hand,

$$D_{Q_n}(x, y) \leq D_{Q_n \cap (\cup_{i=1}^{M-1} C_i)}(x, y) \leq \sum_{i=1}^{M-1} D_{Q_n \cap C_i}(z_i, z_{i+1}).$$

On the other hand, we have

$$\begin{aligned}
\mu\mathcal{D}_{f,\beta}(x, y) + \varepsilon &> \mu \int_{\gamma^*} \frac{1}{f^\beta} + \frac{3\varepsilon}{4} \\
&> \mu \sum_{i=1}^{M-1} \left(\min_{\gamma_i^*} f \right)^{-\beta} |z_{i+1} - z_i| + \frac{\varepsilon}{2} \\
&\geq \mu \sum_{i=1}^{M-1} \left(\min_{C_i} f \right)^{-\beta} |z_{i+1} - z_i| + \frac{\varepsilon}{2} - \frac{\mu M \delta}{m_f^\beta} \varepsilon_2 \\
&> \mu \sum_{i=1}^{M-1} \left(\min_{C_i} f \right)^{-\beta} |z_{i+1} - z_i| + \frac{\varepsilon}{4}.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{P}\left(n^\beta D_{Q_n}(x, y) \geq \mu\mathcal{D}_{f,\beta}(x, y) + \varepsilon\right) &\leq \\
&\leq \mathbb{P}\left(\sum_{i=1}^{M-1} n^\beta D_{Q_n \cap C_i}(z_i, z_{i+1}) \geq \mu \sum_{i=1}^{M-1} \left(\min_{C_i} f \right)^{-\beta} |z_{i+1} - z_i| + \frac{\varepsilon}{4}\right) \\
&\leq \sum_{i=1}^{M-1} \mathbb{P}\left(n^\beta D_{Q_n \cap C_i}(z_i, z_{i+1}) \geq \mu \left(\min_{C_i} f \right)^{-\beta} |z_{i+1} - z_i| + \frac{\varepsilon}{4M}\right) \\
&\leq M \exp(-c_8(m_f n)^{c_2}) \quad \text{for all } n > n_0,
\end{aligned}$$

by Lemma 3.2 (applied to each C_i). Notice that the constant c_8 depends only on δ . This finishes the proof of the lemma. \square

Lemma 3.6 (Lower bound). *In the setting of Theorem 2.3, there exist positive constants c_{13} and n_0 such that*

$$\mathbb{P}\left(n^\beta D_{Q_n}(x, y) < \mu\mathcal{D}_{f,\beta}(x, y) - \varepsilon\right) \leq \exp(-c_{13} n^{c_2}),$$

for all $n > n_0$.

Proof. By Lemma 3.4 we can assume S is bounded (if it is not bounded, we consider $S \cap B(x, a|\gamma|)$, with γ any path from x to y instead of S). Let $\gamma_n = (q_1, \dots, q_{k_n})$ be the minimizing path. For $\delta > 0$, consider the event $E_n = \{\max_{j < k_n} |q_j - q_{j+1}| < \delta\}$. If E_n occurs, there are particles $q_1^*, q_2^*, \dots, q_k^* \in \gamma_n \cap Q_n$ with $\delta < |q_{i+1}^* - q_i^*| < 4\delta$ for $i = 0, 1, 2, \dots, k$, with $q_0^* = x$, $q_{k+1}^* = y$. We can construct this sequence inductively as follows. Denote $\tau_0 = 0$, $q_0^* = x$. For $i \geq 0$, if $|q_i^* - y| < 4\delta$, then $q_{i+1}^* = y$ and we set $k = i + 1$. If not, we choose $q_{i+1}^* = q_{\tau_{i+1}}$ with $\tau_{i+1} = \min\{j > \tau_i : 2\delta < |q_j - q_i^*| < 3\delta\}$. The existence of τ_{i+1} (in case we need to define it) is guaranteed since we are assuming that E_n occurs. With this construction we have $|q_k^* - q_{k-1}^*| = |y - q_{k-1}^*| > |y - q_{k-2}^*| - |q_{k-1}^* - q_{k-2}^*| > 4\delta - 3\delta = \delta$ and hence $\delta < |q_{i+1}^* - q_i^*| < 4\delta$ for every $1 \leq i \leq k - 1$. We will see that there exists a constant K such that $k \leq K$ with overwhelming probability. This would be immediate if we assume that the arc lengths of the minimizing paths are bounded (which is proved in Section 5),

but this assumption is not really necessary at this point as the following argument shows. Notice that

$$D_{Q_n}(x, y) = \sum_{i=0}^k D_{Q_n}(q_i^*, q_{i+1}^*). \quad (3.4)$$

For $\delta_0 > 0$, that will be chosen later, consider the following covering of \bar{S} ,

$$\bar{S} \subset \bigcup_{v \in \mathcal{V}} B(v, \delta_0 n^{-1/d}).$$

Here $\mathcal{V} \subset S$ is chosen such that $\#\mathcal{V} \leq \kappa_3 n$ for some constant $\kappa_3 < \infty$. Let $w_1, w_2, \dots, w_k \in \mathcal{V}$ be such that $q_i^* \in B(w_i, \delta_0 n^{-1/d})$ for every $i \leq k$. For a given $i \leq k$ it holds

$$\begin{aligned} n^\beta D_{Q_n}(q_i^*, q_{i+1}^*) &\geq n^\beta (D_{Q_n}(w_i, w_{i+1}) - D_{Q_n}(w_i, q_i^*) - D_{Q_n}(w_{i+1}, q_{i+1}^*)) \\ &\geq n^\beta D_{Q_n}(w_i, w_{i+1}) - 2(2\delta_0)^\alpha. \end{aligned}$$

