Local Doubling Dimension of Point Sets

Aruni Choudhary * Michael Kerber†

Abstract

We introduce the notion of t-restricted doubling dimension of a point set in Euclidean space as the local intrinsic dimension up to scale t. In many applications information is only relevant for a fixed range of scales. We present an algorithm to construct a hierarchical net-tree up to scale t which we denote as the *net-forest*. We present a method based on *Locality Sensitive Hashing* to compute all near neighbors of points within a certain distance. Our construction of the net-forest is probabilistic, and we guarantee that with high probability, the net-forest is supplemented with the correct neighboring information. We apply our net forest construction to various applications including approximate Rips and Čech complexes and pair decompositions.

1 Introduction

Motivation Often, one wants to perform tasks on data which lives in high dimensional spaces. Typically, algorithms for manipulating such high dimensional data take exponential time with respect to the ambient dimension. This is frequently quoted as the "curse of dimensionality". In many cases, however, practical input instances lie on low dimensional manifolds and a natural question arises as to how do we exploit this structural property to invent computationally feasible algorithms.

A well-established approach is to define a special notion of dimension on a point set to capture its intrinsic dimension. The *doubling dimension* of a point set P is the smallest integer Δ such that every ball centered at any $p \in P$ of radius R is covered by at most 2^{Δ} non-empty balls of radius $R/2$ for any R. For instance, if P is a sample of an affine subspace of dimension d, it holds that $\Delta = \Theta(d)$, and often, $\Delta \ll d$ holds for more general samples of d-manifolds. A common goal is therefore to replace the exponential dependency on d by Δ in the complexity of geometric algorithms.

The concept of *(hierarchical) net-trees* can be visualized as a generalization of quadtrees and allows for the translation of quadtree-based algorithms (which are exponential in d) to instances with small Δ. Technically, a net-tree provides a hierarchy of *nets* which summarize the point set in terms of a clustering scheme at different scales. For n points with doubling dimension Δ, a net-tree can be constructed in expected $2^{O(\Delta)}O(n \log n)$ time, matching the time for constructing a quadtree except for replacing d with Δ [10]. As an application of particular importance, net-trees permit the efficient construction of *well-separated pair decomposition* (WSPD) which have various applications in geometric approximation, such as constructing spanners, finding approximate nearest-neighbors, approximating the diameter and the closest-pair distance.

In some applications, it is natural to investigate only an interval of scales. Applications which use net-trees can easily be adapted to ignore low scales by pruning off the lower part of the tree up to the appropriate scale. We instead concentrate on the more challenging question of upper bounding the range of scales. In such cases, the doubling dimension does not capture the intrinsic complexity of the problem at hand, since it may be defined by a ball that is beyond the range of considered scales. Moreover, the net-tree construction of [10] proceeds in a top-down fashion, considering the high scales of the point set first. Therefore, it suffers from potentially bad large-scale properties of the point set, even when these properties are irrelevant for the given application.

Contributions In this paper, we introduce the concept of t-restricted doubling dimension Δ_t, which is the smallest integer such that any ball centered at any $p \in P$ of radius $R \leq t$ is covered by at most 2^{Δ_t} non-empty balls of radius $R/2$. In this paper, we restrict our attention to the case of point sets in Euclidean space. We present an algorithm to construct a *net-forest*, which contains the relevant data of a net-tree up to scale t. The runtime of the construction depends on Δ_{Ct}, where C is a constant independent of n and d and is defined later on. We hence eliminate the dependence on the doubling dimension Δ. The major geometric primitive of our algorithm is to find all points which are at a distance at most $\Theta(t)$ from a given $p \in P$. We propose an approach based on *Locality Sensitive Hashing* (LSH) from [6]. The LSH based construction of the net-forest yields an expected runtime of $O\left(dn^{1+\rho} \log n(\log n + (14/\rho)^{\Delta_{7t/\rho}})\right)$ where $\rho \in (0,1)$ is a parameter which can be chosen to be as small as desired. Comparing this bound with the full net-tree con-

*Max Planck Institute for Informatics Saarbrücken, Germany, aruni.choudhary@mpi-inf.mpg.de

†Max Planck Institute for Informatics Saarbrücken, Germany, mkerber@mpi-inf.mpg.de

struction, our approach makes sense if $n^\rho \log n \ll 2^{O(\Delta)}$ and $\Delta_{O(t)} \ll \Delta$.

As a consequence of our result, we can construct the part of the WSPD where all pairs are in distance at most $\Theta(t)$, adapting the construction scheme of [10, Sec.5]. That means that any application of WSPD that restricts its attention to low scales can profit from our approach. Our approach also extends to the related concept of *semi-separated pair decomposition* [1].

As a further application, we show how to approximate Rips and Čech complexes using net-forests. These simplicial complexes are standard tools for capturing topological properties of a point cloud. Such a complex depends on a scale parameter; in particular, in the context of persistent homology [7], filtrations are considered, which encode complexes at various scales. In [13], an approximate Rips filtration of size at most $n(\frac{2}{\varepsilon})^{O(k\cdot\Delta)}$ has been constructed using net-trees. "Approximate" means that the exact and approximate filtrations are interleaved in the sense of [4] and therefore yield similar *persistence diagrams*, which summarize the topological properties of the point set. On the other hand, it is common to limit the construction of filtrations to an upper threshold value t. In such a scenario, our results yield a filtration of size $n(\frac{2}{\varepsilon})^{O(k\cdot\Delta_{O(t)})}$, thus replacing the exponential dependence on the doubling dimension by the $O(t)$-restricted doubling dimension. We also show how to approximate Čech complexes up to scale t, with the same size bound as for Rips complexes, building upon the framework of Choudhary et al. [5].

Organization of the paper Section 2 gives a brief overview of doubling spaces and net-trees. We introduce the concept of the restricted version of the doubling dimension in Section 3. In Section 4 we present an algorithm to construct the net-forest up to a desired scale. Our algorithm uses the concept of LSH which we detail in Section 5. In Section 6 we present applications of the net-forest. We summarize our results and conclude in Section 7.

