NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

PhD THESIS

Proximity problems for high-dimensional data

Ioannis D. Psarros

ATHENS

JUNE 2019

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Προβλήματα εγγύτητας για δεδομένα υψηλών διαστάσεων

Ιωάννης Δ. Ψαρρός

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2019

PhD THESIS

Proximity problems for high-dimensional data

Ioannis D. Psarros

SUPERVISOR: Ioannis Z. Emiris, Professor NKUA
THREE-MEMBER ADVISORY COMMITTEE:

Ioannis Z. Emiris, Professor NKUA

Stavros Kolliopoulos, Professor NKUA

Anastasios Sidiropoulos, Assistant Professor UIC

SEVEN-MEMBER EXAMINATION COMMITTEE

Ioannis Z. Emiris,
Professor NKUA

Stavros Kolliopoulos,
Professor NKUA

Anastasios Sidiropoulos,
Assistant Professor UIC

Dimitris Fotakis,
Associate Professor NTUA

Dimitrios Gunopulos,
Professor NKUA

Apostolos Giannopoulos,
Professor NKUA

Aris T. Pagourtzis,
Associate Professor NTUA

Examination Date: June 10, 2019

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Προβλήματα εγγύτητας για δεδομένα υψηλών διαστάσεων

Ιωάννης Δ. Ψαρρός

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Ιωάννης Ζ. Εμίρης, Καθηγητής ΕΚΠΑ
ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:

Ιωάννης Ζ. Εμίρης, Καθηγητής ΕΚΠΑ

Σταύρος Κολλιόπουλος, Καθηγητής ΕΚΠΑ

Αναστάσιος Σιδηρόπουλος, Επίκουρος Καθηγητής ΠΙΣ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Ιωάννης Ζ. Εμίρης,
Καθηγητής ΕΚΠΑ

Σταύρος Κολλιόπουλος,
Καθηγητής ΕΚΠΑ

Αναστάσιος Σιδηρόπουλος,
Επίκουρος Καθηγητής ΠΙΣ

Δημήτρης Φωτάκης,
Αναπληρωτής Καθηγητής ΕΜΠ

Δημήτριος Γουνόπουλος,
Καθηγητής ΕΚΠΑ

Απόστολος Γιαννόπουλος,
Καθηγητής ΕΚΠΑ

Άρης Τ. Παγουρτζής,
Αναπληρωτής Καθηγητής ΕΜΠ

Ημερομηνία Εξέτασης: 10 Ιουνίου 2019

ABSTRACT

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 from mobile phones, GPS sensors, RFID technology and video analysis.

In this thesis, we investigate proximity problems for high-dimensional vectors and polygonal curves. The natural way to measure dissimilarity between two vectors is by evaluating a norm function for the vector difference. Popular examples of such distance functions are the Euclidean distance and the Manhattan distance. Similarly, there exist several well-studied distance functions for polygonal curves, the main example being the Fréchet distance.

The core problem, for both data types, is the nearest neighbor searching problem. Given a set of objects P, we aim for a data structure which supports nearest neighbor queries; a new object q arrives and the data structure returns the most similar object in P. When the data complexity is high, aiming for an exact solution is often futile. This has led researchers to the more tractable task of designing approximate solutions. The largest part of this thesis is devoted to the approximate nearest neighbor problem and the approximate near neighbor problem: given a set of objects P and a radius parameter r, the data structure returns an object in P which is approximately within distance r (if there exists one) from some query object q. Another basic question is that of computing a subset of good representatives for a dataset. This subset often provides with sufficient information for a given computational task, and hence it possibly simplifies existing solutions. Finally, we investigate range systems for polygonal curves: we bound the Vapnik–Chervonenkis dimension for ranges defined by distance functions for curves. These bounds have direct implications in range counting problems and density estimation.

The thesis is organized as follows.

Random projections for proximity search. We introduce a new definition of “low-quality” embeddings for metric spaces [8]. It requires that, for some query point q, there exists an approximate nearest neighbor among the pre-images of the $k > 1$ approximate nearest neighbors in the target space. Focusing on Euclidean spaces, we employ random projections à la Johnson Lindenstrauss in order to reduce the original problem to one in a space of dimension inversely proportional to k. This leads to simple data structures which are space-efficient and also support sublinear queries. By employing properties of certain LSH functions, we exploit a similar mapping to the Hamming space.

Doubling sets and Manhattan distance. Our primary motivation is the approximate nearest neighbor problem in ℓ_1, for pointsets with low intrinsic dimension. Doubling dimension is

a well-established notion which aims to capture the intrinsic dimension of points. Nearest neighbor-preserving embeddings are known to exist for both ℓ_2 and ℓ_1 metrics, as well as for doubling subsets of ℓ_2. We propose a dimension reduction by means of a *near* neighbor-preserving embedding for doubling subsets of ℓ_1 [40].

Approximate r-nets. Nets offers a powerful tool in computational and metric geometry, since they serve as a subset of good representatives: all points are within distance r from some net point and all net points lie at distance at least r from each other. We focus on high-dimensional spaces and present a new randomized algorithm which efficiently computes approximate r-nets with respect to Euclidean distance [19]. Our algorithm follows a recent approach by Valiant in reducing the problem to multi-point evaluation of polynomials.

Proximity search for polygonal curves. We propose simple and efficient data structures [41], based on randomized projections, for a notion of distance between discretized curves, which generalizes both discrete Fréchet and Dynamic Time Warping distance functions. We offer the first data structures and query algorithms for the approximate nearest neighbor problem with arbitrarily good approximation factor, at the expense of increasing space usage and preprocessing time over existing methods.

Proximity search for short query curves. We propose simple and efficient data structures, based on random partitions, for the discrete Fréchet distance, in the short query regime. The data structures are especially efficient when queries are much shorter than the polygonal curves which belong to the dataset. We also study the problem for arbitrary metrics with bounded doubling dimension.

The VC dimension of polygonal curves. The Vapnik-Chervonenkis dimension provides a notion of complexity for set or range systems. We analyze range systems where the ground set is a set of polygonal curves in the Euclidean space and the ranges are metric balls defined by curve dissimilarity measures, such as the Fréchet distance and the Hausdorff distance [36]. Direct implications follow by applying known sampling bounds.

SUBJECT AREA: Computational Geometry

KEYWORDS: Nearest Neighbor, high dimension, polygonal curves

ΠΕΡΙΛΗΨΗ

Η εύρεση όμοιων αντικειμένων είναι ένα γενικό υπολογιστικό πρόβλημα που χρησιμεύει ως υπορουτίνα για πολλά προβλήματα μηχανικής μάθησης όπως η συσταδοποίηση. Με την πρόσφατη αύξηση της διαθεσιμότητας πολύπλοκων συνόλων δεδομένων, αυξήθηκε η ανάγκη για την ανάλυση δεδομένων υψηλών διαστάσεων. Παρομοίως, παρατηρείται αύξηση ενδιαφέροντος στις δομές δεδομένων για επεξεργασία καμπυλών, λόγω της αυξανόμενης διαθεσιμότητας και ποιότητας των δεδομένων τροχιάς από τα κινητά τηλέφωνα, τους αισθητήρες GPS, την τεχνολογία RFID και την ανάλυση βίντεο.

Σε αυτή τη διατριβή, ερευνάμε προβλήματα εγγύτητας για διανύσματα μεγάλης διάστασης και πολυγωνικές καμπύλες. Ο φυσικός τρόπος μέτρησης της ανομοιότητας μεταξύ δύο διανυσμάτων είναι η αποτίμηση μιας συνάρτησης νόρμας για τη διανυσματική διαφορά των δύο διανυσμάτων. Δημοφιλή παραδείγματα τέτοιων συναρτήσεων απόστασης είναι η Ευκλείδεια απόσταση και η απόσταση Μανχάταν. Παρομοίως, υπάρχουν αρκετές καλά μελετημένες συναρτήσεις απόστασης για πολυγωνικές καμπύλες, με κύριο παράδειγμα την απόσταση Fréchet.

Το βασικό πρόβλημα, και για τους δύο τύπους δεδομένων, είναι το πρόβλημα αναζήτησης του κοντινότερου γείτονα. Δεδομένου ενός συνόλου αντικειμένων P, στοχεύουμε σε μια δομή δεδομένων που υποστηρίζει ερωτήματα κοντινότερου γείτονα. Ένα νέο αντικείμενο q δίνεται και η δομή δεδομένων επιστρέφει το ομοιότερο αντικείμενο από το P. Όταν η πολυπλοκότητα των δεδομένων είναι υψηλή, μια λύση με ακρίβεια είναι σπάνια αποδοτική. Αυτό οδήγησε τους ερευνητές στον πιο εύκολο στόχο του σχεδιασμού προσεγγιστικών λύσεων. Το μεγαλύτερο μέρος αυτής της εργασίας είναι αφιερωμένο στο πρόβλημα του προσεγγιστικού κοντινότερου γείτονα και στο πρόβλημα του προσεγγιστικού κοντινού γείτονα: δεδομένου ενός συνόλου αντικειμένων P και μιας παραμέτρου ακτίνας r, η δομή δεδομένων επιστρέφει ένα αντικείμενο στο P (εφόσον υπάρχει) το οποίο είναι κατά προσέγγιση σε απόσταση r από κάποιο αντικείμενο ερώτησης q. Ένα άλλο βασικό ερώτημα είναι αυτό του υπολογισμού ενός υποσυνόλου καλών εκπροσώπων για ένα σύνολο δεδομένων. Αυτό το υποσύνολο παρέχει συχνά επαρκείς πληροφορίες για κάποιο υπολογιστικό πρόβλημα και επομένως απλοποιεί πιθανώς τις υπάρχουσες λύσεις. Τέλος, μελετάμε τους χώρους εύρους για πολυγωνικές καμπύλες: φράσουμε τη διάσταση Vapnik-Chervonenkis για εύρη που ορίζονται από συναρτήσεις απόστασης για καμπύλες. Τα αποτελέσματα αυτά έχουν άμεσες συνέπειες σε προβλήματα μέτρησης εύρους και στην εκτίμηση πυκνότητας.

Η διατριβή έχει δομηθεί ως εξής.

Τυχαίες προβολές για προβλήματα εγγύτητας. Εισάγουμε έναν νέο ορισμό εμβυθίσεων "χαμηλής ποιότητας" για μετρικούς χώρους [8]. Απαιτεί ότι, για κάποιο σημείο ερωτήματος q, υπάρχει ένας προσεγγιστικός κοντινότερος γείτονας μεταξύ των προ-εικόνων των $k > 1$ προσεγγιστικών κοντινότερων γειτόνων στο χώρο προορισμού. Εστιάζοντας σε Ευκλείδειους χώρους, χρησιμοποιούμε τυχαίες προβολές à la Johnson Lindenstrauss προ-

κειμένου να ανάγουμε το αρχικό πρόβλημα σε ένα πρόβλημα όπου η διάσταση του χώρου είναι αντιστρόφως ανάλογη του k. Αυτό οδηγεί σε απλές δομές δεδομένων, οι οποίες είναι αποδοτικές ως προς τον απαιτούμενο χώρο αποθήκευσης και υποστηρίζουν ερωτήματα σε υπογραμμικό χρόνο. Χρησιμοποιώντας ιδιότητες συγκεκριμένων συναρτήσεων LSH, εκμεταλλευόμαστε μια παρόμοια απεικόνιση στον χώρο Hamming.

Χαμηλή εγγενής διάσταση και απόσταση Μανχάταν. Το πρωταρχικό μας κίνητρο είναι το πρόβλημα πλησιέστερου γείτονα στον μετρικό χώρο ℓ_1, για σημεία με χαμηλή εγγενή διάσταση. Η διάσταση διπλασιασμού είναι μια καθιερωμένη έννοια εγγενούς διάστασης των σημείων. Εμβυθίσεις που διατηρούν τον κοντινότερο γείτονα υπάρχουν τόσο για ℓ_2 όσο και για ℓ_1 μετρικές, καθώς και για υποσύνολα του ℓ_2 με χαμηλή διάσταση διπλασιασμού. Προτείνουμε μια τεχνική μείωσης διάστασης που διατηρεί τον *κοντινό* γείτονα για υποσύνολα του ℓ_1 με χαμηλή διάσταση διπλασιασμού [40].

Προσεγγιστικά r-δίκτυα. Τα r-δίκτυα προσφέρουν ένα ισχυρό εργαλείο στην υπολογιστική και τη μετρική γεωμετρία, δεδομένου ότι χρησιμεύουν ως υποσύνολο καλών αντιπροσώπων: όλα τα σημεία βρίσκονται σε απόσταση r από κάποιο σημείο του r-δικτύου και όλα τα κέντρα του r-δικτύου είναι σε απόσταση τουλάχιστον r μεταξύ τους. Εστιάζουμε σε χώρους μεγάλης διαστάσεως και παρουσιάζουμε έναν νέο πιθανοτικό αλγόριθμο ο οποίος υπολογίζει αποτελεσματικά προσεγγιστικά r-δίκτυα σε Ευκλείδειους χώρους [19]. Ο αλγόριθμός μας ακολουθεί μια πρόσφατη προσέγγιση του Valiant για τη αναγωγή του προβλήματος στην αποτίμηση πολλαπλών σημείων πολυωνύμων.

Προβλήματα εγγύτητας για πολυγωνικές καμπύλες. Προτείνουμε απλές και αποτελεσματικές δομές δεδομένων, βασισμένες σε τυχαίες προβολές, για μια έννοια της απόστασης μεταξύ διακριτοποιημένων καμπυλών, η οποία γενικεύει την διακριτή απόσταση Fréchet και την απόσταση Dynamic Time Warping. Προσφέρουμε τις πρώτες δομές δεδομένων για την εύρεση του κοντινότερου γείτονα με αυθαίρετα καλό συντελεστή προσέγγισης, με ταυτόχρονη αύξηση του χώρου σε σχέση με τις υπάρχουσες μεθόδους [41].

Προβλήματα εγγύτητας για καμπύλες επερώτησης μικρού μήκους. Προτείνουμε δομές δεδομένων, βασισμένες σε τυχαίες διαμερίσεις του χώρου, για την διακριτή απόσταση Fréchet όταν καμπύλες επερώτησης είναι μικρού μήκους. Οι δομές δεδομένων είναι ιδιαίτερα αποτελεσματικές όταν τα ερωτήματα είναι πολύ μικρότερα από τις πολυγωνικές καμπύλες που ανήκουν στο σύνολο δεδομένων. Επίσης, μελετάμε το πρόβλημα για αυθαίρετους μετρικούς χώρους με χαμηλή διάσταση διπλασιασμού.

Η VC διάσταση πολυγωνικών καμπυλών. Η διάσταση Vapnik-Chervonenkis παρέχει μια έννοια πολυπλοκότητας για συστήματα συνόλων ή εύρους. Αναλύουμε συστήματα εύρους όπου το βασικό σύνολο είναι ένα σύνολο πολυγωνικών καμπυλών στον Ευκλείδειο χώρο και εύρη είναι μετρικές μπάλες που ορίζονται από συναρτήσεις αποστάσεων για καμπύλες, όπως η απόσταση Fréchet και η απόσταση Hausdorff [36]. Ακολουθούν άμεσες συνέπειες εφαρμόζοντας γνωστά αποτελέσματα δειγματοληψίας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Γεωμετρία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κοντινότερος γείτονας, υψηλή διάσταση, πολυγωνικές καμπύλες

ΣΥΝΟΠΤΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Με τον αυξανόμενο όγκο δεδομένων και καθώς η πρόσβαση σε δεδομένα υψηλής πολυπλοκότητας γίνεται ευκολότερη, δημιουργείται ανάγκη για αλγοριθμικές λύσεις σε βασικά υπολογιστικά προβλήματα, των οποίων η απόδοση θα κλιμακώνει ομαλά με την αύξηση της πολυπλοκότητας των δεδομένων εισόδου. Στην διατριβή αυτή μελετάμε προβλήματα εγγύτητας, προβλήματα δηλαδή στα οποία η είσοδος είναι ένα σύνολο αντικειμένων και υπονοείται μια συνάρτηση ομοιότητας ή απόστασης μεταξύ των αντικειμένων. Ίσως το σημαντικότερο πρόβλημα σε αυτή την περιοχή είναι το πρόβλημα του κοντινότερου γείτονα, στο οποίο πρέπει να σχεδιαστεί δομή δεδομένων η οποία αποθηκεύει ένα σύνολο αντικειμένων και υποστηρίζει την εύρεση του περισσότερο όμοιου αντικειμένου από το αποθηκευμένο σύνολο, σε σχέση με ένα νέο αντικείμενο. Προφανώς η εύρεση του κοντινότερου γείτονα πρέπει να γίνεται σε χρόνο αρκετά μικρότερο από αυτόν που θα απαιτείτο αν ελέγχαμε ένα προς ένα όλα τα αποθηκευμένα δεδομένα. Σχετικός τύπος προβλημάτων είναι αυτός της συσταδοποίησης. Στην συσταδοποίηση, η είσοδος του προβλήματος είναι πάλι ένα σύνολο αντικειμένων και μια συνάρτηση απόστασης και στόχος είναι η διαμέριση των αντικειμένων σε συστάδες: όμοια αντικείμενα πρέπει να ανήκουν στην ίδια συστάδα για την οποία συνήθως υπάρχει κάποιο αντικείμενο που αποτελεί “αντιπρόσωπο”.

Οι βασικοί τύποι δεδομένων που μελετάμε στην διατριβή αυτή, είναι δύο. Ο πρώτος τύπος είναι το απλό διάνυσμα ή σημείο, ή αλλιώς πλειάδα πραγματικών αριθμών. Κάθε συντεταγμένη μπορεί να θεωρηθεί ότι αντιστοιχεί σε ένα διαφορετικό γνώρισμα των δεδομένων. Η υψηλή πολυπλοκότητα των δεδομένων σε αυτό το πλαίσιο μεταφράζεται ώς υψηλή διάσταση των διανυσμάτων, δηλαδή τα δεδομένα έχουν πολλά γνωρίσματα. Ως συναρτήσεις απόστασης θεωρούμε κλασσικά παραδείγματα νορμών και πιο συγκεκριμένα την Ευκλείδεια απόσταση ή την απόσταση Μανχάταν. Ο δεύτερος τύπος δεδομένων που εξετάζεται στην διατριβή είναι η ακολουθία διανυσμάτων η οποία ορίζει μια πολυγωνική καμπύλη. Παραδείγματα τέτοιων δεδομένων είναι οι τροχιές από GPS ή οι χρονοσειρές. Σε αυτή την περίπτωση, οι πιο δημοφιλείς συναρτήσεις απόστασης είναι η Fréchet απόσταση (ή η διακριτή εκδοχή της) και η απόσταση Dynamic Time Warping. Η υψηλή πολυπλοκότητα τέτοιων δεδομένων μεταφράζεται είτε ως υψηλή διάσταση των διανυσμάτων είτε ως μεγάλο μήκος των ακολουθιών.

Καθώς οι βασικές λύσεις για προβλήματα εγγύτητας βασίζονται σε προσεκτικό διαχωρισμό του χώρου, αναπόφευκτα αυτές αποτυγχάνουν όταν η διάσταση είναι υψηλή καθώς ο όγκος του περιβάλλοντος χώρου αυξάνεται εκθετικά. Ως εκ τούτου, βασιζόμαστε σε μεθόδους μείωσης της διάστασης μέσω τυχαίων προβολών. Το κύριο αποτέλεσμα σε αυτή την περιοχή είναι το Johnson-Lindenstrauss λήμμα, το οποίο αποτυπώνει το εξής γεγονός: αν n σημεία στο $\mathbb{R}^d$, προβληθούν σε έναν τυχαίο υπόχωρο διάστασης περίπου $\log n$, τότε με καλή πιθανότητα οι Ευκλείδειες αποστάσεις δεν θα μεταβληθούν πολύ αν εξαιρέσουμε έναν κοινό πολλαπλασιαστικό παράγοντα. Προτείνουμε μια διαφορετική εφαρμογή του λήμματος, προσαρμοσμένη στις ανάγκες του προβλήματος εύρεσης του προσεγγιστικού κοντινότερου γείτονα. Για ένα πρόβλημα συσταδοποίησης σημείων, βασιζόμαστε σε πρόσφατα αποτελέσματα για την εύρεση κοντινότερου ζευγαριού μέσω γρήγορου πολλαπλασιασμού πίνακα και δείχνουμε ότι αντίστοιχες βελτιώσεις μπορούν να επεκταθούν και σε άλλα προβλήματα. Στην περίπτωση των πολυγωνικών καμπυλών, προτείνουμε μεθόδους για την εύρεση προσεγγιστικού κοντινότερου γείτονα και μελετάμε την Vapnik–Chervonenkis (VC) διάσταση για τους ψευδομετρικούς χώρους που ορίζονται από τις αντίστοιχες συναρτήσεις αποστάσεων. Η μελέτη της VC διάστασης βρίσκει εφαρμογές μέσω κλασσικών μεθόδων δειγματοληψίας. Ακολουθεί μια πιο λεπτομερής επισκόπηση των αποτελεσμάτων.

Κοντινότερος γείτονας. Έστω P ένα σύνολο n σημείων σε κάποιο μετρικό χώρο $(\mathcal{M}, \mathsf{d})$. Το πρόβλημα συνίσταται στην δημιουργία μιας δομής δεδομένων τέτοια ώστε, για οποιοδήποτε σημείο ερωτήματος $q \in \mathcal{M}$, η δομή επιστρέφει σημείο $p \in P$ για το οποίο $\mathsf{d}(p,q) \leq \mathsf{d}(p',q)$, για κάθε $p' \in P$. Τότε το σημείο p είναι ένας κοντινότερος γείτονας του q. Συχνά, μια λύση με απόλυτη ακρίβεια για την αναζήτηση κοντινότερου γείτονα απαιτεί απαγορευτικά βαρείς πόρους. Έτσι, οι περισσότερες λύσεις επικεντρώνονται στο λιγότερο απαιτητικό πρόβλημα της εύρεσης του *προσεγγιστικού κοντινότερου γείτονα*. Για οποιοδήποτε μετρικό χώρο $(\mathcal{M}, \mathsf{d})$, και δεδομένου πεπερασμένου συνόλου $P \subset \mathcal{M}$ και πραγματικής παραμέτρου $\epsilon > 0$, ενας $(1+\epsilon)$-προσεγγιστικός κοντινότερος γείτονας σε ένα σημείο επερώτησης $q \in \mathcal{M}$ είναι ένα σημείο $p \in P$ τέτοιο ώστε

$$\mathsf{d}(q,p) \leq (1+\epsilon) \cdot \mathsf{d}(q,p'), \quad \text{for all } p' \in P.$$

Ως εκ τούτου, στοχεύοντας σε μια προσεγγιστική λύση, η απάντηση μπορεί να είναι οποιοδήποτε σημείο του οποίου η απόσταση από το q είναι το πολύ $(1+\epsilon)$ φορές μεγαλύτερη από την απόσταση μεταξύ του q και του πραγματικού κοντινότερου γείτονα.

Το αντίστοιχο πρόβλημα απόφασης (με μάρτυρα) είναι γνωστό ως το πρόβλημα εύρεσης ενός *κοντινού* γείτονα, το οποίο ορίζεται ως εξής.

Ορισμός 1. *Έστω $P \subseteq \mathcal{M}$, με $|P| = n$, όπου $(\mathcal{M}, \mathsf{d})$ κάποιος μετρικός χώρος. Δεδομένου $\epsilon > 0, r > 0$, ζητείται δομή δεδομένων για την οποία, για κάθε επερώτημα $q \in \mathcal{M}$,*

- *αν $\exists p^* \in P$ s.t. $\mathsf{d}(p^*,q) \leq r$, τότε η δομή επιστρέφει οποιοδήποτε $p' \in P$ τ.ω. $\mathsf{d}(p',q) \leq (1+\epsilon) \cdot r$,*
- *αν $\forall p \in P$, $\mathsf{d}(p,q) > (1+\epsilon) \cdot r$, τότε η δομή επιστρέφει “Αποτυχία”.*

Η δομή επιτρέπεται να επιστρέψει είτε ένα σημείο σε απόσταση $\leq (1+\epsilon)r$ είτε το μήνυμα “Αποτυχία”.

Είναι γνωστό ότι το πρόβλημα εύρεσης του προσεγγιστικού κοντινότερου γείτονα μπορεί να λυθεί λύνοντας λογαριθμικά πολλές περιπτώσεις του προβλήματος απόφασης με μάρτυρα [51].

Κοντινότερος γείτονας στον Ευκλείδειο χώρο. Οι ντετερμινιστικές τεχνικές διαμερισμού του χώρου, όπως τα kd-δέντρα, τα BBD-δέντρα και τα Voronoi διαγράμματα, παρέχουν αποδοτικές λύσεις όταν η διάσταση είναι σχετικά χαμηλή αλλά επηρεάζονται από την κατάρα της διαστασιμότητας. Προς επίλυση αυτού του ζητήματος, έχουν προταθεί τυχαιοκρατικές μέθοδοι όπως το Locality Sensitive Hashing (LSH), μια δομή που βασίζεται σε τυχαιοκρατικό κατακερματισμό ώστε κοντινά σημεία να τείνουν να ανήκουν στην ίδια συστάδα κατακερματισμού. Κάποιος μπορεί επίσης να εφαρμόσει το Johnson-Lindenstrauss λήμμα σε συνδυασμό με τεχνικές για χαμηλές διαστάσεις. Το πρόβλημα που προκύπτει είναι είτε ότι η χώρος που απαιτείται από την δομή είναι της τάξης του $\omega(n)$ είτε ο χρόνος επερώτησης είναι $\omega(n)$. Εμείς εστιάζουμε στο σενάριο κατά το οποίο στοχεύουμε σε χώρο $O(dn)$ και παράλληλα σε χρόνο επερώτησης $o(n)$.

Για τον σκοπό αυτό εισάγουμε μια νέα έννοια "χαμηλής ποιότητας" τυχαιοκρατικών εμβυθίσεων και χρησιμοποιούμε τυχαίες προβολές à la Johnson-Lindenstrauss για να ορίσουμε μια συνάρτηση από το ℓ_2^d στο $\ell_2^{d'}$, όπου

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

τέτοια ώστε ένας προσεγγιστικός κοντινότερος γείτονας στον αρχικό χώρο να βρίσκεται ανάμεσα στους k προσεγγιστικούς κοντινότερους γείτονες στον νέο χώρο. Αυτή η παρατήρηση μας επιτρέπει να συνδυάσουμε τις τυχαίες προβολές με την μέθοδο πλέγματος [51], και να κατασκευάσουμε μια τυχαιοκρατική δομή δεδομένων για το πρόβλημα απόφασης με μάρτυρα με βέλτιστο χώρο και υπογραμμικό χρόνο απόκρισης ερωτήματος.

Πιο συγκεκριμένα, μετά την τυχαία προβολή στο $\ell_2^{d'}$, εφαρμόζουμε ένα πλέγμα με μήκος πλευράς κελιού $\epsilon/\sqrt{d'}$ και για κάθε σημείο επερώτησης, εξερευνούμε γειτονικά κελιά που τέμνουν την Ευκλείδεια μπάλα η οποία περιέχει $O(1/\epsilon)^{d'}$ κελιά. Ο αλγόριθμος σταματάει αφού εξετάσει k σημεία που περιέχονται στην μπάλα ή όλα τα κελιά της μπάλας. Θέτοντας κατάλληλα τις παραμέτρους πετυχαίνουμε γραμμικό χρόνο δημιουργίας της δομής, γραμμικό χώρο, και χρόνο επερώτησης $O(dn^{\rho})$, όπου $\rho = 1 - \Theta(\epsilon^2/\log(1/\epsilon))$. Για κάθε σημείο επερώτησης $q \in \mathbb{R}^d$, η κατασκευή της δομής πετυχαίνει με σταθερή πιθανότητα, η οποία μπορεί να ενισχυθεί δημιουργώντας πολλές ανεξάρτητες δομές. Επίσης επεκτείνουμε το αποτέλεσμα για υποσύνολα με χαμηλή *εγγενή* διάσταση.

Χρησιμοποιώντας γνωστές αναγωγές [51], μπορούμε να σχεδιάσουμε λύση για το πρόβλημα του προσεγγιστικού κοντινότερου γείτονα χρησιμοποιώντας την παραπάνω δομή. Η ιδέα όμως μπορεί να εφαρμοστεί απευθείας στο πρόβλημα του προσεγγιστικού κοντινότερου γείτονα, για παράδειγμα χτίζοντας ένα Balanced Box-Decomposition (BBD) δέντρο στο νέο χώρο διάστασης d'. Ο συνδυασμός αυτός επιτυγχάνει πιο αδύναμα φράγματα αλλά μπορεί να βρει πρακτική αξία λόγω της απλότητας του.

Κοντινός γείτονας για μετρικούς χώρους με LSH. Οι παραπάνω ιδέες μπορούν να επεκταθούν για οποιοδήποτε μετρικό χώρο για τον οποίο υπάρχουν LSH συναρτήσεις. Η τυχαία προβολή σε αυτή την περίπτωση ορίζεται από τον αρχικό μετρικό χώρο προς τον χώρο Hamming $(\{0,1\}^{d'}, \|\cdot\|_1)$. Η παρατήρησή αυτή οδηγεί σε μια βελτίωση στον χρόνο επερώτησης, στην περίπτωση του Ευκλείδειου χώρου. Θέτοντας κατάλληλα τις παραμέτρους πετυχαίνουμε γραμμικό χρόνο δημιουργίας της δομής, γραμμικό χώρο, και χρόνο επερώτησης $O(dn^{\rho})$, όπου $\rho = 1 - \Theta(\epsilon^2)$.

Κοντινός γείτονας και μείωση διάστασης για ℓ_1. Η μείωση διάστασης με διατήρηση όλων των αποστάσεων στον μετρικό χώρο ℓ_1, είναι γνωστό ότι αποτελεί δύσκολο εγχείρημα, ακόμα και όταν η διάσταση διπλασιασμού (doubling dimension) των σημείων είναι αρκετά μικρή [66]. Στην διατριβή αυτή μελετάμε απλές τεχνικές μείωσης διάστασης που δεν διατηρούν όλες τις αποστάσεις αλλά διατηρούν πληροφορία που είναι αρκετή για το πρόβλημα του κοντινού γείτονα. Με άλλα λόγια προσφέρουμε μια αναγωγή του προβλήματος εύρεσης κοντινού γείτονα σε υψηλή διάσταση στο αντίστοιχο πρόβλημα σε χαμηλή διάσταση. Για n σημεία στο ℓ_1^d, και για γραμμικό χρόνο προβολής, πετυχαίνουμε διάσταση προβολής πολυωνυμική στο $\log\log n$, όταν η εγγενής διάσταση θεωρείται σταθερή. Παρ'ότι οι συνέπειες του αποτελέσματος δεν περιλαμβάνουν νέα θεωρητικά φράγματα για δομές δεδομένων, η μείωση διάστασης προσφέρει διάφορα πλεονεκτήματα, όπως την μείωση μνήμης που απαιτείται ανά σημείο.

Ένα πρόβλημα συσταδοποίησης. Μελετάμε τα r-δίκτυα (r-nets), ένα χρήσιμο εργαλείο της υπολογιστικής και της μετρικής γεωμετρίας, με πληθώρα εφαρμογών στους προσεγγιστικούς αλγορίθμους. Ένα r-δίκτυο για ένα σημειοσύνολο P στον Ευκλείδειο χώρο $(\mathbb{R}^d, \|\cdot\|_2)$, και για αριθμητική παράμετρο r είναι ένα υποσύνολο $\mathcal{N} \subseteq P$ τέτοιο ώστε οι κλειστές μπάλες ακτίνας $r/2$ με κέντρα τα σημεία του $\mathcal{N}$ είναι ξένες μεταξύ τους, και οι κλειστές μπάλες ακτίνας r με κέντρα τα ίδια σημεία καλύπτουν όλο το P. Ορίζουμε ανάλογα τα προσεγγιστικά r-δίκτυα.

Ορισμός 2. *Δοθέντος ενός σημειοσυνόλου $P \subseteq \mathbb{R}^d$, μιας παραμέτρου ακτίνας $r > 0$ και μιας παραμέτρους προσέγγισης $\epsilon > 0$, ένα $(1+\epsilon)$-προσεγγιστικό r-δίκτυο του P είναι ένα υποσύνολο $\mathcal{N} \subseteq P$ που ικανοποιεί τις ακόλουθες ιδιότητες:*

1. *Για κάθε $p, q \in \mathcal{N}$, $p \neq q$, έχουμε ότι $\|p - q\|_2 \geq r$.*

2. *Για κάθε $p \in P$, υπάρχει ένα $q \in \mathcal{N}$ s.t. $\|p - q\|_2 \leq (1+\epsilon)r$.*

Ο υπολογισμός ενός r-δικτύου μπορεί να γίνει με έναν πολύ απλό τρόπο: θεωρούμε αρχικά όλα τα σημεία του P μη-καλυμμένα, και επαναληπτικά επιλέγουμε κέντρα από το σύνολο ακάλυπτων σημείων και με αυτά καλύπτουμε σημεία από το σύνολο ακάλυπτων σημείων. Η διαδικασία σταματάει όταν καλυφθούν όλα τα σημεία. Όταν η διάσταση είναι χαμηλή, η παραπάνω διαδικασία μπορεί να γίνει με πιο αποδοτικό τρόπο χρησιμοποιώντας πλέγματα και πίνακες κατακερματισμού [49].

