Proximity problems for high-dimensional data

Ioannis Psarros*

National and Kapodistrian University of Athens
ipsarros@di.uoa.gr

Abstract

With the increase of availability of complex datasets, there is a need for algorithmic solutions which scale well as the complexity of the data increases. We focus on proximity problems for high-dimensional vectors and polygonal curves and we present new solutions for the problem of computing approximate nearest neighbors. In Euclidean spaces, we propose and analyse random projections to a very low dimension, aiming for a high-dimensional solution which is also space efficient. For polygonal curves, we design a data structure with arbitrarily small approximation error. In addition, we present a new solution for computing good representatives, when the dataset consists of high-dimensional vectors. Finally, we study range spaces defined by metrics for polygonal curves and we present new bounds on their VC dimension.

1 Introduction

Finding similar objects is a general computational task which serves as a subroutine for many major learning tasks like classification or clustering. With the recent increase of availability of complex datasets, the need for analyzing and handling high-dimensional descriptors has been increased. Likewise, there is a surge of interest into data structures for trajectory processing, motivated by the increasing availability and quality of trajectory data.

Proximity problems in metric spaces of low dimension have been typically handled by methods which discretize the space and hence they are affected by the prominent curse of dimensionality, so called because it refers to the computational hardness of analyzing high-dimensional data. In the past two decades, the increasing need for analyzing high-dimensional data, lead the researchers to devise approximate and randomized algorithms with polynomial dependence on the dimension. Similarly, other complex data such as time series or polygonal curves have been typically handled by approximate or randomized algorithms.

Definition 1 (c-Approximate Nearest Neighbor (c-ANN) problem)**.** *Given a finite set $P \subset M$, a distance function $\mathrm{d}(\cdot,\cdot)$, and an approximation factor $c > 1$, preprocess P into a data structure which supports the following type of queries:*

$$\forall q \in M, \textit{ find } p^* \textit{ such that } \forall p \in P: \ \mathrm{d}(q,p^*) \leq c \cdot \mathrm{d}(q,p).$$

*Dissertation Advisor: Ioannis Z. Emiris, Professor

The corresponding augmented decision problem (with witness) is known as the approximate *near* neighbor problem, defined as follows.

Definition 2 ((c,r)-ANN Problem)**.** *Given a finite set $P \subset M$, a distance function $\mathrm{d}(\cdot,\cdot)$, an approximation factor $c > 1$, and a range parameter r, preprocess P into a data structure which supports the following type of queries:*

- *if $\exists p^* \in P$ s.t. $\mathrm{d}(p^*, q) \leq r$, then return any point $p' \in M$ s.t. $\mathrm{d}(p', q) \leq c \cdot r$,*
- *if $\forall p \in P$, $\mathrm{d}(p, q) > c \cdot r$, then report "Fail".*

The data structure is allowed to return either a point at distance $\leq c \cdot r$ or "Fail".

It is known that one can solve logarithmically many instances of the decision problem with witness to solve the $(1+\epsilon)$-ANN problem [11].

Another problem of interest is that of computing good representatives for a finite metric space. An r-net for a finite metric space (P, d), $|P| = n$ and for numerical parameter r is a subset $\mathcal{N} \subseteq P$ such that the closed $r/2$-balls centered at the points of $\mathcal{N}$ are disjoint, and the closed r-balls around the same points cover all of P. We define approximate r-nets analogously: the closed $r/2$-balls centered at the points of $\mathcal{N}$ are disjoint, and the closed cr-balls around the same points cover all of P, where c denotes the approximation factor.

In all proximity problems, there is an explicit notion of dissimilarity or distance between two input objects. It is natural to define ranges based on the distance function: a range is essentially a pseudo-metric ball. Generally, a *range space* $(X, \mathcal{R})$ (also called *set system*) is defined by a ground set X and a set of ranges $\mathcal{R}$, where each $r \in \mathcal{R}$ is a subset of X. A crucial descriptor of any range space is its VC-dimension. These notions quantify how complex a range space is, and have played foundational roles in machine learning and geometry.

1.1 Related work

In this section, we present previous results on proximity problems in two main settings: normed spaces and polygonal curves.

1.1.1 Normed spaces

Unless otherwise stated, the results concern the case of points in ℓ_2.