2 Background

We fix P to be a finite point set consisting of n points throughout. As mentioned before, we restrict our attention to the Euclidean case $P \subset \mathbb{R}^d$, although some of the presented concepts could be extended to arbitrary metric spaces with some additional effort. In particular, the distance between any two points can be computed in $O(d)$ time in the Euclidean case.

Doubling dimension A *discrete ball* centered at a point $p \in P$ with radius r is the set of points $Q \subseteq P$ which satisfy $\|p - q\| \leq r$ for all $q \in Q$. The *doubling constant* [2, 14] is the smallest integer λ such that for all $p \in P$ and all $r > 0$, the discrete ball centered at p of radius r is covered by at most λ discrete balls of radius $r/2$. The *doubling dimension* Δ of P is $\lceil \log_2 \lambda \rceil$. For example, a point set that is sampled from a k-dimensional subspace has a doubling dimension of $\Theta(k)$, independent of the ambient dimension d. In contrast, the d boundary points of the standard $(d-1)$-simplex form a doubling space of dimension $\lceil \log_2 d \rceil$. Even worse, we can construct a subset of doubling dimension $\Theta(d)$ by placing $2^{\Theta(d)}$ points inside the unit ball in $\mathbb{R}^d$ such that any two points have a distance of at least $3/2$ (the existence of such a point set follows by a simple volume argument). It is NP-hard to calculate the doubling dimension of a metric [9], but it can be approximated within a constant factor [10, Sec.9].

Nets and Net-trees A subset $Q \subseteq P$ is an (α, β)-*net*, denoted by $\mathcal{N}_{\alpha,\beta}$, if all points in P are in distance at most α from some point in Q and the distance between any two points in Q is at least β. Usually, α and β are coupled, that is, $\beta = \Theta(\alpha)$, in which case we talk about a *net at scale* α.

We can represent a nested sequence of nets for increasing scales α using a rooted tree structure, called the *net-tree* [10]. It has n leaves, each representing a point of P, and each internal node has at least two children. Every tree-node v represents the subsets of points given by the sub-tree rooted at v; we denote this set by P_v. Every v has a representative, $\text{rep}_v \in P_v$ that equals the representative of one of its children if v is not a leaf. Moreover, v is associated with an integer $\ell(v)$ called the *level* of v which satisfies $\ell(v) < \ell(\text{parent}(v))$, where $\text{parent}(v)$ is the parent of v in the tree. Finally, each node satisfies the following properties

- *Covering* : $P_v \subseteq \mathbb{B}(\text{rep}_v, \frac{2\tau}{\tau-1} \cdot \tau^{\ell(v)})$
- *Packing* : $P_v \supseteq P \bigcap \mathbb{B}(\text{rep}_v, \frac{\tau-5}{2\tau(\tau-1)} \cdot \tau^{\ell(\text{parent}(v))})$

where $\mathbb{B}(p, r)$ denotes the ball centered at p with radius r and $\tau = 11$.

The covering and packing properties ensure that each node v has at most $\lambda^{O(1)}$ children where λ is the doubling constant for P. Moreover, for any α, a net at scale α can be accessed from the net-tree immediately; see [10, Prop.2.2] for details. A net-tree can be constructed deterministically in time $2^{O(\Delta)}O(dn \log(n \cdot \Phi))$ where Φ represents the *spread* of P, using the greedy clustering scheme of Gonzalez [8] as a precursor to the tree construction. The dependence on the spread can be eliminated by constructing the tree in $2^{O(\Delta)}dn \log n$ time in expectation (the additional factor of d compared to [10] accounts for the fact that we work with the Euclidean metric, and therefore take into account the cost of computing distances in our computational model). The net-tree construction is oblivious to knowing the value of

Δ. One can extract a net at scale ℓ [10, Pro.2.2] by collecting the set of nodes from T satisfying the condition $\mathcal{N}(\ell) = \{\text{rep}_v | \ell(v) < \ell \leq \ell(\text{parent}(v)\}$. The net-tree can be *augmented* to maintain, for each node u, a list of close-by nodes with similar diameter. Specifically, for each node u the data structure maintains the set

$$\text{Rel}(u) := \{v \in T \mid \ell(v) \leq \ell(u) < \ell(\text{parent}(v)) \text{ and } \|\text{rep}_u - \text{rep}_v\| \leq 14\tau^{\ell(u)}\}.$$

Rel(.) is computed during the construction without additional cost.

3 t-restricted doubling dimension

Definition 1 *The* t-restricted doubling constant *of* P *is the smallest positive integer* λ_t *such that all the points in any discrete ball centered at* $p \in P$ *of radius* r *with* $r \leq t$ *are covered by at most* λ_t *non empty balls of radius* $r/2$. *The corresponding* t-restricted doubling dimension Δ_t *is* $\lceil \log \lambda_t \rceil$.

By definition, $\Delta_t \leq \Delta$ for any P. More precisely, Δ_t is zero for t smaller than the closest-pair distance of P, and equals Δ when t is the diameter of P. While Δ for samples from an affine subspace of dimension k is bounded by $\Theta(k)$, this is not generally true for samples of k-manifolds where Δ increases due to curvature. To sketch an extreme example, consider an *almost space-filling* curve γ in $\mathbb{R}^d$ which has distance at most ε to any point of the unit ball, where ε is chosen small enough. We let P be a sufficiently dense sample of γ. While $\Delta_t = 1$ for small values of t, we claim that $\Delta_t = \Theta(d)$ for $t = 1$; indeed, any sparser covering of the unit ball with balls of radius $1/2$ would leave some portion of the ball uncovered, and by construction, γ goes through that uncovered region, so that some point in P is missed.

The "badness" of the previous example stems from the difference between Euclidean and geodesic distance of points lying on a lower-dimensional manifold. A common technique for approximating the geodesic distance is through the *shortest-path metric*: Let $G = (P, E)$ denote the graph whose edges are defined by the pairs of points of Euclidean distance at most t; we call such a graph a t-intersection graph. The distance of two points p and q is then defined as the length of the shortest path from p to q (we assume for simplicity that G is connected). The concept of doubling dimensions extends to any metric space and we let Δ' denote the doubling dimension of P equipped with the shortest path metric. While Δ_t and Δ' appear to be related, Δ' can be much larger than Δ_t: an example is shown in Figure 1. Moreover, using the shortest-path metric raises the question of how to compute shortest path distances efficiently, if the cost of metric queries is taken into account.