Όταν η διάσταση είναι υψηλή στοχεύουμε ξανά σε πολυπλοκότητα πολυωνυμική στην διάσταση. Επίσης η εξάρτηση στο πλήθος σημείων πρέπει να είναι σαφώς μικρότερη από $O(n^2)$, όπου $|P| = n$, αφού τόσο κοστίζει να εξετάσουμε όλες τις αποστάσεις. Μια προσέγγιση η οποία υπολογίζει $(1+\epsilon)$-προσεγγιστικά r-δίκτυα σε υψηλή διάσταση [42], χρησιμοποιεί LSH. Για αρκετά μικρό $\epsilon > 0$, η πολυπλοκότητα χρόνου είναι $\tilde{O}(dn^{2-\Theta(\epsilon)})$, όπου το $\tilde{O}$ κρύβει πολυλογαριθμικούς παράγοντες.

Γενικά πολλά από τα προβλήματα εγγύτητας έχουν επιλυθεί σε υψηλές διαστάσεις μέσω του LSH. Για παράδειγμα το πρόβλημα της εύρεσης του προσεγγιστικού κοντινότερου ζευγαριού ανάμεσα σε n σημεία σε διάσταση d μπορεί να επιλυθεί σε χρόνο $\tilde{O}(dn^{2-\Theta(\epsilon)})$. Πρόσφατα, ο Valiant [77] παρουσίασε έναν αλγόριθμο για τη εύρεση του προσεγγιστικού κοντινότερου ζευγαριού σε χρόνο $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$ ο οποίος δεν βασίζεται στο LSH. Αυτή η διαφορετική προσέγγιση βασίζεται στον γρήγορο υπολογισμό πινάκων για την αποτίμηση πολυωνύμων.

Επεκτείνουμε τον αλγόριθμο του Valiant και υπολογίζουμε προσεγγιστικά r-δίκτυα σε χρόνο $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$, βελτιώνοντας έτσι τον χρόνο του αλγορίθμου που βασίζεται στο LSH, όταν το ϵ είναι αρκετά μικρό. Η βελτίωση αυτή είναι αντίστοιχη της βελτίωσης για το πρόβλημα του κοντινότερου ζευγαριού. Η μελέτη μας ωθείται από τις διάφορες εφαρμογές των r-δικτύων στους προσεγγιστικούς αλγόριθμους [54].

Μια ενιαία αντιμετώπιση προβλημάτων εγγύτητας καμπυλών. Υπάρχουν διάφοροι τρόποι καθορισμού της ανομοιότητας ή της απόστασης μεταξύ δύο καμπυλών. Τα δύο πιο δημοφιλή μέτρα ανισότητας είναι η διακριτή απόσταση Fréchet (Discrete Fréchet distance ή DFD) και η απόσταση Δυναμικής Χρονικής Στρέβλωσης (Dynamic Time Warping ή DTW), οι οποίες είναι ευρέως μελετημένες και εφαρμόζονται σε προβλήματα ταξινόμησης και ανάκτησης για διάφορους τύπους δεδομένων. Το DFD είναι ψευδό-μετρική, σε αντίθεση με το DTW που δεν ικανοποιεί την τριγωνική ανισότητα. Είναι σύνηθες, στις αποστάσεις των καμπυλών, να χρησιμοποιείται η έννοια της διάσχισης (traversal) για δύο καμπύλες. Διαισθητικά, μια διάσχιση αντιστοιχεί σε ένα χρονοδιάγραμμα σύμφωνα με το οποίο διασχίζουμε τις δύο καμπύλες ταυτόχρονα, ξεκινώντας από το πρώτο σημείο κάθε καμπύλης και τελειώνοντας στο τελευταίο σημείο κάθε καμπύλης. Με την πάροδο του χρόνου, η διάσχιση προχωράει σε τουλάχιστον μία από τις δύο καμπύλες. Το DFD είναι η ελάχιστη (ως προς τις διασχίσεις) μέγιστη απόσταση των σημείων κατά την διάσχιση. Το DTW είναι το ελάχιστο (ως προς τις διασχίσεις), άθροισμα των αποστάσεων κατά τη διάσχιση.

Παρουσιάζουμε μια έννοια απόστασης καμπυλών που γενικεύει τις DFD και DTW. Η *ℓ_p-απόσταση* δύο καμπυλών ελαχιστοποιεί, ως προς όλες τις διασχίσεις, την ℓ_p νόρμα του διανύσματος όλων των Ευκλείδειων αποστάσεων μεταξύ σημείων που επισκέπτονται ταυτόχρονα κατά την διάσχιση. Ως εκ τούτου, η DFD αντιστοιχεί στην ℓ_∞-απόσταση πολυγωνικών καμπυλών, και το DTW αντιστοιχεί στην ℓ_1-απόσταση.

Η βασική μας συνεισφορά είναι μια δομή δεδομένων για το πρόβλημα της εύρεσης του προσεγγιστικού κοντινότερου γείτονα για τις ℓ_p-αποστάσεις πολυγωνικών καμπυλών, όταν $1 \leq p < \infty$. Αυτό επεκτείνεται εύκολα για την ℓ_∞-απόσταση καμπυλών λύνοντας για την ℓ_p-απόσταση, όπου το p επιλέγεται να είναι αρκετά μεγάλο. Στόχος μας είναι $1+\epsilon$ προσέγγιση. Τέτοιοι προσεγγιστικοί παράγοντες επιτυγχάνονται για πρώτη φορά, θυσιάζοντας σε χωρικές απαιτήσεις της δομής. Ένα επιπλέον πλεονέκτημα είναι ότι οι μέθοδοι μας λύνουν απευθείας το πρόβλημα του κοντινότερου γείτονα. Παρότι υπάρχουν γνωστές αναγωγές στο πρόβλημα του κοντινού γείτονα για μετρικούς χώρους, δεν είναι γνωστό αν αντίστοιχες αναγωγές μπορούν να λειτουργήσουν σε μη-μετρικές αποστάσεις όπως το DTW.

Συγκεκριμένα, όταν $p > 2$, για n καμπύλες πολυπλοκότητας m, σχεδιάζουμε δομή δεδομένων με χώρο και χρόνο προεπεξεργασίας

$$\tilde{O}\left(n \cdot \left(\frac{d}{p\epsilon} + 2\right)^{O(dm \cdot \alpha_{p,\epsilon})}\right),$$

όπου το $\alpha_{p,\epsilon}$ εξαρτάται μόνο από τα p, ϵ, και ο χρόνος επερώτησης είναι $\tilde{O}(2^{4m} \log n)$.

Δομές δεδομένων για ερωτήματα χαμηλής πολυπλοκότητας. Όταν μελετάμε προβλήματα εγγύτητας για καμπύλες είναι φυσιολογικό να υποθέσουμε ότι οι καμπύλες επερώτησης δεν είναι ίδιας πολυπλοκότητας με αυτές που αποτελούν το σύνολο δεδομένων. Εστιάζουμε στην περίπτωση όπου οι καμπύλες επερώτησης αποτελούνται από μικρότερο πλήθος κορυφών.

Για την διακριτή απόσταση Fréchet στον Ευκλείδειο χώρο, δίνουμε μια τυχαιοκρατική δομή δεδομένων με χώρο $n \cdot O\left(\frac{kd^{3/2}}{\epsilon}\right)^{dk} + O(dnm)$ και χρόνο ερωτήματος σε $O\left(dk\right)$, όπου το k δηλώνει το μήκος της καμπύλης ερωτήματος. Ο αλγόριθμος βασίζεται σε τυχαίες διαμερίσεις του χώρου, και πιο συγκεκριμένα σε διαμερίσεις που δημιουργούνται από τυχαία μετατοπισμένα πλέγματα. Η δομή των δεδομένων μπορεί να γίνει ντετερμινιστική με ελαφρά επιδείνωση της απόδοσης.

Για αυθαίρετους μετρικούς χώρους με χαμηλή διάσταση διπλασιασμού, δίνουμε ανάλογα αποτελέσματα, αλλά η επιτυγχανόμενη απόδοση εξαρτάται από τις υποθέσεις που σχετίζονται με το πως έχουμε πρόσβαση στον εν λόγω μετρικό χώρο. Ο αλγόριθμος βασίζεται και πάλι σε τυχαίες διαμερίσεις του χώρου οι οποίες υλοποιούνται με δομές δεδομένων για προβλήματα εγγύτητας σημείων σε γενικούς μετρικούς χώρους.

Η VC-διάσταση για πολυγωνικές καμπύλες. Ένας *χώρος εύρους* $(X, \mathcal{R})$ (γνωστό επίσης ως *σύστημα συνόλων*) ορίζεται από ένα σύνολο X και ένα σύνολο από σύνολα $\mathcal{R}$, όπου κάθε $r \in \mathcal{R}$ είναι ένα υποσύνολο του X. Ένας βασικός δείκτης για κάθε χώρο εύρους είναι η VC-διάσταση [79]. Η έννοια αυτή ποσοτικοποιεί το πόσο περίπλοκος είναι ένας χώρος εύρους και έχει παίξει θεμελιώδη ρόλο στην μηχανική μάθηση, στις δομές δεδομένων και στην υπολογιστική γεωμετρία.

Η βασική μας συνεισφορά είναι η ανάλυση της VC-διάστασης για χώρους εύρους που ορίζονται από πολυγωνικές καμπύλες. Το σύνολο X αποτελείται από πολυγωνικές καμπύλες με m κορυφές και το σύνολο $\mathcal{R}$ ορίζεται από "μπάλες" της Fréchet απόστασης με κέντρο πολυγωνικές καμπύλες με k κορυφές. Αντίστοιχη ανάλυση γίνεται και για την Hausdorff απόσταση. Ειδικότερα για την απόσταση Fréchet και για την απόσταση Hausdorff για καμπύλες στο επίπεδο δείχνουμε ότι η VC-διάσταση είναι τάξης του $O(k^2 \log(mk))$.

Η ανάλυση μας γίνεται με την διάσπαση του βασικού χώρου εύρους σε επιμέρους απλούστερους χώρους εύρους. Αφού φράξουμε την VC-διάσταση των απλούστερων χώρων και αφού δείξουμε ότι η διάσπαση αυτή είναι ορθή, μπορούμε να συνθέσουμε και να βγάλουμε το επιθυμητό συμπέρασμα για τον αρχικό χώρο εύρους.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Ioannis Emiris for his support and his continuous effort to maintain a workspace that inspires creativity in an otherwise problematic environment, especially due to the economic crisis and its negative effects on academia. His positivity really helped me overcome both academic and non-academic challenges.

I would like to thank Anastasios Sidiropoulos for his hospitality and help during the time I spent in Chicago. I would also like to thank Anne Driemel for inviting me to Bonn and starting a collaboration with her. I am thankful to all of my coauthors and my labmates; I had a great time working with all of them.

I would like to thank my family for all the support. I would like to thank my friends for keeping me sane and balanced, and my partner for being very supportive.

This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Proximity problems

Nearest neighbor searching is a fundamental computational problem with several applications in Computer Science and beyond. The setting is very clear: we need to preprocess a set of objects in a way which assists proximity queries, i.e. when a query object arrives, we should be able to retrieve the most similar object among the set of preprocessed objects. The dissimilarity or distance function typically depends on the context and affects the performance of the solution. 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 from mobile phones, GPS sensors, RFID technology and video analysis.

Definition 1 (Nearest Neighbor (NN) problem)**.** *Given a set of objects P which is a finite subset of some ambient set M, and a distance function* $\mathsf{d}(\cdot,\cdot)$*, preprocess P into a data structure which supports the following type of queries:*

$$\text{for any object } q \text{ in } M, \text{ find } p^* \text{ such that for all } p \text{ in } P: \ \mathsf{d}(q,p^*) \leq \mathsf{d}(q,p).$$

Obviously, a naive linear scan provides a stable and easy-to-implement solution. The problem gets really intriguing when we aim for strictly sublinear query time. Then, we hope that we can exploit properties of the distance function during preprocessing. To simplify things, we may assume that objects live in a metric space, i.e. (M,d) defines a metric. Moreover, we can restrict ourselves to some of the most well-studied metrics, e.g. the Euclidean metric. In particular, for low dimensional Euclidean spaces, we obtain simple solutions. For dimension $d=1$, all points lie on the real line and one can sort them so that any query reduces to a simple binary search. For $d=2$, the solution relies on the notion of *Voronoi Diagram*, one of the most classical structures in Computational Geometry.

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

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

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

Definition 3 ((c, r)-ANN Problem)**.** *Given a finite set $P \subset M$, a distance function* $\mathsf{d}(\cdot, \cdot)$*, 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.* $\mathsf{d}(p^*, q) \leq r$*, then return any point $p' \in M$ s.t.* $\mathsf{d}(p', q) \leq c \cdot r$*,*
- *if $\forall p \in P$,* $\mathsf{d}(p, q) > c \cdot r$*, 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 [51].

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. These notions are very useful since they lead to an economical representation of a pointset, while preserving the structure up to a scale $O(cr)$.

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 [79, 75, 74]. These notions quantify how complex a range space is, and have played foundational roles in machine learning [80, 13], data structures [29], and geometry [50, 26].

Unless otherwise explicitly stated, $\log(\cdot)$ is the logarithm with base 2.

1.2 Related work

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

1.2.1 Normed spaces

This section details results that existed prior to this thesis, and results which appeared concurrently. Unless otherwise stated, the results concern the case of points in ℓ_2.

An exact solution to high-dimensional nearest neighbor search, in sublinear time, requires heavy resources. One notable approach to the problem [69] shows that nearest neighbor queries can be answered in $O(d^5 \log n)$ time, using $O(n^{d+\delta})$ space, for arbitrary $\delta > 0$.

In [16], they introduced the Balanced Box Decomposition (BBD) trees. BBD-trees achieve query time $O(c_d \log n)$ with $c_d \leq d/2\lceil 1 + 6d/\epsilon \rceil^d$, using space in $O(dn)$, and preprocessing time in $O(dn \log n)$. BBD-trees can be used to retrieve the $k \geq 1$ approximate nearest-neighbors at an extra cost of $O(d \log n)$ per neighbor. BBD-trees have proved to be very practical, as well, and have been implemented in software library `ANN`.

Another relevant data structure is the Approximate Voronoi Diagrams (AVD). They are shown to establish a tradeoff between the space complexity of the data structure and the query time it supports [15]. With a tradeoff parameter $2 \leq \gamma \leq \frac{1}{\epsilon}$, the query time is in $O(\log(n\gamma) + 1/(\epsilon\gamma)^{\frac{d-1}{2}})$ and the space in $O(n\gamma^{d-1} \log \frac{1}{\epsilon})$. They are implemented on a hierarchical quadtree-based subdivision of space into cells, each storing a number of representative points, such that for any query point lying in the cell, at least one of the representatives is an approximate nearest neighbor. Further improvements to the space-time trade offs for ANN are obtained in [14].

One might apply the Johnson-Lindenstrauss Lemma and map the points to $O(\epsilon^{-2} \log n)$ dimensions with distortion equal to $1 + \epsilon$ aiming at improving complexity. In particular, AVD combined with the Johnson-Lindenstrauss Lemma have query time polynomial in $\log n$, d and $1/\epsilon$ but require $n^{O(\log(1/\epsilon)/\epsilon^2)}$ space, which is prohibitive if $\epsilon \ll 1$. Notice that we relate the approximation error with the distortion for simplicity.

In high dimensional spaces, classic space partitioning data structures are affected by the curse of dimensionality, as illustrated above. 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). LSH induces a data independent random partition and is dynamic, since it supports insertions and deletions. It relies on the existence of locality sensitive hash functions, which are more likely to map similar objects to the same bucket. The existence of such functions depends on the metric space. In general, LSH requires roughly $O(dn^{1+\rho})$ space and $O(dn^{\rho})$ query time for some parameter $\rho \in (0, 1)$. It has been shown [10] that in the Euclidean case, one can have $\rho = 1/(1 + \epsilon)^2$, which matches the lower bound of hashing algorithms proved in [71]. Lately, it was shown that it is possible to overcome this limitation by switching to a data-dependent scheme which achieves $\rho = \frac{1}{2(1+\epsilon)^2-1} + o(1)$ [12].

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. One approach [73] achieves query time proportional to $dn^{O(1/(1+\epsilon))}$ which is sublinear only when ϵ is large enough. The query time was later improved [10] to $dn^{O(1/(1+\epsilon)^2)}$ which is also sublinear only for large enough ϵ. For comparison, in Theorem 35 we show that it is possible to use near linear space, with query time roughly $O(dn^{\rho})$, where $\rho \approx 1 - \epsilon^2/\log(1/\epsilon)$, achieving sublinear query time even for small values of ϵ.

After the original submission of our paper [8], a better query time of $O(n^{1-4\epsilon^2+O(\epsilon^3)})$ has been established [11]. 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. Heuristics which are related to our method have been successful in practice [76].

Significant amount of work has been done for pointsets with low doubling dimension. For any finite metric space X of doubling dimension $\mathrm{ddim}(X)$, there exists a data structure [52] with expected preprocessing time $O(2^{\mathrm{ddim}(X)} n \log n)$, space usage $O(2^{\mathrm{ddim}(X)} n)$ and query time $O(2^{\mathrm{ddim}(X)} \log n + \epsilon^{-O(\mathrm{ddim}(X))})$. In [58], a new notion of nearest neighbor preserving embeddings has been presented. Moreover, it has been proven that in this context we can achieve dimension reduction which only depends on the doubling dimension of the dataset. Naturally, such an approach can be easily combined with any known data structure for $(1+\epsilon)$-ANN.

Random projection trees [32] have been shown to adapt to pointsets of low doubling dimension. Like kd-trees, every split partitions the pointset into subsets of roughly equal cardinality. Unlike kd-trees, the space is split with respect to a random direction, not necessarily parallel to the coordinate axes. Classic kd-trees also adapt to the doubling dimension of randomly rotated data [81]. However, for both techniques, no related worst-case guarantees about the efficiency of $(1+\epsilon)$-ANN search were given.

In [61], a different notion of intrinsic dimension has been introduced; namely the expansion rate ψ which is formally defined in Subsection 3.3.2. The doubling dimension is a more general notion of intrinsic dimension in the sense that, when a finite metric space has bounded expansion rate, then it also has bounded doubling dimension, but the converse does not hold [48]. Several efficient solutions are known for metrics with bounded expansion rate ψ, including for the problem of exact nearest neighbor. One such solution [63] provides a data-structure which requires $\psi^{O(1)} n$ space and answers queries in $\psi^{O(1)} \ln n$. Moreover, Cover Trees [24] require $O(n)$ space and each query costs $O(\psi^{12} \log n)$ time for exact nearest neighbors. In Theorem 42, we present a data structure for the $(1+\epsilon)$-ANN problem with linear space and $O((\psi^{\log\log\psi}) \cdot d \cdot \log n))$ query time. The result concerns pointsets in d-dimensional Euclidean space.

One related problem is that of computing $(1+\epsilon)$-approximate r-nets. In [52], 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 [42], which uses LSH and requires time $O(n^{1+1/(1+\epsilon)^2+o(1)})$. When ϵ is small enough, we show in Theorem 66 that time complexity can be improved to $O(n^{2-\Theta(\sqrt{\epsilon})})$, without using LSH.

1.2.2 Polygonal curves

The ANN problem has been mainly addressed for datasets consisting of points. Very little is known about distances between curves which, in a sense, are the next more complex type of geometric object. In this thesis, we focus on discrete Fréchet (DFD) and Dynamic

Time Warping (DTW) distance functions.

The first result for DFD by Indyk [55], 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 solution is based on an efficient data structure for ℓ_∞-products of arbitrary metrics, and achieves space and preprocessing time in $O(m^2|X|)^{tm^{1/t}} \cdot n^{2t}$, and query time in $(m \log n)^{O(t)}$. Table 6.1 states these bounds for appropriate $t = 1 + o(1)$, hence a constant approximation factor. 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 [37]. 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 $O(m)$. Our methods can achieve any user-desired approximation factor at the expense of a reasonable increase in the space and time complexities. Furthermore, it is shown that the result establishing an $O(m)$ approximation [37] extends to DTW, whereas the other extreme of the trade-off has remained open. To compare with, we offer the first data structures and query algorithms for $(1+\epsilon)$-ANN with arbitrarily good approximation factor, at the expense of increasing space usage and preprocessing time.

After the publication of our work, a new deterministic data structure [43] was devised, with better query performance.

Notice that all related approaches solve the approximate *near* neighbor problem, which is essentially a decision problem, instead of the optimization $(1+\epsilon)$-ANN. It is known that a data structure for the approximate near neighbor problem can be used as a building block for solving the $(1+\epsilon)$-ANN problem. This procedure has provable guarantees on metrics [51], but it is not clear whether it can be extended to non-metric distances such as the DTW.

1.3 Contribution

1.3.1 Normed spaces

1.3.1.1 Approximate Nearest Neighbors

In Chapter 3, 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 [51], and obtain a randomized data structure with optimal space and sublinear query for the augmented decision problem.

In particular, after a random projection to $\ell_2^{d'}$, we simply employ a grid with cell width $\epsilon/\sqrt{d'}$ and for each query we explore cells inside the approximate Euclidean ball of size $O(1/\epsilon)^{d'}$. The query stops after having examined m candidate points. This is the topic of Section 3.2, and Theorem 35 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 are able to extend our results to doubling subsets of ℓ_2 (see Subsection 3.2.2) by applying our approach to an r-net of the input pointset. The resulting data structure has linear space, preprocessing time which depends on the time required to compute an r-net, and query time $(2/\epsilon)^{O(\mathrm{ddim}(X))}$, where $\mathrm{ddim}(X)$ is the doubling dimension of X.

Our ideas directly extend to the $(1+\epsilon)$-ANN problem, by building a BBD tree in the projected d'-dimensional space. This 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. The main result of Section 3.3 is Theorem 39, which offers a randomized algorithm for the $(1+\epsilon)$-ANN problem with optimal $O(dn)$ space, and query time in $O(dn^{\rho}\log n)$, where $\rho = 1 - \Theta(\epsilon^2/\ln\ln n)$, for $\epsilon \in (0, 1/2]$. The total preprocessing time is $O(dn\log n)$. For each query $q \in \mathbb{R}^d$, the preprocessing phase succeeds with constant probability.

This direct approach is extended to finite subsets of ℓ_2 with bounded expansion rate ψ (see Subsection 3.3.2). The pointset is now mapped to a space of dimension $O(\log\psi)$, and each query costs roughly $O((\psi^{\log\log\psi})d\log n)$.

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 $\tilde{O}(dn^{1-\Theta(\epsilon^2)})$ (see Subsection 3.4).

In Chapter 4, we investigate the problem of reducing the dimension for doubling subsets of ℓ_1. While this embeddability question has a negative answer in general due to known lower bounds [66], 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. We refer to randomized embeddings which satisfy this requirement as *near* neighbor-preserving. In particular, for pointsets with doubling constant λ_P, we show the following:

1. In Theorem 53, we prove that for every $\epsilon \in (0, 1/2)$ and $t \geq 1$, there is a randomized mapping $h : \ell_1^d \to \ell_1^{d'}$ that can be computed in time $\tilde{O}(dn^{1+1/\Omega(t)})$ and is *near* neighbor-preserving for P with distortion $1+6\epsilon$ and probability of correctness $\Omega(\epsilon)$, where
$$d' = (\log\lambda_P \cdot \log(t/\epsilon))^{\Theta(1/\epsilon)}/\zeta(\epsilon).$$

Although the mapping h depends on the pointset, the parameter t is user-defined and therefore provides a trade-off between preprocessing time and target dimension. The term $\zeta(\epsilon)$ depends only on ϵ.

2. In Theorem 56, we show that for every $\epsilon \in (0, 1/2)$, there is a randomized mapping $h' : \ell_1^d \to \ell_1^{d'}$ that can be computed in time $O(dd'n)$ and is *near* neighbor-preserving for P with distortion $1+6\epsilon$ and probability of correctness $\Omega(\epsilon)$, where

$$d' = (\log \lambda_P \cdot \log(d/\epsilon))^{\Theta(1/\epsilon)} / \zeta(\epsilon).$$

In this case, the function h' is oblivious to P and well-defined over the whole space, but the target dimension depends on d. The term $\zeta(\epsilon)$ depends only on ϵ.

1.3.1.2 Approximate Nets

In Chapter 5, we present a new randomized algorithm that computes approximate r-nets in time subquadratic in n and polynomial in the dimension, and improves upon the complexity of the best known algorithm. With probability $1 - o(1)$, our method returns $\mathcal{N} \subseteq X$, which is a $(1+\epsilon)$-approximate r-net of X.

We reduce the problem of computing approximate r-nets for arbitrary vectors (points) under Euclidean distance to the same problem for vectors on $\{-1, 1\}^{O(\log n/\epsilon^2)}$. Then, we extend and simplify Valiant's framework [77] and we compute r-nets in time $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$, thus improving on the exponent of the LSH-based construction [42], when ϵ is sufficiently small. This improvement by $\sqrt{\epsilon}$ in the exponent is the same as the complexity improvement obtained in [77] over the LSH-based algorithm for the approximate closest pair problem.

Our study is motivated by the fact that computing efficiently an r-net leads to efficient approximate solutions for several geometric problems. In particular, our extension of r-nets in high dimensional Euclidean space can be plugged in the framework of [54]. The new framework has many applications, notably the kth nearest neighbor distance problem, which we solve in $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$.

1.3.2 Polygonal curves

1.3.2.1 Approximate Nearest Neighbors

In Chapter 6, 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 (for a formal definition see Definition 5). 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. While a reduction to the near neighbor problem has provable guarantees on metrics [51], we are not aware of an analogous result for non-metric distances such as the DTW.

Specifically, when $p > 2$, we show that there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot \left(\frac{d}{p\epsilon} + 2\right)^{O(dm\cdot\alpha_{p,\epsilon})}\right),$$

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}(2^{4m}\log n)$.

When specialized to DFD and compared to [37], the two methods are only comparable when ϵ is a large enough fixed constant. Indeed, the two space and preprocessing time complexity bounds are equivalent, i.e. they are both exponential in d and m, but our query time is linear instead of being exponential in d.

When $p \in [1,2]$, there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot 2^{O(dm\cdot\alpha_{p,\epsilon})}\right),$$

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}\left(2^{4m}\log n\right)$. This leads to the first approach that achieves $1+\epsilon$ approximation for DTW at the expense of space, preprocessing and query time complexities being exponential in m. Hence our method is best suited when the curve size is small.

In Chapter 7, we focus on DFD, and we provide a solution which is especially efficient in the short query regime. Moreover, we extend our ideas to non-Euclidean spaces: we provide a solution for arbitrary metrics with bounded doubling dimension, and can be accessed through a metric oracle.

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 $O\left(dk\right)$, where k denotes the length of the query curves. This result improves on the (the more general) result of Chapter 6 on DFD, even in the case that queries are of the same complexity as the dataset. It also improves upon [43], when $k \ll m$, and it is comparable otherwise. The data structure can be derandomized with a slight worsening of the performance. For arbitrary doubling metrics, we give analogous results, but the achieved performance depends on the assumptions associated with the metric oracle.

1.3.2.2 Vapnik–Chervonenkis dimension

In Chapter 8, 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. The same holds for our bounds for the range space induced by the Weak Fréchet distance. Our lower bounds show that these bounds are almost tight in both parameters k and m. 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 $\mathbb{R}^2$.

The bounds in the discrete setting hold for ranges of metric balls of all radii and readily extend to ground sets of curves defined in $\mathbb{R}^d$ for $d > 2$. In all cases, the bounds are tight in the dependency on m, the complexity of elements of the ground set.

2. PRELIMINARIES

In this chapter, we formally define basic concepts and we prove preliminary results which will be useful in the subsequent chapters.

2.1 Metrics

While this is not always the case, we may assume that the distance functions of interest satisfy certain properties. This often allows us to prove desirable guarantees for the proposed solutions. Given a set of objects X, a distance function on X is a function $\mathsf{d} : X \times X \mapsto [0, \infty)$. Then, the pair (X, d) defines a *metric* space if for any $x, y, z \in X$, the following conditions are satisfied:

1. $\mathsf{d}(x, y) \geq 0$ (non-negativity)
2. $\mathsf{d}(x, y) = 0 \iff x = y$ (identity of indiscernibles)
3. $\mathsf{d}(x, y) = \mathsf{d}(y, x)$ (symmetry)
4. $\mathsf{d}(x, z) \leq \mathsf{d}(x, y) + \mathsf{d}(y, z)$ (subadditivity or triangle inequality)

A *pseudometric* space is a pair (X, d) which for any $x, y, z \in X$ satisfies

1. $\mathsf{d}(x, y) \geq 0$
2. $\mathsf{d}(x, x) = 0$
3. $\mathsf{d}(x, y) = \mathsf{d}(y, x)$
4. $\mathsf{d}(x, z) \leq \mathsf{d}(x, y) + \mathsf{d}(y, z)$

The difference between a pseudometric and a metric is that in a pseudometric, two distinct objects may have zero distance. *Quasimetric* spaces satisfy all axioms of a metric space with the exception of 3. , the axiom of symmetry. *Ultrametrics* satisfy a stronger version of the triangular inequality: $\mathsf{d}(x, z) \leq \max\{\mathsf{d}(x, y), \mathsf{d}(y, z)\}$.

2.1.1 ℓ_p norms

Metrics in general can be defined on arbitrary sets. A *norm* is defined on some vector space X as follows:

1. $\forall x \in X : \ \|x\| \in [0, \infty)$
2. $\|x\| = 0 \implies x = 0$

3. $\|\alpha x\| = \alpha\|x\|$ for all $\alpha \in \mathbb{R}$

4. $\|x+y\| \leq \|x\| + \|y\|$

Every norm $\|\cdot\|$ defines a metric, in which the distance of points x, y equals $\|x-y\|$. The unit ball of any norm is a symmetric convex body which contains the origin. In addition, any symmetric convex body K defines a norm: $\|x\|_K = \min\{\lambda \geq 0 \mid x \in \lambda K\}$.

For a point $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$ and for $p \in [1, \infty)$, the ℓ_p norm is defined as

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}.$$

We denote by ℓ_p^d the normed space $(\mathbb{R}^d, \|\cdot\|_p)$. When d is not important, we simply use ℓ_p denoting $(\mathbb{R}^d, \|\cdot\|_p)$ for some $d \in \mathbb{N}$.

2.1.2 Distance functions for curves

2.1.2.1 Discrete measures

Let us start with point sequences, which are closely related to curves. For metrics $M_1, \ldots, M_k$, we define the ℓ_p-product of $M_1, \ldots, M_k$ as the metric with domain $M_1 \times \cdots \times M_k$ and distance function

$$d((x_1, \ldots, x_k), (y_1, \ldots, y_k)) = \left(\sum_{i=1}^{k} d_{M_i}^p(x_i, y_i) \right)^{1/p}.$$

It is common, in distance functions of curves, to involve the notion of a traversal for two curves. Intuitively, a traversal corresponds to a time plan for traversing the two curves simultaneously, starting from the first point of each curve and finishing at the last point of each curve. With time advancing, the traversal advances in at least one of the two curves.

Definition 4 (Traversal). *Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $U = u_1, \ldots, u_{m_2}$, a traversal $T = (i_1, j_1), \ldots, (i_t, j_t)$ is a sequence of pairs of indices referring to a pairing of vertices from the two curves such that:*

1. *$i_1, j_1 = 1$, $i_t = m_1$, $j_t = m_2$.*

2. *$\forall (i_k, j_k) \in T : i_{k+1} - i_k \in \{0, 1\}$ and $j_{k+1} - j_k \in \{0, 1\}$.*

3. *$\forall (i_k, j_k) \in T : (i_{k+1} - i_k) + (j_{k+1} - j_k) \geq 1$.*

Now, we define a class of distance functions for polygonal curves. In this definition, it is implied that we use the Euclidean distance to measure distance between any two points. However, the definition easily extends to arbitrary metrics.

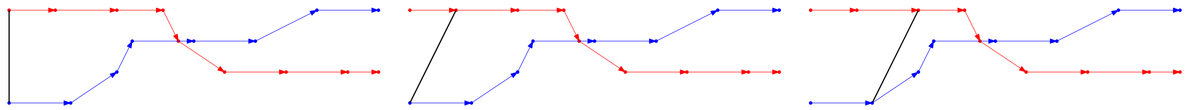

Figure 2.1: The traversal starts from the starting endpoints. Then, it only progresses on the red curve. Then, it progresses on both curves.

Definition 5 (ℓ_p-distance of polygonal curves)**.** *Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $U = u_1, \ldots, u_{m_2}$, we define the ℓ_p-distance between V and U as the following function:*

$$d_p(V, U) = \min_{T \in \mathcal{T}} \left(\sum_{(i_k, j_k) \in T} \|v_{i_k} - u_{j_k}\|_2^p \right)^{1/p},$$

where $\mathcal{T}$ denotes the set of all possible traversals for V and U.