If in addition $\delta_0 < \delta/4$, we have

$$|w_i - w_{i+1}| > |q_i^* - q_{i+1}^*| - |w_i - q_i^*| - |w_{i+1} - q_{i+1}^*| > \delta - \delta/4 - \delta/4 = \delta/2.$$

Let $\Delta = \mu M_f^{-\beta} \delta/8$ and choose δ_0 with $2(2\delta_0)^\alpha < \Delta$. Then

$$\mathbb{P}\left(\min_i n^\beta D_{Q_n}(q_i^*, q_{i+1}^*) < \Delta\right) \leq \mathbb{P}\left(\exists v_1, v_2 \in \mathcal{V}: |v_1 - v_2| > \delta/2 \text{ and } n^\beta D_{Q_n}(v_1, v_2) < 2\Delta\right).$$

Since the number of possible elections of v_1 and v_2 is upper bounded by $(\kappa_3 n)^2$, from Lemma 3.2 we conclude that

$$\mathbb{P}\left(\min_i n^\beta D_{Q_n}(q_i^*, q_{i+1}^*) < \Delta\right) < (\kappa_3 n)^2 \exp(-c_8 (m_f n)^{c_2}).$$

If $n^\beta D_{Q_n}(x, y) < 2\mu m_f^{-\beta} \mathcal{D}_0(x, y)$ and $n^\beta D_{Q_n}(q_i^*, q_{i+1}^*) > \Delta$ for every $i \leq k$, then from (3.4) we obtain $k\Delta < 2\mu m_f^{-\beta} \mathcal{D}_0(x, y)$. Hence, for $K = K(\delta) := 16\delta^{-1} (M_f/m_f)^\beta \mathcal{D}_0(x, y)$,

$$\mathbb{P}(k > K) \leq \exp(-c_8 (M_f n)^{c_2}) + (\kappa_3 n)^2 \exp(-c_8 (m_f n)^{c_2}), \quad (3.5)$$

for n large enough by (3.2).

If we choose $(2\delta_0)^\alpha < (\varepsilon/4K)$, using triangular inequality in (3.4) we get

$$\begin{aligned} n^\beta D_{Q_n}(x, y) &\geq \sum_{i=0}^k n^\beta (D_{Q_n}(w_i, w_{i+1}) - D_{Q_n}(w_i, q_i^*) - D_{Q_n}(w_{i+1}, q_{i+1}^*)) \\ &\geq \sum_{i=0}^k n^\beta (D_{Q_n}(w_i, w_{i+1}) - 2(2\delta_0)^\alpha n^{-\alpha/d}) \\ &\geq \sum_{i=0}^k n^\beta D_{Q_n}(w_i, w_{i+1}) - \varepsilon/2. \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{P}(n^\beta D_{Q_n}(x, y) \leq \mu \mathcal{D}_{f,\beta}(x, y) - \varepsilon) \\ \leq \mathbb{P}\left(\exists v_1, \dots, v_k \in \mathcal{V} \text{ with } k \leq K \text{ and } \frac{\delta}{2} < |v_i - v_{i+1}| < 5\delta \text{ such that} \right. \\ \left. \sum_{i=0}^k n^\beta D_{Q_n}(v_i, v_{i+1}) \leq \mu \mathcal{D}_{f,\beta}(x, y) - \frac{\varepsilon}{2}, E_n\right) + \mathbb{P}(k > K) + \mathbb{P}(E_n^c). \end{aligned}$$

The second term is bounded by (3.5) and Lemma 3.3 gives us an exponential bound for the third one. Let us focus on the first one. Notice that the number of paths (v_1, v_2, \dots, v_k) with $v_i \in \mathcal{V}$ and $k \leq K$ is bounded above by $(\kappa_3 n)^K$. Fix any one these paths and denote

$$M_{f,i} := \sup_{z \in B(v_i, a|\overline{v_i v_{i+1}}|) \cap S} f(z)$$

and consider the events

$$\begin{aligned} A_i &= \{D_{Q_n}(v_i, v_{i+1}) = D_{Q_n \cap B(v_i, a|\overline{v_i v_{i+1}}|)}(v_i, v_{i+1})\} \\ B_i &= \left\{n^\beta D_{Q_n \cap B(v_i, a|\overline{v_i v_{i+1}}|)}(v_i, v_{i+1}) \geq \mu M_{f,i}^{-\beta} |v_i - v_{i+1}| - \frac{\varepsilon}{8K}\right\} \cap A_i. \end{aligned}$$

From Lemma 3.2 and Lemma 3.4 we get that

$$\mathbb{P}(B_i^c) \leq \exp(-c_8 m_f^{c_2} n^{c_2}) + \exp(-c_{11} n^{c_2}) \quad \forall n > n_0, \quad i = 1, 2, \dots, k-1.$$

The constants c_8 , c_{11} and n_0 depend on δ . Now choose $\delta > 0$ such that for $z, z' \in S$ with $|z - z'| < 5(a+1)\delta$ implies $|f^{-\beta}(z) - f^{-\beta}(z')| < \varepsilon_3 = \varepsilon m_f^{2\beta} / (128\mu \mathcal{D}_0(x, y) M_f^\beta)$. Denote r_i the geodesic between v_i and v_{i+1} . We have,

$$\sum_{i=0}^k M_{f,i}^{-\beta} |v_i - v_{i+1}| > \sum_{i=0}^k (1 - \varepsilon_3) \left(\min_{r_i} f\right)^{-\beta} |v_i - v_{i+1}| > \sum_{i=0}^k \left(\min_{r_i} f\right)^{-\beta} |v_i - v_{i+1}| - \frac{\varepsilon}{8\mu}$$