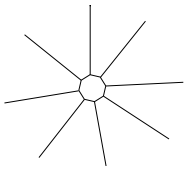

Figure 1: Consider a regular k-gon in the plane with all vertices on the unit circle. To each vertex, attach an "arm" of length 4 in the direction outwards the origin. We call the endpoint of such an arm a *tip*. Let M denote the obtained shape (as depicted). Let P denote a sufficiently dense point sample of M. Let G be the t-intersection graph on P; we choose t appropriately so that the shortest path metric on G approximates the distances on M very closely. Fixing an arbitrary vertex of the k-gon, the furthest tip has a distance of at most $4 + \pi < 8$ on M, so there is a ball of radius less than 8 containing all of P. However, any pair of tips has a distance of more than 8, so no ball of radius less than 4 can contain two tips. It follows that any covering of P with balls of radius less than 4 requires at least one ball per tip. Therefore, the doubling dimension Δ' of this metric is at least $\lceil \log_2 k \rceil$. On the other hand, for the Euclidean metric on a plane, we can easily see that the doubling dimension $\Delta = O(1)$, and therefore, Δ_t is a constant as well.

4 Net-forests

We next define an appropriate data structure for point sets of small t-restricted doubling dimension, where t is a parameter of the construction. Informally, a *net-forest* is the subset of a net-tree obtained by truncating all nodes above scale t. More precisely, it is defined as a collection of net-trees with roots $v_1, \ldots, v_k$ such that the representatives $\text{rep}_{v_1}, \ldots, \text{rep}_{v_k}$ form a (t, t)-net and the point sets $P_{v_1}, \ldots, P_{v_k}$ are disjoint and their union covers P. We define Rel(u) for a node in the forest the same way as for net-trees: it is the set of net-forest nodes that are close to u and have similar diameter. As for net-trees, we call a net-forest *augmented* if each node u is equipped with Rel(u).

Construction Our algorithm for constructing a net-forest is a simple adaptation of the net-tree algorithm: we construct a (t, t)-net of P by clustering the point set and assign each point in P to its closest net-point. Each root in the net-forest represents one of the clusters. We also compute Rel(u) for each root by finding the close-by clusters to u. Having this information, we can simply run the net-tree algorithm from [10] individually on each cluster to construct the net-forest. For augmenting it, we use the top-down traversal strategy as described in [10, Sec.3.4], inferring the neighbors of a node from the neighbors of its parent –since we have set up Rel($\cdot$)

for the roots of the forest, this strategy is guaranteed to detect neighboring nodes even if they belong to different trees of the forest.

Both the initial net construction and the Rel(·)-construction require the following primitive for a point set Q, which we call a *near-neighbor query*: *Given a point $q \in Q$ and a radius r, return a list of points in Q containing exactly the points at distance r or smaller from q).* In the remainder of the section, we give more details on how to compute the net and the associated clusters, and how to find the neighbors for each such cluster, assuming that we have a primitive which can perform near-neighbor queries. In Section 5, we show the implementation of such a primitive.

Net construction We construct the net using a greedy scheme: For any input point, store a pointer $\mathrm{N}(p)$ pointing to the net point assigned to point p. Initially, $\mathrm{N}(p) \leftarrow \varnothing$ for all p. As long as there is a point p with $\mathrm{N}(p) = \varnothing$, we set $\mathrm{N}(p) \leftarrow p$ and query the near-neighbor primitive to get a list of points with distance at most t from p. For any point q in the list we update $\mathrm{N}(q) \leftarrow p$ if either $\mathrm{N}(q) = \varnothing$ or $\|p - q\| < \|N(q) - q\|$. Then we pick the next point p with $\mathrm{N}(p) = \varnothing$.

At the end, the set of points p with $\mathrm{N}(p) = p$ represents the net at scale t and the points q satisfying $\mathrm{N}(q) = p$ constitute p's cluster. The net thus constructed is a (t, t)-net. Moreover, we set $\ell(v) = \lfloor \log_\tau \left(\frac{\tau-1}{2\tau} t\right) \rfloor$ for each root node v.

Computing the Rel(.) set for the roots After computing the net-points and their respective clusters, we need to augment the net-points with neighboring information. Recall that $\mathrm{Rel}(u)$ contains nodes in distance at most $14\tau^{\ell(u)}$ from rep_u. Since we have a (t, t)-net, the level of any root node u satisfies $14\tau^{\ell(u)} \leq 7t$. Hence we need to find neighbors of net-points within $7t$, the minimum distance between any two net-points being more than t. By the doubling property, any root net-node can have at most $\lambda_{7t}^{\log_2 \frac{7t}{t/2}}$ such neighbors which simplifies to $14^{\Delta_{7t}}$. We use the near-neighbor primitive to compute such neighbors.

5 Near-neighbors primitive

We describe the primitive used in the previous section which performs near-neighbor queries. Our approach follows the notion of *locality-sensitive hashing* (LSH) introduced by [11] for the Hamming metric and extended to Euclidean spaces in [6]. LSH is a popular approach to find approximate near-neighbors in high dimensions because of its near linear complexity in n and d.

Locality Sensitive Hashing LSH applies several hash functions on a point set such that close points are more likely to map to the same hash-buckets than points which are sufficiently far apart. A typical application of LSH is the (r, c)-*nearest neighbor* problem: If there exists a point within distance r of the query point q, report some point within distance cr of q, $c > 1$.