The above class of distances for curves includes some widely known distance functions. For instance, $d_\infty(V, U)$ coincides with the DFD of V and U (defined for the Euclidean distance). Moreover $d_1(V, U)$ coincides with DTW for curves V, U.

Remark 6. *The discrete Fréchet distance in an arbitrary metric space defines a pseudo-metric: the triangular inequality is satisfied, but distinct curves may have zero distance. However, for our purposes, it is sufficient to consider the metric space which is naturally induced by that pseudo-metric: two polygonal curves are considered to be equal if their discrete Fréchet distance is zero. This observation allows us to refer to the metric space of polygonal curves under the discrete Fréchet distance.*

2.1.2.2 Continuous distances

Any polygonal curve V with vertices $v_1, \ldots, v_{m_1}$ and edges $\overline{v_1 v_2}, \ldots, \overline{v_{m_1-1} v_{m_1}}$ has a uniform parametrization that allows us to view it as a parametrized curve $v: [0,1] \mapsto \mathbb{R}^d$. Once again, we assume that curves belong to the Euclidean space.

Definition 7 (Directed Hausdorff distance.)**.** *Let X, Y be two subsets of $\mathbb{R}^d$. The directed Hausdorff distance from X to Y is:*

$$\mathrm{d}_{\overrightarrow{H}}(X, Y) = \sup_{u \in X} \inf_{v \in Y} \|u - v\|_2.$$

Definition 8 (Hausdorff distance.)**.** *Let X, Y be two subsets of $\mathbb{R}^d$. The Hausdorff distance between X and Y is:*

$$\mathrm{d}_H(X, Y) = \max\{d_{\overrightarrow{H}}(X, Y), d_{\overrightarrow{H}}(Y, X)\}.$$

Definition 9 (Fréchet distance)**.** *Given two parametrized curves $u, v: [0,1] \mapsto \mathbb{R}^d$, their Fréchet distance is defined as follows:*

$$\mathrm{d}_F(u, v) = \min_{f:[0,1] \mapsto [0,1]} \max_{\alpha \in [0,1]} \|v(a) - u(f(\alpha))\|_2,$$

where f ranges over all continuous and monotone bijections with $f(0) = 0$ and $f(1) = 1$.

Definition 10 (Weak Fréchet distance). *Given two parametrized curves $u, v: [0,1] \mapsto \mathbb{R}^d$, their Weak Fréchet distance is defined as follows:*

$$\mathsf{d}_{wF}(u,v) = \min_{\substack{f:[0,1]\mapsto[0,1] \\ g:[0,1]\mapsto[0,1]}} \max_{\alpha\in[0,1]} \|v(f(\alpha)) - u(g(\alpha))\|_2,$$

where f and g range over all continuous functions (not exclusively bijections) with $f(0)=0$ and $f(1)=1$ and $g(0)=0$ and $g(1)=1$.

2.2 Random projections and dimensionality reduction

In this section, we present basic results and easily-obtained lemmas about random projections.

Theorem 11 ([57]). *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 $\|v\|_2 = 1$:*

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

A simple computation shows the following (see also [58]).

Lemma 12. *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$, and let $D>3$. For any $v \in \mathbb{R}^d$ with $\|v\|_2 = 1$:*

$$\Pr\left[\|Gv\|_2^2 \leq (1/D)\cdot \frac{d'}{d}\right] \leq \left(\frac{3}{D}\right)^{d'}.$$

We also prove concentration inequalities for central absolute moments of the normal distribution. Some of these results may be folklore, and the reasoning is quite similar to the one followed by proofs of the Johnson-Lindenstrauss lemma, e.g. [67]. Notice also that results concerning random projections from ℓ_2 to $\ell_p, p \in [1,2]$ are folklore, but we are also interested in the case $p>2$. In addition, the properties which are required for ANN searching are weaker than the ones which are typically investigated.

The 2-stability property of standard normal variables, along with standard facts about their absolute moments imply the following claim.

Lemma 13. *Let $v \in \mathbb{R}^d$ and let G be $d' \times d$ matrix with i.i.d random variables following $N(0,1)$. Then,*

$$\mathbb{E}\left[\|Gv\|_p^p\right] = c_p \cdot d' \cdot \|v\|_2^p,$$

where $c_p = \frac{2^{p/2}\cdot\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$ is a constant depending only on $p>1$.

Proof. Let $g = (X_1, \ldots, X_d)$ be a vector of random variables which follow $N(0,1)$ and any vector $v \in \mathbb{R}^d$. The 2-stability property of gaussian random variables implies that $\langle g, v\rangle \sim N(0, \|v\|_2^2)$. Recall the following standard fact for central absolute moments of $Z \sim N(0,\sigma^2)$:

$$\mathbb{E}[|Z|^p] = \sigma^p \cdot \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.$$

Hence,

$$\mathbb{E}\left[\|Gv\|_p^p\right] = \mathbb{E}\left[\sum_{i=1}^{d'} |\langle g_i, v\rangle|^p\right] = d' \cdot \|v\|_2^p \cdot \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}.$$

□

In the following lemma, we give a simple upper bound on the moment generating function of $|X|^p$, where $X \sim N(0,1)$.

Lemma 14. *Let $X \sim N(0,\sigma^2)$, $p \geq 1$, and $t > 0$, then $\mathbb{E}[exp(-t|X|^p)] \leq exp(-t\mathbb{E}[|X|^p] + t^2\mathbb{E}[|X|^{2p}])$.*

Proof. We use the easily verified fact that for any $x \leq 1$, $exp(x) \leq 1 + x + x^2$ and the standard inequality $1 + x \leq \mathrm{e}^x$, for all $x \in \mathbb{R}$.

$$\mathbb{E}\left[\mathrm{e}^{-t|X|^p}\right] \leq 1 - t \cdot \mathbb{E}\left[|X|^p\right] + t^2 \cdot \mathbb{E}\left[|X|^{2p}\right] \leq \mathrm{e}^{-t\mathbb{E}[|X|^p] + t^2\mathbb{E}[|X|^{2p}]}.$$

□

Lemma 15. *Let $X \sim N(0,1)$. Then, there exists constant $C > 0$ s.t. for any $p \geq 1$, $\mathbb{E}[|X|^{2p}] \leq C \cdot 2^p \cdot \mathbb{E}[|X|^p]^2$.*

Proof. In the following, we denote by $f(p) \approx g(p)$ the fact that there exist constants $0 < c < C$ s.t. for any $p > 1$, $f(p) \leq C \cdot g(p)$ and $f(p) \geq c \cdot g(p)$. In addition, $f(p) \gtrsim g(p)$ means that $\exists C > 0$ s.t. $\forall p > 1$, $C \cdot f(p) \geq g(p)$. In the following we make use of the Stirling approximation and standard facts about moments of normal variables.

$$\mathbb{E}\left[|X|^{2p}\right] = \frac{2^p \cdot \Gamma\left(\frac{2p+1}{2}\right)}{\sqrt{\pi}} \approx (2p-1)!! = \frac{(2p)!}{2^p \cdot p!} \approx \left(\frac{(2p)^{2p}}{\mathrm{e}}\right) \cdot \sqrt{p} \cdot \frac{1}{2^p \cdot \left(\frac{p}{\mathrm{e}}\right)^p} \approx \frac{2^p p^p \sqrt{p}}{\mathrm{e}^p} \approx 2^p \cdot p!$$

$$\mathbb{E}\left[|X|^p\right]^2 \approx ((p-1)!!)^2 \gtrsim \left(2^{p/2+1/2} \cdot \left(\frac{(p/2+1/2)}{\mathrm{e}}\right)^{(p/2+1/2)}\right)^2 \approx \frac{p^{p+1}}{\mathrm{e}^{p+1}} \gtrsim p!$$

□

The following lemma is the main ingredient of our embedding, since it provides us with a lower tail inequality for one projected vector.

Lemma 16. *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$, s.t. $\|v\|_2 = 1$. For appropriate constant $c' > 0$, for $p \geq 1$ and $\delta \in (0,1)$,*

$$\mathsf{P}r[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}[\|Gv\|_p^p]] \leq \mathsf{e}^{-c' \cdot 2^{-p} \cdot d' \cdot \delta^2}.$$

Proof. For $X \sim N(0,1)$ and any $t > 0$,

$$\mathsf{P}r\left[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}\left[\|Gv\|_p^p\right]\right] \leq \mathbb{E}\left[\mathsf{e}^{-t|X|^p}\right]^{d'} \cdot \mathsf{e}^{(t(1-\delta)d' \cdot \mathbb{E}[|X|^p])} \leq$$

$$\leq \mathsf{e}^{d'(-t \cdot \mathbb{E}[|X|^p] + t^2 \cdot C \cdot 2^p \cdot \mathbb{E}[|X|^p]^2 + t \cdot (1-\delta) \cdot \mathbb{E}[|X|^p])}.$$

The last inequality derives from Claim 15. Now, we set $t = \frac{\delta}{2 \cdot C \cdot 2^p \cdot \mathbb{E}[|X|^p]}$. Hence,

$$\mathsf{P}r[\|Gv\|_p^p \leq (1-\delta) \cdot \mathbb{E}[\|Gv\|_p^p]] \leq \mathsf{e}^{-c' \cdot 2^{-p} \cdot d' \cdot \delta^2},$$

for some constant $c' > 0$. □

Standard properties of ℓ_p norms imply a loose upper tail inequality.

Corollary 17. *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$. Let $p \geq 2$. Then, for constant $C > 0$,*

$$\mathsf{P}r\left[\|Gv\|_p \geq (1+\epsilon)\|v\|_2\sqrt{d'}\right] \leq \mathsf{e}^{-C \cdot d' \cdot \epsilon^2}.$$

Proof. Since $p \geq 2$, we have that $\forall x \in \mathbb{R}^d$ $\|x\|_p \leq \|x\|_2$. Hence, by Theorem 11,

$$\mathsf{P}r[\|Gv\|_p \geq (1+\epsilon)\|v\|_2\sqrt{d'}] \leq \mathsf{P}r[\|Gv\|_2 \geq (1+\epsilon)\|v\|_2\sqrt{d'}] \leq \mathsf{e}^{-C \cdot d' \cdot \epsilon^2}.$$

□

However, an improved upper tail inequality can be derived when $p \in [1,2]$.

Lemma 18. *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$ and consider vector $v \in \mathbb{R}^d$. Let $p \in [1,2]$. Then, for constant $C > 0$,*

$$\mathsf{P}r\left[\|Gv\|_p \geq (3 \cdot c_p \cdot d')^{1/p}\|v\|_2\right] \leq \mathsf{e}^{-C \cdot d'}.$$

Proof. Let $X \sim N(0,1)$.

$$\mathbb{E}\left[\mathsf{e}^{|X|^p/3}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathsf{e}^{|x|^p/3 - x^2/2} \mathsf{d}x \leq \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{+\infty} \mathsf{e}^{x^2/3 - x^2/2} \mathsf{d}x = \sqrt{3}.$$

Now, assume wlog $\|v\|_2 = 1$,

$$\mathsf{P}r\left[\|Gv\|_p^p \geq 3 \cdot \mathbb{E}\left[\|Gv\|_p^p\right]\right] \leq \mathbb{E}\left[\mathsf{e}^{|X|^p/3}\right]^{d'} \cdot \mathsf{e}^{-d' \cdot \mathbb{E}[|X|^p]} \leq \mathsf{e}^{-d'(c_p - 2/3)} \leq \mathsf{e}^{-d'/10},$$

where $c_p = \frac{2^{p/2} \cdot \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$. □

2.3 Doubling dimension and nets

In this section, we define basic notions about doubling metrics and nets.

Definition 19 (Doubling constant). *Consider any metric space (X, d_X) and let $B(p,r) = \{x \in X \mid \mathsf{d}_X(x,p) \leq r\}$. The* doubling constant *of X, denoted λ_X, is the smallest integer λ_X such that for any $p \in X$ and $r > 0$, the ball $B(p,r)$ can be covered by λ_X balls of radius $r/2$ centered at points in X.*

The *doubling dimension* of (X, d_X) is defined to be equal to $\log \lambda_X$. Nets play an important role in the study of embeddings, as well as in designing efficient data structures for doubling metrics. They are generally subsets of the original sets, which satisfy the following: no two points in the net are within distance r of each other, and for every point in the original set there exists a net point within distance r. Figure 2.2 illustrates this notion. In the following we introduce the notion of c-approximate r-nets.

Definition 20 (Approximate nets). *For $c \geq 1$, $r > 0$ and metric space (V, d_V), a c-approximate r-net of V is a subset $\mathcal{N} \subseteq V$ such that no two points of $\mathcal{N}$ are within distance r of each other, and every point of V lies within distance at most $c{\cdot}r$ from some point of $\mathcal{N}$.*

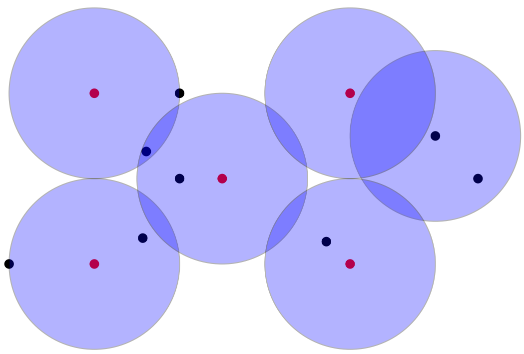

Figure 2.2: r**-nets.**

Theorem 21. *Let $P \subset \ell_1^d$ consisting of n points. Then, for any $c > 0$, $r > 0$, one can compute a c-approximate r-net of P in time $\tilde{O}(dn^{1+1/c'})$, where $c' = \Omega(c)$. The result is correct with high probability. The algorithm also returns the assignment of each point of P to the point of the net which covers it.*

Proof. We employ basic ideas from [51]. An analogous result in ℓ_2 is stated in [42]. First, we assume $r = 1$, since we are able to re-scale the point set. Now, we consider a randomly shifted grid with side-length 2. The probability that two points $p, q \in P$ fall into the same grid cell, is at least $1 - \|p-q\|_1/2$. For each non-empty grid cell we snap points to a grid: each coordinate is rounded to the nearest multiple of $\delta = 1/10dc$. Then, coordinates are multiplied by $1/\delta$ and each point $x = (x_1, \ldots, x_d) \in [2\delta]^d$ is mapped to $\{0,1\}^{2d/\delta}$ by a function G as follows: $G(x) = (g(x_1), \ldots, g(x_d))$, where $g(z)$ is a binary string of z ones followed by $2/\delta - z$ zeros. For any two points p, q in the same grid cell, let $f(p)$,$f(q)$ be the two binary strings obtained by the above mapping. Notice that,

$$\|f(p) - f(q)\|_1 \in (2/\delta) \cdot \|p - q\|_1 \pm 1.$$

Hence,

$$\|p - q\|_1 \leq 1 \implies \|f(p) - f(q)\|_1 \leq (2/\delta) + 1,$$
$$\|p - q\|_1 \geq c \implies \|f(p) - f(q)\|_1 \geq (2/\delta) \cdot c - 1.$$

Now, we employ the LSH family of [51], for the Hamming space. After standard concatenation, we can assume that the family is $(\rho, c'\rho, n^{-1/c'}, n^{-1})$-sensitive, where $\rho = (2/\delta) + 1$ and $c' = \Omega(c)$. Let $\alpha = n^{-1/c'}$ and $\beta = n^{-1}$.

Notice that for the above two-level hashing table we obtain the following guarantees. Any two points $p, q \in P$, such that $\|p-q\|_1 \leq 1$, fall into the same bucket with probability $\geq \alpha/2$. Any two points $p, q \in P$, such that $\|p - q\|_1 \geq c$, fall into the same bucket with probability $\leq \beta$.

Finally, we independently build $k = \Theta(n^{1/c'} \log n)$ hashtables as above, where the random hash function is defined as a concatenation of the function which maps points to their grid cell id and one LSH function. We pick an arbitrary ordering $p_1, \ldots, p_n \in P$. We follow a greedy strategy in order to compute the approximate net. We start with point p_1, and we add it to the net. We mark all (unmarked) points which fall at the same bucket with p_1, in one of the k hashtables, and are at distance $\leq cr$. Then, we proceed with point p_2. If p_2 is unmarked, then we repeat the above. Otherwise, we proceed with p_3. The above iteration stops when all points have been marked. Throughout the procedure, we are able to store one pointer for each point, indicating the center which covered it.

Correctness. The probability that a good pair p, q does not fall into the same bucket for any of the k hashtables is $\leq (1 - \alpha/2)^k \leq n^{-10}$. Hence, with high probability, the packing property holds, and the covering property holds because the above algorithm stops when all points are marked.

Running time. The time to build the k hashtables is $k \cdot n = \tilde{O}(n^{1+1/c'})$. Then, at most n queries are performed: for each query, we investigate k buckets and the expected number of false positives is $\leq k \cdot n^2 \cdot \beta = \tilde{O}(n^{1+1/c'})$. Hence, if we stop after having seen a sufficient amount of false positives, we obtain time complexity $\tilde{O}(n^{1+1/c'})$ and the covering property holds with constant probability. We can repeat the above procedure $O(\log n)$ times to obtain high probability of success. $\square$

2.4 Range spaces and Vapnik–Chervonenkis dimension

Each range space can be defined as a pair of sets $(X, \mathcal{R})$, where X is the *ground set* and $\mathcal{R}$ is the *range set*. Let $(X, \mathcal{R})$ be a range space. For $Y \subseteq X$, we denote:

$$\mathcal{R}_{|Y} = \{r \cap Y \mid r \in \mathcal{R}\}.$$

If $\mathcal{R}_{|Y}$ contains all subsets of Y, then Y is *shattered* by $\mathcal{R}$.

Definition 22 (Vapnik-Chernovenkis dimension [79])**.** *The Vapnik-Chernovenkis dimension (VC dimension) of $(X, \mathcal{R})$ is the maximum cardinality of a shattered subset of X.*

Definition 23 (Shattering dimension)**.** *The shattering dimension of $(X, \mathcal{R})$ is the smallest δ such that, for all m,*

$$\max_{\substack{B \subset X \\ |B|=m}} |\mathcal{R}_{|B}| = O(m^{\delta}).$$

It is well-known [13, 50] that for a range space $(X, \mathcal{R})$ with VC-dimension ν and shattering dimension δ that $\nu \leq O(\delta \log \delta)$ and $\delta = O(\nu)$. So bounding the shattering dimension and bounding the VC-dimension are asymptotically equivalent within a log factor.

Definition 24 (Dual range space)**.** *Given a range space $(X, \mathcal{R})$, for any $p \in X$, we define*

$$\mathcal{R}_p = \{r \mid r \in \mathcal{R}, p \in r\}.$$

The dual range space of $(X, \mathcal{R})$ is the range space $(\mathcal{R}, \{\mathcal{R}_p \mid p \in X\})$.

It is a well-known fact that if a range space has VC dimension ν, then the dual range space has VC dimension $\leq 2^{\nu+1}$ (see e.g. [50]).

It is also known [25] that the composition ranges formed as the k-fold union or intersection of ranges from a range space with bounded VC-dimension ν induces a range space with VC-dimension $O(\nu k \log k)$, and this was recently shown that this is tight for even some simple range spaces such as those defined by halfspaces [31].

3. RANDOM PROJECTIONS WITH FALSE POSITIVES

Deterministic space partitioning techniques, such as kd-trees, BBD-trees and approximate Voronoi diagrams, perform well in solving $(1+\epsilon)$-ANN when the dimension is relatively low, but are affected by the curse of dimensionality. To address this issue, randomized methods have been proposed, such as Locality Sensitive Hashing (LSH), which are more efficient when the dimension is high. One might try applying the celebrated Johnson-Lindenstrauss Lemma, followed by standard space partitioning techniques, but the properties of the projected pointset are too strong for designing an efficient $(1+\epsilon)$-ANN search method when aiming for near-linear storage.

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.

After dimension reduction, we store points in a uniform grid of side length $\epsilon/\sqrt{d'}$, where d' is the reduced dimension. Given a query, we explore cells intersecting the unit ball around the query. This data structure requires linear space, and query time in $O(dn^{\rho})$, $\rho \approx 1 - \epsilon^2/\log(1/\epsilon)$, where n denotes input cardinality and d space dimension. Bounds are improved for doubling subsets via r-nets. A small improvement on the exponent ρ can be achieved by employing certain LSH functions to define a mapping to the Hamming space.

Organization. Section 3.1 introduces our embeddings to dimension lower than predicted by the Johnson-Linderstrauss Lemma. Section 3.2 states our main result for the (c, r)-ANN problem in ℓ_2 and an extension to doubling subsets of ℓ_2. Section 3.3 states a weaker, yet practical result on c-ANN in ℓ_2, and an extension to pointsets with bounded expansion rate. Section 3.4 extends the results to the case of LSH-able metrics, and includes a slightly improved result for the Euclidean space. We conclude with a summary of our results.

In the sequel, the approximation factor c is equal to $1+\epsilon$, for some $\epsilon \in (0, 1/2]$.

3.1 Randomized Embeddings with slack

This section examines standard dimensionality reduction techniques and extends them to approximate embeddings optimized to our setting.

In [1], they consider non-oblivious embeddings from finite metric spaces with small dimension and distortion, while allowing a constant fraction of all distances to be arbitrarily distorted. In [23], they present non-oblivious embeddings for the ℓ_2 case, which preserve distances in local neighborhoods. In [45], they provide a non-oblivious embedding which preserves distances up to a given scale and the target dimension mainly depends on ddim(X) with no dependence on $|X|$. In general, embeddings based on probabilistic partitions are not oblivious. In [21], they solve ANN in ℓ_p spaces, for $2 < p < \infty$, by oblivious embeddings to ℓ_∞ and ℓ_2.

But, it is not obvious how to use a non-oblivious embedding in the scenario in which we preprocess a dataset and we expect a query to arrive. Therefore we focus on oblivious embeddings.

Let us now revisit the classic Johnson-Lindenstrauss Lemma:

Proposition 25. [59] *For any set $X \subset \mathbb{R}^d$, $\epsilon \in (0, 1/2]$ there exists a distribution over linear mappings $f : \mathbb{R}^d \longrightarrow \mathbb{R}^{d'}$, where $d' = O(\log |X|/\epsilon^2)$, such that for any $p, q \in X$,*

$$(1 - \epsilon)\|p - q\|_2^2 \leq \|f(p) - f(q)\|_2^2 \leq (1 + \epsilon)\|p - q\|_2^2.$$

In the initial proof [59], they show that this can be achieved by orthogonally projecting the pointset on a random linear subspace of dimension d'. In [72], they provide a proof based on elementary probabilistic techniques. In [57], they prove that it suffices to apply a gaussian matrix G on the pointset. G is a $d \times d'$ matrix with each of its entries independent random variables given by the standard normal distribution $N(0, 1)$. Instead of a gaussian matrix, we can apply a matrix whose entries are independent random variables with uniformly distributed values in $\{-1, 1\}$ [2], or even independent random variables with uniform subgaussian tails [68].

However, it has been realized that this notion of randomized embedding is stronger than what is required for c-ANN. The following has been introduced in [58] and focuses on the distortion of the nearest neighbor.

Definition 26. *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 27. *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 27, we propose to weaken Definition 26 as follows.

Definition 28. *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 $S = \{f(p_1), \dots, f(p_k)\}$ is a solution to c-kANNs for q under Assumption 27, 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 11 and Lemma 12.

Remark 29. *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 30. *Under the notation of Definition 28, there exists a randomized mapping $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ which satisfies Definition 28 for*

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

$\epsilon \in (0, 1/2]$, 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 11. 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 27). 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 11,

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

where $C' > 0$ is a constant, which is slightly different than the one that appears in Lemma 11 (since we aim for distortion factor $1/(1+\epsilon)$ instead of $(1-\epsilon)$). The event of failure is defined as the disjunction of two events:

$$N \geq k \;\vee\; \|f(u)\|_2 \geq (\beta/c), \tag{3.1}$$

and its probability is at most equal to

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

by applying again Theorem 11. 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 assume that the random experiment succeeds, and let $S = \{f(p_1), \dots, f(p_k)\}$ be a solution of the c'-kANNs problem in the projected space, given by a data-structure which satisfies Assumption 27. It holds that $\forall f(x) \in f(X)\backslash S'$, $\|f(x)-f(q)\|_2 > \|f(p_k)-f(q)\|_2/c'$, 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 28 for $D = (1+\epsilon)^2$ and the theorem is established. $\square$

Theorem 30 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.

3.2 Approximate Near Neighbor

This section combines the ideas developed in Section 3.1 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 31. [51, 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 32 ((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 [51] 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 [51], 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,

$$V_2^d(R) \leq \frac{2\pi^{d/2}}{d \cdot \Gamma(d/2)} R^d = \frac{2\pi^{d/2}}{d(d/2-1)!} R^d \leq \frac{2\pi^{d/2}}{(d/2)!} R^d \leq \frac{2\pi^{d/2}}{e(d/(2e))^{d/2}} R^d \leq \frac{2^{d+1}(18)^{d/2}}{\epsilon^d e} \leq \frac{9^d}{\epsilon^d}.$$

Hence, $C' \leq 9$.

Theorem 33. *There exists a data structure for Problem 32 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 30 for the Approximate Near Neighbor Problem.

Theorem 34. *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 30. The proof is indeed the same. □

3.2.1 Finite subsets of ℓ_2

We are about to show what Theorems 33 and 34 imply for the (c, r)-ANN problem.

Theorem 35. *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. □

3.2.2 The case of doubling subsets of ℓ_2

In this section, we apply our ideas to pointsets with bounded doubling dimension, in order to obtain non-linear randomized embeddings for the (c, r)-ANN problem.

Now, let $X \subset \mathbb{R}^d$ s.t. $|X| = n$ and X has doubling constant $\lambda_X = 2^{\mathrm{ddim}(X)}$. Consider also $S_i \subseteq X$ with diameter $2r_i$. Then we need $\lambda_X^{\log \frac{8r_i}{\epsilon}}$ tiny balls $b_\epsilon \subseteq X$ of diameter $\epsilon/4$ in order to cover S_i. We can assume that $r = 1$, since we can scale X. The idea is that we first compute $X' \subseteq X$ which satisfies the following two properties:

- $\forall p, q \in X'$ $\|p - q\|_2 > \epsilon/8$,
- $\forall q \in X$ $\exists p \in X'$ s.t. $\|p - q\|_2 \leq \epsilon/8$.

This is an r-net for X for $r = \epsilon/8$. The obvious naive algorithm computes X' in $O(n^2)$ time. Better algorithms exist for the case of low dimensional Euclidean space [49]. Approximate r-nets can be also computed in time $2^{O(\mathrm{ddim}(X))} n \log n$ for doubling metrics [52] , assuming that the distance can be computed in constant time.

Then, for X' we know that each $S_i \subseteq X'$ contains $\leq \lambda_X^{\log \frac{8r_i}{\epsilon}}$ points, since $X' \subseteq X \implies \mathrm{ddim}(X') \leq \mathrm{ddim}(X)$.

Theorem 36. *The (c^3, r)-ANN problem in $\mathbb{R}^d$ reduces to checking the solution set of the (c, cr)-kANNs problem in $\mathbb{R}^{d'}$, where $d' = O(\text{ddim}(X))$ and $k = (2/\epsilon)^{O(\text{ddim}(X))}$, by a randomized algorithm which succeeds with constant probability. Preprocessing costs an additional of $O(n^2)$ time and the delay in query time is proportional to $d \cdot k$.*

Proof. Once again we proceed in the same spirit as in the proof of Theorem 30.

Let X' be an $\epsilon/8$-net of X. Let $r_i = 2^{i+3}(1+\epsilon)$ for $i \geq 0$ and let $B_p(r) \subseteq X'$ denote the points of X' lying in the closed ball centered at p with radius r. We assume $0 < \epsilon \leq 1/2$ and we define:

$$N_{close} = |\,\{x \in X \,:\, \|x\|_2 \in [(1+\epsilon)^2, r_1) \,\wedge\, \|f(x)\|_2 \leq 1+\epsilon\}\,|,$$

$$N_{far} = |\,\{x \in X \,:\, \|x\|_2 \geq r_1 \,\wedge\, \|f(x)\|_2 \leq 1+\epsilon\}\,|.$$

We make use of Lemma 12.

$$\mathbb{E}[N_{far}] \leq \sum_{i=2}^{\infty} |B_p(r_i)| \cdot \left(\frac{3}{r_{i-1}}\right)^{d'} \leq \sum_{i=2}^{\infty} \lambda_X^{\log(16 r_i/\epsilon)} \cdot \left(\frac{1}{2^i}\right)^{d'} \leq \lambda_X^{O(\log(2/\epsilon))} \cdot \sum_{i=2}^{\infty} \frac{\lambda_X^i}{2^{i \cdot d'}} =$$

$$\overset{d' \geq \Omega(\log \lambda_X)}{=} 2^{O(\text{ddim}(X)\log(2/\epsilon))} = \left(\frac{2}{\epsilon}\right)^{O(\text{ddim}(X))}.$$

In addition,

$$\mathbb{E}[N_{close}] \leq \lambda_X^{O(\log(1/\epsilon))} \cdot \exp(-d' \cdot \epsilon^2 \cdot C) \leq \lambda_X^{O(\log(1/\epsilon))} = \left(\frac{2}{\epsilon}\right)^{O(\text{ddim}(X))},$$

where $C > 0$ is a constant, which is slightly different than the one that appears in Lemma 11 (since we aim for distortion factor $1/(1+\epsilon)$ instead of $(1-\epsilon)$). The number of grid cells of sidewidth $\epsilon/\sqrt{d'}$ intersected by a ball of radius 1 in $\mathbb{R}^{d'}$ is also $(2/\epsilon)^{O(\text{ddim}(X))}$. Notice, that if there exists a point in X which lies at distance 1 from q, then there exists a point in X' which lies at distance $1+\epsilon/8$ from q. Finally the probability that the distance between the query point q and one approximate near neighbor gets arbitrarily expanded is less than $\lambda_X^{-\Theta(\epsilon^2)}$. □

Now using the above ideas we obtain a data structure for the (c, r)-ANN problem.

Theorem 37. *There exists a data structure which solves the (c, r)-ANN problem which requires space and preprocessing time $O(dn)$ and the query costs*

$$d\left(\frac{2}{\epsilon}\right)^{O(\text{ddim}(X))}.$$

For fixed $q \in \mathbb{R}^d$, the building process of the data structure succeeds with constant probability.

3.3 Approximate Nearest Neighbor Search

This section combines tree-based data structures which solve c-kANNs with the results of Section 3.1, in order to obtain a randomized data structure which solves c-ANN. The main result of this section does not rely on an efficient reduction from the (c, r)-ANN problem, and hence it is simpler to implement. On the other hand, the obtained bounds are weaker than those of Section 3.2.

3.3.1 Finite subsets of ℓ_2

This subsection examines the general case of finite subsets of ℓ_2.

BBD-trees [16] require $O(dn)$ space, and allow computing k points, which are $(1+\epsilon)$-approximate nearest neighbors, in time $O((\lceil 1+6\frac{d}{\epsilon}\rceil^d + k)d\log n)$. The preprocessing time is $O(dn\log n)$. Notice, that BBD-trees satisfy Assumption 27.

The algorithm for the c-kANNs search visits cells in increasing order with respect to their distance from the query point q. If the current cell lies at distance more than r_k/c, where r_k is the current distance to the kth nearest neighbor, the search terminates. We apply the random projection for distortion $D = c = 1+\epsilon$, thus relating approximation error to the allowed distortion; this is not required but simplifies the analysis.

Moreover, $k = n^\rho$; the formula for $\rho < 1$ is determined below. Our analysis then focuses on the asymptotic behavior of the term $O(\lceil 1+6\frac{d'}{\epsilon}\rceil^{d'} + k)$.

Lemma 38. *With the above notation, for fixed $\epsilon \in (0,1)$, there exists $k > 0$ s.t., it holds that $\lceil 1+6\frac{d'}{\epsilon}\rceil^{d'} + k = O(n^\rho)$, where $\rho = 1 - \Theta(\epsilon^2/\log\log n) < 1$.*

Proof. Recall that $d' \le \frac{\tilde{c}}{\epsilon^2}\ln\frac{n}{k}$ for some appropriate constant $\tilde{C} > 0$. Since $(\frac{d'}{\epsilon})^{d'}$ is a decreasing function of m, we need to choose k s.t. $(\frac{d'}{\epsilon})^{d'} = \Theta(k)$. Let $k = n^\rho$. It is easy to see that $\lceil 1+6\frac{d'}{\epsilon}\rceil^{d'} \le \left(c'\frac{d'}{\epsilon}\right)^{d'}$, for some appropriate constant $C' \in (1,7)$. Then, by substituting d', k we obtain:

$$\ln\left(C'\frac{d'}{\epsilon}\right)^{d'} = \frac{\tilde{C}(1-\rho)}{\epsilon^2}\ln\left(\frac{\tilde{C}C'(1-\rho)\ln n}{\epsilon^3}\right)\ln n. \tag{3.2}$$

We assume $\epsilon \in (0,1)$ is a fixed constant. Hence, it is reasonable to assume that $\frac{1}{\epsilon} < \ln n$. Substituting $\rho = 1 - \frac{\epsilon^2}{2\tilde{C}(\epsilon^2+\ln(C'\ln n))}$ into equation (3.2), the exponent of n is bounded as follows:

$$\frac{\tilde{C}(1-\rho)}{\epsilon^2}\ln\left(\frac{\tilde{C}C'(1-\rho)\ln n}{\epsilon^3}\right) =$$

$$= \frac{\tilde{C}}{2\tilde{C}(\epsilon^2+\ln(C'\ln n))}\cdot\left(\ln(C'\ln n) + \ln\frac{1}{\epsilon} - \ln\left(2\epsilon^2 + 2\ln(C'\ln n)\right)\right) < \rho.$$

□

Notice that

$$d' = O\left(\frac{\log n}{\epsilon^2 + \log\log n}\right).$$

Combining Theorem 30 with Lemma 38 yields the following theorem.