Since the boundary of S is C^1 , we can control the geodesic distance by the Euclidean distance uniformly in \bar{S} . More precisely, for each $x \in S$ there exists $\delta_x > 0$ such that $B(x, \delta_x) \subset S$ and consequently $\mathcal{D}(x, y) = |x - y|$ for all $y \in B(x, \delta_x)$. If $x \in \partial S$, since the boundary of S is C^1 , we have $\mathcal{D}_0(x, y) = |x - y| + o(|x - y|)$. Then, by compactness of \bar{S} , we can choose $\delta > 0$ such that $|v_i - v_{i+1}| > (1 - \varepsilon_4) \mathcal{D}_0(v_i, v_{i+1})$ with $\varepsilon_4 = \varepsilon_3/10$. If we call (r_1, r_2, \dots, r_k) the concatenation of the geodesics r_1, r_2, \dots, r_k , we have

$$\sum_{i=0}^k \left(\min_{r_i} f\right)^{-\beta} |v_i - v_{i+1}| > \int_{(r_1, r_2, \dots, r_k)} \frac{1}{f^\beta} - \frac{\varepsilon}{8\mu}.$$

Then,

$$\begin{aligned}
& \mathbb{P} \left(\sum_{i=0}^k n^\beta D_{Q_n}(v_i, v_{i+1}) \leq \mu \mathcal{D}_{f,\beta}(x, y) - \frac{\varepsilon}{2}, k \leq K, E_n \right) \\
& \leq \mathbb{P} \left(\sum_{i=0}^k \mu M_{f,i}^{-\beta} |v_i - v_{i+1}| - \frac{\varepsilon k}{8K} \leq \mu \mathcal{D}_{f,\beta}(x, y) - \frac{\varepsilon}{2}, k \leq K, E_n, \bigcap_{i=0}^k B_i \right) + \sum_{i=0}^k \mathbb{P}(B_i^c) \\
& \leq \mathbb{P} \left(\mu \int_{(r_1, \dots, r_k)} \frac{1}{f^\beta} \leq \mu \mathcal{D}_{f,\beta}(x, y) - \frac{\varepsilon}{8}, k \leq K, E_n, \bigcap_{i=0}^k B_i \right) + \sum_{i=0}^k \mathbb{P}(B_i^c). \tag{3.6}
\end{aligned}$$

Since

$$\int_{(r_1, \dots, r_k)} \frac{1}{f^\beta} \geq \mathcal{D}_{f,\beta}(x, y),$$

the first term in (3.6) is zero. Combining all these facts, we get

$$\begin{aligned}
\mathbb{P} \left(n^\beta D_{Q_n}(x, y) \leq \mu \mathcal{D}_{f,\beta}(x, y) - \varepsilon \right) & \leq \mathbb{P}(k \geq K) + \mathbb{P}(E_n^c) + \sum_{\substack{v_1, \dots, v_k \in \mathcal{V} \\ |v_i - v_{i+1}| > \delta/2}} \sum_{i=0}^k \mathbb{P}(B_i^c) \\
& \leq \exp(-c_8(M_f n)^{c_2}) + (\kappa_3 n)^2 \exp(-c_8(m_f n)^{c_2}) \\
& \quad + \exp(-c_9 n) \\
& \quad + (\kappa_3 n)^K (\exp(-c_8 m_f^2 n^{c_2}) + \exp(-c_{11} n^{c_2})) \\
& \leq \exp(-c_{13} n^{c_2}),
\end{aligned}$$

for every $n \geq n_0$ if c_{13} and n_0 are chosen adequately. This concludes the proof of the lemma and Theorem 2.3. \square

4. MANIFOLDS

We now consider the case in which the data is supported on a (possibly lower dimensional) manifold. We consider a manifold \mathcal{M} that is the image of an isometric transformation from the closure of an open connect set of \mathbb{R}^d . The proof is based on the fact that a d -dimensional manifold is locally equivalent to \mathbb{R}^d and that if in addition \mathcal{M} is smooth enough, then geodesic and Euclidean distances are similar locally.

We consider $S \subset \mathbb{R}^d$ an open connected set and a diffeomorphism $\phi : \bar{S} \mapsto \mathcal{M} := \phi(\bar{S}) \subset \mathbb{R}^D$, with $d < D$. Let $J_\phi(z) \in \mathbb{R}^{D \times d}$ be the Jacobian matrix of ϕ defined by

$$(J_\phi(z))_{ij} = \frac{\partial \phi_i}{\partial z_j}(z).$$

We assume that ϕ is an isometric transformation, i.e. for every $z \in S$ and $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$ tangent to \mathcal{M} at $\phi(z)$ we have

$$(J_\phi(z)\mathbf{v})^t (J_\phi(z)\mathbf{w}) = \mathbf{v}^t \mathbf{w},$$

which is equivalent to $J_\phi(z)^T J_\phi(z) = \mathbb{I}_d$. Here \mathbb{I}_d is the identity matrix in $\mathbb{R}^{d \times d}$. If \mathcal{M} is compact, then for every $\varepsilon_0 > 0$ there exists $\delta_0 > 0$ such that

$$(1 - \varepsilon_0)|\phi^{-1}(x) - \phi^{-1}(y)| < |x - y| < (1 + \varepsilon_0)|\phi^{-1}(x) - \phi^{-1}(y)|, \quad (4.1)$$

if $|x - y| < \delta_0$.

We first need to extend Lemma 3.3 to manifolds. The proof is straightforward and we omit it.