In high dimensional spaces, classic space partitioning data structures are affected by the curse of dimensionality. This means that, when the dimension increases, either the query time or the required space increases exponentially. An important method conceived for high dimensional data is Locality Sensitive Hashing (LSH). In general, LSH requires roughly $O(dn^{1+\rho})$ space and $O(dn^{\rho})$ query time for some parameter $\rho \in (0,1)$. Lately, it was shown that a data-dependent scheme achieves $\rho = \frac{1}{2(1+\epsilon)^2-1} + o(1)$ [4].

For practical applications, memory consumption is often a limitation. Most of the previous work in the (near) linear space regime $dn^{1+o(1)}$ focuses on the case that ϵ is greater than 0 by a constant term. After the original submission our paper [2], a query time of $O(n^{1-4\epsilon^2+O(\epsilon^3)})$ has been established [3]. The bound has been shown to be optimal for a large class of data structures. Despite the fact that our algorithm is sub-optimal, it is simpler and easier to implement.

Significant amount of work has been done for pointsets with low doubling dimension. In [15], a new notion of nearest neighbor preserving embeddings has been presented. It has been proven that in this context we can achieve dimension reduction which only depends on the doubling dimension of the dataset. Such an approach can be combined with any known data structure for $(1+\epsilon)$-ANN.

One related problem is that of computing $(1+\epsilon)$-approximate r-nets. In [12], they show that an approximate net hierarchy for an arbitrary finite metric X, such that $|X| = n$, can be computed in $O(2^{\mathrm{ddim}(X)} n \log n)$. This is satisfactory when doubling dimension is constant, but requires a vast amount of resources when it is high. In the latter case, one approach is that of [10], which uses LSH and requires time $O(n^{1+1/(1+\epsilon)^2+o(1)})$.

1.1.2 Polygonal curves

For polygonal curves, we focus on discrete Fréchet (DFD) and Dynamic Time Warping (DTW) distance functions.

The first result for DFD by Indyk [13], defined by any metric $(X, d(\cdot,\cdot))$, achieved approximation factor $O((\log m + \log\log n)^{t-1})$, where m is the maximum length of a curve, and $t > 1$ is a trade-off parameter. The data structure achieves space and preprocessing time in $O(m^2|X|)^{tm^{1/t}} \cdot n^{2t}$, and query time in $(m \log n)^{O(t)}$. It is not clear whether the approach may achieve a $1+\epsilon$ approximation factor by employing more space.

More recently, a new data structure was devised for the DFD of curves in Euclidean spaces [7]. The approximation factor is $O(d^{3/2})$. The space required is $O(2^{4md} n \log n + mn)$ and each query costs $O(2^{4md} m \log n)$. They also provide a trade-off between performance, and the approximation factor. At the other extreme of this trade-off, they achieve space in $O(n \log n + mn)$, query time in $O(m \log n)$ and approximation factor $O(m)$. Furthermore, it is shown that the result establishing an $O(m)$ approximation [7] extends to DTW.

1.2 Contribution

1.2.1 Normed spaces

Approximate Nearest Neighbors. In [2], we introduce a notion of "low-quality" randomized embeddings and we employ standard random projections à la Johnson-Lindenstrauss in order to define a mapping from ℓ_2^d to $\ell_2^{d'}$, for $d' = O\left(\epsilon^{-2} \cdot \log\left(\frac{n}{k}\right)\right)$, such that an approximate nearest neighbor of the query lies among the pre-images of k approximate nearest neighbors in the projected space. This observation allows us to combine random projections with the bucketing method [11], and obtain a randomized data structure with optimal space and sublinear query for the augmented decision problem. The main result states that there exists a randomized data structure for the $(1+\epsilon, r)$-ANN problem, with linear space, linear preprocessing time, and query time $O(dn^{\rho})$, where $\rho = 1 - \Theta(\epsilon^2/\log(1/\epsilon))$. For each query $q \in \mathbb{R}^d$, preprocessing succeeds with constant probability, and can be amplified by repetition. We extend our results to doubling subsets of ℓ_2.

Our ideas directly extend to the $(1+\epsilon)$-ANN problem, but it achieves bounds which are weaker than the ones obtained through the $(1+\epsilon, r)$-ANN solution, but

the algorithm is very simple and quite interesting in practice, since reducing $(1+\epsilon)$-ANN to $(1+\epsilon, r)$-ANN is nontrivial and typically avoided in implementations.