Theorem 39. *Given n points in $\mathbb{R}^d$, there exists a randomized data structure which requires $O(dn)$ space and reports an $(1+\epsilon)$-approximate nearest neighbor in time*

$$O(dn^\rho \log n), \text{ where } \rho \leq 1 - \Theta(\epsilon^2/\log\log n) < 1.$$

The preprocessing time is $O(dn \log n)$. For each query $q \in \mathbb{R}^d$, the preprocessing phase succeeds with any constant probability.

Proof. The space required to store the dataset is $O(dn)$. The space used by BBD-trees is $O(d'n)$ where d' is defined in Lemma 38. We also need $O(dd')$ space for the matrix A as specified in Theorem 30. Hence, since $d' < d$ and $d' < n$, the total space usage is bounded above by $O(dn)$.

The preprocessing consists of building the BBD-tree which costs $O(d'n \log n)$ time and sampling A. We sample in time $O(dd')$, a $d \times d'$ matrix where its elements are independent random variables with the standard normal distribution $N(0,1)$. Since $d' = O(\log n)$, the total preprocessing time is bounded by $O(dn \log n)$.

For each query we use A to project the point in time $O(dd')$. Next, we compute its $d' = n^\rho$ approximate nearest neighbors in time $O(d'n^\rho \log n)$ and we check these neighbors with their d-dimensional coordinates in time $O(dn^\rho)$. Hence, each query costs $O(d \log n + d'n^\rho \log n + dn^\rho) = O(dn^\rho \log n)$ because $d' = O(\log n)$, $d' = O(d)$. Thus, the query time is dominated by the time required for ϵ-kANNs search and the time to check the returned sequence of k approximate nearest neighbors. □

To be more precise, the probability of success, which is the probability that the random projection succeeds according to Theorem. 30, is at least constant and can be amplified to high probability of success with repetition. Notice that the preprocessing time for BBD-trees has no dependence on ϵ.

3.3.2 Finite subsets of ℓ_2 with bounded expansion rate

This subsection models some structure that the data points may have so as to obtain tighter bounds.

The bound on the dimension d' obtained in Theorem 30 is quite pessimistic. We expect that, in practice, the space dimension needed in order to have a sufficiently good projection is less than what Theorem 30 guarantees. Intuitively, we do not expect to have instances where all points in X, which are not approximate nearest neighbors of q, lie at distance $\approx (1+\epsilon)\mathrm{d}(q, X)$. To this end, we consider the case of pointsets with bounded expansion rate.

Definition 40. *Let M be a metric space and $X \subseteq M$ be a finite pointset and let $B_p(r) \subseteq X$ denote the points of X lying in the closed ball centered at p with radius r. We say that X has (τ, ψ)-expansion rate if and only if, $\forall p \in M$ and $r > 0$,*

$$|B_p(r)| \geq \tau \implies |B_p(2r)| \leq \psi \cdot |B_p(r)|.$$

Theorem 41. *Under the notation of Definition 28, there exists a randomized mapping $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ which satisfies Definition 28 for dimension $d' = O(\log \psi)$, locality parameter $k = O(\tau\psi^3)$, distortion $D = (1+\epsilon)^2$ and constant probability of success, for pointsets with (τ, ψ)-expansion rate.*

Proof. We proceed in the same spirit as in the proof of Theorem 30.

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

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

where A is a matrix chosen from a distribution as in Theorem 11. Without loss of generality the query point q lies at the origin and its nearest neighbor u lies at distance 1 from q. Let r_0 be the distance to the $\tau-$th nearest neighbor, excluding neighbors at distance $\leq (1+\epsilon)^2$. For $i > 0$, let $r_i = 6 \cdot r_{i-1}$. Notice also that $r_0 \geq (1+\epsilon)^2$.

We distinguish the set of bad candidates according to whether they correspond to “close” of “far” points in the initial space. More precisely,

$$N_{close} = |\,\{x \in X \,:\, \|x\|_2 \in [r_0, r_1) \,\wedge\, \|f(x)\|_2 \leq \beta\}\,|,$$

$$N_{far} = |\,\{x \in X \,:\, \|x\|_2 \geq r_1 \,\wedge\, \|f(x)\|_2 \leq \beta\}\,|,$$

where $\beta = 1+\epsilon$. Clearly, by Theorem 11, and for $d' \geq \Omega(\log \psi)$,

$$\mathbb{E}[N_{close}] \leq \psi \cdot \tau \cdot \exp(-d' \cdot \epsilon^2 \cdot C') = O(\psi \cdot \tau),$$

where $C' > 0$ is a constant, which is slightly different than the one that appears in Lemma 11 (since we aim for distortion factor $1/(1+\epsilon)$ instead of $(1-\epsilon)$). and similarly by Lemma 12,

$$\mathbb{E}[N_{far}] \leq \sum_{i=1}^{\infty} \psi^{i+3}\tau \cdot \left(\frac{1}{2}\right)^{d' \cdot i} \leq \tau \cdot \psi^3 \sum_{i=1}^{\infty} \psi^i \left(\frac{1}{2^i}\right)^{d'} = O(\tau \cdot \psi^3).$$

Finally, using Markov’s inequality, we obtain constant probability of success. ☐

Employing Theorem 41 we obtain a result analogous to Theorem 39 which is weaker than those in [63, 24] but underlines the fact that our scheme shall be sensitive to structure in the input data, for real world assumptions.

Theorem 42. *Given n points in ℓ_2^d with (τ, ψ)-expansion rate, there exists a randomized data structure which requires $O(dn)$ space and reports an $(1+\epsilon)^3$-approximate nearest neighbor in time*

$$O((\psi^{\log(\log \psi/\epsilon)} + \tau \cdot \psi^3\,)\, d \,\log n)).$$

The preprocessing time is $O(dn \log n)$. For each query $q \in \mathbb{R}^d$, the preprocessing phase succeeds with constant probability.

Proof. We combine the embedding of Theorem 41 with the BBD-trees. Then,

$$O\Big(\Big(\frac{\sqrt{d'}}{\epsilon}\Big)^{d'}\Big) = O\Big(\Big(\frac{\log\psi}{\epsilon}\Big)^{\log\psi}\Big),$$

and the number of approximate nearest neighbors in the projected space is

$$k = O(\tau \cdot \psi^3).$$

This establishes the result. $\square$

3.4 On LSHable metrics

An important approach for proximity problems today is Locality Sensitive Hashing (LSH). It has been designed precisely for problems in high dimension. The LSH method is based on the idea of using hash functions designed so that it is more probable to map nearby points to the same bucket.

Definition 43. *Take reals $r_1 < r_2$ and $p_1 > p_2 > 0$. We call a family F of hash functions (p_1, p_2, r_1, r_2)-sensitive for a metric space $\mathcal{M}$ if, for any $x, y \in \mathcal{M}$, and h distributed uniformly in F, it holds:*

- $\mathsf{d}_{\mathcal{M}}(x, y) \leq r_1 \implies Pr[h(x) = h(y)] \geq p_1$,
- $\mathsf{d}_{\mathcal{M}}(x, y) \geq r_2 \implies Pr[h(x) = h(y)] \leq p_2$.

We start our presentation with an idea applicable to any metric admitting an LSH-based construction, aka LSH-able metric. Then, we study some classical LSH families which are also simple to implement.

The algorithmic idea is to apply a random projection from any LSH-able metric to the Hamming hypercube. Given an LSH family of functions F for some metric space, we uniformly select d' hash functions, where d' is specified later. The nonempty buckets defined by any hash function are randomly mapped to $\{0, 1\}$, with equal probability for each bit.

In particular, the random projection works as follows. We first sample $h_1 \in F$. We denote by $h_1(P)$ the image of P under h_1, which is a set of nonempty buckets. Now each nonempty bucket $x \in h_1(P)$ is mapped to $\{0, 1\}$: with probability $1/2$, set $f_1(x) = 0$, otherwise set $f_1(x) = 1$.

This is repeated d' times, and eventually for $p \in \mathcal{M}$, we compute the function

$$f(p) = (f_1(h_1(p)), \ldots, f_{d'}(h_{d'}(p))),$$

where $f : P \to \{0, 1\}^{d'}$.

Thus, points are projected to the Hamming cube of dimension d' and we obtain binary strings serving as keys for buckets containing the input points. The query algorithm projects a given point, and tests points assigned to the same or nearby vertices on the hypercube. To achieve the desired complexities, it suffices to choose $d' = \log n$.

The main lemma below describes the general ANN data structure whose complexity and performance depends on the LSH family that we assume is available. The proof details the data structure construction.

Lemma 44 (Main). *Given a (p_1, p_2, r, cr)-sensitive hash family F for some metric $(\mathcal{M}, \mathrm{d}_\mathcal{M})$ and input dataset $P \subseteq \mathcal{M}$, there exists a data structure for the (c, r)-ANN problem with space $O(dn)$, time preprocessing $O(dn)$, and query time $O(dn^{1-\delta} + n^{H((1-p_1)/2)})$, where*

$$\delta = \delta(p_1, p_2) = \frac{(p_1 - p_2)^2}{(1 - p_2)} \cdot \frac{\log e}{4},$$

where e *denotes the basis of the natural logarithm, and $H(\cdot)$ is the binary entropy function. The bounds hold assuming that computing* $\mathrm{d}_\mathcal{M}(.)$ *and computing the hash function cost $O(d)$. Given some query $q \in \mathcal{M}$, the building process succeeds with constant probability.*

Proof. The first step is a random projection to the Hamming space of dimension d', for d' to be specified in the sequel. We first sample $h_1 \in F$. We denote by $h_1(P)$ the image of P under h_1, which is a set of nonempty buckets. Now each nonempty bucket $x \in h_1(P)$ is mapped to $\{0, 1\}$: with probability $1/2$, set $f_1(x) = 0$, otherwise set $f_1(x) = 1$.

This is repeated d' times, and eventually for $p \in \mathcal{M}$, we compute the function

$$f(p) = (f_1(h_1(p)), \ldots, f_{d'}(h_{d'}(p))),$$

where $f : P \to \{0, 1\}^{d'}$. Now, observe that

$$\mathrm{d}_\mathcal{M}(p, q) \leq r \implies \mathbb{E}[\|f_i(h_i(p)) - f_i(h_i(q))\|_1] \leq 0.5(1 - p_1),\ i = 1, \ldots, d' \implies$$

$$\implies \mathbb{E}[\|f(p) - f(q)\|_1] \leq 0.5 \cdot d' \cdot (1 - p_1),$$

$$\mathrm{d}_\mathcal{M}(p, q) \geq cr \implies \mathbb{E}[\|f_i(h_i(p) - f_i(h_i(q))\|_1] \geq 0.5(1 - p_2),\ i = 1, \ldots, d' \implies$$

$$\implies \mathbb{E}[\|f(p) - f(q)\|_1] \geq 0.5 \cdot d' \cdot (1 - p_2).$$

We distinguish two cases.

First, consider the case $\mathrm{d}_\mathcal{M}(p, q) \leq r$. Let $\mu = \mathbb{E}[\|f(p) - f(q)\|_1]$. Then,

$$\mathrm{P}r[\|f(p) - f(q)\|_1 \geq \ \mu] \leq \frac{1}{2},$$

since $\|f(p) - f(q)\|_1$ follows the binomial distribution.

Second, consider the case $\mathrm{d}_\mathcal{M}(p, q) \geq cr$. By standard Chernoff bounds, $Pr[\|f(p) - f(q)\|_1 \leq \frac{1-p_1}{1-p_2} \cdot \mu] \leq \exp(-0.5 \cdot \mu \cdot (p_1 - p_2)^2/(1 - p_2)^2) \leq \exp(-d' \cdot (p_1 - p_2)^2/4(1 - p_2))$.

After mapping the query $q \in \mathcal{M}$ to $f(q)$ in the d'-dimensional Hamming space we search for all "near" Hamming vectors $f(p)$ s.t. $\|f(p) - f(q)\|_1 \leq 0.5 \cdot d' \cdot (1 - p_1)$. This search costs $\binom{d'}{1} + \binom{d'}{2} + \cdots + \binom{d'}{\lfloor d' \cdot (1-p_1)/2 \rfloor} \leq O(d' \cdot 2^{d' \cdot H((1-p_1)/2)})$, where $H(\cdot)$ is the binary entropy function. The inequality is obtained from standard bounds on binomial coefficients, e.g. [70]. Now, the expected number of points $p \in P$, for which $\mathrm{d}_{\mathcal{M}}(p, q) \geq cr$ but are mapped "near" q is $\leq n \cdot \exp(-d' \cdot (p_1 - p_2)^2/4(1 - p_2)))$. If we set $d' = \log n$, we obtain expected query time

$$O(n^{H((1-p_1)/2))} + dn^{1-\delta}),$$

where

$$\delta = \frac{(p_1 - p_2)^2}{(1 - p_2)} \cdot \frac{\log \mathrm{e}}{4}.$$

If we stop searching after having seen, say $10n^{1-\delta}$ points for which $\mathrm{d}_{\mathcal{M}}(p, q) \geq cr$, then we obtain the same time with constant probability of success. Notice that "success" translates to successful preprocessing for a fixed query $q \in \mathcal{M}$. The space required is $O(dn)$. $\square$

The value of δ could be somewhat larger, but we have used simplified Chernoff bounds to keep our exposition simple.

Discussion on parameters. We set the dimension $d' = \log n$ (which denotes the binary logarithm), since it minimizes the expected number of candidates under the linear space restriction. We note that it is possible to set $d' < \log n$ and still have sublinear query time. This choice of d' is interesting in practical applications since it improves space requirement. The number of candidate points is set to $n^{1-\delta}$ for the purposes of Lemma 44 and under worst case assumptions on the input.

3.4.1 The ℓ_2 case

3.4.1.1 Project on random lines

Let p, q two points in $\mathbb{R}^d$ and η the distance between them. Let $w > 0$ be a real parameter, and let t be a random number distributed uniformly in the interval $[0, w]$. In [33], they present the following LSH family. For $p \in \mathbb{R}^d$, consider the random function

$$h(p) = \left\lfloor \frac{\langle p, v \rangle + t}{w} \right\rfloor, \quad p, v \in \mathbb{R}^d, \tag{3.3}$$

where v is a vector randomly distributed with the d-dimensional normal distribution. This function describes the projection on a random line, where the parameter t represents the random shift and the parameter w the discretization of the line. For this LSH family, the probability of collision is

$$\alpha(\eta, w) = \int_{t=0}^{w} \frac{2}{\sqrt{2\pi}\eta} \exp\Big(-\frac{t^2}{2\eta^2}\Big)\Big(1 - \frac{t}{w}\Big)dt.$$

Lemma 45. *Given a set of n points $P \subseteq \mathbb{R}^d$, there exists a data structure for the (c,r)-ANN problem under the Euclidean metric, requiring space $O(dn)$, time preprocessing $O(dn)$, and query time $O(dn^{1-\delta} + n^{0.9})$, where*

$$\delta \geq 0.03\,(c-1)^2.$$

Given some query point $q \in \mathbb{R}^d$, the building process succeeds with constant probability.

Proof. In the sequel we use the standard Gauss error function, denoted by $erf(\cdot)$. For probabilities p_1, p_2, it holds that

$$p_1 = \alpha(1,w) = \int_{t=0}^{w} \frac{2}{\sqrt{2\pi}} \text{exp}\Big(-\frac{t^2}{2}\Big)\Big(1-\frac{t}{w}\Big)dt = erf\Big(\frac{w}{\sqrt{2}}\Big) - \sqrt{\frac{2}{\pi}}\frac{1}{w}\Big(1-\text{exp}\,(-\frac{w^2}{2})\Big),$$

and also that

$$p_2 = \alpha(c,w) = \int_{t=0}^{w} \frac{2}{\sqrt{2\pi}c} \text{exp}\Big(-\frac{t^2}{2c^2}\Big)\Big(1-\frac{t}{w}\Big)dt = erf\Big(\frac{w}{\sqrt{2}c}\Big) - \sqrt{\frac{2}{\pi}}\frac{c}{w}\Big(1-\text{exp}\,(-\frac{w^2}{2c^2})\Big).$$

The LSH scheme is parameterized by w. One possible value is $w = 3$, as we have checked on a computer algebra system. On the other hand, $w = c$ gives similar results, and they are simpler to obtain. In particular, we have

$$p_1 - p_2 = erf\Big(\frac{c}{\sqrt{2}}\Big) - \sqrt{\frac{2}{\pi}}\frac{1}{c}\Big(1-\text{exp}\,(-\frac{c^2}{2})\Big) - erf\Big(\frac{1}{\sqrt{2}}\Big) + \sqrt{\frac{2}{\pi}}\Big(1-\text{exp}\,(-\frac{1}{2})\Big).$$

We shall prove that, given $w = c$, for $c \in (1,2]$, it holds that $p_1 - p_2 > \frac{5(c-1)}{21}$. Let us define

$$g(c) = p_1 - p_2 - \frac{5(c-1)}{21} = erf\Big(\frac{c}{\sqrt{2}}\Big) - erf\Big(\frac{1}{\sqrt{2}}\Big) -$$

$$-\sqrt{\frac{2}{\pi}}\frac{1}{c}\Big(1-\text{exp}\,(-\frac{c^2}{2})\Big) + \sqrt{\frac{2}{\pi}}\Big(1-\text{exp}\,(-\frac{1}{2})\Big) - \frac{5(c-1)}{21},$$

$c \in (1,2]$. Using elementary calculus, it is easy to show that $g(c)$ is concave over $c \in (1,2]$. Also, $g(1) = 0$ and $g(2) > 0$, thus $\forall c \in (1,2],\ \ g(c) > 0$ and consequently $p_1 - p_2 > \frac{5(c-1)}{21}$. In addition, $w = c$ implies $1 - p_2 = 1 - erf\big(\frac{1}{\sqrt{2}}\big) + \sqrt{\frac{2}{\pi}}\big(1-\text{exp}(-\frac{1}{2})\big) < 0.64$, and $H\big(\frac{1-p_1}{2}\big) < 0.9$. Hence, for $w = c$ and $c \in (1,2]$, $\delta > 0.03(c-1)^2$. $\square$

3.4.1.2 Hyperplane LSH

This section reduces the Euclidean ANN to an instance of ANN for which the points lie on a unit sphere. The latter admits an LSH scheme based on partitioning the space by randomly selected halfspaces.

In Euclidean space $\mathbb{R}^d$, let us assume that the dimension is $d = O(\log n \cdot \log\log n)$, since one can project points à la Johnson-Lindenstrauss [72], and preserve pairwise distances up to multiplicative factors of $1 \pm o(1)$. Then, we partition $\mathbb{R}^d$ using a randomly shifted grid, with cell edge of length $O(\sqrt{d}) = O((\log n \cdot \log\log n)^{1/2})$. Any two points $p, q \in \mathbb{R}^d$ for which $\|p-q\|_2 \leq 1$ lie in the same cell with constant probability. Let us focus on the set of points lying inside one cell. This set of points has diameter bounded by $O((\log n \cdot \log\log n)^{1/2})$. Now, a reduction of [77], reduces the problem to an instance of ANN for which all points lie on a unit sphere $\mathbb{S}^{d-1}$, and the search radius is roughly $r' = \Theta((\log n \cdot \log\log n)^{-1/2})$. These steps have been also used in [11], as a data-independent reduction to the spherical instance.

Let us now consider the LSH family introduced in [28]. Given n unit vectors $P \subset \mathbb{S}^{d-1}$, we define, for each $q \in \mathbb{S}^{d-1}$, hash function $h(q) = sign\langle q, v\rangle$, where v is a random unit vector. Obviously, $\Pr[h(p) = h(q)] = 1 - \frac{\theta(p,q)}{\pi}$, where $\theta(p,q)$ denotes the angle formed by the vectors $p \neq q \in \mathbb{S}^{d-1}$. Instead of directly using the family of [28], we employ its amplified version, obtained by concatenating $d' \approx 1/r'$ functions $h(\cdot)$, each chosen independently and uniformly at random from the underlying family. The amplified function $g(\cdot)$ shall be fully defined in the proof below. This procedure leads to the following.

Lemma 46. *Given a set of n points $P \subset \mathbb{R}^d$, there exists a data structure for the (c,r)-ANN problem under the Euclidean metric, requiring space $O(dn)$, time preprocessing $O(dn)$, and query time $O(dn^{1-\delta} + n^{0.91})$, where*

$$\delta \geq 0.05 \cdot \left(\frac{c-1}{c}\right)^2.$$

Given some query $q \in \mathbb{R}^d$, the building process succeeds with constant probability.

Proof. We exploit the reduction described above that translates the Euclidean ANN to a spherical instance of ANN with search radius $r' = \Theta((\log n \cdot \log\log n)^{-1/2})$. The latter is handled by a hyperplane LSH scheme based on [28] as detailed immediately below.

Let us denote by F the aforementioned LSH family of [28]. We build a new (amplified) family of functions $G_{d'} = \{g(x) = (h_1(x), \ldots, h_{d'}(x)) : i = 1 \ldots d',\ h_i \in F)\}$. Now, obviously, for any two unit vectors $p \neq q$, we have

$$\Pr_{g \in G}[g(p) = g(q)] = \left(1 - \frac{\theta(p,q)}{\pi}\right)^{d'}.$$

Hence, $\|p-q\|_2 \leq r' \implies 2\sin\left(\frac{\theta(p,q)}{2}\right) \leq r' \implies \theta(p,q) \leq 2\arcsin\left(\frac{r'}{2}\right) = \theta_r$, which defines θ_r.
Moreover, $\|p-q\|_2 \geq cr' \implies 2\sin\left(\frac{\theta(p,q)}{2}\right) \geq cr' \implies \theta(p,q) \geq 2\arcsin\left(\frac{cr'}{2}\right)$.
By using elementary calculus, it is easy to prove that $2\arcsin\left(\frac{cr'}{2}\right) \geq 2c \cdot \arcsin\left(\frac{r'}{2}\right) \implies \theta(p,q) \geq c \cdot \theta_r$. Hence, for $d' = \lfloor \pi/\theta_r \rfloor$ and since $r' = \Theta((\log n \cdot \log\log n)^{-1/2}) \implies \theta_r = o(1)$,

$$p_1 = \Pr[g(p) = g(q) \mid \|p-q\|_2 \leq r] \geq \left(1 - \frac{\theta_r}{\pi}\right)^{d'} \geq \exp\left(-\frac{\pi}{(\pi - \theta_r)}\right) \geq \frac{1}{\mathrm{e}^{1+o(1)}},$$

$$p_2 = \Pr[g(p) = g(q) \mid \|p-q\|_2 \geq c \cdot r] \leq \left(1 - \frac{c \cdot \theta_r}{\pi}\right)^{d'} \leq \exp\left(-\frac{c \cdot \theta_r}{\pi} \cdot \left(\frac{\pi}{\theta_r} - 1\right)\right) \leq \frac{1}{c \cdot \mathrm{e}^{1-o(1)}}.$$

Now applying Lemma 44 yields

$$\delta \geq \frac{1}{\mathsf{e}^{2+o(1)}} \cdot \Big(1 - \frac{\mathsf{e}^{o(1)}}{c}\Big)^2 \cdot \frac{1}{1-(c\cdot \mathsf{e})^{-1}} \cdot \frac{\log(\mathsf{e})}{4} \geq 0.059 \cdot \Big(1 - \frac{1}{c}\Big)^2, \quad \text{for } c \in (1,2].$$

The space required is $O(dn + nd')) = O(dn)$. Notice also that $H(\frac{1-p_1}{2}) \leq 0.91$. $\square$

The data structure of Lemma 46 provides slightly better query time than that of Lemma 45, when c is small enough.

3.4.2 The ℓ_1 case

In this section, we study the (c,r)-ANN problem under the ℓ_1 metric. The dataset consists again of n points $P \subset \mathbb{R}^d$ and the query point is $q \in \mathbb{R}^d$.

For this case, let us consider the following LSH family, introduced in [9]. A point p is hashed as follows:

$$h(p) = \Big(\Big\lfloor \frac{p_1+t_1}{w} \Big\rfloor, \Big\lfloor \frac{p_2+t_2}{w} \Big\rfloor, \ldots, \Big\lfloor \frac{p_d+t_d}{w} \Big\rfloor \Big),$$

where $p = (p_1, p_2, \ldots, p_d)$ is a point in P, $w = \alpha r$, and the t_i are drawn uniformly at random from $[0, \ldots, w)$. Buckets correspond to cells of a randomly shifted grid.

Now, in order to obtain a better lower bound, we employ an amplified hash function, defined by concatenation of $d' = \alpha$ functions $h(\cdot)$ chosen uniformly at random from the above family.

Lemma 47. *Given a set of n points $P \subseteq \mathbb{R}^d$, there exists a data structure for the (c,r)-ANN problem under the ℓ_1 metric, requiring space $O(dn)$, time preprocessing $O(dn)$, and query time $O(dn^{1-\delta} + n^{0.91})$, where*

$$\delta \geq 0.05 \cdot \Big(\frac{c-1}{c}\Big)^2.$$

Given some query point $q \in \mathbb{R}^d$, the building process succeeds with constant probability.

Proof. We denote by F the previously introduced LSH family of [9], which is $(1-\frac{1}{\alpha}, 1-\frac{c}{c+\alpha}, 1, c)$-sensitive. We build the amplified family of functions

$$G_{d'} = \{g(x) = (h_1(x), \ldots, h_{d'}(x)) : i = 1, \ldots, d', \ h_i \in F)\}.$$

Setting $\alpha = d' = \log n$, we have:

$$p_1 = \Big(1 - \frac{1}{\alpha}\Big)^{d'} = \Big(1 - \frac{1}{\log n}\Big)^{\log n} \geq \Big(\exp\Big(-\frac{1}{\log n - 1}\Big)\Big)^{\log n} \geq \frac{1}{\mathsf{e}^{1+o(1)}},$$

$$p_2 = \Big(1 - \frac{c}{\alpha + c}\Big)^{d'} = \Big(1 - \frac{c}{\log n + c}\Big)^{\log n}.$$

Table 3.1: Juxtaposition of our results with previous and concurrent results on the linear-space regime.

	Space	Query
Entropy-based LSH [73]	$\tilde{O}(dn)$	$dn^{O((1+\epsilon)^{-1})}$
Entropy-based LSH [10]	$\tilde{O}(dn)$	$dn^{O((1+\epsilon)^{-2})}$
Theorem 35	$\tilde{O}(dn)$	$dn^{1-\Theta(\epsilon^2/\log(1/\epsilon))}$
Lemma 45	$\tilde{O}(dn)$	$dn^{1-\Theta(\epsilon^2)}$
LSH tradeoffs [11]	$\tilde{O}(dn)$	$O(dn^{(2(1+\epsilon)^2-1)/(1+\epsilon)^4})$

Hence,

$$p_2 \geq \exp(-c) \geq \frac{1}{\mathrm{e}\cdot(2c-1)},$$

and

$$p_2 \leq \exp\Big(-\frac{c}{1+\frac{c}{\log n}}\Big) = \exp\Big(-\frac{c}{1+o(1)}\Big) \leq \exp\big(-c+o(1)\big) \leq \frac{\mathrm{e}^{o(1)}}{\mathrm{e}c}.$$

Therefore, for n large enough, it holds that

$$\delta = \frac{\big(p_1-p_2\big)^2}{\big(1-p_2\big)}\cdot\frac{\log\mathrm{e}}{4} \geq \frac{1}{\mathrm{e}^{2+o(1)}}\cdot\frac{(1-\frac{1}{c})^2}{1-\frac{1}{\mathrm{e}(2c-1)}}\cdot\frac{\log\mathrm{e}}{4} \geq 0.055\cdot(1-\frac{1}{c})^2, \quad \text{for } c\in(1,2].$$

Notice that $H((1-p_1)/2) \leq 0.91$. □

3.5 Summary

In this section, we presented (c, r)-ANN data structures on the linear-space regime with sublinear query time for any $c > 1$, and polynomial dependence. As it is shown in Table 3.1, previously, most results in this regime were non-trivial only when c was a large enough constant. After the original submission of our paper [8], a better query time of $O(n^{1-4\epsilon^2+O(\epsilon^3)})$ has been established [11]. The bound has been shown to be optimal for a large class of data structures. Despite the fact that our algorithms are sub-optimal, they are simpler and easier to implement.

4. NEAR-NEIGHBOR PRESERVING DIMENSION REDUCTION FOR DOUBLING SUBSETS OF ℓ_1

In this chapter we focus on the $(1+\epsilon, r)$-ANN problem for subsets of ℓ_1 with bounded doubling dimension. It is known that dimension reduction in ℓ_1 cannot be achieved in the same generality as in ℓ_2, even assuming that the pointset is of low doubling dimension [66]: there are arbitrarily large n-point subsets $P \subseteq \ell_1$ which are doubling with constant 6, such that every embedding with distortion D of P into $\ell_1^{d'}$ requires dimension $n^{\Omega(1/D^2)}$. Aiming for more restrictive guarantees, e.g. preserving distances within some pre-defined range, is a relevant workaround. Then, dimension reduction techniques for doubling subsets of ℓ_p, $p \in [1,2]$, exist [22], but they rely on partition algorithms which require the whole pointset to be known in advance. Hence, applicability of such techniques is quite limited and, specifically, it is not clear whether they can be used in an online setting where query points are not known beforehand.

The main result in the context of randomized embeddings for dimension reduction in ℓ_1^d is the following theorem, which exploits the 1-stability property of Cauchy random variables and provides an asymmetric guarantee: The probability of non-contraction is high, but the probability of non-expansion is constant. Nevertheless, this asymmetric property is sufficient for proximity search.

Theorem 48 (Theorem 5, [56]). *For any $\epsilon \leq 1/2$, $\delta > 0$, $\epsilon > \gamma > 0$ there is a probability space over linear mappings $f : \ell_1^d \rightarrow \ell_1^{d'}$, where $d' = (\ln(1/\delta))^{1/(\epsilon-\gamma)}/\zeta(\gamma)$, for a function $\zeta(\gamma) > 0$ depending only on γ, such that for any pair of points $p, q \in \ell_1^d$:*

$$\Pr\left[\, \|f(p)-f(q)\|_1 \leq (1-\epsilon)\, \|p-q\|_1 \,\right] \leq \delta, \Pr\left[\, \|f(p)-f(q)\|_1 \geq (1+\epsilon)\, \|p-q\|_1 \,\right] \leq \frac{1+\gamma}{1+\varepsilon}.$$

Note that the embedding is defined as $f(u) = Au/T$, where A is a $d' \times d$ matrix with each element being an i.i.d. Cauchy random variable. In addition, T is a scaling factor defined as the expectation of a sum of truncated Cauchy variables, such that $T = \Theta(d' \log(d'/\epsilon))$ (see Lemma 5 in [56]).

In this chapter, we establish two non-linear *near* neighbor-preserving embeddings for doubling subsets of ℓ_1^d. We use a definition which is essentially a modified version of the nearest neighbor preserving embedding of [58]:

Definition 49 (Near-neighbor preserving embedding). *Let (Y, d_Y), (Z, d_Z) be metric spaces and $X \subseteq Y$. A distribution over mappings $f : Y \rightarrow Z$ is a* near-neighbor preserving embedding *with range $r > 0$, distortion $D \geq 1$ and probability of correctness $\mathcal{P} \in [0,1]$ if, $\forall \alpha \geq 1$ and $\forall q \in Y$, if $x \in X$ is such that $d_Y(x,q) \leq r$, then with probability at least $\mathcal{P}$,*

- $d_Z(f(x), f(q)) \leq D \cdot r$,
- $\forall p \in X : \; d_Y(p,q) \geq D \cdot \alpha \cdot r \implies d(f(p), f(q)) \geq \alpha \cdot r.$

Both embeddings consist of two basic components. First, we represent the pointset P with an ϵ-covering set, and then we apply a random linear projection à la Indyk [56] to that set, using Cauchy variables.

The role of the covering set is to exploit the doubling dimension of P. In the analogous result for ℓ_2 [58], no representative sets were used; the mapping was just a random linear projection of P. In the case of ℓ_1 however, a similar analysis of a linear projection with Cauchy variables without these representative sets seems to be impossible, since the Cauchy distribution is heavy tailed.

In Theorem 53, we consider c-approximate r-nets as a covering set. Inspired by the algorithm of [42] for ℓ_2, we design an algorithm that computes a c-approximate r-net in ℓ_1 in subquadratic –but superlinear– time. On the other hand, Theorem 56 relies on randomly shifted grids, which can be computed in linear time, but are inferior to nets in terms of capturing the doubling dimension of the pointset.