Lemma 4.1. *Assume $\mathcal{M} \subset \mathbb{R}^D$ is a C^1 d -dimensional manifold. Let $Q_n = \{q_1, \dots, q_n\}$ be independent random points with common density f . For $\alpha > 1$ and $x, y \in \mathcal{M}$, let (q_1, \dots, q_{k_n}) be the minimizing path. Given $\delta > 0$, there exists a positive constant c_{14} such that*

$$\mathbb{P}\left(\max_{i < k_n} |q_i - q_{i+1}| > \delta\right) \leq \exp(-c_{14} n).$$

Proof of Theorem 2.7. Given Q_n , we consider $\tilde{Q}_n = \phi^{-1}(Q_n)$, $\tilde{x} = \phi^{-1}(x)$, $\tilde{y} = \phi^{-1}(y)$. The points in \tilde{Q}_n are independent, with common density $g : S \rightarrow \mathbb{R}_{\geq 0}$ given by

$$g(z) = f(\phi(z)) \sqrt{\det(J_\phi(z)^t J_\phi(z))} = f(\phi(z)).$$

Given $\varepsilon_0 > 0$, let δ_0 be as in (4.1). Then for every path (q_1, q_2, \dots, q_k) in \mathcal{M} with $|q_i - q_{i+1}| < \delta_0$ we have

$$(1 - \varepsilon_0)^\alpha \sum_{i=1}^{k-1} |\tilde{q}_{i+1} - \tilde{q}_i|^\alpha < \sum_{i=1}^{k-1} |q_{i+1} - q_i|^\alpha < (1 + \varepsilon_0)^\alpha \sum_{i=1}^{k-1} |\tilde{q}_{i+1} - \tilde{q}_i|^\alpha.$$

Then, on the event $\{n^\beta D_{\tilde{Q}_n}(\tilde{x}, \tilde{y}) < 2\mu m_f^{-\beta} \mathcal{D}_0(x, y)\}$ we can choose ε_0 small enough to guarantee

$$|n^\beta D_{Q_n}(x, y) - n^\beta D_{\tilde{Q}_n}(\tilde{x}, \tilde{y})| < \frac{\varepsilon}{2}$$

On the other hand, since ϕ is an isometry it holds

$$\mathcal{D}_{f,\beta}(x, y) = \inf_{\gamma \subset \mathcal{M}} \int_\gamma \frac{1}{f^\beta} = \inf_{\sigma \subset S} \int_\sigma \frac{1}{g^\beta} = \mathcal{D}_{g,\beta}(\tilde{x}, \tilde{y}).$$

Finally,

$$\mathbb{P}\left(|n^\beta D_{Q_n}(x, y) - \mu \mathcal{D}_{f,\beta}(x, y)| > \varepsilon\right) \leq \mathbb{P}\left(|n^\beta D_{\tilde{Q}_n}(\tilde{x}, \tilde{y}) - \mu \mathcal{D}_{g,\beta}(\tilde{x}, \tilde{y})| > \frac{\varepsilon}{2}\right) \quad (4.2)$$

$$+ \mathbb{P}\left(n^\beta D_{\tilde{Q}_n}(\tilde{x}, \tilde{y}) < 2\mu f_{\min}^{-\beta} \mathcal{D}_0(x, y)\right). \quad (4.3)$$

$$+ \mathbb{P}\left(\max_{i < k_n} |q_i - q_{i+1}| > \delta_0\right). \quad (4.4)$$

We bound (4.2) by means of Theorem 2.3. Lemma 3.2 is used to bound (4.3) and Lemma 4.1 to bound (4.4), which concludes the proof. \square

5. THE ARC LENGTH OF GEODESICS

In this section we show a bound for the arc length of geodesics. We think this result is of independent interest. We then prove that microscopic geodesics converge to macroscopic ones.

Proof of Proposition 2.9. Denote $r_n := r_{Q_n, \alpha}(x, y)$ and $(q_1, \dots, q_{k_n}) := r_n$ the particles that form the minimizing path. Notice that k_n is the number of particles in r_n . From Hölder's inequality we have

$$|\bar{r}_n| \leq k_n^{(\alpha-1)/\alpha} D_{Q_n}(x, y)^{1/\alpha},$$

Then,

$$\begin{aligned} \mathbb{P}(|\bar{r}_n| > \ell) &\leq \mathbb{P}\left(n^\beta D_{Q_n}(x, y) (k_n n^{-1/d})^{\alpha-1} > \ell |\bar{r}_n|^{\alpha-1}\right) \\ &\leq \mathbb{P}\left(n^\beta D_{Q_n}(x, y) \left(\frac{k_n}{n^{1/d} |\bar{r}_n|}\right)^{\alpha-1} > \ell\right) \\ &\leq \mathbb{P}\left(n^\beta D_{Q_n}(x, y) > 2\mu m_f^{-\beta} \mathcal{D}_0(x, y)\right) + \mathbb{P}\left(\frac{k_n}{n^{1/d} |\bar{r}_n|} > \left(\ell/2\mu m_f^{-\beta} \mathcal{D}_0(x, y)\right)^{1/(\alpha-1)}\right). \end{aligned}$$

The first term can be bounded by means of (3.2). To bound the second one we will show the existence of positive constants c_{15}, c_{16}, c_{17} , with c_{17} depending only on δ , such that

$$\mathbb{P}\left(\frac{k_n}{n^{1/d} |\bar{r}_n|} > c_{15}\right) \leq c_{16} \exp(-c_{17} n^{1/d}). \quad (5.1)$$