However, for our construction we wish to solve the following problem: report *all* points within distance r of the query point. We need the LSH oracle for two steps in our construction: constructing the net at scale t and computing the Rel(·) for the root-nodes. We show that both these steps requires a runtime sub-quadratic in n by a slight modification of the method presented in [6]. We repeat some of their definitions for clarity:

Definition 2 *A family of hash functions $\mathcal{H} = \{h : S \to U\}$ is called* (r_1, r_2, p_1, p_2)-sensitive *if for all $a, b \in S$, the following holds for $p_1 \geq p_2$ and $r_1 \leq r_2$:*

- *if $\|a - b\| \leq r_1, P_{r_1} = P[h(a) = h(b)] \geq p_1$*
- *if $\|a - b\| \geq r_2, P_{r_2} = P[h(a) = h(b)] \leq p_2$*

We amplify the gap between P_{r_1} and P_{r_2} by concatenating k such hash functions, creating the family of hash functions $\mathcal{G} = \{g : S \to U^k\}$ such that $g(x) = (h_1(x), h_2(x), ..., h_k(x))$. For $g(x)$, we have the modified properties:

- if $\|a - b\| \leq r_1, P[g(a) = g(b)] \geq p_1^k$
- if $\|a - b\| \geq r_2, P[g(a) = g(b)] \leq p_2^k$

We describe our near-neighbor primitive next: The input is a point set Q with n points and a distance $r > 0$. As a pre-processing step, we choose l hash functions $g_1, \cdots, g_l$ uniformly at random from $\mathcal{G}$ [6, Sec.3]. and hash each $p \in Q$ to the buckets $g_i(p) \forall i \in [1, l]$. Given a query point $q \in Q$ (we *only* consider near neighbor queries for points within Q), we check for any point p in bucket $g_i(q)$ whether the distance to q is at most r. We output such points as the near-neighbors of q.

We need to specify the parameters of LSH in the above description. Most importantly, we have to ensure that, with high probability, the output contains all points in distance r from q. Moreover, we want the buckets to be of small size so that the primitive does not have to filter out too many false positives.

The performance of the LSH scheme depends upon a parameter $\rho \in (0, 1)$ which appears as an exponent of n in the runtime. We choose the parameters p_1, p_2, r_1 and r_2 of the hashing scheme such that $\rho = \frac{\log p_1}{\log p_2} \approx \frac{r_1}{r_2}$ [6, Sec.4]. In the following parts of the section, we let $\rho = \frac{r_1}{r_2}$.

Lemma 1 *Let $r_1 := r$ and $r_2 := r/\rho$, $k := \lceil -\log_{p_2} n \rceil$ and $l := \lceil 2n^\rho \ln \frac{n}{\sqrt{\delta}} \rceil$ with an arbitrarily small constant δ. The near-neighbor primitive has the following properties:*

(i) With probability at least $1 - \delta$, all points in distance at most r are reported for all query points.

(ii) For any query point q, the aggregate expected size of all buckets $g_1(q), ..., g_l(q)$ is at most $l(\widetilde{C}+1)$, where $\widetilde{C}$ is the number of points in Q with distance at most r_2 to q.

(iii) The pre-processing runtime is $O(dnkl)$ and the expected query runtime for a point is $O(dl(k+\widetilde{C}))$, where $\widetilde{C}$ is defined as in (ii).

We defer the proof to Appendix A.

Net-forest construction using LSH We analyze the complexity of our net-forest construction from Section 4 with our near-neighbor primitive in Lemma 5 under Appendix B. The primitive is used in the construction of the (t,t)-net, where we find the near-neighbors in distance at most t for a subset of points that form the net in the end. That means, we initialize the primitive with $r \leftarrow t$ and $Q \leftarrow P$.

The second appearance of the near-neighbor primitive is in the construction of the Rel(.) sets for the roots of the net forest. Recall that the roots are represented by the net-points constructed before; let M denote the set of net-points and $|M|m$.We simply have to find all pairs of points of distance at most $7t$ among the net-points; and to do so we call the near neighbor primitive with $r \leftarrow 7t$ and $Q \leftarrow M$ for all $q \in M$ (see Lemma 6).

Theorem 2 *The expected time for constructing the net-forest using LSH is*

$$O\left(dn^{1+\rho}\log n\left(\log n + \left(\frac{14}{\rho}\right)^{\Delta_{7t/\rho}}\right)\right).$$

We defer the details of the proof to Appendix B.

We see how the choice of ρ affects the complexity bound: For ρ very close to zero, we get a almost linear complexity in n, to the price that we have to consider larger balls in our algorithm and thus increase the restricted doubling dimension.

6 Applications

Well-Separated Pair Decomposition A pair of net-tree nodes (u,v) is *ε-well-separated* if it satisfies $\max\{\mathrm{diam}(P_u), \mathrm{diam}(P_v)\} \leq \varepsilon\|\mathrm{rep}_u - \mathrm{rep}_v\|$, where $\mathrm{diam}(P_u)$ denotes the *diameter* of the points stored in u. Informally speaking, all pairs of points (p,q) with $p \in P_u$, $q \in P_v$ have a similar distance to each other if (u,v) is well-separated. An *ε-well-separated pair decomposition* (ε-WSPD) [3] is a collection of ε-well-separated pairs such that for any pair $(p,q) \in P \times P$, there exists a well-separated pair (u,v) such that $p \in P_u$ and $q \in P_v$; we say that such a pair (p,q) is *covered* by (u,v).

An ε-WSPD of size $n\varepsilon^{-O(\Delta)}$ can be computed in time $d\left(2^{O(\Delta)}n\log n + n(1/\varepsilon)^{O(\Delta)}\right)$ [10]. A WSPD considers pairs over all scales of distance, because it has to cover any pair of points. Our structure only requires that all pairs of points in distance at most t are covered. We call the resulting structure a *t-restricted ε-WSPD*.

We construct the t-restricted ε-WSPD as follows: We start by constructing the corresponding augmented net forest; let $u_1, \ldots, u_m$ be its roots. Since we know the Rel$(\cdot)$ set for any root, we can identify pairs (u_i, u_j) such that u_i is in Rel(u_j) and vice versa (this also includes pairs where $u_i = u_j$). For any such pair, we call `genWSPD`(u_i, u_j) from [10, Sec.5], which simply traverses the sub-trees until it finds well-separated pairs. We output the union of all pairs generated in this way.

Theorem 3 *For $0 < \varepsilon < 1$ and $t > 0$, our algorithm computes a t-restricted ε-WSPD of size $n\varepsilon^{-O(\Delta_{7t})}$ in expected time $NF + dn\varepsilon^{-O(\Delta_{7t})}$, where NF is the complexity for computing the net-forest from Theorem 2.*

We defer the proof to Appendix C.