Finally, we are able to define a mapping from any metric which admits an LSH family of functions to the Hamming space. Using this mapping, we achieve improved query time in $\tilde{O}(dn^{1-\Theta(\epsilon^2)})$.

In [8], we study the problem of reducing the dimension for doubling subsets of ℓ_1. While this embeddability question has a negative answer in general, we show that one can reduce the dimension considerably when focused on the (c, r)-ANN problem. The main requirement is that the dimension reduction preserves enough information for reducing the (c, r)-ANN problem in a high dimensional space to the (c, r)-ANN problem in a much lower dimensional space.

Approximate Nets. In [5], we present a new randomized algorithm that computes $(1+\epsilon)$-approximate r-nets in time time $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$. This improves upon the complexity of the best known algorithm, when ϵ is sufficiently small.

1.2.2 Polygonal curves

Approximate Nearest Neighbors. In [9], we study the $(1+\epsilon)$-ANN problem for polygonal curves. We present a notion of distance between two polygonal curves, which generalizes both DFD and DTW. The *ℓ_p-distance* of two curves minimizes, over all traversals, the ℓ_p norm of the vector of all Euclidean distances between paired points. Hence, DFD corresponds to ℓ_∞-distance of polygonal curves, and DTW corresponds to ℓ_1-distance of polygonal curves.

Our main contribution is an $(1+\epsilon)$-ANN data structure for the ℓ_p-distance of curves, when $1 \leq p < \infty$. This easily extends to ℓ_∞-distance of curves by solving for the ℓ_p-distance, for a sufficiently large value of p. Our target are methods with approximation factor $1+\epsilon$. Such approximation factors are obtained for the first time, at the expense of larger space or time complexity. Moreover, a further advantage is that our methods solve $(1+\epsilon)$-ANN directly instead of requiring to reduce it to near neighbor search.

We also focus on DFD, and we provide a solution which is especially efficient in the short query regime. For the Euclidean space, we give a randomized data structure with space in $n \cdot O\left(\frac{kd^{3/2}}{\epsilon}\right)^{dk} + O(dnm)$ and query time in $O(dk)$, where k denotes the length of the query curves. The data structure can be derandomized with a slight worsening of the performance. We give analogous results for arbitrary doubling metrics.

Vapnik-Chervonenkis dimension. In [6], we analyze the VC dimension of range spaces defined by polygonal curves. To the best of our knowledge, the results presented here are the first for this problem. For Discrete Hausdorff or Fréchet balls defined on point sets (resp. point sequences) in $\mathbb{R}^d$ we show that the VC dimension is at most near-linear in k, the complexity of the ball centers that define the ranges, and at most logarithmic in m, the size of the point sets of the ground set. For the Fréchet distance, where the ground set X are continuous polygonal curves in $\mathbb{R}^2$ we show an upper bound that is quadratic in k and logarithmic in m. These initial bounds assume a fixed radius of the metric balls that define the ranges $\mathcal{R}$. The same holds for the Hausdorff distance, where the ground set are sets of line segments in $\mathbb{R}^2$.

2 Randomized Embeddings with slack

We introduce a new notion of embedding for metric spaces requiring that, for some query, there exists an approximate nearest neighbor among the pre-images of its $k > 1$ approximate nearest neighbors in the target space. In Euclidean spaces, we employ random projections à la Johnson-Lindenstrauss to a dimension inversely proportional to k. In other words, we allow k false positives, meaning that at most k far points will appear as near neighbors in the projected space.

Let us now revisit one form of the classic Johnson-Lindenstrauss Lemma:

Theorem 3 ([14])**.** *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$. There exists a constant $C > 0$, such that for any $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$:*

- $\Pr\left[\|Gv\|_2^2 \leq (1-\epsilon) \cdot \frac{d'}{d}\right] \leq \exp\left(-Cd'\epsilon^2\right),$
- $\Pr\left[\|Gv\|_2^2 \geq (1+\epsilon) \cdot \frac{d'}{d}\right] \leq \exp\left(-Cd'\epsilon^2\right).$

In the initial proof [16], they show that this can be achieved by orthogonally projecting the pointset on a random linear subspace of dimension d'. Instead of a gaussian matrix, we can apply a matrix whose entries are independent random variables with uniformly distributed values in $\{-1, 1\}$ [1], or even random variables with uniform subgaussian tails [17].