Then, if we take $\ell \geq K c_{15}^{\alpha-1}$, we can conclude (2.4). The proof of (5.1) is similar to the one of Lemma 3 in [8]. Hereafter we include the adaptation of that proof to our context. Let us consider a covering \mathcal{C} of \mathbb{R}^d by closed cubes C of edge size $\varepsilon = \varepsilon_0 n^{-1/d}$ and vertices in $\varepsilon_0 n^{-1/d} \mathbb{Z}^d$. That is, if $C \in \mathcal{C}$, then $C = z + [0, \varepsilon_0 n^{-1/d}]^d$ for some $z \in \varepsilon_0 n^{-1/d} \mathbb{Z}^d$. Let $m_n = \#\{C \in \mathcal{C} : C \cap \bar{r}_n \neq \emptyset\}$. We say that two cubes (cells) C and C' are adjacent if they share a face and we denote that $C \sim C'$. We call (C_1, \dots, C_m) a *path of cells* of length m if $C_j \sim C_{j+1}$ for every $j = 1, \dots, m-1$. Let us consider the event

$$E_n^m = \left\{ \text{There exist a path } (C_1, \dots, C_m) \text{ with } \#\bigcup_{j=1}^m C_j \cap Q_n \geq \frac{m}{2d} \right\}.$$

Given m cells C_1, C_2, \dots, C_m , it is clear that $\#\bigcup_{j=1}^m C_j \cap Q_n$ is stochastically bounded by a random variable $V_m \sim \text{Poisson}(m\varepsilon_0^d M_f)$. By means of Chernoff bounds we get for $\theta \in \mathbb{R}$

that

$$\begin{aligned}
\mathbb{P}\left(\#\bigcup_{j=1}^m C_j \cap Q_n \geq \frac{m}{2d}\right) &\leq \mathbb{P}\left(V_m \geq \frac{m}{2d}\right) \\
&= \mathbb{P}\left(e^{\theta V_m} \geq e^{\theta \frac{m}{2d}}\right) \\
&\leq \exp\left(-\theta \frac{m}{2d}\right) \mathbb{E}\left(e^{\theta V_m}\right) \\
&= \exp\left(-\theta \frac{m}{2d} + m\varepsilon_0^d M_f(e^\theta - 1)\right).
\end{aligned}$$

The total number of paths of cells of length m with $x \in C_1$ is bounded above by $(2d)^m$. Then,

$$\mathbb{P}(E_n^m) \leq (2d \exp(-\theta/2d) \exp(\varepsilon_0^d M_f(e^\theta - 1)))^m.$$

Choosing $\theta > 0$ such that $(2d)e^{-\theta/2d} < e^{-1/2}$ and $\varepsilon_0 > 0$ such that $e^{\varepsilon_0^d M_f(e^\theta - 1)} < 2$, we obtain $\mathbb{P}(E_n^m) \leq e^{-m}$. Notice that any (particle) path from x to y must intersect at least $\kappa_4 \varepsilon_0^{-1} |x - y| n^{1/d}$ cells, for some geometric constant $\kappa_4 > 0$ that depends on d . Let

$$F_n = \left\{ \frac{m_n}{2d} \leq k_n \right\} \subset \bigcup_{m \geq \frac{\kappa_4}{\varepsilon_0} |x-y| n^{1/d}} E_n^m.$$

Then,

$$\mathbb{P}(F_n) \leq \sum_{m=\lfloor \frac{\kappa_4}{\varepsilon_0} |x-y| n^{1/d} \rfloor}^{\infty} \mathbb{P}(E_n^m) \leq e(1 - e^{-1})^{-1} e^{-\frac{\kappa_4}{\varepsilon_0} |x-y| n^{1/d}}.$$

Let $(C_1, C_2, \dots, C_{m_n})$ be the path of cells intersected by \bar{r}_n sorted according to r_n . That is, let $(\gamma_n(t))_{0 \leq t \leq |\bar{r}_n|}$ be the parametrization by arc length of the polygonal through (q_1, \dots, q_{k_n}) with $\gamma_n(0) = x$, $\gamma_n(|\bar{r}_n|) = y$. Then the cell-path is defined by

$$C_1 \ni x, \quad \tau_0 = 0, \quad C_j \neq C_{j-1}, \quad C_j \ni \gamma(\tau_j) \text{ with } \tau_j = \inf\{t > \tau_{j-1} : \gamma(t) \notin C_{j-1}\}$$

If F_n^c occurs, then there are at least $m_n/3d$ indices i for which d divides i , $i + d - 1 < m_n$ and $C_j \cap Q_n = \emptyset$ for all j with $i \leq j < i + d$. For each of these indices, there is a straight line that passes completely through d adjacent cells C_j and consequently crosses $d + 1$ different hyperplanes of the grid $\varepsilon \mathbb{Z}^d$. Using the Pigeonhole principle, we conclude that the straight line passes through two parallel hyperplanes separated by at least ε , that is, each line segment of \bar{r}_n that passes completely through d contiguous empty cells contributes at least ε to the length $|\bar{r}_n|$. In other words, $|\bar{r}_n| \geq \frac{m_n}{3d} \varepsilon$. Then

$$k_n \leq \frac{m_n}{2d} \leq \frac{3}{2\varepsilon_0} n^{1/d} |\bar{r}_n| \quad \text{in } F_n^c.$$

Choosing

$$c_{15} \geq \frac{3}{2\varepsilon_0}, \quad c_{16} \geq e(1 - e^{-1})^{-1}, \quad c_{17} \leq \frac{\kappa_4 \delta}{\varepsilon_0} \leq \frac{\kappa_4}{\varepsilon_0} |x - y|$$

we get (5.1). We conclude the proof by taking $c_3(\delta) = \min\{c_8(\delta), c_{17}(\delta)\}$ and from $c_2 < 1/d$. \square

We are ready to prove Corollary 2.10.