Semi-Separated Pair Decomposition(SSPDs) SSPDs [1] are a related concept to the WSSDs with some advantages. We briefly describe them and present our *t-restricted* version of SSPDs in Appendix D.

Approximating Simplicial Complexes For a point set P, let $\mathrm{rad}(Q)$ denote the radius of the minimum enclosing ball of $Q \subseteq P$. The Čech complex on P at scale r is defined as: $\check{C}ech(r) := \{Q \subseteq P \mid \mathrm{rad}(Q) \leq r\}$. The Rips complex is defined as: $\mathrm{Rips}(r) := \{Q \subseteq P \mid \mathrm{diam}(Q) \leq 2r\}$. Rips and Čech complexes are standard constructions for topological analysis of point clouds; more precisely, one constructs a sequence of Rips or Čech complexes for growing r (called a *filtration*) and tracks the evolution of topological features in the process. This gives rise to the *persistence diagram* [7], a multi-scale summary of the topological properties of the point cloud. However, a major computational drawback of this approach is that both Rips and Čech complexes become prohibitively large when P has high ambient dimension. A common remedy is to bound the maximal dimension k of the simplices constructed in the filtration (bounding the filtration size to $O(n^{k+1})$) and/or bound the maximal scale r which is considered.

Several recent works have come up with *approximate Rips (Čech) filtrations* of significantly smaller size. In this case, "approximate" means that the persistence diagram of the exact and approximate filtrations are ε-close in *interleaving distance* [4]. Sheehy [13] showed how to compute an $(\frac{1}{1-2\varepsilon})$-approximate Rips filtration of size $n(\frac{2}{\varepsilon})^{O(k\cdot\Delta)}$ in expected time $d2^{O(\Delta)}n\log n + n(\frac{2}{\varepsilon})^{O(\Delta)}$, where k denotes the maximal dimension of the constructed approximation. His algorithm works by creating simplicial complexes on a hierarchically sparsified point set, obtained through the net-tree. In [5], Choudhary et

al. gave an alternative algorithm for $(1+\varepsilon)$-approximate Čech filtrations with the same guarantees and running time, building upon the framework from [12].

For the case that the highest scale in the construction is bounded for t, our results imply that Δ can be replaced with Δ_{7t} in the size bound above.

Theorem 4 *An approximate filtration of size at most $n(\frac{2}{\varepsilon})^{O(k\cdot\Delta_{7t})}$ can be computed for both* Rips *and* Čech *filtrations for the range $[0,t]$.*

Proof. For Rips-complexes, this follows directly by applying the algorithm of [13, Sec.10] on a net forest of scale t. Since $w = \{u \cup \mathrm{Rel}(u)\}$ captures all edges of length at most $5t > t$ between a point $p \in P_u$ and $q \in P_w$, the algorithm is able to compute the $E(p)$ sets [13, Sec.10] successfully, thereby yielding a filtration of size at most $n(\frac{2}{\varepsilon})^{O(k\cdot\Delta_{7t})}$.

To prove the same for Čech complexes, we need to apply our net-forests to a generalization of WSPDs (called well-separated simplicial decomposition) first; we postpone the details to Appendix E. □

Approximating the t-doubling dimension One can approximate Δ_t for any point set P up to a constant factor by constructing a net-forest T of scale t over P. Let x denote the maximum out-degree of any node in T. Then $\log x$ is a constant-factor approximation of Δ_t. This follows directly from the arguments of [10, Sec.9].

7 Conclusion and future work

In this paper we presented an algorithm to construct a hierarchical net-forest up to a certain scale and applied it to the construction of WSPDs, SSPDs, and approximate Rips and Čech complexes. Finding more applications tailored for the t-restricted doubling dimension is an appealing research direction we would like to look into.

Acknowledgement This research is supported by the Max Planck Center for Visual Computing and Communication.

References

[1] Mohammad A. Abam and Sariel Har-Peled. New constructions of SSPDs and their applications. *Comput. Geom. Theory Appl.*, 45(5-6):200–214, July 2012.

[2] Patrice Assouad. Plongements Lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Societe Mathematique de France*, 111:429–448, 1983.

[3] Paul B. Callahan and S. Rao Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *J. ACM*, 42(1):67–90, January 1995.

[4] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, pages 237–246, 2009.

[5] Aruni Choudhary, Michael Kerber, and R. Sharathkumar. Approximate Čech complexes in low and high dimensions. `http://people.mpi-inf.mpg.de/~achoudha/Files/Papers/AppCech.pdf`.

[6] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pages 253–262, 2004.

[7] Herbert Edelsbrunner and John Harer. *Computational Topology. An Introduction.* American Mathematical Society, 2010.

[8] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.

[9] Lee-Ad Gottlieb and Robert Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Discrete Math.*, 27(4):1759–1769, 2013.

[10] Sariel Har-Peled and Manor Mendel. Fast construction of nets in low dimensional metrics, and their applications. In *SIAM J. Comput*, pages 150–158, 2005.

[11] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, 1998.

[12] Michael Kerber and R. Sharathkumar. Approximate Čech complexes in low and high dimensions. In *International Symposium on Algorithms and Computation*, pages 666–676, 2013.

[13] Donald R. Sheehy. Linear-size approximations to the Vietoris-rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013.

[14] Kunal Talwar. Bypassing the embedding: Algorithms for low dimensional metrics. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 281–290, 2004.

A Proof of Lemma 1

Proof. First we bound the expected aggregate size of the buckets. A bucket contains "close" points which are in distance at most r_2 from q and "far" points which are further away. However, since the probability of a far point falling in the same bucket as q is at most p_2^k, the expected size of a single bucket is at most $\widetilde{C}+np_2^k \leq \widetilde{C}+1$ by our choice of k. Since there are l buckets, (ii) is satisfied.