To bound the distortion incurred by the randomized projection, we exploit the 1-stability property of the Cauchy distribution. To this end, we prove a concentration bound for sums of independent Cauchy variables. To overcome the technical difficulties associated with the heavy tails of the Cauchy distribution, we study sums of *square roots* of Cauchy variables, where in [56], Indyk considers sums of *truncated* Cauchy variables instead. Although our concentration bound is rather weak, it is sufficient for our purposes and its analysis is much simpler compared to Indyk's.

Organization. Section 4.1 establishes a concentration bound on sums of independent Cauchy variables. Section 4.2, achieves dimensionality reduction by means of representing the pointset by a carefully chosen net, while Section 4.3 employs randomly shifted grids for the same task. We conclude with discussion of results and implications.

4.1 Concentration bounds for Cauchy variables

In this section, we prove some basic properties of the Cauchy distribution, which serves as our main embedding tool.

Let $C_{\mathcal{D}}$ denote the Cauchy distribution with density $c(x) = (1/\pi)/(1+x^2)$. One key property of the Cauchy distribution is the so-called 1-stability property: Let $v = (v_1, \ldots, v_{d'}) \in \mathbb{R}^{d'}$ and $X_1, \ldots, X_{d'}$ be i.i.d. random variables following $C_{\mathcal{D}}$, then $\sum_{j=1}^{d'} X_i v_i$ is distributed as $X\|v\|_1$, where $X \sim C_{\mathcal{D}}$.

The Cauchy distribution has undefined mean. However, for $0 < q < 1$, the mean of the q-th power of a Cauchy random variable can be defined. More specifically, for some $X \sim C_{\mathcal{D}}$ we have

$$\mathbb{E}\left[|X|^{1/2}\right] = \frac{2}{\pi}\int_0^\infty \frac{\sqrt{x}}{1+x^2}\,\mathsf{d}x = \frac{2}{\pi}\frac{\pi}{\sqrt{2}} = \sqrt{2}.$$

The following lemma provides a bound for the moment-generating function of $|X|^{1/2}$.

Lemma 50. *Let $X \sim C_{\mathcal{D}}$. Then for any $\beta > 1$:*

$$\mathbb{E}\left[\exp\left(-\beta|X|^{1/2}\right)\right] \le \frac{2}{\beta}.$$

Proof. For any constant β,

$$\int_0^1 \mathrm{e}^{-\beta x^{1/2}}\, \mathrm{d}x = \frac{2}{\beta^2}\left(1 - \frac{\beta+1}{\mathrm{e}^{\beta}}\right).$$

Then, for any $\beta > 1$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\beta|X|^{1/2}\right)\right] &= \int_{-\infty}^{\infty} \mathrm{e}^{-\beta|x|^{1/2}} \cdot c(x)\, \mathrm{d}x = \frac{2}{\pi}\int_0^{\infty} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\, \mathrm{d}x = \\
&= \frac{2}{\pi}\int_0^{1} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\, \mathrm{d}x + \frac{2}{\pi}\int_1^{\infty} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\, \mathrm{d}x \le \\
&\le \frac{2}{\pi}\int_0^{1} \mathrm{e}^{-\beta x^{1/2}}\, \mathrm{d}x + \frac{2}{\pi}\int_1^{\infty} \mathrm{e}^{-\beta} \cdot \frac{1}{1+x^2}\, \mathrm{d}x = \\
&= \frac{2}{\pi}\cdot\frac{2}{\beta^2}\left(1 - \frac{\beta+1}{\mathrm{e}^{\beta}}\right) + \frac{1}{2\mathrm{e}^{\beta}} \le \frac{4}{\pi\beta^2} + \frac{1}{2\mathrm{e}^{\beta}} \le \frac{2}{\beta}.
\end{aligned}$$

□

Let $S := \sum_{j=1}^{d'} |X_j|$ where each X_j is an i.i.d. Cauchy variable. To prove concentration bounds for S, we study the sum $\tilde{S} := \sum_{j=1}^{d'} |X_j|^{1/2}$. By known bounds, $S \le \tilde{S}^2 \le d' \cdot S$ hence, for any $t > 0$,

$$\mathsf{P}r[S \le t] \le \mathsf{P}r[\tilde{S} \le \sqrt{td'}]. \tag{4.1}$$

We use the bound on the moment-generating function, to prove a Chernoff-type concentration bound for $\tilde{S}$, which by Eq. (4.1) translates into a concentration bound for S.

Lemma 51. *For every $D > 1$,*

$$\mathsf{P}r\left[\tilde{S} \le \frac{\mathbb{E}[\tilde{S}]}{D}\right] \le \left(\frac{10}{D}\right)^{d'}.$$

Proof. Since X_j's are independent, $\mathbb{E}[\tilde{S}] = \sqrt{2}d'$. Then, by Lemma 50 and Markov's inequality, for any $\beta > 1$, it follows that

$$\begin{aligned}
\mathsf{P}r\left[\tilde{S} \le \frac{\mathbb{E}[\tilde{S}]}{D}\right] &= \mathsf{P}r\left[\exp(-\beta\tilde{S}) \ge \exp\left(-\beta\cdot\frac{\mathbb{E}[\tilde{S}]}{D}\right)\right] \le \\
&\le \frac{\mathbb{E}[\exp(-\beta\tilde{S})]}{\exp(-\beta\mathbb{E}[\tilde{S}]/D)} = \frac{\mathbb{E}[\exp(-\beta|X_j|^{1/2})]^{d'}}{\exp(-\beta\sqrt{2}d'/D)} \le \left(\frac{2}{\beta}\right)^{d'} \cdot \mathrm{e}^{\sqrt{2}\beta d'/D}.
\end{aligned}$$

Setting $\beta = D$ completes the proof. □

4.2 Net-based dimension reduction

In this section we describe the dimension reduction mapping for ℓ_1 via r-nets. Let $P \subset \ell_1^d$ be a set of n points with doubling constant λ_P. For some point $x \in \mathbb{R}^d$ and $r > 0$, we denote by $B_1(x, r)$ the ℓ_1-ball of radius r around x. The embedding is non-linear and is carried out in two steps.

First, we compute a c-approximate (ϵ/c)-net $\mathcal{N}$ of P with the algorithm of Theorem 21. Moreover, the algorithm assigns each point of P to the point of $\mathcal{N}$ which covered it. Let $g : P \to \mathcal{N}$ be this assignment. In the second step, for every $s \in \mathcal{N}$ and any query point $q \in \ell_1^d$, we apply the linear map of Theorem 48. That is, $f(s) = As/T$, where A is a $d' \times d$ matrix with each element being an i.i.d. Cauchy random variable. Recall that value $T = \Theta(d' \log(d'/\epsilon))$. By the 1-stability property of the Cauchy distribution, $f(s)$ is distributed as $\|s\|_1 \cdot (Y_1, \dots, Y_{d'})$, where each Y_j is i.i.d. and $Y_j \sim C_{\mathcal{D}}$. Hence, $\|f(s)\|_1 = \|s\|_1 \cdot S$ where $S := \sum_j |Y_j|$.

We define the embedding to be $h = f \circ g$. We apply h to every point in P, and f to any query point q. It is clear from the properties of the net that g incurs an additive error of $\pm\epsilon$ on the distance between q and any point in P, so it is sufficient to consider the distortion of f.

Our analysis consists of studying separately the following disjoint subsets of $\mathcal{N}$: Points that lie at distance at most D_0 from the query and points that lie at distance at least D_0, for some $D_0 > 1$ chosen appropriately. For the former set, we directly apply Theorem 48, as it has bounded diameter.

The next lemma guarantees the low distortion for points of the latter set, namely those that are sufficiently far from the query. We consider the sum of the square roots of each $|Y_j|$, i.e., $\tilde{S} = \sum_j |Y_j|^{1/2}$, in order to employ the tools of Section 4.1.

Lemma 52. *Fix a query point $q \in \ell_1^d$. For any $\epsilon \leq 1/2$, $c \geq 1$, $\delta \in (0,1)$, there exists $D_0 = O(\log(d'/\epsilon))$ such that for $d' = \Theta\left(\log^2 \lambda_P \cdot \log(c/\epsilon) + \log(1/\delta)\right)$, with probability at least $1 - \delta$,*

$$\forall s \in \mathcal{N} : \ \|s - q\|_1 \geq D_0 \implies \|f(s) - f(q))\|_1 \geq 4.$$

Proof. Assume wlog that the query point is the origin $(0, \dots, 0)$. For some $D_0 > 1$, we define the following subsets of $\mathcal{N}$:

$$N_i = \{s \in \mathcal{N} \mid D_i \leq \|s\|_1 < D_{i+1}\}, \ D_i = 2^{2i} D_0, \ i = 0, 1, 2, \dots$$

By the definition of doubling constant and the fact that two points of $\mathcal{N}$ lie at distance at least ϵ, $|N_i|$ is at most $\lambda_P^{\lceil \log(4cD_{i+1}/\epsilon) \rceil} \leq \lambda_P^{4\log(cD_{i+1}/\epsilon)}$. Therefore, by the union bound, and Eq. (4.1):

$$\Pr\left[\exists i \exists s \in N_i : \|f(s)\|_1 \leq \frac{4\|s\|_1}{D_i}\right] = \Pr\left[\exists i \exists s \in N_i : S \leq \frac{4T}{D_i}\right] \leq$$
$$\leq \sum_{i=0}^{\infty} |N_i| \Pr\left[\tilde{S} \leq \frac{\sqrt{4d'T}}{\sqrt{D_i}}\right] = \sum_{i=0}^{\infty} |N_i| \Pr\left[\tilde{S} \leq \mathbb{E}[\tilde{S}] \cdot \sqrt{\frac{2T}{d' 2^{2i} D_0}}\right].$$

By Lemma 51, for $D_0 = \lceil 800T/d' \rceil = \Theta(\log(d'/\epsilon))$ and $d' > 4 \cdot \log\lambda_P \cdot \log(cD_0/\epsilon) + 2\log(2\lambda_P/\delta)$:

$$\sum_{i=0}^{\infty} |N_i| \mathsf{P}r\left[\tilde{S} \leq \frac{\mathbb{E}[\tilde{S}]}{10 \cdot 2^{i+1}}\right] \leq \sum_{i=0}^{\infty} \lambda_P^{4\log(cD_{i+1}/\epsilon)} \left(\frac{1}{2^{i+1}}\right)^{d'} = \sum_{i=0}^{\infty} \frac{2^{\log(\lambda_P)(4\log(cD_0/\epsilon)+2i+2)}}{2^{d'(i+1)}} \leq$$

$$\leq \sum_{i=0}^{\infty} \frac{2^{\log(\lambda_P)\cdot 4\log(cD_0/\epsilon)} \cdot 2^{2\log(\lambda_P)(i+1)}}{2^{(4\cdot\log\lambda_P\cdot\log(cD_0/\epsilon))(i+1)} \cdot 2^{2\log(2\lambda_P/\delta))(i+1)}} \leq$$

$$\leq \sum_{i=0}^{\infty} 2^{-2\log(2/\delta))(i+1)} = \sum_{i=0}^{\infty} \left(\frac{\delta^2}{4}\right)^i - 1 = \frac{\delta^2}{4-\delta^2} \leq \delta.$$

Finally, for some large enough constant C, we demand that

$$d' > C\left(\log\lambda_P \cdot \log(c\log d'/\epsilon) + \log(1/\delta)\right) > 4 \cdot \log\lambda_P \cdot \log(cD_0/\epsilon) + 2\log(2\lambda_P/\delta)$$

which is satisfied for $d' = \Theta\left(\log^2\lambda_P \cdot \log(c/\epsilon) + \log(1/\delta)\right)$. □

Theorem 53. *Let $P \subset \ell_1^d$ such that $|P| = n$. For any $\epsilon \in (0, 1/2)$ and $c \geq 1$, there is a non-linear randomized embedding $h = f \circ g : \ell_1^d \to \ell_1^{d'}$, where $d' = (\log\lambda_P \cdot \log(c/\epsilon))^{\Theta(1/\epsilon)}/\zeta(\epsilon)$, for a function $\zeta(\epsilon) > 0$ depending only on ϵ, such that, for any $q \in \ell_1^d$, if there exists $p^* \in P$ such that $\|p^* - q\|_1 \leq 1$, then, with probability $\Omega(\epsilon)$:*

$$\|h(p^*) - f(q)\|_1 \leq 1 + 3\epsilon, \forall p \in P: \ \|p - q\|_1 > 1 + 9\epsilon \implies \|h(p) - f(q)\|_1 > 1 + 3\epsilon.$$

Set P can be embedded in time $\tilde{O}(dn^{1+1/\Omega(c)})$, and any query $q \in \ell_1^d$ can be embedded in time $O(dd')$.

Proof. Let f, g be the mappings defined in the beginning of the section and $D_0 = \Theta(\log(d'/\epsilon))$. Assume wlog for simplicity that $q = 0^d$. Then, by Lemma 52 for $d' = \Theta\left(\log^2\lambda_P \cdot \log(c/\epsilon)\right)$, with probability at least $1 - \epsilon/5$, we have:

$$\forall p \in P: \ \|p - q\|_1 \geq D_0 + \epsilon \implies \|h(p) - f(q)\|_1 \geq 4.$$

By Theorem 48, for $\gamma = \epsilon/10$ and $\delta = \epsilon/(5\lambda_P^{8\log(cD_0/\epsilon)})$, with probability at least $1 - \epsilon/5$, we get:

$$\forall p \in P: \ \|p - q\|_1 \in (1 + 9\epsilon, D_0 + \epsilon) \implies \|h(p) - f(q)\|_1 > (1 + 8\epsilon)(1 - \epsilon) \geq 1 + 3\epsilon.$$

Moreover,

$$\mathsf{P}r\left[\|h(p^*) - f(q)\|_1 \leq 1 + 3\epsilon\right] \geq 1 - \frac{1 + \epsilon/10}{1 + \varepsilon} \geq 1 - (1 - \epsilon/2).$$

Then, the target dimension needs to satisfy the following inequality:

$$d' \geq \frac{\left(\ln(5\lambda_P^{8\log(cD_0/\epsilon)}/\epsilon)\right)^{2/\epsilon}}{\zeta(\epsilon)} = \frac{\left(\Theta(\log\log d' \cdot \log\lambda_P + \log\lambda_P \cdot \ln(c/\epsilon))\right)^{2/\epsilon}}{\zeta(\epsilon)}.$$

Hence, for $d' = (\log\lambda_P \cdot \log(c/\epsilon))^{\Theta(1/\epsilon)}/\zeta(\epsilon)$, we achieve a total probability of success in $\Omega(\epsilon)$, which completes the proof. □

4.3 Dimension reduction based on randomly shifted grids

In this section, we explore some properties of randomly shifted grids, and we present a simplified embedding which consists of a first step of snapping points to a grid, and a second step of randomly projecting grid points.

Let $w > 0$ and t be chosen uniformly at random from the interval $[0, w]$. The function

$$h_{w,t}(x) = \left\lfloor \frac{x-t}{w} \right\rfloor$$

induces a random partition of the real line into segments of length w. Hence, the function

$$g_w(x) = (h_{w,t_1}(x_1), ..., h_{w,t_d}(x_d)),$$

for $t_1, \ldots, t_d$ independent uniform random variables in the interval $[0, w]$, induces a randomly shifted grid in $\mathbb{R}^d$. For a set $X \subseteq \mathbb{R}^d$, we denote by $g_w(X)$, the image of X on the randomly shifted grid points defined by g_w. For some $x \in \mathbb{R}^d$ and $r > 0$, the number of grid cells of $g_w(\ell_1^d)$ that $B_1(x, r)$ intersects per axis is independent, and in expectation is $1+2r/w$. Then, the expected total number of grid cells that $B_1(x, r)$ intersects is $(1+2r/w)^d$.

Now let $P \subset \ell_1^d$ be a set of n points with doubling constant λ_P and $q \in \ell_1^d$ a query point. For $w = \epsilon/d$, the ℓ_1-diameter of each cell is ϵ and therefore $g_w(P)$ is an ϵ-covering set of P.

Lemma 54. *Let $R > 1$ and $P' := B_1(q, R) \cap P$. Then, for $w = \epsilon/d$*

$$\mathbb{E}\left[|g_w(P')|\right] \leq 8\lambda_P^{2\log(dR/\epsilon)}.$$

Proof. By the doubling constant definition, there exists a set of balls of radius ϵ/d^2 centered at points in P', of cardinality at most $\lambda_P^{2\log(dR/\epsilon)}$ which covers P'. For each ball, the expected number of intersecting grid cells is $(1 + 2/d)^d \leq \mathrm{e}^2$. The lemma follows by linearity of expectation. □

The next lemma shows that, with constant probability, the growth on the number of representatives, as we move away from q, is bounded.

Lemma 55. *Let $\{D_i\}_{i\in\mathbb{N}}$ be a sequence of radii such that, for any i, $D_{i+1} = 4D_i$. Let A_i be the points of $g_w(P)$ within distance $D_{i+1} = 2^{2(i+1)}D_0$ from q. Then, with probability at least $1/3$,*

$$\forall i \in \{-1, 0, \ldots\} : |A_i| \leq 4^{i+3}\lambda_P^{2\log(dD_{i+1}/\epsilon)}.$$

Proof. By Lemma 54, $\mathbb{E}[|A_i|] \leq 8\lambda_P^{2\log(dD_{i+1}/\epsilon)}$ for every $i \in \{-1, 0, \ldots\}$. Then, a union bound followed by Markov's inequality yields

$$\mathrm{P}r\left[\exists i \in \{0, 1, \ldots\} : |A_i| \geq 4^{i+1}\mathbb{E}[|A_i|]\right] \leq 1/3.$$

In addition,

$$\mathrm{P}r\left[|A_{-1}| \geq 4\mathbb{E}[|A_i|]\right] \leq 1/4.$$ □

Theorem 56. *Let $P \subset \ell_1^d$ such that $|P| = n$. For any $\epsilon \in (0, 1/2)$, there is a non-linear randomized embedding $h' : \ell_1^d \to \ell_1^{d'}$, where $d' = (\log \lambda_P \cdot \log(d/\epsilon))^{\Theta(1/\epsilon)} / \zeta(\epsilon)$, for a function $\zeta(\epsilon) > 0$ depending only on ϵ, such that for any $q \in \ell_1^d$, if there exists $p^* \in P$ such that $\|p^* - q\|_1 \leq 1$, then with probability $\Omega(\epsilon)$,*

$$\|h'(p^*) - f(q)\|_1 \leq 1 + 3\epsilon, \forall p \in P : \ \|p - q\|_1 > 1 + 9\epsilon \implies \|h'(p) - f(q)\|_1 > 1 + 3\epsilon.$$

Any point can be embedded in time $O(dd')$.

Proof. We follow the same reasoning as in the proof of Theorem 53. The embedding is $h' = f \circ g_{\epsilon/d}$, where f is the randomized linear map defined in Section 4.2. As before, we apply h' to every point in P, and only f to queries. The randomly shifted grid incurs an additive error of ϵ in the distances between q and P.

Assume wlog that $q = 0^d$ and let A_i be the points of $g_{\epsilon/d}(P)$ within distance $D_{i+1} = 2^{2(i+1)} D_0$ from q. Hence, by Lemma 55,

$$\Pr\left[\exists i \exists s \in A_i : \|f(s)\|_1 \leq \frac{4\,\|s\|_1}{D_i}\right] \leq \sum_{i=0}^{\infty} |A_i| \Pr\left[S \leq \frac{4T}{D_i}\right] \leq$$

$$\leq \sum_{i=0}^{\infty} 4^{i+3} \lambda_P^{2\log(dD_{i+1}/\epsilon)} \Pr\left[\tilde{S} \leq \frac{\sqrt{4d'T}}{\sqrt{D_i}}\right].$$

As in Lemma 52, for $D_0 = \lceil 800T/d' \rceil = \Theta(\log(d'/\epsilon))$, $d' \geq 20 \log \lambda_P \cdot \log\left(\frac{dD_0}{\epsilon\delta}\right)$ and $\delta = \epsilon/5$,

$$\sum_{i=0}^{\infty} 4^{i+3} \lambda_P^{2\log(dD_{i+1}/\epsilon)} \Pr\left[\tilde{S} \leq \frac{\sqrt{4d'T}}{\sqrt{D_i}}\right] \leq \sum_{i=0}^{\infty} \frac{2^{2i+6+2\log\lambda_P[\log(dD_0/\epsilon)+2(i+1)]}}{2^{d'(i+1)}} \leq \epsilon/5.$$

Hence, for $d' = \Omega\left((\log^2 \lambda_P \cdot \log(d/\epsilon)\right)$, with probability at least $1 - \epsilon/5$, we have:

$$\forall p \in P : \ \|p - q\|_1 \geq D_0 + \epsilon \implies \|h'(p) - f(q)\|_1 \geq 4.$$

Now, we are able to use Theorem 48 for points which are at distance at most $D_0 + \epsilon$ from q, and the near neighbor. By Lemma 55, with constant probability, the number of grid points at distance $\leq D_0 + \epsilon$, is at most $32 \cdot \lambda_P^{4\log(dD_0/\epsilon)}$. Hence, by Theorem 48, for $\gamma = \epsilon/10$ and $\delta = \epsilon/(160\lambda_P^{4\log(dD_0/\epsilon)})$, with probability at least $1 - \epsilon/5$, it holds:

$$\forall p \in P : \ \|p - q\|_1 \in (1 + 9\epsilon, D_0 + \epsilon) \implies \|h'(p) - f(q)\|_1 > 1 + 3\epsilon.$$

Moreover, with probability at least $\epsilon/2$, we obtain:

$$\|h'(p^*) - f(q)\|_1 \leq 1 + 3\epsilon.$$

As in Theorem 53, the target dimension needs to satisfy the following:

$$d' \geq \frac{\left(\ln(160\lambda_P^{4\log(dD_0/\epsilon)}/\epsilon)\right)^{2/\epsilon}}{\zeta(\epsilon)}.$$

Hence, for $d' = (\log \lambda_P \cdot \log(d/\epsilon))^{\Theta(1/\epsilon)} / \zeta(\epsilon)$ we achieve total probability of success $\Omega(\epsilon)$. □

Table 4.1: Comparison with related dimension reduction results.

Comments	Target dimension	Time
[56], Nearest-Neighbor preserving, ℓ_1	$d' = (\log n)^{\Theta(1/\epsilon)}/\zeta(\epsilon)$	$O(dd'n)$
[58], Nearest-Neighbor preserving, ℓ_2	$d' = \log(1/\epsilon)\log\lambda_P/\epsilon^2$	$O(dd'n)$
Theorem 53	$d' = (\log\lambda_P \cdot \log(\mathbf{c}/\epsilon))^{\Theta(1/\epsilon)}/\zeta(\epsilon)$	$\tilde{O}(dn^{1+1/\Omega(c)})$
Theorem 56	$d' = (\log\lambda_P \cdot \log(\mathbf{d}/\epsilon))^{\Theta(1/\epsilon)}/\zeta(\epsilon)$	$O(dd'n)$

4.4 Summary and algorithmic implications.

In Table 4.1, we show a comparison of our results with previous results on dimension reduction for proximity search. Previous results focus on different scenarios: either subsets of ℓ_1 without any assumption on the doubling dimension, or doubling subsets of ℓ_2.

Our results show that efficient dimension reduction for doubling subsets of ℓ_1 is possible, in the context of ANN. In particular, these results imply efficient sketches, meaning that one can solve $(1+\epsilon, r)$-ANN with minimal storage per point. Dimension reduction also serves as a problem reduction from a high-dimensional hard instance to a low-dimensional easy instance. Since the algorithms presented in this chapter are quite simple, they should also be of practical interest: they easily extend the scope of any implementation which has been optimized to solve the problem in low dimension, so that it may handle high-dimensional data.

Our embedding can be combined with the bucketing method of [51] for the $(1+\epsilon, r)$-ANN problem in ℓ_1^d. For instance, setting $c = \log n$ in Theorem 53, yields preprocessing time $dn^{1+o(1)}$, space $n^{1+o(1)}$ and query time $O(d)\cdot(\log\lambda_P \cdot \log\log n)^{O(1/\epsilon)}$ assuming that the doubling dimension is a fixed constant. This improves upon existing results: the query time of [63] depends on the aspect ratio of the dataset, while the data structures of [52, 30] support queries with time complexity which depends exponentially on the doubling dimension. However, it is worth noting that one could potentially improve the results of [63, 52, 30] in the special case of ℓ_1, by employing ANN data structures with fast query time, in order to accelerate the traversal of the net-tree. Hence, while our result gives a simple framework for exploiting the intrinsic dimension of doubling subsets of ℓ_1, it is unlikely that it shall improve upon simple variants of previous results in terms of complexity bounds.

5. APPROXIMATE NETS IN HIGH DIMENSIONS

We study r-nets, a powerful tool in computational and metric geometry, with several applications in approximation algorithms. We focus on the ℓ_2^d metric, in the high-dimensional regime. This chapter is essentially a simplified exposition of [19].

An r-net for a finite metric space (X, d), $|X| = n$ and for numerical parameter r is a subset $\mathcal{N} \subseteq X$ 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 X. We define approximate r-nets analogously (see Definition 20). We restate the definition for the special case of finite subsets of ℓ_2^d.

Definition 57. *Given a pointset $X \subseteq \mathbb{R}^d$, a distance parameter $r \in \mathbb{R}$ and an approximation parameter $\epsilon > 0$, a $(1+\epsilon)r$-net of X is a subset $\mathcal{N} \subseteq X$ s.t. the following properties hold:*

1. *(packing) For every $p, q \in \mathcal{N}$, $p \neq q$, we have that $\|p-q\|_2 \geq r$.*
2. *(covering) For every $p \in X$, there exists a $q \in \mathcal{N}$ s.t. $\|p-q\|_2 \leq (1+\epsilon)r$.*

A simple reduction, which is also utilized in [5] and shares its main idea with results of Section 3.4 allows us to focus on the space $\{-1,1\}^{O(\log n/\epsilon^2)}$. The reduction is based on the randomized embedding described in Section 3.4 (but to a higher dimension) $f : X \mapsto \{0,1\}^{O(\log n/\epsilon^2)}$ such that with high probability the following holds: $\forall p, q \in X$, if $\|p-q\|_2 \leq r$ then $\|f(p)-f(q)\|_1 \leq r'$ and if $\|p-q\|_2 \geq (1+2\epsilon)r$ then $\|f(p)-f(q)\|_1 \geq (1+\epsilon)r'$. Moreover, $r' = 1/2 + O(\epsilon)$. Then, translating binary coordinates to sign coordinates is trivial.

Organization. Section 5.1 discusses the main results, and Section 5.2 shows implications.

5.1 Points in $\{-1,1\}^d$ under inner product

In this section, we resolve the problem of computing nets for subsets of $\{-1,1\}^d$. Using the fact that the Euclidean norms of all vectors in our new space are equal to d, we can define the new notion of ρ-nets with respect to their inner product.

Definition 58. *For any $X \subset \{-1,1\}^d$, an approximate ρ-net for $(X, \langle\cdot,\cdot\rangle)$, with additive approximation parameter $\epsilon > 0$, is a subset $C \subseteq X$ which satisfies the following properties:*

- *for any two $p \neq q \in C$, $\langle p, q\rangle < \rho$, and*
- *for any $x \in X$, there exists $p \in C$ s.t. $\langle x, p\rangle \geq \rho - \epsilon$.*

The algorithm follows the recipe of [77], later also explored in [5]. The main observation is that finding the correlations between points in $\{-1,1\}^d$ can be reduced to a polynomial multi-point evaluation problem, which can be solved by fast matrix multiplication. A high-level description follows.

High-level description of net algorithm.

- Compute part of the net greedily; the remaining set is "sparse".
- For suitable $\phi(\cdot)$ compute $f(X)$ and $f'(X)$ s.t.

$$\forall x, y \in X: \ \langle f(x), f'(y) \rangle \approx \phi(\langle x, y \rangle).$$

- Arbitrary partition of X: $P_1, \ldots, P_m$.
- For any $x \in X$:
  - For any part P_i:
    * compute

$$\sum_{y \in P_i} \langle f(x), f'(y) \rangle \approx \sum_{y \in P_i} \phi(\langle x, y \rangle) \approx \bigvee_{y \in P_i} [\langle x, y \rangle \geq d/2 + \epsilon d] \, .$$

    * decide: is x correlated with some vector in P_i?

We need $\phi(\cdot)$ s.t. $\frac{\phi(d/2+\epsilon d)}{\phi(d/2)}$ as large as possible. To that end, we use the Chebyshev polynomial which is known to satisfy nice threshold properties.

Definition 59 (Chebyshev Polynomials)**.** *An explicit expression for the qth Chebyshev polynomial of the first kind is the following:*

$$T_q(x) = \sum_{k=0}^{\lfloor q/2 \rfloor} \binom{q}{2k} (x^2 - 1)^k x^{q-2k}.$$

Fact 60. *Let $T_q(x)$ denote the qth Chebyshev polynomial of the first kind, then the following hold:*

- *The leading coefficient $= 2^{q-1}$.*
- *All roots of $T_q(x)$ are real and within $[-1, 1]$.*
- *For $x \in [-1, 1]$, $|T_q(x)| \leq 1$.*
- *For $\delta \in (0, 1/2]$, $T_q(1 + \delta) \geq \frac{1}{2} e^{q\sqrt{\delta}}$.*

Valiant's result [77] includes a double randomized embedding $f, f' : \{-1, 1\}^d \mapsto \{-1, 1\}^{d'}$ which aims for the following property: $\langle f(x), f'(y) \rangle \approx T_q(\langle x, y \rangle)$. We refer to this algorithm as `Chebyshev Embedding` and state the formal guarantees associated with it in the following theorem.

Theorem 61 ([77]). *Let $Y, Y' \in \{-1,1\}^{d' \times n}$ be the matrices output by algorithm `Chebyshev Embedding` on input $X \in \{-1,1\}^{d \times n}$, integers q, d'. With probability $1 - o(1)$ over the randomness in the construction of Y, Y', for all $i, j \in [n]$,*

$$\langle Y_i, Y_j' \rangle \in T_q\left(2\frac{\langle X_i, X_j \rangle}{d}\right) \cdot d' \cdot 2^{-3q+1} \pm \sqrt{d'} \log n$$

where T_q is the degree-q Chebyshev polynomial of the first kind. The algorithm runs in time $O(d' \cdot n \cdot q)$.

Corollary 62. *Let $Y = [y_1, \ldots, y_n]$, $Y' = [y_1', \ldots, y_n']$ be the matrices output by algorithm "Chebyshev Embedding" on input $X \in \{-1,1\}^{d \times n}$, $q = \log\log n$, $d' = \log^9 n$. With probability $1 - o(1)$, for all pairs i, j, the following holds:*

- $\langle x_i, x_j \rangle \in [-d/2, d/2] \implies |\langle y_i, y_j' \rangle| \leq 10 \log^6 n$,
- $\langle x_i, x_j \rangle \geq d/2 + \epsilon d \implies \langle y_i, y_j' \rangle \geq \left(0.1 \cdot \log^{\sqrt{\epsilon}} n\right) \cdot \log^6 n$.

Lemma 63. *Let $Y, Y' \in \{-1,1\}^{d' \times n}$ be the output of the algorithm in Corollary 62. Consider set of indices $J \subset [n]$ and the d'-variate polynomial $F_J(y) = \sum_{j \in J} \langle y, y_j' \rangle^q$ of degree q. Set $q = 0.1 \cdot \frac{\log n}{\log d'} = 0.1 \cdot \frac{\log n}{9 \log\log n}$ assuming q is even. Then, there exists an $\alpha = n^{O(1)}$ such that,*

- $\forall j \in J : |\langle y, y_j' \rangle| \leq 10 \log^6 n \implies F_J(y) \leq |J| \cdot \alpha$
- $\exists j \in J : |\langle y, y_j' \rangle| \geq \left(0.1 \cdot \log^{\sqrt{\epsilon}} n\right) \cdot \log^6 n \implies F_J(y) \geq \alpha \cdot n^{\sqrt{\epsilon}/100}$*, for large enough n.*

Proof. The statement holds by a simple calculation on the bounds derived by Corollary 62. □

Hence, we can partition $[n]$ (equivalently input set X) into $n^{1-\sqrt{\epsilon}/100}$ parts which correspond to $n^{1-\sqrt{\epsilon}/100}$ polynomials. Each polynomial has $\leq n^{0.1}$ monomials.

To evaluate the $n^{1-\sqrt{\epsilon}/100}$ polynomials, we employ fast rectangular matrix multiplication.