The following has been introduced in [15] and focuses on the distortion of the nearest neighbor.

Definition 4. *Let (Y, d_Y), (Z, d_Z) be metric spaces and $X \subseteq Y$. A distribution over mappings $f : Y \to Z$ is a* nearest-neighbor preserving embedding *with distortion $D \geq 1$ and probability of correctness $P \in [0,1]$ if, $\forall \epsilon \geq 0$ and $\forall q \in Y$, with probability at least P, when $x \in X$ is such that $f(x)$ is an c-ANN of $f(q)$ in $f(X)$, then x is a $(D \cdot c)$-approximate nearest neighbor of q in X.*

Let us now consider a closely related problem. While in c-ANN we search one point which is approximately nearest, in the k approximate nearest neighbors problem, or c-kANNs, we seek an approximation of the k nearest points, in the following sense. Let X be a set of n points in $\mathbb{R}^d$, let $q \in \mathbb{R}^d$ and $1 \leq k \leq n$. The problem consists in reporting a sequence $S = \{p_1, \ldots, p_k\}$ of k distinct points such that the i-th point p_i is an c-approximation to the i-th nearest neighbor of q. Furthermore, the following assumption is satisfied by the search routine of certain tree-based data structures, such as BBD-trees.

Assumption 5. *The c-kANNs search algorithm visits a set S' of points in X. Let $S = \{p_1, \ldots, p_k\}$ be the k nearest points to the query in S'. We assume that for all $x \in X \setminus S'$ and $y \in S$, $d(x,q) > d(y,q) \cdot c$.*

Assuming the existence of a data structure which solves c-kANNs and satisfies Assumption 5, we propose to weaken Definition 4 as follows.

Definition 6. *Let (Y, d_Y), (Z, d_Z) be metric spaces and $X \subseteq Y$. A distribution over mappings $f : Y \mapsto Z$ is a* locality preserving embedding *with distortion $D \geq 1$, probability of correctness $P \in [0,1]$ and locality parameter k if, $\forall c \geq 1$*

and $\forall q \in Y$*, with probability at least* P*, when* $S = \{f(p_1), \ldots, f(p_k)\}$ *is a solution to c-kANNs for* q *under Assumption 5, then there exists* $f(x) \in S$ *such that* x *is a* $(D \cdot c)$*-approximate nearest neighbor of* q *in* X*.*

According to this definition we can reduce the problem of c-ANN in dimension d to the problem of computing k approximate nearest neighbors in dimension $d' < d$.

We employ the Johnson-Lindenstrauss dimensionality reduction technique and, more specifically, Theorem 3.

Remark 7. *In the statements of our results, we use the term* $(1+\epsilon)^2$ *or* $(1+\epsilon)^3$ *for the sake of simplicity. Notice that we can replace* $(1+\epsilon')^2$ *by* $1+\epsilon$ *just by rescaling* $\epsilon' \leftarrow \epsilon/4 \implies (1+\epsilon')^2 \leq 1+\epsilon$*, when* $\epsilon < 1/2$*.*

We are now ready to prove the main theorem of this section.

Theorem 8. *Under the notation of Definition 6, there exists a randomized mapping* $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ *which satisfies Definition 6 for*

$$d' = O\left(\epsilon^{-2} \log \frac{n}{k}\right),$$

$\epsilon \in (0, 1/2]$*, distortion* $D = (1+\epsilon)^2$ *and probability of success* $2/3$*.*