For (i), fix two points $q_1, q_2 \in Q$ such that $\|q_1 - q_2\| \leq r_1$. We have to ensure that $g_j(q_1) = g_j(q_2)$ for some $j \in \{1, \ldots, l\}$; this implies that q_1 will be reported for query point q_2, and vice versa for at least one of the l buckets. The probability for $g_j(q_1) = g_j(q_2)$ for a fixed j is at least p_1^k, which is $p_1^{-\log_{p_2} n} = n^{-\rho}$. Hence the probability that $g_j(q_1) \neq g_j(q_2)$ holds for all $j \in \{1, \ldots, l\}$ is at most $(1 - n^{-\rho})^l$ because we choose the hash functions uniformly at random. There are at most n^2 point pairs within distance at most r_1. By the union bound, the probability that at least one such pair maps into different buckets is at most $n^2(1 - n^{-\rho})^l$. Now we can bound

$$\begin{aligned} n^2(1 - n^{-\rho})^l &= n^2(1 - n^{-\rho})^{2n^\rho \ln \frac{n}{\sqrt{\delta}}} \\ &= n^2(1 - \frac{1}{n^\rho})^{n^\rho \ln \frac{n^2}{\delta}} \leq n^2 e^{-\ln \frac{n^2}{\delta}} = \delta, \end{aligned}$$

where we used the fact that $(1 - 1/x)^x \leq 1/e$ for all $x \geq 1$. It follows that the probability that all pairs of points in distance at most r_1 fall in at least one common bucket is at least $1 - \delta$. This implies (i).

It remains to show (iii): in the pre-processing step, we have to compute $k \cdot l$ hash functions for n points. Computing the hash value for a point p, $h_i(p)$ takes $O(d)$ time [6, Sec.3.2]. For a query, we have to identify the buckets to consider in $O(dkl)$ time and then iterate through the (expected) $l(\widetilde{C} + 1)$ candidates (using (ii)), spending $O(d)$ for each. □

B Proof of Theorem 2

To prove Theorem 2, we first need to prove two supporting Lemmas, which we do next.

Lemma 5 *The expected time to construct the (t, t)-net using LSH is*

$$O\Big(dn^{1+\rho} \log n \Big(\log n + \left(\frac{2}{\rho}\right)^{\Delta_{t/\rho}} \Big)\Big).$$

Proof. We consider the time spent on all near-neighbor queries: Let the resulting net consist of $m \leq n$ points. This implies that the algorithm proceeds in m rounds and queries the near-neighbors of m points. Let $\widetilde{C}_i$ denote the number of points in distance t/ρ from the i-th query point. By Lemma 1, the total complexity for all near-neighbor queries is:

$$O\left(ndkl + \sum_{i=1}^{m} dl(k + \widetilde{C}_i)\right) = O(ndkl + dl \sum_{i=1}^{m} \widetilde{C}_i)$$

We only need to bound the sum of the $\widetilde{C}_i$. For that, we fix some $q \in P$ and count in how many sets $\widetilde{C}_i$ may it appear. Let p_i denote the net-point chosen in the i-th iteration. We call such a net-point *close* to q if the distance to q is at most t/ρ. By definition, the net-points close to q lie in a ball of radius t/ρ centered at q. Since any pair of net-points has a distance of more than t, any ball of radius $t/2$ can contain at most one close net point. Following the definition of the t-restricted doubling dimension, the number of such net-points can be at most $\lambda_{t/\rho}^{\log_2 \frac{t/\rho}{t/2}}$ which simplifies to $\left(\frac{2}{\rho}\right)^{\Delta_{t/\rho}}$. It follows that

$$\sum_{i=1}^{m} \widetilde{C}_i \leq n \left(\frac{2}{\rho}\right)^{\Delta_{t/\rho}}.$$

Hence, we get the claimed running time, observing that $k = O(\log n)$ and $l = O(n^\rho \log n)$ by Lemma 1. All additional operations in the net construction besides the calls of the primitive are dominated by that complexity. □

Lemma 6 *Computing the* Rel(.) *sets using LSH takes expected time*

$$O\left(dn^{1+\rho} \log n \left(\log n + \left(\frac{14}{\rho}\right)^{\Delta_{7t/\rho}}\right)\right).$$

Proof. The proof is analogous to the proof of Lemma 5: Let $\widetilde{C}_i$ (for $i = 1, \ldots, m$) denote the number of net-points in distance at most $\frac{7t}{\rho}$ to the i-th net point. The same packing argument as in the previous Lemma shows that any $\widetilde{C}_i$ can be at most $(\frac{14}{\rho})^{\Delta_{7t/\rho}}$, so that their sum is bounded by $m(\frac{14}{\rho})^{\Delta_{7t/\rho}}$. Analogous to the proof of Lemma 5, we can thus bound the runtime to be as required, noting that $m \leq n$. □

B.1 Proof of Lemma 2

Proof. Using Lemma 5 and Lemma 6, constructing the net and its Rel(·) sets are within the complexity bound. Constructing a single net-tree for a node containing n_i points takes time at most $2^{14\Delta_{2t}} n_i \log n_i$ (the factor of 14 in the exponent can be seen by a careful analysis of [10, Sec.3.4]). Constructing individual net-trees for the clusters takes time: $\sum_{i=1}^{m} 2^{14\Delta_{2t}} dn_i \log n_i$ Since $\sum_{i=1}^{m} n_i = n$, the above runtime simplifies to be at most $2^{14\Delta_{2t}} dn \log n$. Augmenting the net-forest takes time $dn2^{14\Delta_{7t}}$ [10, Sec.3.4]. The runtimes for the latter steps are dominated by the Rel(·) and net construction for sufficiently large values of n. □

C Proof of Theorem 3

Proof. For correctness, any pair of nodes generated is ε-well-separated by definition. For the relaxed covering property, consider a pair (p, q) of points in distance at most t. There are roots u_1, u_2 in the net-forest with $p \in P_{u_1}$ and $q \in P_{u_2}$. Since the diameter of u_1 and u_2 is at most $2t$, the distance of rep_{u_1} and rep_{u_2} is at most $5t < 7t$. Therefore, $u_2 \in \text{Rel}(u_1)$ (and vice versa), and there will be a pair generated that covers (p, q).