Theorem 64 (Coppersmith '97). *For any positive $\gamma > 0$, provided that $\beta < 0.29$, the product of a $k \times k^\beta$ with a $k^\beta \times k$ matrix can be computed in time $O(k^{2+\gamma})$.*

Theorem 65. *Let $X \subseteq \{-1,1\}^d$, $|X| = n$, $\epsilon > 0$, and assume that $|x, y \in X \mid \langle x, y \rangle \geq \rho| \leq t$, where $\rho = 1/2 + \Theta(\epsilon)$. We can compute a (ρ, ϵ)-approximate net, as defined in Definition 58, in time $n^{2-O(\sqrt{\epsilon})} + dtn^{O(\sqrt{\epsilon})}$. The algorithm succeeds with probability $1 - o(1)$.*

Proof. We need to multiply a $n^{1-\sqrt{\epsilon}/100} \times n^{0.1}$ matrix with a $n^{0.1} \times n$ matrix. Equivalently, we perform $n^{\sqrt{\epsilon}/100}$ fast rectangular matrix multiplications in time:

$$n^{\sqrt{\epsilon}/100} \cdot n^{(1-\sqrt{\epsilon}/100)\cdot(2+\gamma)} \leq n^{2-\sqrt{\epsilon}/100+\gamma} \leq n^{2-\sqrt{\epsilon}/200},$$

by setting γ to be a sufficiently small multiple of $\sqrt{\epsilon}$. Then, there are at most t "heavy" elements, each one corresponding to $n^{O(\sqrt{\epsilon})}$ points: we visit all of them in a bruteforce manner. □

Theorem 66. *Let $X \subseteq \{-1,1\}^d$, $|X| = n$, $\epsilon > 0$. We can compute a (ρ, ϵ)-approximate net, as defined in Definition 58, in time $n^{2-O(\sqrt{\epsilon})} + dn^{1.5+O(\sqrt{\epsilon})}$.*

Proof. The complete algorithm consists of a first step which aims to compute a subset of the net greedily. The remaining set of uncovered points has the desired property that it is "sparse".

Repeat $n^{0.5}$ times:

- Choose a column x_i uniformly at random.
- $C \leftarrow C \cup \{x_i\}$.
- Delete column i from matrix X.
- Delete each column k from matrix X s.t. $|\langle x_i, x_k\rangle| \geq \rho$.

We perform $n^{0.5}$ iterations and for each, we compare the inner products between the randomly chosen vector and all other vectors. Hence, the time needed is $O(dn^{1.5})$.

In the following, we denote by X_i the number of vectors which have "large" magnitude of the inner product with the randomly chosen point in the ith iteration. Towards proving correctness, suppose first that $\mathbb{E}[X_i] > 2n^{0.5}$ for all $i = 1, \ldots n^{0.5}$. The expected number of vectors we delete in each iteration of the algorithm is more than $2n^{0.5} + 1$. So, after $n^{0.5}$ iterations, the expected total number of deleted vectors will be greater than n. This means that if the hypothesis holds for all iterations we will end up with a proper net.

Finally, the proof is complete after invoking Theorem 65. □

5.2 Applications and Future work

The main result of Section 5.1 is an algorithm for computing approximate r-nets in high dimensions. Another set of particular interest, is the set of "far" points, that is points which do not have any neighbor at distance $\leq r$. This is obviously a subset of any r-net. We remark that throughout the execution of the algorithm described in Section 5.1, we can mark points which are approximately far. We denote this modified algorithm by `DelFar` with input set X, radius parameter r, and approximation parameter $\epsilon > 0$. This algorithm outputs $X \setminus S$, for a set S such that,

$$\{x \in X \mid \forall y \in X \|y - x\| \geq (1+\epsilon)r\} \subseteq S \subseteq \{x \in X \mid \forall y \in X \|y - x\| \geq r\}.$$

In [54], they design an approximation scheme, which solves various distance optimization problems. Their algorithm works by randomly sampling a point and computing the distance to its nearest neighbor. Let this distance be r. Then they rely on the existence of an efficient decider for the problem: assuming that r is not a good guess, then if r is too small then an r-net is computed, and if r is too large then `DelFar` is computed. In both cases, the computation proceeds with a subset of the initial set and selects a new random value for r.

We apply our algorithms to the problem of approximating the *kth nearest neighbor distance*.

Definition 67. *Let $X \subset \mathbb{R}^d$ be a set of n points, approximation error $\epsilon > 0$, and let $d_1 \leq \ldots \leq d_n$ be the nearest neighbor distances. The problem of computing an $(1+\epsilon)$-approximation to the kth nearest neighbor distance asks for a pair $x, y \in X$ such that $\|x - y\| \in [(1-\epsilon)d_k, (1+\epsilon)d_k]$.*

Now we present an approximate decider for the problem above. This procedure combined with the framework of [54], results in an efficient solution for this problem in high dimension.

`kth NND Decider`
Input: $X \subseteq \mathbb{R}^d$, constant $\epsilon \in (0, 1/2]$, integer $k > 0$.
Output: An interval for the optimal value $f(X, k)$.

- Call $\texttt{DelFar}(X, \frac{r}{1+\epsilon/4}, \epsilon/4)$ and store its output in W_1.
- Call $\texttt{DelFar}(X, r, \epsilon/4)$ and store its output in W_2.
- Do one of the following:
  - If $|W_1| > k$, then output "$f(X, k) < r$".
  - If $|W_2| < k$, then output "$f(X, k) > r$".
  - If $|W_1| \leq k$ and $|W_2| \geq k$, then output "$f(X, k) \in [\frac{r}{1+\epsilon/4}, \frac{1+\epsilon/4}{r}]$".

Theorem 68 ([19] Theorem 4.1). *Given a pointset $X \subseteq \mathbb{R}^d$, one can compute a $(1+\epsilon)$-approximation to the k-th nearest neighbor in $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$, with probability $1 - o(1)$.*

To the best of our knowledge, this is the best high dimensional solution for this problem, when ϵ is sufficiently small. Setting $k = n$ and applying Theorem 68 one can compute the *farthest nearest neighbor* in $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})})$ with high probability.

Concerning future work, let us start with the problem of finding a greedy permutation. A permutation $\Pi =< \pi_1, \pi_2, \cdots >$ of the vertices of a metric space $(X, \|\cdot\|)$ is a *greedy permutation* if each vertex π_i is the farthest in X from the preceding vertices $\Pi_{i-1} =< \pi_1, \ldots, \pi_{i-1} >$. The computation of r-nets is closely related to that of the greedy permutation.

The k-center clustering problem asks the following: given a set $X \subseteq \mathbb{R}^d$ and an integer k, find the smallest radius r such that X is contained within k balls of radius r. Our algorithm can be plugged into the framework of [54] to achieve a $(4+\epsilon)$ approximation for the k-center problem in time $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})}$. By [42], a simple modification of our net construction implies an algorithm for the $(1+\epsilon)$ approximate greedy permutation in time $\tilde{O}(dn^{2-\Theta(\sqrt{\epsilon})}\log\Phi)$ where Φ denotes the spread of the pointset. Then, approximating the greedy permutation implies a $(2+\epsilon)$ approximation algorithm for k-center clustering problem. We expect that one can avoid any dependencies on Φ.

6. APPROXIMATE NEAREST NEIGHBORS FOR POLYGONAL CURVES

Our first contribution is a simple data structure for the $(1+\epsilon)$-ANN problem in ℓ_p-products of finite subsets of ℓ_2^d, for any constant p. The key ingredient is a random projection from points in ℓ_2 to points in ℓ_p. Although this has proven a relevant approach for $(1+\epsilon)$-ANN of pointsets, it is quite unusual to employ randomized embeddings from ℓ_2 to ℓ_p, $p > 2$, because such norms are considered "harder" than ℓ_2 in the context of proximity searching. After the random projection, the algorithm "vectorizes" all point sequences. The original problem is then translated to the $(1+\epsilon)$-ANN problem for points in $\ell_p^{d'}$, for $d' \approx d \cdot m$ to be specified later, and can be solved by simple bucketing methods in space $\tilde{O}\left(d'n \cdot (1/\epsilon)^{d'}\right)$ and query time $\tilde{O}(d' \log n)$, which is very efficient when $d \cdot m$ is low.

Then, we present a notion of distance between two polygonal curves, which generalizes both DFD and DTW (for a formal definition see Definition 5). 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 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, where p is sufficiently large. 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. While a reduction to the near neighbor problem has provable guarantees on metrics [51], we are not aware of an analogous result for non-metric distances such as the DTW.

Specifically, when $p > 2$, there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot \left(\frac{d}{p\epsilon} + 2\right)^{O(dm \cdot \alpha_{p,\epsilon})}\right),$$

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}(2^{4m} \log n)$.

When specialized to DFD and compared to [37], the two methods are only comparable when ϵ is a large enough fixed constant. Indeed, the two space and preprocessing time complexity bounds are equivalent, i.e. they are both exponential in d and m, but our query time is linear instead of being exponential in d.

When $p \in [1, 2]$, there exists a data structure with space and preprocessing time in

$$\tilde{O}\left(n \cdot 2^{O(dm \cdot \alpha_{p,\epsilon})}\right),$$

where $\alpha_{p,\epsilon}$ depends only on p, ϵ, and query time in $\tilde{O}\left(2^{4m} \log n\right)$. This leads to the first approach that achieves $1+\epsilon$ approximation for DTW at the expense of space, preprocessing

Table 6.1: Summary of previous results compared to this chapter's. The result of [55] holds for arbitrary metrics and X denotes the domain set of the input metric. All results except [55] are randomized. All previous results are tuned to optimize the approximation factor. The parameters ρ_u, ρ_q satisfy $(1+\epsilon)\sqrt{\rho_q}+\epsilon\sqrt{\rho_u} \geq \sqrt{1+2\epsilon}$.

	Space	Query	Approx.	Comments
DFD	$O(m^2\|X\|)^{m^{1-o(1)}} \times O(n^{2-o(1)})$	$(m \log n)^{O(1)}$	$O(1)$	det. [55]
	$\tilde{O}(2^{4md}n)$	$\tilde{O}(2^{4md} \log n)$	$O(d^{3/2})$	ℓ_2^d [37]
	$\tilde{O}(n) \times \left(\frac{d}{\log m}+2\right)^{O(dm^{1+1/\epsilon} \log(1/\epsilon))}$	$\tilde{O}(dm^{1+1/\epsilon} \cdot 2^{4m} \log n)$	$1+\epsilon$	ℓ_2^d, Thm 74
DTW	$\tilde{O}(mn)$	$O(m \log n)$	$O(m)$	ℓ_2^d [37]
	$\tilde{O}(n) \times 2^{O(m \cdot d \log(1/\epsilon))}$	$\tilde{O}(d \cdot 2^{4m} \log n)$	$1+\epsilon$	ℓ_2^d, Thm 75
	$\tilde{O}\left(2^{4m}n^{1+\rho_u}\right)$	$\tilde{O}\left(2^{4m}n^{\rho_q}\right)$	$1+\epsilon$	ℓ_2^d, Thm 76

and query time complexities being exponential in m. Hence our method is best suited when the curve size is small.

Our results for DTW and DFD are summarized in Table 6.1 and juxtaposed to existing approaches in [37, 55].

Organization. The rest of this chapter is structured as follows. In Section 6.1, we present a data structure for $(1+\epsilon)$-ANN in ℓ_p-products of ℓ_2, which is of independent interest. In Section 6.2, we employ this result to address the ℓ_p-distance of curves. We conclude with future work.

6.1 ℓ_p-products of ℓ_2

In this section, we present a simple data structure for $(1+\epsilon)$-ANN in ℓ_p-products of finite subsets of ℓ_2. Recall that the ℓ_p-product of $X_1, \ldots, X_m$, which are finite subsets of ℓ_2, is a metric space with ground set $X_1 \times X_2 \times \cdots \times X_m$ and distance function:

$$\mathrm{d}((x_1,\ldots,x_m),(y_1,\ldots,y_m)) = \| \|x_1-y_1\|_2, \ldots, \|x_m-y_m\|_2 \|_p = \left(\sum_{i=1}^{m} \|x_i-y_i\|_2^p\right)^{1/p}.$$

For $(1+\epsilon)$-ANN, the algorithm first randomly embeds points from ℓ_2 to ℓ_p. For this purpose, we build upon results which are probably folklore and the reasoning is quite similar to the one followed by proofs of the Johnson-Lindenstrauss lemma, e.g. [67]. Then, it is easy to translate the original problem to $(1+\epsilon)$-ANN in ℓ_p for large vectors corresponding to point sequences.

We now present our main results concerning $(1+\epsilon)$-ANN for ℓ_p-products of ℓ_2. First, we show that a simple random projection maps points from ℓ_2^d to $\ell_p^{d'}$, where $d' = \tilde{O}(d)$, without arbitrarily contracting norms. The probability of failure decays exponentially with d'. For our purposes, there is no need for an almost isometry between norms. Hence, our efforts focus on proving lower tail inequalities which imply that, with good probability, no far neighbor corresponds to an approximate nearest neighbor in the projected space.

We now prove bounds concerning the contraction of distances of the embedded points. Our proof builds upon the inequalities developed in Section 2.2.

Theorem 69. *Let G be a $d' \times d$ matrix with i.i.d. random variables following $N(0,1)$. Then,*

- *if $2 < p < \infty$ then,*

$$\Pr\left[\exists v \in \mathbb{R}^d : \ \|Gv\|_p \leq \frac{(c_p \cdot d')^{1/p}}{1+\epsilon} \cdot \|v\|_2\right] \leq O\left(\frac{d'^{\frac{1}{2}-\frac{1}{p}}}{p\epsilon} + 2\right)^d \cdot \mathrm{e}^{-c' \cdot 2^{-p} \cdot d' \cdot (p\epsilon/(2+p\epsilon))^2},$$

- *if $p \in [1,2]$ then,*

$$\Pr\left[\exists v \in \mathbb{R}^d : \ \|Gv\|_p \leq \frac{(c_p \cdot d')^{1/p}}{1+\epsilon} \cdot \|v\|_2\right] \leq O\left(\frac{1}{\epsilon}\right)^d \cdot \mathrm{e}^{-c' \cdot d' \cdot (p\epsilon/(2+p\epsilon))^2},$$

where $c' > 1$ is a constant, $\epsilon \in (0, 1/2)$.

Proof. By Lemma 16:

$$\Pr\left[\|Gv\|_p^p \leq \frac{c_p \cdot d'}{(1+\epsilon)^p} \cdot \|v\|_2^p\right] \leq \Pr\left[\|Gv\|_p^p \leq \frac{c_p \cdot d'}{1+p\epsilon/2} \cdot \|v\|_2^p\right] \leq \mathrm{e}^{-c' \cdot d' \cdot (p\epsilon/(2+p\epsilon))^2}.$$

In order to bound the probability of contraction among all distances, we argue that it suffices to use the strong bound on distance contraction, which is derived in Lemma 16, and the weak bound on distance expansion from Corollary 17 or Lemma 18, for a δ-dense set $N \subset \mathbb{S}^{d-1}$ for δ to be specified later. First, a simple volumetric argument [51] shows that there exists $N \subset \mathbb{S}^{d-1}$ s.t. $\forall x \in \mathbb{S}^{d-1}\ \exists y \in N\ \|x-y\|_2 \leq \delta$, and $|N| = O\left(1/\delta\right)^d$.

We first consider the case $p > 2$. From now on, we assume that for any $u \in N$, $\|Gu\|_p \geq (c_p \cdot d')^{1/p}/(1+\epsilon)$ and $\|Gu\|_p \leq 2\sqrt{d'}$ which is achieved with probability

$$\geq 1 - O\left(\frac{1}{\delta}\right)^d \cdot \mathrm{e}^{-c' \cdot 2^{-p} \cdot d' \cdot (p\epsilon/(2+p\epsilon))^2}.$$

Now let x be an arbitrary vector in $\mathbb{R}^d$ s.t. $\|x\|_2 = 1$. Then, there exists $u \in N$ s.t. $\|x-u\|_2 \leq \delta$. Also, by the triangular inequality we obtain the following,

$$\|Gx\|_p \leq \|Gu\|_p + \|G(x-u)\|_p = \|Gu\|_p + \|x-u\|_2 \left\|G\frac{(x-u)}{\|x-u\|_2}\right\|_p \leq \|Gu\|_p + \delta \left\|G\frac{(x-u)}{\|x-u\|_2}\right\|_p. \tag{6.1}$$

Let $M = \max_{x \in \mathbb{S}^{d-1}} \|Gx\|_p$. The existence of M is implied by the fact that $\mathbb{S}^{d-1}$ is compact and $x \mapsto \|x\|_p$, $x \mapsto Gx$ are continuous functions. Then, by plugging M into (6.1),

$$M \leq \|Gu\|_p + \delta M \implies M \leq \frac{\|Gu\|_p}{1-\delta} \leq \frac{2\sqrt{d'}}{1-\delta},$$

where the last inequality is implied by Corollary 17. Again, by the triangular inequality,

$$\|Gx\|_p \geq \|G_u\|_p - \|G(x-u)\|_p \geq \frac{(c_p \cdot d')^{1/p}}{1+\epsilon} - \frac{2\delta\sqrt{d'}}{1-\delta} \geq \frac{1-\epsilon/2}{1+\epsilon} \cdot (c_p \cdot d')^{1/p},$$

for $\delta \leq \frac{\epsilon \cdot (c_p \cdot d')^{1/p}}{2\sqrt{d'} + \epsilon \cdot (c_p \cdot d')^{1/p}}$.

Notice now that

$$\frac{1}{\delta} = O\left(\frac{d'^{1/2-1/p}}{p\epsilon}\right) + 1.$$

In the case $p \in [1,2]$, we are able to use a better bound on the distance expansion; namely Lemma 18. We now assume that for any $u \in N$, $\|Gu\|_p \geq (c_p \cdot d')^{1/p}/(1+\epsilon)$ and $\|Gu\|_p \leq (3 \cdot c_p \cdot d')^{1/p}$ which is achieved with probability

$$\geq 1 - O\left(\frac{1}{\delta}\right)^d \cdot \mathrm{e}^{-c' \cdot d' \cdot (p\epsilon/(2+p\epsilon))^2}.$$

Once again, we use inequality (6.1) to obtain:

$$M \leq \frac{\|Gu\|_p}{1-\delta} \leq \frac{(3 \cdot c_p \cdot d')^{1/p}}{1-\delta} \implies$$

$$\implies \|Gx\|_p \geq \|Gu\|_p - \|Gx - Gu\|p \geq (c_p \cdot d')^{1/p}\left(\frac{1}{1+\epsilon} - \frac{3^{1/p} \cdot \delta}{1-\delta}\right) \implies$$

$$\implies \|Gx\|_p \geq (c_p \cdot d')^{1/p} \cdot \frac{1-\epsilon/2}{1+\epsilon},$$

for $\delta \leq \epsilon/(6(1+\epsilon)+\epsilon) = \Omega(\epsilon)$. □

Theorem 69 implies that the $(1+\epsilon)$-ANN problem for ℓ_p products of ℓ_2 translates to the $(1+\epsilon)$-ANN problem for ℓ_p products of ℓ_p. The latter easily translates to the $(1+\epsilon)$-ANN problem in $\ell_p^{d'}$. One can then solve the approximate near neighbor problem in $\ell_p^{d'}$, by approximating $\ell_p^{d'}$ balls of radius 1 with a regular grid with side length $\epsilon/(d')^{1/p}$. Each approximate ball is essentially a set of $O(1/\epsilon)^{d'}$ cells [51]. Building not-so-many approximate near neighbor data structures for various radii leads to an efficient solution for the $(1+\epsilon)$-ANN problem [51].

Theorem 70. *There exists a data structure which solves the $(1+\epsilon)$-ANN problem for point sequences in ℓ_p-products of ℓ_2, and satisfies the following bounds on performance:*

- *If $p \in [1,2]$, then space usage and preprocessing time is in*

$$\tilde{O}(dmn) \times \left(\frac{1}{\epsilon}\right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

query time is in $\tilde{O}(dm \log n)$, and $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2+p\epsilon)^2 \cdot (p\epsilon)^{-2}$.

- *If $2 < p < \infty$, then space usage and preprocessing time is in*

$$\tilde{O}(dmn) \times \left(\frac{d}{p\epsilon} + 2\right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

query time is in $\tilde{O}\left(dm \cdot 2^p \log n\right)$, and $\alpha_{p,\epsilon} = 2^p \cdot \log(1/\epsilon) \cdot (2+p\epsilon)^2 \cdot (p\epsilon)^{-2}$.

We assume $\epsilon \in (0, 1/2]$. The probability of success is $\Omega(\epsilon)$ and can be amplified to $1-\delta$, by building $\Omega(\log(1/\delta)/\epsilon)$ independent copies of the data-structure.

Proof. Let $\delta_{p,\epsilon} = p\epsilon/(2+p\epsilon)$. We first consider the case $p > 2$. We employ Theorem 69 and we map point sequences to point sequences in $\ell_p^{d'}$, for

$$d' = \Theta\left(\frac{d \cdot 2^p \cdot \log \frac{d}{p\epsilon}}{\delta_{p,\epsilon}^2}\right).$$

Hence, Theorem 69 implies that,

$$\Pr\left[\exists v \in \mathbb{R}^d : \ \|Gv\|_p \leq \frac{(c_p \cdot d')^{1/p}}{1+\epsilon} \cdot \|v\|_2\right] \leq \epsilon/10.$$

Then, by concatenating vectors, we map point sequences to points in $\ell_p^{d'm}$.

Now, fix query point sequence $Q = q_1, \ldots, q_m \in \left(\mathbb{R}^d\right)^m$ and its nearest neighbor $U_* = u_1, \ldots, u_m \in \left(\mathbb{R}^d\right)^m$. By a union bound, the probability of failure for the embedding is at most

$$\Pr\left[\exists v \in \mathbb{R}^d : \|Gv\|_p \leq \frac{(c_p \cdot d')^{1/p}}{1+\epsilon}\|v\|_2\right] + \Pr\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p \leq (1+\epsilon)^p c_p d' \sum_{i=1}^m \|u_i - q_i\|_2^p\right].$$

We know that the first probability is $\leq \epsilon/2$. Hence, we now bound the second probability. Notice that

$$\mathbb{E}\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p\right] = \sum_{i=1}^m \mathbb{E}\left[\|G(u_i - q_i)\|_p^p\right] = c_p \cdot d' \sum_{i=1}^m \|u_i - q_i\|_2^p.$$

By Markov's inequality, we obtain,

$$\Pr\left[\sum_{i=1}^m \|Gu_i - Gq_i\|_p^p \leq (1+\epsilon)^p \cdot c_p \cdot d' \sum_{i=1}^m \|u_i - q_i\|_2^p\right] \leq (1+\epsilon)^{-p}.$$

Hence, the total probability of failure is $\frac{1+\epsilon/10}{(1+\epsilon)^p}$. In the projected space, we build AVDs[51]. The total space usage, and the preprocessing time is

$$\tilde{O}(dmn) \times O(1/\epsilon)^{d'm} = \tilde{O}(dmn) \times \left(\frac{d}{p\epsilon} + 2\right)^{O(m \cdot d \cdot 2^p \cdot \log(1/\epsilon)/\delta^2_{p,\epsilon})}.$$

The query time is $\tilde{O}(dm2^p \log n)$. The probability of success can be amplified by repetition. By building $\Theta\left(\frac{\log(1/\delta)}{\epsilon}\right)$ data structures as above, the probability of failure becomes δ.

The same reasoning is valid in the case $p \in [1,2]$, but it suffices to set

$$d' = \Theta\left(\frac{d \log \frac{1}{\epsilon}}{\delta^2_{p,\epsilon}}\right).$$

□

When $p \in [1,2]$, we can also utilize "high-dimensional" solutions for ℓ_p and obtain data structures with complexities polynomial in $d \cdot m$. Combining Theorem 69 with the data structure of [11], we obtain the following result.

Theorem 71. *There exists a data structure which solves the $(1+\epsilon)$-ANN problem for point sequences in ℓ_p-products of ℓ_2, $p \in [1,2]$, and satisfies the following bounds on performance: space usage and preprocessing time is in $\tilde{O}(n^{1+\rho_u})$, and the query time is in $\tilde{O}(n^{\rho_q})$, where ρ_q, ρ_u satisfy:*

$$(1+\epsilon)^p \sqrt{\rho_q} + ((1+\epsilon)^p - 1)\sqrt{\rho_u} \geq \sqrt{2(1+\epsilon)^p - 1}$$

We assume $\epsilon \in (0, 1/2]$. The probability of success is $\Omega(\epsilon)$ and can be amplified to $1-\delta$, by building $\Omega(\log(1/\delta)/\epsilon)$ independent copies of the data-structure.

Proof. We proceed as in the proof of Theorem 70. We employ Theorem 69 and by Markov's inequality, we obtain,

$$\Pr\left[\sum_{i=1}^{m} \|Gv_i - Gu_i\|_p^p \leq (1+\epsilon)^p \cdot c_p \cdot d' \sum_{i=1}^{m} \|v_i - u_i\|_2^p\right] \leq (1+\epsilon)^{-p}.$$

Then, by concatenating vectors, we map point sequences to points in $\ell_p^{d'm}$, where $d' = \tilde{O}(d)$. For the mapped points in $\ell_p^{d'm}$, we build the LSH-based data structure from [11] which succeeds with high probability $1-o(1)$. By independence, both the random projection and the LSH-based structure succeed with probability $\Omega(\epsilon) \times (1-o(1)) = \Omega(\epsilon)$. □

6.2 Polygonal Curves

In this section, we show that one can solve the $(1+\epsilon)$-ANN problem for the class of ℓ_p-distance functions defined on polygonal curves, as in Definition 5. Since this class is related to ℓ_p-products of ℓ_2, we invoke results of Section 6.1, and we show an efficient data structure for the case of short curves, i.e. when m is relatively small compared to the other complexity parameters.

The class of ℓ_p-distances for polygonal curves includes some widely known distance functions. For instance, $d_\infty(V,U)$ coincides with the DFD of V and U (defined for the Euclidean distance). Moreover $d_1(V,U)$ coincides with DTW for curves V, U.

Theorem 72. *Suppose that there exists a randomized data structure for the $(1+\epsilon)$-ANN problem in ℓ_p products of ℓ_2, with space in $S(n)$, preprocessing time $T(n)$ and query time $Q(n)$, with probability of failure less than 2^{-4m-1}. Then, there exists a data structure for the $(1+\epsilon)$-ANN problem for the ℓ_p-distance of polygonal curves, $1 \leq p < \infty$, with space in $m \cdot (4\mathrm{e})^{m+1} \cdot S(n)$, preprocessing time $(4\mathrm{e})^{m+1} \cdot T(n)$ and query time $(4\mathrm{e})^{m+1} \cdot Q(n)$, where m denotes the maximum length of a polygonal curve, and the probability of failure is less than $1/2$.*

Proof. We denote by X the input dataset. Given polygonal curves $V = v_1, \ldots, v_{m_1}$, $Q = q_1, \ldots, q_{m_2}$, and traversal T, one can define $V_T = v_1, \ldots, v_l$, $Q_T = q_1, \ldots, q_l$, sequences of l points (allowing consecutive duplicates) s.t. $\forall k$, $v_{i_k} = V_T[k]$ and $q_{j_k} = Q_T[k]$, if and only if $(i_k, j_k) \in T$.

One traversal of V, Q is uniquely defined by its length $l \in \{\max(m_1, m_2), \ldots, m_1 + m_2\}$, the set of indices $A = \{k \in \{1, \ldots, l\} \mid i_{k+1} - i_k = 0$ and $j_{k+1} - j_k = 1\}$ for which only Q is progressing and the set of indices $B = \{k \in \{1, \ldots, l\} \mid i_{k+1} - i_k = 1$ and $j_{k+1} - j_k = 1\}$ for which both Q and V are progressing. We can now define $V_{l,A,B}$, $Q_{l,A,B}$ to be the corresponding sequences of l points. In other words if l, A, B corresponds to traversal T, $V_{l,A,B} = V_T$, $Q_{l,A,B} = Q_T$. Observe that it is possible that curve V is not compatible with some triple l, A, B.

We build one $(1+\epsilon)$-ANN data structure, for ℓ_p products of ℓ_2, for each possible l, A, B. Each data structure contains at most $|X|$ point sequences which correspond to curves that are compatible to l, A, B. We denote by $m = \max(m_1, m_2)$. The total number of data structures is upper bounded by

$$\sum_{l=m}^{2m} \sum_{t=0}^{m} \binom{l}{t} \cdot \binom{l-t}{m-t} \leq \sum_{l=m}^{2m} \sum_{t=0}^{m} \binom{l}{t} \cdot \binom{l}{m-t} = \sum_{l=m}^{2m} \binom{2l}{m} \leq \sum_{l=m}^{4m} \binom{l}{m} = \binom{4m+1}{m+1} \leq$$

$\leq (4\mathrm{e})^{m+1}$. For any query curve Q, we create all possible combinations of l, A, B and we perform one query per $(1+\epsilon)$-ANN data structure. We report the best answer. The probability that the building of one of the $\leq (4\mathrm{e})^{m+1}$ data structures is not successful is less than $1/2$ due to a union bound. $\square$

We now investigate applications of the above results, to the $(1+\epsilon)$-ANN problem for some popular distance functions for curves.

Discrete Fréchet Distance. DFD is naturally included in the distance class of Definition 5 for $p = \infty$. However, Theorem 72 is valid only when p is bounded. To overcome this issue, p is set to a suitable large value.

Lemma 73. *Let $V = v_1, \ldots, v_{m_1} \in \mathbb{R}^d$ and $U = u_1, \ldots, u_{m_2} \in \mathbb{R}^d$ be two polygonal curves. Then for any traversal T of V and U:*

$$(1+\epsilon)^{-1} \cdot \left(\sum_{(i_k,j_k)\in T} \|v_{i_k} - u_{j_k}\|^p \right)^{1/p} \leq \max_{(i_k,j_k)\in T} \|v_{i_k} - u_{j_k}\| \leq \left(\sum_{(i_k,j_k)\in T} \|v_{i_k} - u_{j_k}\|^p \right)^{1/p},$$

for $p \geq \log(|T|) / \log(1+\epsilon)$.

Proof. For any $x \in \mathbb{R}^{|T|}$, it is known that $\|x\|_\infty \leq \|x\|_p \leq (|T|)^{1/p} \|x\|_\infty$. □

Theorem 74. *There exists a data structure for the $(1+\epsilon)$-ANN problem for the DFD of curves, with space and preprocessing time*

$$\tilde{O}(dm^2 n) \times \left(\frac{d}{\log m} + 2 \right)^{O(m^{1+1/\epsilon} \cdot d \cdot \log(1/\epsilon))},$$

and query time $\tilde{O}(dm^{1+1/\epsilon} \cdot 2^{4m} \log n)$, where m denotes the maximum length of a polygonal curve, and $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. We combine Theorem 72 with Theorem 70 for $p \geq \log m / \log(1+\epsilon) \geq \epsilon^{-1} \log m$. Notice that in order to plug the data structure of Theorem 70 into Theorem 72 we need to amplify the probability of success to $1 - 2^{-4m-1}$. Hence, the data structure for the $(1+\epsilon)$-ANN problem for ℓ_p-products of ℓ_p needs space and preprocessing time

$$\tilde{O}(dm^2 n) \times \left(\frac{d}{p\epsilon} + 2 \right)^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

and each query time costs $O(dm^2)$, where $\alpha_{p,\epsilon} = 2^p \cdot \log(1/\epsilon) \cdot (2 + p\epsilon)^2 \cdot (p\epsilon)^{-2}$. Now, substituting p and invoking Theorem 72 completes our proof.

□

Dynamic Time Warping. DTW corresponds to the ℓ_1-distance of polygonal curves as defined in Definition 5. Now, we combine Theorem 72 with each of the Theorems 70 and 71.

Theorem 75. *There exists a data structure for the $(1+\epsilon)$-ANN problem for DTW of curves, with space and preprocessing time*

$$\tilde{O}(dm^2 n) \times \left(\frac{1}{\epsilon} \right)^{O(m \cdot d \cdot \epsilon^{-2})},$$

and query time $\tilde{O}(d \cdot 2^{4m} \log n)$, where m denotes the maximum length of a polygonal curve, and $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. We first amplify the probability of success for the data structure of Theorem 70 to $1 - 2^{-4m-1}$. Hence, the data structure for the $(1+\epsilon)$-ANN problem for ℓ_1-products of ℓ_1 needs space and preprocessing time

$$\tilde{O}(dm^2 n) \times 2^{O(m \cdot d \cdot \alpha_{p,\epsilon})},$$

and each query time costs $O(dm^2)$, where $\alpha_{p,\epsilon} = \log(1/\epsilon) \cdot (2+\epsilon)^2 \cdot (\epsilon)^{-2}$. We plug this data structure into Theorem 72. □

Theorem 76. *There exists a data structure for the $(1+\epsilon)$-ANN problem for DTW of curves, with space and preprocessing time $\tilde{O}(2^{4m} n^{1+\rho_u})$, and the query time is in $\tilde{O}(2^{4m} n^{\rho_q})$, where ρ_q, ρ_u satisfy:*

$$(1+\epsilon)\sqrt{\rho_q} + \epsilon\sqrt{\rho_u} \geq \sqrt{1+2\epsilon}.$$

We assume $\epsilon \in (0, 1/2]$. The data structure succeeds with probability $1/2$, which can be amplified by repetition.

Proof. First amplify the probability of success for the data structure of Theorem 71 to $1 - 2^{-4m-1}$, by building independently $\tilde{O}(m)$ such data structures. We plug the resulting data structure into Theorem 72. □

6.3 Conclusion

Thanks to the simplicity of the approach, it should be easy to implement it and should have practical interest. We plan to apply it to real scenarios with data from road segments or time series.