Proof. Let X be a set of n points in $\mathbb{R}^d$ and consider map

$$f : \mathbb{R}^d \mapsto \mathbb{R}^{d'} : v \mapsto \sqrt{d/d'} \cdot G\, v,$$

where G is a matrix chosen from a distribution as in Theorem 3. Without loss of generality the query point q lies at the origin and its nearest neighbor u lies at distance 1 from q. We denote by $c' \geq 1$ the approximation ratio guaranteed by the assumed data structure (see Assumption 5). That is, the assumed data structure solves the c'-kANNs problem. Let N be the random variable whose value indicates the number of false positives, that is

$$N = |\{x \in X : \|x\|_2 > \gamma \wedge \|f(x)\|_2 \leq \beta\}|,$$

where we define $\beta = c'(1+\epsilon)$, $\gamma = c'(1+\epsilon)^2$. Hence, by Lemma 3,

$$\mathbb{E}[N] \leq n \cdot \exp(-Cd' \cdot \epsilon^2),$$

where $C > 0$ is a constant. The event of failure is defined as the disjunction of two events:

$$N \geq k \ \vee \ \|f(u)\|_2 \geq (\beta/c), \tag{1}$$

and its probability is at most equal to

$$\Pr[N \geq k] + \exp\left(-Cd'\epsilon^2\right),$$

by applying again Theorem 3. Now, we set $d' = \Theta\left(\log(\frac{n}{k})/\epsilon^2\right)$ and we bound these two terms. Hence, there exists d' such that

$$d' = O\left(\epsilon^{-2} \cdot \log \frac{n}{k}\right)$$

and with probability at least 2/3, the following two events occur:

$$\|f(q) - f(u)\|_2 \leq (1+\epsilon)\|u-q\|_2,$$

$$|\{p \in X | \|p-q\|_2 > c(1+\epsilon)^2\|u-q\|_2 \implies \|f(q)-f(p)\|_2 \leq c(1+\epsilon)\|u-q\|_2\}| < k.$$

Let us consider the case when the random experiment succeeds, and let $S = \{f(p_1), \ldots, f(p_k)\}$ be a solution of the c'-kANNs problem in the projected space, given by a data-structure which satisfies Assumption 5. It holds that $\forall f(x) \in f(X) \setminus S'$, $\|f(x) - f(q)\|_2 > \|f(p_k) - f(q)\|_2/c'$, where S' is the set of all points visited by the search routine.

If $f(u) \in S$, then S contains the projection of the nearest neighbor. If $f(u) \notin S$, then if $f(u) \notin S'$ we have the following:

$$\|f(u) - f(q)\|_2 > \|f(p_k) - f(q)\|_2/c \implies \|f(p_k) - f(q)\|_2 < c(1+\epsilon)\|u-q\|_2,$$

which means that there exists at least one point $f(p^*) \in S$ s.t. $\|q - p^*\|_2 \leq c'(1+\epsilon)\|u-q\|_2$. Finally, if $f(u) \notin S$ but $f(u) \in S'$ then

$$\|f(p_k) - f(q)\|_2 \leq \|f(u) - f(q)\|_2 \implies \|f(p_k) - f(q)\|_2 \leq (1+\epsilon)\|u-q\|_2,$$

which means that there exists at least one point $f(p^*) \in S$ s.t. $\|q - p^*\|_2 \leq c'(1+\epsilon)^2\|u-q\|_2$.

Hence, f satisfies Definition 6 for $D = (1+\epsilon)^2$ and the theorem is established. □

Theorem 8 essentially translates the c-ANN problem to the c-kANNs problem. While this is convenient in practice, better bounds can be achieved when working with the (c, r)-ANN problem.

2.1 Approximate Near Neighbor

This section combines the ideas developed in Section 2 with a simple, auxiliary data structure, namely the grid, yielding an efficient solution for the augmented decision (c, r)-ANN problem. In the following, the $\tilde{O}(\cdot)$ notation hides factors polynomial in $1/\epsilon$ and $\log n$.

The data structure succeeds if it indeed answers the approximate decision problem for query q. Building a data structure for the Approximate Nearest Neighbor Problem reduces to building several data structures for the decision (c, r)-ANN problem. For completeness, we include the corresponding theorem.

Theorem 9. [11, Theorem 2.9] *Let P be a given set of n points in a metric space, and let $c = 1 + \epsilon > 1$, $f \in (0, 1)$, and $\gamma \in (1/n, 1)$ be prescribed parameters. Assume that we are given a data structure for the (c, r)-ANN that uses space $S(n, c, f)$, has query time $Q(n, c, f)$, and has failure probability f. Then there exists a data structure for answering $c(1 + O(\gamma))$-NN queries in time $O(\log n)Q(n, c, f)$ with failure probability $O(f \log n)$. The resulting data structure uses $O(S(n, c, f)/\gamma \cdot \log^2 n)$ space.*

A natural generalization of the (c, r)-ANN problem is the k-Approximate Near Neighbors Problem, denoted (c, r)-kANNs.