For the size bound, we can use the same charging argument as in [10, Sec.5]. We can additionally ensure by our construction that in all doubling arguments, the radius of the balls in question is at most $7t$ and therefore replace the doubling dimension by Δ_{7t} in the bound. The running time follows because the number of recursive calls of `genWSPD` is proportional to the output size, and we spend $O(d)$ time per recursion step. □

D Semi-Separated Pair Decomposition

A pair of net-tree nodes (u, v) is called *ε-semi-separated* if $\min(diam_u, diam_v) \leq \varepsilon \|\text{rep}_u - \text{rep}_v\|$. An *$\varepsilon$-semi-separated pair decomposition* (ε-SSPD) is a collection of ε-semi-separated pairs such that for any pair of points $(p, q) \in P \times P$ there exists a semi-separated pair (u, v) such that $p \in P_u$ and $q \in P_v$; we say that such a pair (p, q) is *covered* by (u, v). This is a weaker notion than the WSPD, because it only requires that the *smaller* diameter of the participating sets be small compared to the distance between them. The advantage of using a SSPD over a WSPD is reduced weight: The weight of a WSPD W is defined as $w = \sum_{(x,y)\in W}(|x| + |y|)$. The weight of any WSPD can be $\Omega(n^2)$ in the worst case [1]. In contrast, it is possible to construct a SSPD with near linear weight, where the weight in question is an analogous definition to the one used for WSPDs. An ε-SSPD of expected weight $O(\varepsilon^{-O(\Delta)} n \log n)$ can be calculated in $O(\varepsilon^{-O(\Delta)} n \log n)$ expected time [1].

We adapt the algorithm of [1, Sec.4] to compute a t-restricted SSPD, which requires that only pairs of points in distance at most t be covered. The construction is as follows: first we construct the net-forest at scale t. Let $u_1, \cdots, u_m$ denote the root nodes of the net-forest. We invoke the algorithm of [1, Sec.4] on each of the sets $\{u_i \cup \text{Rel}(u_i)\}$. We output the union of the pairs generated by each invocation. The SSPD generated this way also ensures coverage of points which are at most $5t$ apart and hence, covers all pairs of points at most t apart. Moreover, the pairs involved in the SSPD have the additional property that their diameter is at most $16t$.

Theorem 7 *For $0 < \varepsilon < 1$, our algorithm computes a t-restricted ε-SSPD of P of expected weight $(\frac{2}{\varepsilon})^{O(\Delta_{16t})} O(n \log n)$ in $(\frac{2}{\varepsilon})^{O(\Delta_{16t})} O(n \log n)$ expected time, after computing the net forest at scale t.*

Proof. To prove correctness, consider points p and q such that $\|p - q\| \leq 5t$. Let u_i and u_j be the root nodes of the net-forest covering p and q respectively. By a simple application of the triangle inequality, $\|\text{rep}_{u_i} - \text{rep}_{u_j}\| \leq 7t$. Hence, $u_i \in \text{Rel}(u_j)$ and vice versa. This ensures that the algorithm creates a semi-separated pair which covers the pair (p, q). For the diameter, let us say that we work with the set $P_m = \{u \cup \text{Rel}(u)\}$ and wish to upper bound $\text{diam}(P_m)$. For any $u_i, u_j \in \text{Rel}(u)$ covering points a and b respectively, it follows from the triangle inequality that $\|a - b\| \leq 16t$. Hence the claim about the diameter is true.

Next we bound the weight of the SSPD. Consider the individual invocations of the algorithm of [1] after constructing the net-forest. Let N_i denote the number of points in $u_i \cup \text{Rel}(u_i)$. Then the total weight of the SSPD is $\sum_{i=1}^{m} O(\varepsilon^{-O(\Delta_{16t})} N_i \log N_i) \leq O(\varepsilon^{-O(\Delta_{16t})} \log n) \sum_{i=1}^{m} N_i$. To bound the sum $S = \sum_{i=1}^{m} N_i$, consider any point p covered by u_i. This point participates in at most $2^{O(\Delta_{7t})}$ invocations of the algorithm, since that is precisely the maximum possible size of the Rel(.) set. This implies that the contribution of point p to S is at most $2^{O(\Delta_{7t})}$. Hence the sum S is at most $2^{O(\Delta_{7t})} n$, and bound follows. A similar argument bounds the runtime as well. □

E Approximate Čech Complexes

Well Separated Simplicial Decomposition The concept of well-separated simplicial decomposition (WSSDs) of point sets, introduced by Kerber and Sharathkumar [12] and extended to doubling spaces by Choudhary et al [5], generalizes the concept of WSPD to larger tuples. A $(k+1)$-tuple $(v_0, v_1, \ldots, v_k)$ is called *ε-well separated* if each v_i is a node of the net-tree and for any ball $\mathbb{B}$ which contains at least one point of each v_i, it holds that

$$v_0 \cup v_1 \cup .... \cup v_k \subseteq (1 + \varepsilon)\mathbb{B}$$

where $(1+\varepsilon)\mathbb{B}$ denotes a ball with same center as $\mathbb{B}$ and radius multiplied by $(1 + \varepsilon)$. An (ε, k)-WSSD is a set of ε-well-separated tuples of length $(k + 1)$ such that any k-simplex is covered by some tuple. In [5], an (ε, k)-WSSD of size $n(2/\varepsilon)^{O(\Delta \cdot k)}$ is constructed in expected time $d\left(2^{O(\Delta)} n \log n + n(2/\varepsilon)^{O(\Delta \cdot k)}\right)$.

Similar as before, we define a *t-restricted (ε, k)-WSSD* to be a collection of ε-well-separated tuples such that each k-simplex that fits into a ball of radius t is covered by a tuple. The statement is equivalent to the condition that the radius of the smallest minimum enclosing ball containing points from each node of the tuple is at most t.

Construction of the t-restricted WSSD We describe the algorithm to construct the t-restricted (ε, k)-WSSD and prove its correctness and runtime. In this appendix, we will heavily rely on the notations, algorithms, and results presented in [5]. The algorithm proceeds iteratively; for $k = 1$, we construct a $(2t)$-restricted $\varepsilon/2$-WSPD using the algorithm from Section 6. To construct Γ_{k+1} from Γ_k, we iterate over the tuples $\gamma \in \Gamma_k$. We use the scheme of [5, Sec.3], computing an approximate meb of γ and then exploring ancestors of v_0 and their descendants at appropriate levels. The only complication arises when the algorithm requests for an ancestor higher than root of the tree of v_0. In such a case, our algorithm uses the root as the ancestor. In the following lemma, we will show that with this approach, we still cover all simplices with meb radius of at most t.