The key ingredient of our approach is a randomized embedding from ℓ_2 to ℓ_p which is the first step to the $(1+\epsilon)$-ANN solution for ℓ_p-products of ℓ_2. The embedding is essentially a gaussian projection and it exploits the 2-stability property of normal variables, along with standard properties of their tails. We expect that a similar result can be achieved for ℓ_p-products of ℓ_q, where $q \in [1, 2)$. One related result for $(1+\epsilon)$-ANN [22], provides dimension reduction for ℓ_q, $q \in [1, 2)$.

7. APPROXIMATE NEAR NEIGHBORS FOR SHORT QUERY CURVES UNDER THE DISCRETE FRÉCHET DISTANCE

In this chapter, we study data structures for queries under the discrete Fréchet distance in the short queries regime. In this scenario, the dataset consists of polygonal curves of length at most m, but the queries are of length $k < m$. We base our solution on the $O(k)$-approximate data structure proposed by Driemel and Silvestri [38] and achieve a $(1+\epsilon)$-approximation with little computational overhead. Our main idea is to handle queries in two stages. After the input is snapped to a (coarse) randomly shifted grid, each bucket of the hash table is refined further using (finer) ϵ-grids. For the discrete Fréchet distance, the data structure improves upon our (more general) result of Chapter 6 even for the case $k = m$.

Finally, we show that our techniques generalize to variants of the discrete Fréchet distance that are derived from other metrics. When the underlying metric is a doubling metric, we can use net-trees instead of ϵ-grids. This incurs a slight increase in query time since we cannot simply snap the query to the grid and instead use a lookup table.

We use $\mathbb{X}_m^d = \left(\mathbb{R}^d\right)^m$ and treat the elements of this set as ordered sets of points in $\mathbb{R}^d$ of size m called polygonal curves. In the metric case, we assume a metric space $(\mathcal{M}^m, \mathrm{d}_m)$, write a curve p with m vertices as $p = p_1, \ldots, p_m$ and denote the space of all curves by $\mathcal{M}^m$. For any polygonal curve p, $V(p)$ denotes the set of its vertices.

Organization. In Section 7.1, we show our results for polygonal curves. In Section 7.2, we extend our ideas to metric spaces of bounded doubling dimension.

7.1 ANN for short query curves in Euclidean spaces

In this section, we present efficient data structures for the $(1+\epsilon, r)$-ANN problem, for polygonal curves under the discrete Fréchet distance d_{dF} in Euclidean spaces. We further assume that $r = 1$ since we can uniformly scale the ambient space.

Randomly shifted grids constitute the main ingredient of our algorithm. It has been previously observed [38] that randomly shifted grids induce a good partition of the space of curves: with good probability, near curves pass through the same sequence of cells and hence they belong to the same part. Let $\delta > 0$ and z chosen uniformly at random from the interval $[0, \delta]$. The function $h_{\delta,z}(x_i) = \lfloor \delta^{-1}(x_i - z) \rfloor$ induces a random partition of the line. Hence, for any vector $x = (x_1, \ldots, x_d)$, the function $g_{\delta,z}(x) = (h_{\delta,z}(x_1), ..., h_{\delta,z}(x_d))$, induces a randomly shifted grid. Notice that, for our purposes, it suffices to use the same random variable for all coordinates. It is easy to bound the probability that a set with bounded diameter is entirely contained in a cell.

For any set X, $diam(X)$ denotes the diameter of X. We begin with simple technical lemmas and then we proceed to our main theorems.

Lemma 77. *Let $X \subseteq \mathbb{R}^d$ be a set such that $diam(X) \leq \Delta$. Then,*

$$\mathsf{P}r_z\left[\exists x \in X\ \exists y \in X:\ g_{\delta,z}(x) \neq g_{\delta,z}(y)\right] \leq \frac{d\Delta}{\delta}.$$

Proof. Let $a, b \in \mathbb{R}$ such that $|a-b| \leq \Delta$. Then,

$$\mathsf{P}r_z\left[\left\lfloor \frac{a-z}{\delta} \right\rfloor \neq \left\lfloor \frac{b-z}{\delta} \right\rfloor\right] \leq \frac{\Delta}{\delta}.$$

Hence, by a union bound over all coordinates:

$$\mathsf{P}r_z\left[\exists x \in X\ \exists y \in X:\ g_{\delta,z}(x) \neq g_{\delta,z}(y)\right] \leq \frac{d\Delta}{\delta}.$$

□

The same argument extends to k sets of bounded diameter.

Lemma 78. *Let $X_1, \ldots, X_k \subseteq \mathbb{R}^d$ be k sets such that $\forall i \in [k]:\ diam(X_i) \leq \Delta$.*

$$\mathsf{P}r_z\left[\exists X_i\ \exists x \in X_i\ \exists y \in X_i:\ g_{\delta,z}(x) \neq g_{\delta,z}(y)\right] \leq \frac{dk\Delta}{\delta}.$$

Proof. The statement holds by Lemma 77 and a union bound over all sets. □

Lemma 79. *For any two curves $p \in \mathbb{X}_m^d$ and $q \in \mathbb{X}_k^d$, let $X_1^T, \ldots, X_l^T$ be a sequence of subsets of $V(p) \cup V(q)$, where X_i^T denotes the ith disconnected component of an optimal traversal T. If $\mathsf{d}_{dF}(p,q) \leq 1$, then for $\delta = 4dk$:*

$$\mathsf{P}r_z\left[\exists i \in [d]\ \exists x \in X_i\ \exists y \in X_i:\ g_{\delta,z}(x) \neq g_{\delta,z}(y)\right] \leq \frac{1}{2}.$$

Proof. Lemma 78, and the fact that for any $i \in [k]$ $diam(X_i^T) \leq 2$, imply the result. □

The following lemma indicates that the optimal traversal between two polygonal curves $p \in \mathbb{X}_m^d$ and $q \in \mathbb{X}_k^d$, $k \leq m$, can be viewed as a matching between $V(p)$ and $V(q)$.

Lemma 80 (Lemma 3 [38])**.** *For any two curves $p \in \mathbb{X}_{m_1}^d$ and $q \in \mathbb{X}_{m_2}^d$, there always exists an optimal traversal T with the following two properties:*

(i) T consists of at most $k = \min\{m_1, m_2\}$ disconnected components.

(ii) Each component is a star, i.e., all edges of this component share a common vertex.

Hence, by a union bound, we are able to bound the probability of splitting one of the k disconnected components with a random partition induced by a randomly shifted grid with side-length $\Theta(kd)$. Furthermore, we can precompute and store solutions for polygonal curves realized by the grid points of a refined grid of side-length $\Theta(\epsilon/\sqrt{d})$, and use these solutions to answer any query, after snapping its vertices to the grid.

Theorem 81. *Given as input a set of n polygonal curves $P \subset \mathbb{X}_m^d$, and an approximation parameter $\epsilon > 0$, there exists a randomized data structure with space in $n \cdot O\left(\frac{kd^{3/2}}{\epsilon}\right)^{kd} + O(dnm)$, preprocessing time in $dnmk \cdot O\left(\frac{kd^{3/2}}{\epsilon}\right)^{kd}$, and query time in $O(dk)$, for the $(1+\epsilon, r)$-ANN problem under the discrete Fréchet distance. For any query curve $q \in \mathbb{X}_k^d$, the preprocessing algorithm succeeds with constant probability.*

Proof. For any vector $x = (x_1, \ldots, x_d)$, we define the random function

$$g_{\delta,z}(x) = \left(\left\lfloor \frac{x_1 - z}{\delta} \right\rfloor, \ldots, \left\lfloor \frac{x_1 - z}{\delta} \right\rfloor \right),$$

where z is a random variable following the uniform distribution in $[0, \delta]$, and $\delta = 2dk$. We also define

$$g_{w,\cdot}(x) = \left(\left\lfloor \frac{x_1}{w} \right\rfloor, \ldots, \left\lfloor \frac{x_1}{w} \right\rfloor \right),$$

where $w = \epsilon/(2\sqrt{d})$. The preprocessing algorithm:

(a) Input: n polygonal curves $P \subset \mathbb{X}_m^d$.

(b) For each curve $p \in P$, assign a key vector $\in \mathbb{Z}^k$ which is defined by the sequence of cells induced by $g_{\delta,z}$, which are stabbed by p. The curves which stab more than k cells are not stored. If the number of stabbed cells is less than k, then for the last coordinates we use a special character indicating emptiness.

(c) Store curves in a hashtable: each bucket corresponds to a key vector (as described in (b)).

(d) Let $C_1, \ldots, C_t$ be the sequence of cells which corresponds to a given bucket: compute the solutions for all curves of complexity k which are defined by points in $g_{w,\cdot}(C_1), \ldots, g_{w,\cdot}(C_t)$ (and respect the ordering).

(e) Store the solutions (as indices) in a new hashtable: one new hashtable per bucket of (c). Any curve within distance $1 + \epsilon/2$ is considered an appropriate near neighbor.

The query algorithm:

(i) Input: query curve $q \in \mathbb{X}_k^d$.

(ii) Hash the curve twice: first by $g_{\delta,z}(\cdot)$, and then by $g_{w,\cdot}(\cdot)$. Report the answer.

Storage. We use perfect hashing to store the curves. There are at most n non-empty buckets which contain curves. For each such bucket, we precompute and store (approximate) answers for all possible queries. The number of possible queries which are compatible with a given sequence of k cells is upper bounded by:

$$\sum_{\substack{t_1+\ldots+t_k=k\\ \forall i:\ t_i\geq 0\\ t_1\geq 1, t_k\geq 1}} \prod_{i=1}^{k} \left(\frac{4d^{3/2}k}{\epsilon}\right)^{t_i d} \leq \sum_{\substack{t_1+\ldots+t_k=k\\ \forall i:\ t_i\geq 0}} \left(\frac{4d^{3/2}k}{\epsilon}\right)^{kd} = \binom{2k-1}{k} \cdot \left(\frac{4d^{3/2}k}{\epsilon}\right)^{kd} \leq \left(\frac{16d^{3/2}k}{\epsilon}\right)^{kd}.$$

Hence there are $n\cdot O(d^{3/2}k\epsilon^{-1})^{kd}$ indices to store. Indices refer to the input set of polygonal curves which are stored in $O(dnm)$.

Preprocessing time. For each data curve, we compute the real distance to all possible queries. Hence, the total preprocessing time is $dnmk \cdot O\left(\frac{kd^{3/2}}{\epsilon}\right)^{kd}$.

Query time. $O(kd)$ because of perfect hashing.

Correctness. By Lemma 79, we have that if $\mathrm{d}_{dF}(p,q)\leq 1$, then p,q lie at the same bucket with probability $\geq 1/2$. Now, let any two points $x,y\in\mathbb{R}^d$, and let x' be the image of x in $G_{\epsilon/2\sqrt{d}}$. If $\|x-y\|_2\leq 1$, then $\|x'-y\|_2 \leq \|x-x'\|_2+\|x-y\|_2 \leq 1+\epsilon/2$. Similarly, If $\|x-y\|_2 > 1+\epsilon$ then $\|x-y\|_2 > 1+\epsilon/2$. □

One may notice that the above data structure requires limited randomness. In fact, there is only one random variable which is used for the randomly shifted grid. As a consequence, the data structure can be easily derandomized.

Theorem 82. *Given as input a set of n polygonal curves $P\subset\mathbb{X}_m^d$, and an approximation parameter $\epsilon>0$, there exists a deterministic data structure with space in $O\left(dnm\right)+\left(d^{3/2}nk\epsilon^{-1}\right)\times O\left(\frac{kd^{3/2}}{\epsilon}\right)^{kd}$, preprocessing time in $O\left(d^{5/2}nmk\epsilon^{-1}\right)\times O\left(\frac{kd^{3/2}}{\epsilon}\right)^{kd}$, and query time in $O\left(\frac{k^2d^{5/2}}{\epsilon}\right)$, for the $(1+\epsilon,r)$-ANN problem under the discrete Fréchet distance, for query curves in $\mathbb{X}_k^d$.*

Proof. The data structure is essentially a derandomized version of the data structure of Theorem 81. First we snap all points to a grid with side-length $\Theta(\epsilon/\sqrt{d})$. This introduces an additive error of $\Theta(\epsilon)$. Then, instead of applying a randomly shifted grid, we build several shifted grids; one for each interesting value of z. After having discretized the coordinates, there are $O(d^{3/2}k/\epsilon)$ such values. □

7.2 ANN for short query curves in doubling spaces

In this section, we consider an arbitrary metric space $(\mathcal{M},\mathrm{d}_{\mathcal{M}})$. We assume the existence of a constant-time oracle that gives us access to the metric space. We refer to the two computational models relevant for our work as follows:

- *black-box model* ([30, 53, 63]): there exists a constant-time *distance oracle* for the metric space that reports the pairwise distance for any two points,

- *weakly explicit model* ([17]): there exists a distance oracle and a *doubling oracle* for the metric space. Given any ball in the metric space $\mathcal{M}$, the doubling oracle returns in time $\lambda_{\mathcal{M}}$ a covering with $\lambda_{\mathcal{M}}$ balls of half the radius.

Note that for any finite set $X \subset \mathcal{M}$, $\lambda_X \leq \lambda_{\mathcal{M}}$. We present two data structures for the (c, r)-ANN problem of polygonal curves in arbitrary doubling metric spaces, under the discrete Fréchet distance. The dataset consists of curves in $\mathcal{M}^m$ and queries belong to $\mathcal{M}^k$. Once again, we aim for polynomial dependence on m. The first data structure achieves $O(k)$ approximation in the black-box model when the doubling dimension is constant, and the second one achieves $(1+\epsilon)$ approximation in the weakly explicit model.

The high-level idea of our solution is very similar to the one of Section 7.1. We use nets, in order to discretize the input space, and a net-hierarchy which allows for a fast implementation of a Δ-bounded-diameter random partition. Such partitions are quite common in the literature (see e.g. [50], Chapter 26). The random partition of points naturally extends to a random partition of curves by considering k-tuples of parts. Then, we use perfect hashing and we build a look-up table where the set of non-empty buckets realizes the partition (each bucket contains only these curves which belong to a certain part). Now, any two curves which fall into the same bucket are Δ-near, and by carefully adjusting the parameters, this already provides with an $O(k)$ approximation. Furthermore, assuming the existence of a doubling oracle for the ambient space, we can precompute $(1+\epsilon)$-approximate answers to all possible queries. To answer a query, we use the net-hierarchy to efficiently compute the corresponding part and then we retrieve the answer from the look-up table.

7.2.1 Net Hierarchies

We now introduce the main algorithmic tool of this section. Our data structure is based on the notion of net-trees.

Definition 83 (Net-tree [53]). *Let $P \subset \mathcal{M}$ be a finite set. A* net-tree *of P is a tree T whose set of leaves is P. We denote by $P_v \subseteq P$ the set of leaves in the subtree rooted at a vertex $v \in T$. Associate with each vertex v a point $rep_v \in P_v$. Internal vertices have at least two children. Each vertex v has a level $\ell(v) \in \mathbb{Z} \cup \{-\infty\}$. The levels satisfy $\ell(v) < \ell(\overline{p}(v))$, where $\overline{p}(v)$ is the parent of v in T. The levels of the leaves are $-\infty$. Let τ be some large enough constant, say $\tau = 11$. We require the following properties from T:*

- Covering property: *For every vertex $v \in T$:*

$$P_v \subset b_{\mathcal{M}}\left(rep_v, \frac{2\tau}{\tau-1} \cdot \tau^{\ell(v)}\right).$$

- Packing property: *For every nonroot vertex $v \in T$,*

$$b_{\mathcal{M}}\left(rep_v, \frac{\tau-5}{2(\tau-1)} \cdot \tau^{\ell(\overline{p}(v))-1}\right) \cap P \subset P_v.$$

- Inheritance property: *For every nonleaf vertex $u \in T$, there exists a child $v \in T$ of u such that $rep_u = rep_v$.*

Theorem 84 (Theorem 3.1 [53])**.** *Given a set P of n points in $\mathcal{M}$, one can construct a net-tree for P in $\lambda_P^{O(1)} n \log n$ expected time.*

Enhancing the net-tree so that it supports several auxiliary operations leads to the following theorem.

Theorem 85 (Theorem 4.4 [53])**.** *Given a set P of n points in a metric space $\mathcal{M}$, one can construct a data-structure for answering $(1+\epsilon)$-ANN queries (where the quality parameter ϵ is provided together with the query). The query time is $\lambda_P^{O(1)} \log n + \epsilon^{-O(\log \lambda_P)}$, the expected preprocessing time is $\lambda_P^{O(1)} n \log n$, and the space used is $\lambda_P^{O(1)} n$.*

Definition 86 (Pruned net-tree)**.** *Given some pruning parameter $w > 0$, we define the pruned net-tree to be a net-tree as in Definition 83 which is pruned as follows: for any $v \in T$ such that $P_v \subset b_{\mathcal{M}}(rep_v, w)$, we delete all points in P_v, except for rep_v which remains as the single leaf of v.*

We present a data structure for the range search problem on nets, which is entirely based on [53]. We note that in order to keep the presentation simple, we make use of the main results there in a black-box manner, but a more straightforward solution is likely attainable.

Theorem 87. *Let $X \subset \mathcal{M}$, where $(\mathcal{M}, \mathsf{d}_{\mathcal{M}})$ is a metric space, and X is the set of n leaves in a pruned net-tree T with pruning parameter w (i.e. X is a $\Omega(w)$-net). There exists a data structure with input X which supports the following type of range queries:*

- *given $q \in \mathcal{M}$, $r > 0$, report $b_{\mathcal{M}}(q, r) \cap X$.*

The expected preprocessing time is $\lambda_X^{O(1)} n \log n$, the space consumption is $\lambda_X^{O(1)} n$ and the query time is $\lambda_X^{O(1)} \log n + \lambda_X^{O(\log(r/w))}$.

Proof. We build a data structure as in Theorem 85, and we are able to find a 2-approximate nearest neighbor of q in time $\lambda_X^{O(1)} \log n$, with expected preprocessing time in $\lambda_X^{O(1)} n \log n$ and space in $\lambda_X^{O(1)} n$. This point is denoted by q'. By the triangular inequality, it suffices to seek for the points of $b_{\mathcal{M}}(q, r) \cap X$ in $b_{\mathcal{M}}(q', 3r) \cap X$.

In order to perform a range query for a leaf q', we invoke an auxiliary data structure from [53] (see Section 3.5), which, for any query node v, allows us to find all points U within radius $r' = O(\tau^{\ell(v)})$ that are roughly at the same level, i.e $\forall u \in U: \ \ell(u) \leq \ell(v) < \ell(\overline{p}(u))$. This can be done by maintaining appropriate lists of size $\lambda_X^{O(1)}$, while building the net-tree, and it does not affect asymptotically the construction of the net-tree. By the packing property of pruned net-trees, we can retrieve all leaves within distance $O(r)$ from q' in time $\lambda_X^{O(\log(r/w))}$. $\square$

7.2.2 A data structure for curves

Our data structure is based on a quite standard random partition method which has been used repeatedly in the literature, especially in results concerning metric embeddings. We use this method in order to obtain a partition of the curves with the desired property that near curves probably belong to the same part. For any set X, $diam(X)$ denotes the diameter of X.

`partition`$(X \subset \mathcal{M}, \Delta > 0)$

- Set random permutation of X: $x_1, x_2, \ldots, x_n$.
- Set $C_0 \leftarrow \emptyset$.
- Set ordered set $\mathcal{P} \leftarrow \emptyset$.
- Choose uniformly at random $R \in [\Delta/4, \Delta/2]$.
- **For** $i = 1, \ldots, n$:
  - Set $C_i \leftarrow \{p \in X \mid \mathrm{d}_{\mathcal{M}}(x_i, p) \leq r\} \cup C_{i-1}$, where $C_{i-1} \subseteq X$ is the set of covered points in the $(i-1)$th iteration.
  - Set $P_i \leftarrow C_i \setminus C_{i-1}$. $\mathcal{P} \leftarrow \mathcal{P} \cup \{P_i\}$.
- **Return** the permutation $x_1, x_2, \ldots, x_n$, and indices to corresponding parts according to $\mathcal{P}$.

The following lemma describes the performance of the above partition scheme. Typically, similar guarantees discussed in the literature concern only points participating in the procedure (e.g. Lemma 26.7 [50]), while we need to take into account a query point which is not known in advance. To that end, we include a proof for completeness.

Lemma 88. *Let $(\mathcal{M}, \mathrm{d}_{\mathcal{M}})$ be a metric space, $X \subset \mathcal{M}$ a finite subset, and let $\mathcal{P}$ be the random partition generated by* `partition`(X, Δ). *For any $x \in X$, let $\mathcal{P}(x)$ be the part to which x has been assigned. Then, the following hold:*

- *For any $P \in \mathcal{P}$, $diam(P) \leq \Delta$.*
- *Let $q \in \mathcal{M}$ and let $x_j \in X$ be such that $j = \min\{i \mid \mathrm{d}_{\mathcal{M}}(q, x_i) \leq R\}$. Then, if $b_{\mathcal{M}}(q,t) \cap X \neq \emptyset$ and $t \leq \Delta/8$,*

$$\mathrm{P}r[b_{\mathcal{M}}(q,t) \cap X \not\subseteq \mathcal{P}(x_j)] \leq \frac{8t}{\Delta} \ln\left(|b_{\mathcal{M}}(q,\Delta) \cap X|\right).$$

Proof. Since $R \leq \Delta/2$, obviously $\forall P \in \mathcal{P} : \ diam(P) \leq \Delta$.

Let $m = |b_{\mathcal{M}}(q,\Delta) \cap X|$ and let $p_1, \ldots, p_m$ be the points in $b_{\mathcal{M}}(q,\Delta) \cap X$ which are ordered in increasing distance from q. The probability that a certain point p_i serves as the first center for a cluster that intersects (but does not include) $b_{\mathcal{M}}(q,t)$ is upper bounded by the probability that $R \in [d_{\mathcal{M}}(p_i,q)-t, d_{\mathcal{M}}(p_i,q)+t]$ and p_i appears before $p_1,\ldots,p_{i-1}$ in the permutation, since otherwise one of the previous clusters would have intersected (and possibly covered) $b_{\mathcal{M}}(q,t)$. Formally,

$$\mathsf{P}r[\exists x \in X \mid b_{\mathcal{M}}(q,t)\cap \mathcal{P}(x) \neq \emptyset \text{ and } b_{\mathcal{M}}(q,t)\cap X \not\subseteq \mathcal{P}(x)] \leq$$

$$\leq \sum_{i=1}^{m} \mathsf{P}r[R \in d_{\mathcal{M}}(p_i,q)\pm t]\cdot \frac{1}{i} \leq \frac{8t}{\Delta}\ln m.$$

Finally, since $b_{\mathcal{M}}(q,t)\cap X \neq \emptyset$ and $t \leq \Delta/8$, there exists at least one point which serves as a center for a cluster containing $b_{\mathcal{M}}(q,t)$.

□

Lemma 89. *Given as input parameters $\Delta > 0$, a pruned net-tree T with pruning parameter w, where X is the set of n leaves in T, $\mathtt{partition}(X,\Delta)$ can be implemented to run in $\lambda_X^{O(1)} n\cdot \log n + n\cdot \lambda_X^{O(\log(\Delta/w))}$ time.*

Proof. By Theorem 87, we can build a data structure which supports range queries: given a point $q \in \mathcal{M}$, $R \in [0,\Delta/2]$, we are able to report $\{x \in X \mid \mathsf{d}_{\mathcal{M}}(q,x) \leq R\}$ in time $\lambda_X^{O(1)}\log n + \lambda_X^{O(\log(R/w))} \leq \lambda_X^{O(1)}\log n + \lambda_X^{O(\log(\Delta/w))}$. Hence, for any point x_i, we cover and mark points which had not been covered before, and since we need to consider at most n points, the total amount of time needed is $\lambda_X^{O(1)} n\cdot\log n + n\cdot\lambda_X^{O(\log(\Delta/w))}$.

□

Now, for a partition which is obtained by `partition` (actually for any partition), each polygonal curve in $\mathcal{M}^m$ stabs at most m distinct parts. Using Theorem 87, we are able to build a data structure on the centers of the partition. Then, recovering the part that some point belongs to, is easy: we perform a Δ-range query for the given point and then we examine all $\leq \lambda_X^{O(\log(\Delta/w))}$ points inside this range.

Theorem 90. *Given as input a set of n polygonal curves $P \subset \mathcal{M}^m$ in the black-box model, there exists a randomized data structure for the $(O(\rho),r)$-ANN problem under the discrete Fréchet distance, with space in $\lambda_X^{O(1)}nm$, expected preprocessing time in $n\cdot m\cdot\left(\lambda_X^{O(\log\rho)} + \lambda_X^{O(1)}\log(nm)\right)$, and query time in $k\cdot\left(\lambda_X^{O(\log\rho)} + \lambda_X^{O(1)}\log(nm)\right)$, where $X := \bigcup_{p\in P} V(p)$, and $\rho := \rho(\lambda_X,k) \in O(k\log\lambda_X)$. For any query curve $q \in \mathcal{M}^k$, the preprocessing algorithm succeeds with constant probability.*

Proof. Preprocessing. Let r' be the ANN radius search parameter, and let $r := 4r'/3$. First, we build a pruned net-tree on $X := \bigcup_{p\in P} V(p)$. A net-tree can be built in expected time $\lambda_X^{O(1)}nm\log(nm)$ by [53]. Then, we transform it to a pruned net-tree T with pruning parameter $w := r/4$, by visiting at most all nodes and checking which ones should be deleted. We then build the data structure of Theorem 87 and we run the algorithm of Lemma 89 with input X, $\Delta = 100\cdot r\cdot(k\log\lambda_X)\log(k\log\lambda_X)$. The output consists of an ordered set of points and the partition.

We store P in a hashtable as follows. First we compute one vector of indices per curve indicating the corresponding parts. By Theorem 87, this costs $m \cdot \left(\lambda_X^{O(\log(\Delta/w))} + \lambda_X^{O(1)} \log(nm)\right)$ time for each curve. If one polygonal curve stabs more than k parts, we discard it. If it stabs less than k parts, we use a special character for the remaining coordinates. The polygonal curves are then stored in a hashtable: each bucket is assigned to a key vector of dimension k. Any non-empty bucket corresponds to $\leq k$ parts, of diameter $\leq \Delta$.

Storage. We store a net tree, which requires $\lambda_X^{O(1)} nm$ space, and a hashtable with at most n non-empty buckets containing indices to curves.

Query. For any query $q \in \mathcal{M}^k$, we perform k Δ-range queries on the leaves of T. For each of the k vertices, we explore points within distance Δ, in order to find which point is the first in the permutation used in `partition`, that covers it. Hence we compute the corresponding key vector in time $k \cdot \left(\lambda_X^{O(\log(\Delta/w))} + \lambda_X^{O(1)} \log(nm)\right)$. We have access to the bucket in $O(k)$ time, and we report any data curve stored in that bucket.

Correctness. We claim that the above data structure solves the $(O(\Delta/r)), 3r/4)$ANN problem. The choice of our pruning parameter implies that if there is a point in the original pointset within distance $3r/4$ from some query point, then there is a leaf in the net-tree within distance r. In order to prove that the approximation factor holds, we make use of Lemma 80, and the fact that the pruning step only induces constant multiplicative error. This implies that if $\mathsf{d}_{dF}(p, q) \leq 3r/4$ then there exists an optimal traversal which consists of k components and each component can be covered by a ball of radius r centered at a point of $X \cup V(q)$. By Lemma 88, the probability that `partition` splits one component is at most

$$\frac{8r}{\Delta} \ln \lambda_X \cdot \log \frac{8\Delta}{r} \leq \frac{8}{100k} \cdot \frac{\log\left(800(k \log \lambda_X) \cdot \log(k \log \lambda_X)\right)}{\log(k \log \lambda_X)} \leq \frac{8}{100k} \cdot \frac{10 + 2\log\left((k \log \lambda_X)\right)}{\log(k \log \lambda_X)}$$

$\leq 99/(100k)$, and by a union bound the probability that q is separated from its near neighbor is constant. □

Theorem 91. *Given as input a set of n polygonal curves $P \subset \mathcal{M}^m$ in the weakly explicit model, and an approximation parameter $\epsilon > 0$, there exists a randomized data structure for the $(1+\epsilon, r)$-ANN problem under the discrete Fréchet distance, with space in $\lambda_X^{O(1)} nm + \lambda_{\mathcal{M}}^{O(k \cdot \log \rho)} n$, expected preprocessing time in $\lambda_X^{O(1)} nm \log(nm) + \lambda_{\mathcal{M}}^{O(k \cdot \log \rho)} \cdot nmk$, and query time in $k \cdot \left(\lambda_{\mathcal{M}}^{O(\log \rho)} + \lambda_X^{O(1)} \log(nm)\right)$, where $X := \bigcup_{p \in P} V(p)$, and*

$$\rho := \rho(\lambda_X, k, \epsilon) \in O\left(\epsilon^{-1} \cdot k \cdot (\log \lambda_X) \cdot \log(1/\epsilon)\right).$$

For any query curve $q \in \mathcal{M}^k$, the preprocessing algorithm succeeds with constant probability.

Proof. Preprocessing. The first preprocessing step is similar to the one applied in the proof of Theorem 90. We build a pruned net-tree T on $X := \bigcup_{p \in P} V(p)$, with pruning parameter $w=\epsilon r$, in expected time $\lambda_X^{O(1)} nm\log(nm)$. We then build the data structure of Theorem 87 and we run the algorithm of Lemma 89 with input X, and

$$\Delta = 100r\cdot(k\log\lambda_X\log(1/\epsilon))\cdot\log\left(k\log\lambda_X\log(1/\epsilon)\right).$$

We compute one vector of indices per curve indicating the corresponding parts. This costs $m\cdot\left(\lambda_X^{O(\log(\Delta/w))}+\lambda_X^{O(1)}\log(nm)\right)$ time for each curve. If one polygonal curve stabs more than k parts, we discard it. If it stabs less than k parts, we use a special character for the remaining coordinates. The polygonal curves are then stored in a hashtable: each bucket is assigned to a key vector of dimension k. Any non-empty bucket corresponds to $\leq k$ parts, of diameter $\leq\Delta$. The weakly explicit model assumes that we are able to access points which ϵr-cover a ball of radius r in $\lambda_{\mathcal{M}}^{O(\log(1/\epsilon))}$ time. Given a sequence of k pointsets which ϵr-cover the whole bucket, we precompute and store the answers for all possible approximate queries. The number of possible queries which are compatible with a given sequence of k parts is: $\leq$

$$\sum_{\substack{t_1+\ldots+t_k=k\\ \forall i:\ t_i\geq 0\\ t_1\geq 1, t_k\geq 1}}\prod_{i=1}^{k}\lambda_{\mathcal{M}}^{t_i\log(\Delta/w)}=\sum_{\substack{t_1+\ldots+t_k=k\\ \forall i:\ t_i\geq 0}}\lambda_{\mathcal{M}}^{k\log(\Delta/w)}=\binom{2k-1}{k}\cdot\lambda_{\mathcal{M}}^{k\log(\Delta/w)}\leq\lambda_{\mathcal{M}}^{O(k\log(\Delta/w))}.$$

Storage. We store a net-tree in $\lambda_X^{O(1)}nm$. We also store a hashtable with at most n non-empty buckets, which correspond to different parts. For each bucket/part we store a hashtable with $\leq\lambda_{\mathcal{M}}^{O(k\log(\Delta/w))}$ non-empty buckets, one for each approximate query.

Query. For any query $q\in\mathcal{M}^k$, we perform k Δ-range queries on the leaves of T. For any point $x\in V(q)$, we explore points within distance Δ, in order to find which point is the first in the permutation used in `partition`, which also covers x. Hence, we compute the corresponding key vector in time $k\cdot\left(\lambda_X^{O(\log(\Delta/w))}+\lambda_X^{O(1)}\log(nm)\right)$. Then, we have access to the bucket in $O(k)$ time, and we locate the representative sequence of points in $k\cdot\lambda_{\mathcal{M}}^{O(\log(\Delta/w))}$ time.

*Correctness.*We claim that the data structure solves the $(1+\Theta(\epsilon),(1-2\epsilon)r)$-ANN problem. In order to prove correctness, we make use of Lemma 80 and the fact that approximating the input dataset by the net, only induces $\Theta(\epsilon r)$ additive error. This implies that if $\mathrm{d}_{dF}(p,q)\leq(1-2\epsilon)r$ then there exists an optimal traversal which consists of k components and each component can be covered by a ball of radius r centered at a point of $X\cup V(q)$. The probability that `partition` splits one component is at most

$$\frac{8r}{\Delta}\ln\lambda_X\cdot\log\frac{\Delta}{\epsilon r}\leq\frac{8}{100k\log(1/\epsilon)}\cdot\frac{\log\left(100\epsilon^{-1}(k\log\lambda_X\log(1/\epsilon))\cdot\log(k\log\lambda_X\log(1/\epsilon))\right)}{\log(k\log\lambda_X\log(1/\epsilon))}$$
$$\leq\frac{8}{100k\log(1/\epsilon)}\cdot\frac{7+\log(1/\epsilon)+2\log\left((k\log\lambda_X\log(1/\epsilon))\right)}{\log(k\log\lambda_X\log(1/\epsilon))}\leq\frac{9}{10k},$$

and by a union bound the probability that q is separated from its approximate near neighbor is $\leq 1/10$. □

8. VAPNIK–CHERVONENKIS DIMENSION FOR POLYGONAL CURVES

A crucial descriptor of any range space is its VC-dimension [79, 75, 74] and related shattering dimension, which we define formally below. These notions quantify how complex a range space is, and have played foundational roles in machine learning [80, 13], data structures [29], and geometry [50, 26]. For instance, the specific task of bounding these complexity parameters have is critical for tasks as diverse as neural networks [13, 62], art-gallery problems [78, 44, 64], and kernel density estimation [60].