Proof of Corollary 2.10. We first need to define a topology in the space of curves contained in S . Let \mathcal{S} be the set of continuous and rectifiable curves in S . For $\gamma, \gamma' \in \mathcal{S}$ define

$$d_{\mathcal{S}}(\gamma, \gamma') = \min_{\substack{h \in P_{\gamma} \\ g \in P_{\gamma'}}} \max_{t \in [0,1]} |h(t) - g(t)|.$$

Here $P_{\gamma} = \{h: [0,1] \rightarrow S, h \text{ is a parametrization of } \gamma\}$. Notice that $d_{\mathcal{S}}(\gamma, \gamma') < \delta$ implies $\gamma \subset B(\gamma', \delta)$ and $\gamma' \subset B(\gamma, \delta)$. For every $\ell > 0$, the set $\{\gamma \in \mathcal{S}: |\gamma| \leq \ell\}$ is compact with respect to this metric, [13, Lemma 3]. Observe also that the map $\gamma \mapsto \int_{\gamma} f^{-\beta}$ is continuous from \mathcal{S} to \mathbb{R} .

For $\varepsilon_4 > 0$, we will see that the event $d_{\mathcal{S}}(\bar{r}_n, \gamma^*) \geq \varepsilon_4$ occurs finitely many times. Since γ^* is the unique minimizer, there exist $\varepsilon_5 > 0$ such that

$$\int_{\gamma^*} \frac{1}{f^{\beta}} + \varepsilon_5 < \inf_{d_{\mathcal{S}}(\gamma, \gamma^*) \geq \varepsilon_4} \int_{\gamma} \frac{1}{f^{\beta}}.$$

Given $\varepsilon > 0$, by means of Theorem 2.3 with $S = B(\gamma, \delta)$ and the compactness of $\{|\gamma| < \ell^*\}$ we get the existence of $\delta > 0$ such that for all γ with $|\gamma| < \ell^*$

$$\mathbb{P} \left(\left| n^{\beta} D_{Q_n \cap B(\gamma, \delta)}(x, y) - \mu \int_{\gamma} \frac{1}{f^{\beta}} \right| > \varepsilon \right) < \exp(-c_{18} n^{c_2}), \quad (5.2)$$

for some constant $c_{18} > 0$. Take $\varepsilon = \varepsilon_5/2$ and δ_5 such that (5.2) holds. From the compactness of bounded sets of \mathcal{S} we get the existence of a finite number of curves $\gamma^1, \gamma^2, \dots, \gamma^m \in S \setminus \{\gamma: d_{\mathcal{S}}(\gamma, \gamma^*) < \varepsilon_4\}$ such that for every $\gamma \subset S$ continuous and rectifiable, with arc length bounded by ℓ^* and such that $d_{\mathcal{S}}(\gamma, \gamma^*) \geq \varepsilon_4$, there exists γ^j with $d_{\mathcal{S}}(\gamma, \gamma^j) < \min\{\varepsilon_4, \delta_5\}$. Then

$$\begin{aligned} \mathbb{P} \left(d_{\mathcal{S}}(\bar{r}_n, \gamma^i) < \delta \right) &\leq \mathbb{P} \left(n^{\beta} D_{Q_n}(x, y) = n^{\beta} D_{Q_n \cap B(\gamma^i, \delta_5)}(x, y) \right) \\ &\leq \mathbb{P} \left(\mu \int_{\gamma^*} \frac{1}{f^{\beta}} + \frac{\varepsilon_5}{2} > n^{\beta} D_{Q_n}(x, y) = n^{\beta} D_{Q_n \cap B(\gamma^i, \delta_5)}(x, y) > \mu \int_{\gamma^i} \frac{1}{f^{\beta}} - \frac{\varepsilon_5}{2} \right) \\ &+ \mathbb{P} \left(\left| n^{\beta} D_{Q_n \cap B(\gamma^i, \delta_5)}(x, y) - \mu \int_{\gamma^i} \frac{1}{f^{\beta}} \right| > \frac{\varepsilon_5}{2} \right) \\ &+ \mathbb{P} \left(\left| n^{\beta} D_{Q_n}(x, y) - \mu \int_{\gamma^*} \frac{1}{f^{\beta}} \right| > \frac{\varepsilon_5}{2} \right). \end{aligned}$$

The first term is zero and the last two terms decay exponentially fast as $n \rightarrow \infty$. By Borel-Cantelli's lemma, the event $\{\bar{r}_n \subset S \setminus \{\gamma: d_{\mathcal{S}}(\gamma, \gamma^*) < \varepsilon_4\}\}$ occurs finitely many times with probability one, as we wanted to prove. \square

6. RESTRICTION TO NEAREST NEIGHBORS

In this section we prove that if we restrict ourselves to paths composed by k nearest neighbors, the sample Fermat distance remains unchanged with high probability when $k \approx \log n$. This reduces the computational cost from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log^2 n)$.

Proof of Proposition 2.12. Recall that Given $k \geq 1$ and $q \in Q_n$, we denote the k -th nearest neighbor of q by $q^{(k)}$ and we denote $\mathcal{N}_k(z) = \{q^{(1)}, q^{(2)}, \dots, q^{(k)}\}$ the set of k -nearest neighbors of q .