Definition 10 ((c, r)-kANNs Problem)**.** *Let $X \subset \mathbb{R}^d$ and $|X| = n$. Given $c > 1$, $r > 0$, build a data structure which, for any query $q \in \mathbb{R}^d$:*

- *if* $|\{p \in X \mid \|q-p\|_2 \leq r\}| \geq k$, *then report* $S \subseteq \{p \in X \mid \|q-p\|_2 \leq c \cdot r\}$ *s.t.* $|S| = k$,
- *if* $a := |\{p \in X \mid \|q-p\|_2 \leq r\}| < k$, *then report* $S \subseteq \{p \in X \mid \|q-p\|_2 \leq c \cdot r\}$ *s.t.* $a \leq |S| \leq k$.

The following algorithm is essentially the bucketing method which is described in [11] and concerns the case $k = 1$. We define a uniform grid of side length $\epsilon/\sqrt{d}$ on $\mathbb{R}^d$. Clearly, the distance between any two points belonging to one grid cell is at most ϵ. Assume $r = 1$. For each ball $B_q = \{x \in \mathbb{R}^d \mid \|x-q\|_2 \leq r\}$, $q \in \mathbb{R}^d$, let $\overline{B_q}$ be the set of grid cells that intersect B_q.

In [11], they show that $|\overline{B_q}| \leq (C'/\epsilon)^d$. Hence, the query time is the time to compute the hash function, retrieve near cells and report the k neighbors:

$$O(d + k + (C'/\epsilon)^d).$$

The required space usage is $O(dn)$.

Furthermore, we are interested in optimizing this constant C'. The bound on $|\overline{B_q}|$ follows from the following fact:

$$|\overline{B_q}| \leq V_2^d(R),$$

where $V_2^d(R)$ is the volume of the ball with radius R in ℓ_2^d, and $R = \frac{2\sqrt{d}}{\epsilon}$. Now,

$$V_2^d(R) \leq \frac{2\pi^{d/2}}{d \cdot \Gamma(d/2)} R^d = \frac{2\pi^{d/2}}{d(d/2-1)!} R^d \leq \frac{2\pi^{d/2}}{(d/2)!} R^d \leq$$

$$\leq \frac{2\pi^{d/2}}{e(d/(2e))^{d/2}} R^d \leq \frac{2^{d+1}(18)^{d/2}}{\epsilon^d e} \leq \frac{9^d}{\epsilon^d}.$$

Hence, $C' \leq 9$.

Theorem 11. *There exists a data structure for Problem 10 with required space* $O(dn)$ *and query time* $O\left(d + k + (\frac{9}{\epsilon})^d\right)$.

The following theorem is an analogue of Theorem 8 for the Approximate Near Neighbor Problem.

Theorem 12. *The* $((1+\epsilon)^2 c, r)$*-ANN problem in* $\mathbb{R}^d$ *reduces to checking the solution set of the* $(c, (1+\epsilon)r)$*-kANNs problem in* $\mathbb{R}^{d'}$*, where* $d' = O\left(\log\left(\frac{n}{k}\right)/\epsilon^2\right)$*, by a randomized algorithm which succeeds with constant probability. The delay in query time is proportional to* $d \cdot k$.

Proof. The theorem can be seen as a direct implication of Theorem 8. The proof is indeed the same.

Let X be a set of n points in $\mathbb{R}^d$ and consider map

$$f : \mathbb{R}^d \mapsto \mathbb{R}^{d'} : v \mapsto \sqrt{d/d'} \cdot G\ v,$$

where G is a matrix chosen from a distribution as in Theorem 3. Let $u \in X$ a point at distance 1 from q and assume without loss of generality that lies at the origin. Let N be the random variable whose value indicates the number of false positives, that is

$$N = |\{x \in X : \|x\|_2 > \gamma \wedge \|f(x)\|_2 \leq \beta\}|,$$

where we define $\beta = c(1+\epsilon)$, $\gamma = c(1+\epsilon)^2$. Hence, by Theorem 3,

$$\mathbb{E}[N] \leq n \cdot \exp\left(-Cd'\epsilon^2\right).$$

The probability of failure is at most equal to

$$\Pr[N \geq k] + \exp\left(-Cd'\epsilon^2\right),$$

by applying again Theorem 3. Now, we bound these two terms for $d' = \Theta\left(\log \frac{n}{k}\right)/\epsilon^2$. With probability at least 2/3, these two events occur:

- $\|f(q) - f(u)\|_2 \leq (1+\epsilon)$.
- $|\{p \in X | \|p-q\|_2 > c(1+\epsilon)^2 \implies \|f(q) - f(p)\|_2 \leq c(1+\epsilon)\}| < k$.

□

2.1.1 Finite subsets of ℓ_2

We are about to show what Theorems 11 and 12 imply for the (c, r)-ANN problem.

Theorem 13. *There exists a data structure for the (c, r)-ANN problem with $O(dn)$ required space and preprocessing time, and query time $\tilde{O}(dn^\rho)$, where $\rho = 1 - \Theta(\epsilon^2/\log(1/\epsilon)) < 1$.*

Proof. For $k = \Theta\left(n^\rho\right)$,

$$\left(\frac{9}{\epsilon}\right)^{d'} + dk \leq O(dn^\rho)$$

Since, the data structure succeeds only with probability 9/10, it suffices to build it $O(\log n)$ times in order to achieve high probability of success. □

3 Conclusions

In this thesis, we investigated proximity problems for high-dimensional vectors and polygonal curves. We proposed algorithms and data structures for various important problems such as the approximate nearest neighbor problem and the problem of computing nets. Most of the techniques analyzed in this thesis are actually simple and can be easily implemented. Hence, we believe that the rigorous arguments presented in this thesis have the potential to lead to practical innovations.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] Evangelos Anagnostopoulos, Ioannis Z. Emiris, and Ioannis Psarros. Randomized embeddings with slack and high-dimensional approximate nearest neighbor. *ACM Trans. Algorithms*, 14(2):18:1–18:21, 2018.

[3] Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. In *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2017. Also as arxiv.org/abs/1608.03580.

[4] Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In *the Proc. 47th ACM Symp. Theory of Computing*, STOC'15, 2015.

[5] Georgia Avarikioti, Ioannis Z. Emiris, Loukas Kavouras, and Ioannis Psarros. High-dimensional approximate r-nets. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 16–30, 2017.

[6] Anne Driemel, Jeff M. Phillips, and Ioannis Psarros. The VC dimension of metric balls under Fréchet and Hausdorff distances. In *Proc. 35st International Symposium on Computational Geometry*, 2019.

[7] Anne Driemel and Francesco Silvestri. Locality-sensitive hashing of curves. In *Proc. 33rd Intern. Symposium on Computational Geometry*, pages 37:1–37:16, 2017.

[8] Ioannis Z. Emiris, Vasilis Margonis, and Ioannis Psarros. Near neighbor preserving dimension reduction for doubling subsets of ℓ_1. *CoRR*, abs/1902.08815, 2019.

[9] Ioannis Z. Emiris and Ioannis Psarros. Products of euclidean metrics and applications to proximity auestions among curves. In *34th International Symposium on Computational Geometry, SoCG 2018, Budapest, Hungary*, volume 99 of *LIPIcs*, pages 37:1–37:13, 2018.

[10] David Eppstein, Sariel Har-Peled, and Anastasios Sidiropoulos. Approximate greedy clustering and distance selection for graph metrics. *CoRR*, abs/1507.01555, 2015.

[11] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

[12] Sariel Har-Peled and Manor Mendel. Fast construction of nets in low dimensional metrics, and their applications. In *Proc. 21st Annual Symp. Computational Geometry*, SCG'05, pages 150–158, 2005.

[13] Piotr Indyk. Approximate nearest neighbor algorithms for frechet distance via product metrics. In *Proc. 18th Annual Symp. on Computational Geometry*, SoCG '02, pages 102–106, New York, NY, USA, 2002. ACM.

[14] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annual ACM Symp. Theory of Computing*, STOC'98, pages 604–613, 1998.

[15] Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3), 2007.

[16] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. 26:189–206, 1984.

[17] Jivr'i Matouvsek. On variants of the johnson-lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.