Lemma 8 *The algorithm computes a t-restricted (ε, k)-WSSD.*

Proof. We show by induction that with modified ancestor search, we still cover all simplices with meb radius at most t. For $k = 1$, the correctness of the algorithm follows from Theorem 3 in Section 6, Let Γ_{k-1} cover all $(k-1)$-simplices γ which satisfy $\mathrm{rad}(\gamma) \leq t$. Consider any k-simplex $\sigma = (m_0, \ldots, m_k)$ with $\mathrm{rad}(\sigma) \leq t$. From [5, Lem.9], there exists a point (say m_k) such that $m_k \in 2\mathrm{meb}(\sigma')$ where $\sigma' := \sigma \setminus \{m_k\}$ and $2\mathrm{meb}(\sigma')$ represents a ball with twice the radius and the same center as $\mathrm{meb}(\sigma')$. Since σ' is a $(k-1)$-simplex and $\mathrm{rad}(\sigma') \leq \mathrm{rad}(\sigma) \leq t$, it is covered by some k-tuple $\gamma = (v_0, \ldots, v_{k-1}) \in \Gamma_{k-1}$. To prove correctness, we show that when our algorithm reaches tuple γ, it produces a $(k+1)$-tuple (γ, x) such that $m_k \in P_x$ which implies that the simplex σ is covered by the $(k+1)$-tuple (γ, x).

When handling γ, the algorithm searches for an ancestor of v_0 at an appropriate scale. If this ancestor is found within the tree of v_0 in the net-forest, the arguments from [5, Lem.12] carry over to ensure that a suitable x is found. So let us assume that the algorithm chooses the root of the tree of the net-forest that v_0 lies in. Call that root node a_0. The algorithm considers all nodes in $\mathrm{Rel}(a_0)$ and creates new tuples with their descendants. Moreover, the net-forest contains a leaf representing the point m_k; let a' denote the root of its tree. It suffices to show that $a' \in \mathrm{Rel}(a_0)$. Since $\mathrm{rad}(\sigma) \leq t$, the distance of m_0 and m_k is at most $2t$. Moreover, the distance of m_0 to rep_{a_0} is at most t, because the representatives of the roots form a (t,t)-net. The same holds for m_k and a'. Using triangle inequality, the distance of rep_{a_0} and $\mathrm{rep}_{a'}$ is at most $4t$. This implies that $a' \in \mathrm{Rel}(a_0)$. □

Lemma 9 *The size of the computed t-restricted (ε, k)-WSSD Γ_k is $n(\frac{2}{\varepsilon})^{O(\Delta_{7t} \cdot k)}$.*

Proof. The proof of [5, Lem.13] carries over directly – indeed, we can replace all occurrences of Δ by Δ_{7t}. This comes from the fact that a node u has at most $14^{\Delta_{7t}}$ nodes in $\mathrm{Rel}(u)$, and for any node in $\mathrm{Rel}(u)$ we reach descendants of a level of at most $O(\log(2/\varepsilon))$ smaller then u (see the proof of [5, Lem.13] for details). Since every node in the net-forest has at most $2^{O(\Delta_t)}$ children, we create at most

$$14^{\Delta_{7t}} \left(\frac{2}{\varepsilon}\right)^{O(\Delta_t)} = \left(\frac{2}{\varepsilon}\right)^{O(\Delta_{7t})}$$

tuples in Γ_k from a tuple in Γ_{k-1}. With that, the bound can be proved by induction. □

Lemma 10 *Computing a t-restricted (ε)-WSSD takes time $nd(2/\varepsilon)^{O(\Delta_{7t} \cdot k)}$ after computing the net forest at scale t.*

Proof. The proof is analogous to [5, Lem.14], plugging in the running time for t-restricted ε-WSPD from Theorem 11 and the size bound from Lemma 9. □

Computing the approximate Čech filtration We use the scheme of [5, Sec.4] to construct the $(1+\varepsilon)$-approximate filtration on the t-restricted WSSD. The original construction works without modification. Using the notation from [5, Sec.4]., for any WST $\sigma = (v_0, v_1, \ldots, v_k)$ with $\ell(v_i) \leq h$, we add $\sigma' = (\mathrm{vcell}(v_0, h), \mathrm{vcell}(v_1, h), \ldots, \mathrm{vcell}(v_k, h))$ to $\mathcal{A}_\alpha$ if $\mathrm{rad}(\sigma') \leq \theta_\Delta$. The only potential problem with the t-restricted case is that such a vcell() might be a node higher than a root of the net-forest. This cannot happen, however, since h is chosen such that

$$\frac{2\tau}{\tau - 1}\tau^h \leq \frac{\varepsilon}{7}\alpha.$$

Since $\alpha \leq t$ and $\varepsilon \leq 1$, we have that

$$h < \lfloor \log_\tau \frac{\tau - 1}{2\tau} t \rfloor = \ell(u)$$

for any root u in the net-forest.

Theorem 11 *A t-restricted (ε, k)-WSSD of size $n(\frac{2}{\varepsilon})^{O(\Delta_{7t} \cdot k)}$ can be computed in time*

$$NF + nd\left(\frac{2}{\varepsilon}\right)^{O(\Delta_{7t} \cdot k)},$$

where NF is the complexity for computing the net-forest from Theorem 2. Within the same time bound, we can construct a sequence of approximation complexes $(\mathcal{A}_\alpha)_{\alpha \in [0,t]}$ of size $n(\frac{2}{\varepsilon})^{O(\Delta_{7t} \cdot k)}$ whose persistence module is an $(1+\varepsilon)$-approximation (in the sense that the two modules are interleaved *[4]) of the* truncated Čech filtration $(\mathcal{C}_\alpha)_{\alpha \in [0,t]}$.

The claim follows directly from Lemmas 8, 9 and 10 and the preceding construction.