Given two points $z_1, z_2 \in S$ we define

$$A_{z_1, z_2}^\alpha = \{z \in S : |z_1 - z|^\alpha + |z_2 - z|^\alpha < |z_2 - z_1|^\alpha\}.$$

There exists a constant $\delta > 0$, that depends only on α such that $B((z_1 + z_2)/2, \delta|\overline{z_1 z_2}|) \subset A_{z_1, z_2}^\alpha$. Let q_1, q_2, \dots, q_{k_n} be the optimal path and define

$$k^* = \min \{k \in \mathbb{N} : q_{i+1} \in \mathcal{N}_k(q_i) \text{ for all } i < k_n\}.$$

We need to prove

$$\mathbb{P}(k^* > c_4 \log(n/\varepsilon) + c_5) < \varepsilon.$$

Notice that for every $1 \leq i \leq k_n$, $A_{q_i, q_{i+1}}^\alpha \cap Q_n = \emptyset$ since if it is nonempty we can construct a path with lower cost than the minimizing path. For $k \in \mathbb{N}$, define the random variable

$$\mathbf{s}_k = \sup \left\{ s : \text{there exists a ball } B_s \text{ with radius } s \text{ that contains at least } k \right. \\ \left. \text{particles and another ball } B_{\delta s} \subset B_s \text{ with radius } \delta s \text{ and } B_{\delta s} \cap Q_n = \emptyset \right\},$$

and $A_k = \{\mathbf{s}_k > 0\}$. Here we use the convention $\sup \emptyset = 0$. Since $q_{i+1} = (q_i)^{(k)}$ implies A_k , we have

$$\{k^* \geq k\} \subset \bigcup_{j=k}^{\infty} A_j. \quad (6.1)$$

Define

$$\underline{\mathbf{s}} = \frac{1}{3} \left(\frac{k}{2M_f n} \right)^{1/d}, \quad \bar{\mathbf{s}} = 2\sqrt{d} \left(\frac{2k}{m_f n} \right)^{1/d},$$

Clearly $\underline{\mathbf{s}} < \bar{\mathbf{s}}$ and

$$\mathbb{P}(A_k) = \mathbb{P}(0 < \mathbf{s}_k < \underline{\mathbf{s}}) + \mathbb{P}(\mathbf{s}_k > \bar{\mathbf{s}}) + \mathbb{P}(\mathbf{s}_k \in [\underline{\mathbf{s}}, \bar{\mathbf{s}}]). \quad (6.2)$$

We proceed to bound each term in (6.2).

$$\begin{aligned} \mathbb{P}(0 < \mathbf{s}_k < \underline{\mathbf{s}}) &\leq \mathbb{P}(\exists \text{ a ball } B_{\underline{\mathbf{s}}} \subset S \text{ with radius } \underline{\mathbf{s}} \text{ with at least } k \text{ particles}) \\ &\leq \mathbb{P}(\exists \text{ a cube } C_{2\underline{\mathbf{s}}} \subset S \text{ of edge size } 2\underline{\mathbf{s}} \text{ with at least } k \text{ particles}). \end{aligned}$$

Consider the family \mathcal{C} of cubes $C \subset \mathbb{R}^d$ with edge size $3\underline{\mathbf{s}}$ and vertices in $\underline{\mathbf{s}}\mathbb{Z}^d$. Notice that the number of elements in $\mathcal{C}_S = \{C \cap S : C \in \mathcal{C}\}$ is bounded above by $\kappa_S^1 n/k$, for some constant κ_S^1 that depends on the diameter of S . On the other hand, any cube with edge size $2\underline{\mathbf{s}}$ is strictly contained in a cube $C \in \mathcal{C}$. The number of particles in $C \cap S$ is a Poisson random variable with parameter bounded above by $3^d \underline{\mathbf{s}}^d M_f n = k/2$. Then,

$$\mathbb{P}(0 < \mathbf{s}_k < \underline{\mathbf{s}}) \leq \kappa_S^1 \frac{n}{k} e^{-\theta_1 k},$$

for some positive constant θ_1 . Next,

$$\begin{aligned} \mathbb{P}(\mathbf{s}_k > \bar{\mathbf{s}}) &\leq \mathbb{P}(\exists \text{ a ball } B_{\bar{\mathbf{s}}} \subset S \text{ with radius } \bar{\mathbf{s}} \text{ with } k \text{ particles}) \\ &\leq \mathbb{P}\left(\exists \text{ a cube } C_{\bar{\mathbf{s}}/\sqrt{d}} \subset S \text{ with edge size } \bar{\mathbf{s}}/\sqrt{d} \text{ with at most } k \text{ particles}\right). \end{aligned}$$

Now we consider the family \mathcal{C}' of cubes $C \subset \mathbb{R}^d$ with edge size $\bar{s}/(2\sqrt{d})$ and vertices in $(\bar{s}/(2\sqrt{d}))\mathbb{Z}^d$. The number of elements in $\mathcal{C}'_S = \{C \in \mathcal{C}' : C \subset S\}$ is bounded above by $\kappa_S^2 n/k$. If there is a cube $C_{\bar{s}/\sqrt{d}}$ with at most k particles, then there is $C \in \mathcal{C}'_S$ with at most k particles. The number of particles in C is Poisson with parameter at least $\bar{s}^d m_f n / (2^d d^{d/2}) = 2k$. Then

$$\mathbb{P}(\mathfrak{s}_k > \bar{s}) \leq \kappa_S^2 \frac{n}{k} e^{-\theta_2 k},$$

for some positive constant θ_2 . Finally

$$\begin{aligned} \mathbb{P}(\underline{\delta\mathfrak{s}} \leq \mathfrak{s}_k \leq \bar{s}) &\leq \mathbb{P}(\exists \text{ ball } B_{\delta\bar{s}} \subset S \text{ with radius } \delta\bar{s} \text{ and } B_{\delta\bar{s}} \cap Q_n = \emptyset) \\ &\leq \mathbb{P}\left(\exists \text{ cube } C_{\delta\bar{s}/\sqrt{d}} \subset S \text{ with edge size } \delta\bar{s}/\sqrt{d} \text{ and } C_{\delta\bar{s}/\sqrt{d}} \cap Q_n = \emptyset\right). \end{aligned}$$

We proceed as before but now with the grid $(\delta\bar{s}/2\sqrt{d})\mathbb{Z}^d$. There is at most $\kappa_S^3 n/k$ cubes with vertices in the grid and nonempty intersection with S , the number of particles in a cube is Poisson with intensity no greater than $\underline{s}^d M_f n / (2^d d^{d/2}) = k / (2^{d+1} 3^d d^{d/2})$. Then,

$$\mathbb{P}(\underline{s} \leq \mathfrak{s}_k \leq \bar{s}) \leq \kappa_S^3 \frac{n}{k} e^{-\theta_3 k},$$

with $\theta_3 = (2^{d+1} 3^d d^{d/2})^{-1}$. We conclude that

$$\mathbb{P}(A_k) \leq \kappa_S \frac{n}{k} e^{-\theta k},$$

for $\theta = \min\{\theta_1, \theta_2, \theta_3\}$ and $\kappa_S = \kappa_S^1 + \kappa_S^2 + \kappa_S^3$. By (6.1) we get

$$\mathbb{P}(k^* \geq k) \leq \sum_{j=k}^{\infty} \kappa_S \frac{n}{j} e^{-\theta j} \leq \kappa_S \frac{n}{k} (1 - e^{-\theta})^{-1} e^{-\theta k} < \kappa_S n (1 - e^{-\theta})^{-1} e^{-\theta k}.$$

So, we can guarantee $\mathbb{P}(k^* \geq k) < \varepsilon$ if

$$k > \frac{1}{\theta} \log \left(\frac{\kappa_S n}{1 - e^{-\theta}} \frac{1}{\varepsilon} \right).$$

This concludes the proof. □

ACKNOWLEDGMENTS

We want to thank Daniel Carando, Gabriel Larotonda, and Chuck Newman for enlightening conversations and acknowledge the team at Aristas, especially Yamila Barrera and Alfredo Umfurer, for useful discussions and implementation of algorithms. We also thank Steven Damelin and Daniel Mckenzie for private communications that helped us to clarify our respective contributions.

REFERENCES

- [1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory, ICDT 2001*. Springer Berlin Heidelberg, 2001.
- [2] Morteza Alamgir and Ulrike Von Luxburg. Shortest path distance in random k-nearest neighbor graphs. *arXiv preprint arXiv:1206.6381*, 2012.

- [3] Avleen S Bijral, Nathan Ratliff, and Nathan Srebro. Semi-supervised learning with density based distances. *arXiv preprint arXiv:1202.3702*, 2012.
- [4] A. Carpio, L. L. Bonilla, J. C. Mathews, and A. R. Tannenbaum. Fingerprints of cancer by persistent homology. *bioRxiv*, 2019.
- [5] Hong Chang and Dit-Yan Yeung. Robust path-based spectral clustering with application to image segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 278–285. IEEE, 2005.
- [6] Jose A. Costa and Alfred O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Process.*, 52(8):2210–2221, 2004.
- [7] Jose A. Costa and Alfred O. Hero, III. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and analysis of shapes*, Model. Simul. Sci. Eng. Technol., pages 231–252. Birkhäuser Boston, Boston, MA, 2006.
- [8] C. D. Howard and C. M. Newman. Euclidean models of first-passage percolation. *Probability Theory and Related Fields*, 108(2):153–170, 1997.
- [9] C. Douglas Howard and Charles M. Newman. Geodesics and spanning trees for Euclidean first-passage percolation. *Ann. Probab.*, 29(2):577–623, 2001.
- [10] Sung Jin Hwang, Steven B. Damelin, and Alfred O. Hero, III. Shortest path through random points. *Ann. Appl. Probab.*, 26(5):2791–2823, 2016.
- [11] Daniel Mckenzie and Steven Damelin. Power weighted shortest paths for clustering euclidean data. *Foundations of Data Science*, 1(3):307, 2019.
- [12] J. Moller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. CRC Press, 2003.
- [13] S. B. Myers. Arcs and geodesics in metric spaces. *Transactions of the American Mathematical Society*, 57(2):217–227, 1945.
- [14] Alon Orlitsky et al. Estimating and computing density based distance metrics. In *Proceedings of the 22nd international conference on Machine learning*, pages 760–767. ACM, 2005.
- [15] Mathew D Penrose, Joseph E Yukich, et al. Limit theory for point processes in manifolds. *The Annals of Applied Probability*, 23(6):2161–2211, 2013.
- [16] Michalis Potamias, Francesco Bonchi, Carlos Castillo, and Aristides Gionis. Fast shortest path distance estimation in large networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, pages 867–876, New York, NY, USA, 2009. ACM.
- [17] Facundo Sapienza, Pablo Groisman, and Matthieu Jonckheere. Weighted geodesic distance following Fermat’s principle. In *International Conference on Learning Representation*, 2018.
- [18] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

IMAS-CONICET, DEPARTAMENTO DE MATEMÁTICA, FAC. CS. EXACTAS Y NATURALES, UNIVERSIDAD DE BUENOS AIRES, ARGENTINA AND NYU-ECNU INSTITUTE OF MATHEMATICAL SCIENCES AT NYU SHANGHAI

E-mail address: `pgroisma@dm.uba.ar`

IMAS-CONICET AND INTITUTO DE CÁLCULO, FAC. CS. EXACTAS Y NATURALES, UNIVERSIDAD DE BUENOS AIRES.

E-mail address: `mjonckhe@dm.uba.ar`

ARISTAS S.R.L. AND DEPARTAMENTO DE MATEMÁTICA, FAC. CS. EXACTAS Y NATURALES, UNIVERSIDAD DE BUENOS AIRES, BUENOS AIRES, ARGENTINA.

E-mail address: `f.sapienza@aristas.com.ar`