The last five years have seen a surge of interest into data structures for trajectory processing under the Fréchet distance, manifested in a series of publications [34, 47, 35, 4, 82, 20, 39, 27, 38, 18, 41]. Partially motivated by the increasing availability and quality of trajectory data from mobile phones, GPS sensors, RFID technology and video analysis [65, 83, 46]. Initial results in this line of research, such as the approximate range counting data structure by de Berg, Gudmundsson and Cook [34], use classical data structuring techniques. Afshani and Driemel extended their results and in addition showed lower bounds on the space-query-time trade-off in this setting [4]. In particular, they showed a lower bound which is exponential in the complexity of the curves for exact range searching. In 2017, ACM SIGSPATIAL, the premier conference for geographic information science, devoted their software challenge (GIS CUP) to the problem of range searching under the Fréchet distance [82]. Spurring further developments, the most recent results explore the use of heuristics and randomization, such as locality-sensitive hashing.

The Fréchet distance is a popular distance measure for curves. Intuitively, it can be defined using the metaphor of a person walking a dog, where the person follows one curve and the dog follows the other curve, and throughout their traversal they are connected by a leash of fixed length. The Fréchet distance corresponds to the length of the shortest dog leash that permits a traversal in this fashion. The Fréchet distance is very similar to the Hausdorff distance for sets, which is defined as the minimal maximum distance of a pair of points, one from each set, under all possible matchings between the two sets. The difference between the two distance measures is that the Fréchet distance requires the matching to adhere to the ordering of the points along the curve. Both distance measures allow flexible associations between parts of the input elements which sets them apart from classical ℓ_p distances and makes them so suitable for trajectory data under varying speeds.

Our contribution in this chapter is a comprehensive analysis of the Vapnik-Chervonenkis dimension of the corresponding range spaces. In particular, we analyze the asymmetric case: the ground set consists of polygonal curves of complexity m, and the ranges are defined by polygonal curves of complexity k. The resulting VC dimension bounds, while being interesting in their own right, have a plethora of applications through the implied sampling bounds.

Organization. In Section 8.1, we state basic definitions. Section 8.2 provides an overview of the results obtained in this chapter. In Section 8.3, we summarize our approach and we present our first results for the simple discrete setting. Section 8.4 states our results for the weak Fréchet distance, Section 8.5 extends our results to the Fréchet distance and Section 8.6 is dedicated to the Hausdorff distance.

8.1 Preliminaries

In this section, we formally define primitives, which are repeatedly used throughout the chapter.

Geometric primitives. For any $p \in \mathbb{R}^2$ we denote by $C_r(p)$ the circle of radius r, centered at p. For any $p \in \mathbb{R}^2$ we denote by $D_r(p)$ the disk of radius r, centered at p. For any two points $s, t \in \mathbb{R}^2$, we denote by $\overline{st}$ the line segment from s to t. For any two points $s, t \in \mathbb{R}^2$, we define the stadium *centered* at $\overline{st}$, $B_r(s,t) = \left\{x \in \mathbb{R}^2 \mid \exists p \in \overline{st}\ \|p-x\|_2 \leq r\right\}$. For any two points $s, t \in \mathbb{R}^2$, we define $L_r(s,t) = \left\{x \in \mathbb{R}^2 \mid \exists p \in \ell(\overline{st})\ \|p-x\|_2 \leq r\right\}$. Finally, for any two points $s, t \in \mathbb{R}^2$, we define the rectangle *centered* at $\overline{st}$: $R_r(\overline{st}) = \text{conv}\{s-u, s+u, t+u, t-u\}$ and $u \in \mathbb{R}^2$ s.t. $\langle t-s, u\rangle = 0$ and $\|u\|_2 = r$. For a set A, we denote by ∂A the boundary of A, e.g. $C_r(p) = \partial D_r(p)$.

We also need to define the ball for pseudometric spaces.

Definition 92. *Let (M, d) be a pseudometric space. We define the ball of radius r and center p, under the distance measure* d*, as the following set:*

$$b_{\mathsf{d}}(p,r) = \{x \in M \mid \mathsf{d}(x,p) \leq r\},$$

where $p \in M$.

8.2 Our Results

Table 8.1 shows an overview of our bounds.

While the VC dimension bounds for the Hausdorff metric balls may seem like an easy implication of composition theorems for VC dimension [25, 31], we still find two things about these techniques remarkable. First, for Fréchet variants, there are $\Theta(2^k 2^m)$ valid alignment paths in the free space diagram. And one may expect that these may materialize in the size of the composition theorem. Yet by a simple analysis of the shattering dimension, we show that they do not. Second, the VC dimension only has logarithmic dependence on the size m of the curves in the ground set, rather than a polynomial dependence one would obtain by simple application of composition theorems (even ignoring the alignment path issue). This difference has important implications in analyzing real data sets where we can query with simple curves (small k), but may not have a small bound on the size of the curves in the data set (large m).

Table 8.1: Our results on the VC dimension of range space $(X, \mathcal{R})$. In the first column we distinguish between X consisting of *discrete* point sequences vs. X consisting of *continuous* polygonal curves. The ground set X consists of polygonal curves of complexity m and the range set $\mathcal{R}$ consists of balls centered at polygonal curves of complexity k. Additional upper bounds on the range space under the directed Hausdorff distance are stated in Theorems 117 and 118.

<table>
<tr><th>X, m</th><th>$\mathcal{R}$, k</th><th>Upper bound</th><th>Lower bound</th></tr>
<tr><td rowspan="2">discrete
$(d = 2)$</td><td>Hausdorff</td><td rowspan="3">$O(k \log(km))$ (Theorems 93,94,99)</td><td rowspan="5">$(d \geq 2)$
$\Omega(\max(k, \log m))$
(Theorem 127)</td></tr>
<tr><td>Fréchet</td></tr>
<tr><td rowspan="3">cont.
$(d = 2)$</td><td>weak Fréchet</td></tr>
<tr><td>Fréchet</td><td rowspan="2">$O(k^2 \log(km))$ (Theorems 106,119)</td></tr>
<tr><td>Hausdorff</td></tr>
</table>

8.3 Our Approach

Our methods use the fact that both the Fréchet distance and the Hausdorff distance are determined by one of a discrete set of events, where each event involves a constant number of simple geometric objects. For example, it is well known that the Hausdorff distance between two discrete sets of points is equal to the distance between two points from the two sets. The corresponding event happens as we consider a value $\delta > 0$ increasing from 0 and we record which points of one set are contained in which balls of radius δ centered at points from the other set. The same phenomenon is true for the discrete Fréchet distance between two point sequences. In particular, the so-called free-space matrix which can be used to decide whether the discrete Fréchet distance is smaller than a given value δ encodes exactly the information about which pairs of points have distance at most δ. The basic phenomenon remains true for the continuous versions of the two distance measures if we extend the set of simple geometric objects to include line segments and if we also consider triple intersections. Each type of event can be translated into a range space of which we can analyze the VC dimension. Together, the concatenation of the range spaces encodes the information about which curves lie inside which metric balls in the form of a set system. This representation allows us to prove bounds on the VC dimension of metric balls under these distance measures.

We now prove our upper bounds in the discrete setting. Let $\mathbb{X}_m = (\mathbb{R}^2)^m$; we treat the elements of this set as ordered sets of points in $\mathbb{R}^2$ of size m. The range spaces that we consider in this section are defined over the ground set $\mathbb{X}_m$ and the range set of balls under either the Hausdorff or the Discrete Fréchet distance. The proofs in the proceeding sections all follow the basic idea of the proof in the discrete setting.

Theorem 93. *Let $(\mathbb{X}_m, \mathcal{R}_{H,k})$ be the range space with $\mathcal{R}_{H,k}$ the set of all balls under the Hausdorff distance centered at sets in $\mathbb{X}_k$. The VC dimension is $O(k \log(km))$.*

Proof. Let $\{S_1, \dots, S_t\} \subseteq \mathbb{X}_m$ and $S = \bigcup_i S_i$; we define S so that it ignores the ordering with each S_i and is a single set of size tm. Any intersection of a Hausdorff ball with $\{S_1, \dots, S_t\}$ is uniquely defined by a set $\{D_1 \cap S, \dots, D_k \cap S\}$, where $D_1, \dots, D_k$ are disks in $\mathbb{R}^2$.

Consider the range space $(\mathbb{R}^2, \mathcal{D})$, where $\mathcal{D}$ is the set of disks in the plane. We know that the shattering dimension is 3 [50]. Hence,

$$\max_{S \subseteq \mathbb{R}^2, |S|=tm} |\mathcal{D}_{|S}| = O((tm)^3).$$

This implies that $|\{\{D_1 \cap S, \ldots, D_k \cap S\} \mid D_1, \ldots, D_k \text{ are disks in } \mathbb{R}^2\}| \leq O((tm)^{3k})$, and hence[1],

$$2^t \leq 2^{O(k \log(tm))} \implies t = O\left(k \log(km)\right). \qquad \square$$

Theorem 94. *Let $(\mathbb{X}_m, \mathcal{R}_{dF,k})$ be the range space with $\mathcal{R}_{dF,k}$ the set of all balls under the Discrete Fréchet distance centered at polygonal curves in $\mathbb{X}_k$. The VC dimension is $O\left(k \log(km)\right)$.*

Proof. Let $\{S_1, \ldots, S_t\} \subseteq X$ and $S = \bigcup_i S_i$. Any intersection of a Discrete Fréchet ball with $\{S_1, \ldots, S_t\}$ is uniquely defined by a sequence $D_1 \cap S, \ldots, D_k \cap S$, where $D_1, \ldots, D_k$ are disks in $\mathbb{R}^2$. The number of such sequences can be bounded by $O((tm)^{3k})$ as in the proof of Theorem 93. Enforcing that a sequence contains a valid alignment path only reduces the number of possible distinct sets formed by t curves, and it can be determined using these intersections and the two orderings of $D_1, \ldots, D_k$ and of vertices within some $S_j \in \mathbb{X}_m$. $\square$

8.4 Weak Fréchet distance

In this section we prove our upper bounds for the Weak Fréchet distance. Let $\mathbb{W}_m$ be the set of polygonal curves of complexity m; for each $s \in \mathbb{W}_m$, we associate an ordered set of vertices $V(s)$ and an ordered set of edges $E(s)$. We consider the range space $(\mathbb{W}_m, \mathcal{R}_{wF})$, where $\mathcal{R}_{wF}$ is the set of all balls under the Weak Fréchet distance.

8.4.1 Some useful lemmas

Lemma 95. *Consider the range space $(X, \mathcal{R})$, where $X = \mathbb{R}^2$ and $\mathcal{R}$ is the set of the form $\{B_r(s,t) \mid r \geq 0,\ s,t \in \mathbb{R}^2\}$. The shattering dimension of this range space is $O(1)$.*

Proof. Let $Y \subset X$ s.t. $|Y| = n$ and let $\mathcal{D}$ be the set of all disks in $\mathbb{R}^2$. Let $D_r(s) = \{x \in \mathbb{R}^2 \mid \|x - s\|_2 \leq r\}$ and $D_r(t) = \{x \in \mathbb{R}^2 \mid \|x - t\|_2 \leq r\}$. Consider any intersection $S = B_r(s,t) \cap Y$. We can assume that S contains a point q at distance exactly r from the segment $\overline{st}$ (otherwise decrease r). Then, S is uniquely defined by the intersections $D_r(s) \cap Y$, $D_r(t) \cap Y$ and $D_r(p) \cap Y$, where $\|p - q\|_2 = r$, $p \in \overline{st}$. Hence, $|\mathcal{R}_{|Y}| \leq |\mathcal{D}_{|Y}|^3 = O(n^9)$. $\square$

[1] for $u > \sqrt{e}$ if $x/\ln(x) \leq u$ then $x \leq 2u \ln u$. Hence, if $tm/\log(tm) \leq km$, then $tm = O(km \log(km))$.

Corollary 96. *Let $X = \{B_r(s,t) \mid r \geq 0,\ s,t \in \mathbb{R}^2\}$. Consider the range space $(X, \mathcal{R})$, where $\mathcal{R} = \{\mathcal{R}_p \mid p \in \mathbb{R}^2\}$ and $\mathcal{R}_p = \{r \in X \mid p \in r\}$. The shattering dimension of this range space is $O(1)$.*

Proof. The range space $(X, \mathcal{R})$ is the dual of the range space from Lem. 95. □

8.4.2 Representation in terms of predicates

It is known that the Fréchet distance between two polygonal curves can be attained, either at a distance between their endpoints, at a distance between a vertex and a line supporting an edge, or at the common distance of two vertices with a line supporting an edge. In this sense, our representation of the ball of radius r under the Fréchet distance is based on the following predicates.[2] Let $s \in \mathbb{W}_m$ with vertices $s_1, \ldots, s_m$ and $q \in \mathbb{W}_k$ with vertices $q_1, \ldots, q_k$.

P_1 *(Endpoints (start))* This predicate returns true if and only if $\|s_1 - q_1\|_2 \leq r$.

P_2 *(Endpoints (end))* This predicate returns true if and only if $\|s_m - q_k\|_2 \leq r$.

P_3 *(Vertex-edge (horizontal))* Given an edge of s, $\overline{s_j s_{j+1}}$, and a vertex q_i of q, this predicate returns true iff there exist a point $p \in \overline{s_j s_{j+1}}$, such that $\|p - q_i\|_2 \leq r$.

P_4 *(Vertex-edge (vertical))* Given an edge of q, $\overline{q_i q_{i+1}}$, and a vertex s_j of s, this predicate returns true iff there exist a point $p \in \overline{q_i q_{i+1}}$, such that $\|p - s_j\|_2 \leq r$.

P_5 *(Monotonicity (horizontal))* Given two vertices of s, s_j and s_t with $j < t$ and an edge of q, $\overline{q_i q_{i+1}}$, this predicate returns true if there exist two points p_1 and p_2 on the *line* supporting the directed edge, such that p_1 appears before p_2 on this line, and such that $\|p_1 - s_j\|_2 \leq r$ and $\|p_2 - s_t\|_2 \leq r$.

P_6 *(Monotonicity (vertical))* Given two vertices of q, q_i and q_t with $i < t$ and an directed edge of s, $\overline{s_j s_{j+1}}$, this predicate returns true if there exist two points p_1 and p_2 on the *line* supporting the directed edge, such that p_1 appears before p_2 on this line, and such that $\|p_1 - q_i\|_2 \leq r$ and $\|p_2 - q_t\|_2 \leq r$.

Lemma 97 (Lemma 9, [3]). *Given the truth values of all predicates $(P1) - (P6)$ of two curves s and q for a fixed value of r, one can determine if $d_F(s,q) \leq r$.*

Predicates $P_1 - P_4$ are sufficient for representing metric balls under the weak Fréchet distance. We include a proof for the sake of completeness.

Lemma 98. *Given the truth values of all predicates $(P_1) - (P_4)$ of two curves s and q for a fixed value of r, one can determine if $\mathrm{d}_{wF}(s,q) \leq r$.*

[2] This representation was earlier derived in the context of data structures for range searching under the Fréchet distance (see [4, 3]). We repeat the relevant definitions and lemmas here.

Proof. Alt and Godau [7] describe an algorithm for computing the Weak Fréchet distance which can be used here. In particular, one can construct an edge-weighted grid graph on the cells (edge-edge pairs) of the parametric space of the two polygonal curves and subsequently compute a bottleneck-shortest path from the pair of first edges to the pair of last edges along the two curves. We can use edge weights in $\{0,1\}$ to encode if the corresponding vertex-edge pair has distance at most r, as given by the predicates P_3 and P_4. If and only if there exists a bottleneck shortest path of cost 0 and the endpoint conditions are satisfied (as given by the predicates P_1 and P_2), the Weak Fréchet distance between q and s is at most r. □

8.4.3 Representation as a range space

Predicates $P_1 - P_4$ can be directly translated into simple range spaces. Consider any two polygonal curves $s \in \mathbb{W}_m$ and $q \in \mathbb{W}_k$. In order to encode the intersection of polygonal curves with metric balls, we will make use of the following sets:

- $P_1^r(q,s) = D_r(q_1) \cap V(s)$,
- $P_2^r(q,s) = D_r(q_k) \cap V(s)$,
- $P_3^r(q,s) = \{B_r(s_i, s_{i+1}) \cap V(q) \mid \overline{s_i s_{i+1}} \in E(s)\}$,
- $P_4^r(q,s) = \{B_r(q_i, q_{i+1}) \cap V(s) \mid \overline{q_i q_{i+1}} \in E(q)\}$.

8.4.4 VC dimension bound

Theorem 99. *Let $\mathcal{R}_{wF}$ be the set of balls under the Weak Fréchet metric centered at polygonal curves in $\mathbb{W}_k$. The VC dimension of $(\mathbb{W}_m, \mathcal{R}_{wF})$ is $O(k \log(km))$.*

Proof. If S is a set of t polygonal curves of complexity m, the set $\{s \in S \mid \mathrm{d}_{wF}(s,q) \leq r\}$ is uniquely defined by the sets

$$\bigcup_{s\in S} P_1^r(q,s), \bigcup_{s\in S} P_2^r(q,s), \bigcup_{s\in S} P_3^r(q,s), \bigcup_{s\in S} P_4^r(q,s).$$

Notice that the number of all possible sets $\bigcup_{r\geq 0}\bigcup_{s\in S} P_1^r(q,s)$ is bounded by the shatter function for the range space of points and disks and it is $(tm)^{O(1)}$. The same holds for the number of all possible sets $\bigcup_{r\geq 0}\bigcup_{s\in S} P_2^r(q,s)$.

The number of all possible sets $\bigcup_{r\geq 0}\bigcup_{s\in S} P_3^r(q,s)$ and the number of all possible sets $\bigcup_{r\geq 0}\bigcup_{s\in S} P_4^r(q,s)$ are both bounded by $(tm)^{O(k)}$ by Lemma 95 and Corollary 96 respectively. Hence, $2^t \leq 2^{O(k \log(tm))} \implies t = O(k \log(km))$. □

8.5 The Fréchet distance

In this section we prove our upper bounds for the Fréchet distance. Let $\mathbb{W}_m$ be the set of polygonal curves of complexity m; for each $s \in \mathbb{W}_m$, we associate an ordered set of vertices $V(s)$ and an ordered set of edges $E(s)$. We consider the range space $(\mathbb{W}_m, \mathcal{R}_F^r)$, where $\mathcal{R}_F^r$ denotes the set of all balls, of radius r, under the Fréchet distance.

8.5.1 Some useful lemmas

Lemma 100. *Fix $r \geq 0$. Consider the range space $(X, \mathcal{R})$, where $X = \mathbb{R}^2$ and $\mathcal{R}$ is the set of the form $\{L_r(s,t) \mid s,t \in \mathbb{R}^2\}$. The shattering dimension of this range space is $O(1)$.*

Proof. The VC dimension of halfspaces in $\mathbb{R}^2$ is $O(1)$, which also bounds its shattering dimension. Each $L_r(s,t)$ coincides with the intersection of two parallel halfspaces which define the set of points at distance $\leq r$ from $\ell(\overline{st})$. Hence, the shattering dimension is $O(1)$. $\square$

Corollary 101. *Fix $r \geq 0$. Let $X = \{L_r(s,t) \mid s,t \in \mathbb{R}^2\}$. Consider the range space $(X, \mathcal{R})$, where $\mathcal{R} = \{\mathcal{R}_p \mid p \in \mathbb{R}^2\}$ and $\mathcal{R}_p = \{r \in X \mid p \in r\}$. The shattering dimension of this range space is $O(1)$.*

Proof. The range space $(X, \mathcal{R})$ is the dual of the range space from Lemma 100. $\square$

Lemma 102. *Consider the range space $(X, \mathcal{R})$, where $X = \mathbb{R}^2$ and $\mathcal{R}$ is the set of the form $\{A(\theta_1, \theta_2) \mid \theta_1, \theta_2 \in [0, 2\pi]\}$, where*

$$A(\theta_1, \theta_2) = \left\{x \in \mathbb{R}^2 \mid \theta(x) \in [\theta_1, \theta_2]\right\},$$

and $\theta(x)$ denotes the angle of vector x. The shattering dimension of this range space is $O(1)$.

Proof. When $|\theta_1 - \theta_2| \leq \pi$, each set $A(\theta_1, \theta_2)$ coincides with the intersection of two halfspaces crossing the origin. If $|\theta_1 - \theta_2| \in [\pi, 2\pi]$, then $A(\theta_1, \theta_2)$ coincides with the union of two halfspaces crossing the origin. Hence, the shattering dimension is $O(1)$. $\square$

Corollary 103. *Let $X = \{A(\theta_1, \theta_2) \mid \theta_1, \theta_2 \in [0, 2\pi]\}$, where*

$$A(\theta_1, \theta_2) = \left\{x \in \mathbb{R}^2 \mid \theta(x) \in [\theta_1, \theta_2]\right\},$$

and $\theta(x)$ denotes the angle of vector x. Consider the range space $(X, \mathcal{R})$, where $\mathcal{R} = \{\mathcal{R}_p \mid p \in \mathbb{R}^2\}$ and $\mathcal{R}_p = \{r \in X \mid p \in r\}$. The shattering dimension of this range space is $O(1)$.

Proof. The range space $(X, \mathcal{R})$ is the dual of the range space from Lemma 102. $\square$

8.5.2 Representation in terms of predicates

We use the predicates $P_1 - P_6$ from Section 8.4. Correctness follows from Lemma 97. For encoding the monotonicity predicates P_5 and P_6, we repeat the definitions from [4, 3].

Let a_1, a_2 be the vertices and let ℓ be the line supporting the directed edge e of a monotonicity predicate P_5 (respectively, P_6). Let points b_1, b_2 be $C_r(a_1) \cap C_r(a_2)$.

(d) The line ℓ intersects the circle of radius r centered at a_1.

(e) The line ℓ intersects the circle of radius r centered at a_2.

(f) The angle between the translation vector $(a_2 - a_1)$ and the edge e is at most $\frac{\pi}{2}$.

(h) The line ℓ passes in between the two points b_1 and b_2

(i) The angle of ℓ is contained in the range of angles of tangents of the circular arc between b_1 and b_2 of the circle of radius r centered at a_1.

Lemma 104 (Lemma 16, [3])**.** *Given the truth values of the predicates (d)-(i) one can determine the truth value of the predicate P_5 (respectively, P_6). Moreover, the predicate P_5 (respectively, P_6) is true if and only if the clause $(d \wedge e \wedge f) \vee (h \vee (d \wedge e \wedge i))$ is true.*

8.5.3 Representation as a range space

Now, consider any two polygonal curves s and q. In addition to the sets $P_1^r(q,s), \ldots, P_4^r(q,s)$ which were defined in Section 8.4.3, we need to define sets which describe predicates P_5, P_6. We invoke Lemma 104 to show that our sets are sufficient in order to determine whether $\mathrm{d}_F(s,q) \leq r$ or $\mathrm{d}_F(s,q) > r$. The new sets are defined as follows:

- $P_{d\wedge e}(q,s) = \{L_r(q_i, q_{i+1}) \cap V(s) \mid \overline{q_i q_{i+1}} \in E(q)\}$
- $P'_{d\wedge e}(q,s) = \{L_r(s_i, s_{i+1}) \cap V(q) \mid \overline{s_i s_{i+1}} \in E(s)\}$
- $P_f(q,s) = \{\{x \in \mathbb{R}^2 \mid \langle q_{i+1} - q_i, x\rangle \geq 0\} \cap \tilde{V}(s)\}$, where $\tilde{V}(s) = \{s_k - s_j \mid k > j,\ s_k, s_j \in V(s)\}$
- $P'_f(q,s) = \{\{x \in \mathbb{R}^2 \mid \langle s_{i+1} - s_i, x\rangle \geq 0\} \cap \tilde{V}(q)\}$, where $\tilde{V}(q) = \{q_k - q_j \mid k > j,\ q_k, q_j \in V(q)\}$
- $P_h(q,s) = \{h^+(\overline{q_i q_{i+1}}) \cap V_r^*(s) \mid \overline{q_i q_{i+1}} \in E(q)\} \cup \{\ell(\overline{q_i q_{i+1}}) \cap V_r^*(s) \mid \overline{q_i q_{i+1}} \in E(q)\}$, where $h^+(\overline{q_i q_{i+1}})$ denotes the right-side halfspace which is supported by the directed edge $\overline{q_i q_{i+1}}$, $V_r^*(s) = \bigcup_{k>j} C_r(s_k) \cap C_r(s_j)$
- $P'_h(q,s) = \{h^+(\overline{s_i s_{i+1}}) \cap V_r^*(q) \mid \overline{s_i s_{i+1}} \in E(s)\} \cup \{\ell(\overline{s_i s_{i+1}}) \cap V_r^*(q) \mid \overline{s_i s_{i+1}} \in E(s)\}$,where $h^+(\overline{s_i s_{i+1}})$ denotes the right-side halfspace which is supported by the directed edge $\overline{s_i s_{i+1}}$, $V_r^*(q) = \bigcup_{k>j} C_r(q_k) \cap C_r(q_j)$

- $P_i(q,s) = \{A(\theta_1(q_k,q_j),\theta_2(q_k,q_j)) \cap \tilde{E}(s) \mid k > j,\ q_k, q_j \in V(q)\}$, where $[(\theta_1(q_k,q_j),\theta_2(q_k,q_j)]$ defines the range of angles of tangents of the circular arc between the two points of $C_r(q_k) \cap C_r(q_j)$, and $\tilde{E}(s) = \{s_{i+1} - s_i \mid s_i, s_{i+1} \in E(s)\}$. If $C_r(q_k) \cap C_r(q_j) = \emptyset$, then we define $A(\theta_1(q_k,q_j),\theta_2(q_k,q_j)) = \emptyset$

- $P_i'(q,s) = \{A(\theta_1(s_k,s_j),\theta_2(s_k,s_j)) \cap \tilde{E}(q) \mid k > j,\ s_k, s_j \in V(s)\}$, where $[(\theta_1(s_k,s_j),\theta_2(s_k,s_j)]$ defines the range of angles of tangents of the circular arc between the two points of $C_r(s_k) \cap C_r(s_j)$, and $\tilde{E}(q) = \{q_{i+1} - q_i \mid q_i, q_{i+1} \in E(q)\}$. If $C_r(q_k) \cap C_r(q_j) = \emptyset$, then we define $A(\theta_1(q_k,q_j),\theta_2(q_k,q_j)) = \emptyset$

Lemma 105. *Let s be a polygonal curve in $\mathbb{W}_m$ with vertices $s_1,\ldots,s_m$ and q be a polygonal curve in $\mathbb{W}_k$ with vertices $q_1,\ldots,q_k$. Fix any $r \geq 0$. The following sets are sufficient in order to determine whether $\mathsf{d}_F(s,q) \leq r$ or $\mathsf{d}_F(s,q) > r$:*

$$P_1^r(q,s), P_2^r(q,s), P_3^r(q,s), P_4^r(q,s), P_{d\wedge e}(q,s), P'_{d\wedge e}(q,s), P_f(q,s), P'_f(q,s), P_h(q,s),$$
$$P'_h(q,s), P_i(q,s), P'_i(q,s).$$

Proof. Sets $P_1^r,\ldots,P_4^r$ correspond to high level predicates $(P_1),\ldots,(P_4)$ from Lemma 97.

We will now use Lemma 104, to show that for any $s_j, s_k \in V(s)$ s.t. $j < k$ and assuming that $C_r(s_j) \cap C_r(s_k) = \{a,b\}$, the outcome of the high-level monotonicity predicate $P_5(s_j, s_k, \overline{q_iq_{i+1}})$ is uniquely defined by the above-mentioned sets.

By Lemma 104, we have that $P_5(s_j, s_k, \overline{q_iq_{i+1}})$ is true iff one of the following is true:

- $[(s_j, s_k \in L_r(q_i,q_{i+1})) \wedge (\langle q_{i+1} - q_i, s_k - s_j\rangle \geq 0)]$,
- $[((a \in h^+(\overline{q_iq_{i+1}}) \wedge b \notin h^+(\overline{q_iq_{i+1}})) \vee (a \notin h^+(\overline{q_iq_{i+1}}) \wedge b \in h^+(\overline{q_iq_{i+1}}))) \vee (a, b \in \ell(\overline{q_iq_{i+1}}))]$,
- $[(s_j, s_k \in L_r(q_i,q_{i+1})) \wedge (\langle q_{i+1} - q_i, s_k - s_j\rangle \geq 0) \wedge (s_k - s_j \in A(\theta_1(q_k,q_j),\theta_2(q_k,q_j)))]$.

Notice that if $|C_r(s_j) \cap C_r(s_k)| \leq 1$, then the predicate is equivalent to

$$[(s_j, s_k \in L_r(q_i,q_{i+1})) \wedge (\langle q_{i+1} - q_i, s_k - s_j\rangle \geq 0)] \vee [C_r(s_j) \cap C_r(s_k) \cap \ell(\overline{q_iq_{i+1}}) \neq \emptyset].$$

Similarly for P_6.

□

8.5.4 VC dimension bound

The associated VC dimension is quadratic in k because sets P_h and P'_h are defined with respect to $V_r^*(q)$ which may include all $O(k^2)$ pairs of vertices in q.

Theorem 106. *Let $\mathcal{R}_F^r$ be the set of all balls of radius r, under the Fréchet distance, centered at polygonal curves in $\mathbb{W}_k$. The VC dimension of $(\mathbb{W}_m, \mathcal{R}_F^r)$ is $O\left(k^2 \log(km)\right)$.*

Proof. Due to Lemma 105, if $S \subset \mathbb{W}_m$ is a set of t polygonal curves and $q \in \mathbb{W}_k$, the set $\{s \in S \mid \mathsf{d}_F(s,q) \le r\}$ is uniquely defined by the sets

$$\bigcup_{s\in S} P_1^r(q,s), \bigcup_{s\in S} P_2^r(q,s), \bigcup_{s\in S} P_3^r(q,s), \bigcup_{s\in S} P_4^r(q,s), \bigcup_{s\in S} P_{d\wedge e}(q,s), \bigcup_{s\in S} P'_{d\wedge e}(q,s),$$

$$\bigcup_{s\in S} P_f(q,s), \bigcup_{s\in S} P'_f(q,s), \bigcup_{s\in S} P_h(q,s), \bigcup_{s\in S} P'_h(q,s), \bigcup_{s\in S} P_i(q,s), \bigcup_{s\in S} P'_i(q,s).$$

As in the proof of Theorem 99, the number of all possible sets $\left(\bigcup_{s\in S} P_1(q,s), \bigcup_{s\in S} P_2(q,s), \bigcup_{s\in S} P_3(q,s), \bigcup_{s\in S} P_4(q,s)\right)$ is bounded by $(tm)^{O(k)}$. Now, by Lemma 100 and Corollary 101 we are able to bound the number of all possible sets

$$\left(\bigcup_{s\in S} P_{d\wedge e}(q,s), \bigcup_{s\in S} P'_{d\wedge e}(q,s)\right),$$

which is also in $(tm)^{O(k)}$.

The shattering dimension of the range space implied by $\bigcup_{s\in S} P_f(q,s)$ is $O(1)$, since each range is a halfspace. Its dual corresponds to the set $\bigcup_{s\in S} P'_f(q,s)$ and also has shattering dimension of $O(1)$. The number of all possible sets $\left(\bigcup_{s\in S} P_f(q,s), \bigcup_{s\in S} P'_f(q,s)\right)$ is bounded by $(tm)^{O(k^2)}$, because $|\tilde{V}(q)| = \Theta(k^2)$.

The same arguments apply to the range space implied by $\bigcup_{s\in S} P_h(q,s)$. The shattering dimension of this range space is $O(1)$, since each range is a halfspace, and the same holds for its dual which corresponds to $\bigcup_{s\in S} P'_h(q,s)$. The number of all possible sets $\left(\bigcup_{s\in S} P_h(q,s), \bigcup_{s\in S} P'_h(q,s)\right)$ is bounded by $(tm)^{O(k^2)}$, because $|\tilde{V}_r^*(q)| = \Theta(k^2)$.

Finally by Lemma 102 and Corollary 103, we are able to bound the number of all possible sets $\left(\bigcup_{s\in S} P_i(q,s), \bigcup_{s\in S} P'_i(q,s)\right)$ by $(tm)^{O(k^2)}$. Hence,

$$2^t \le 2^{O(k^2 \log(tm))} \implies t = O\left(k^2 \log(km)\right).$$

□

8.6 The Hausdorff distance

In this section we prove our upper bounds for the Hausdorff distance. Let $\mathbb{W}_m$ be the set of polygonal curves[3] of complexity m; for each $s \in \mathbb{W}_m$, we associate an ordered set of vertices $V(s)$ and an ordered set of edges $E(s)$. We consider the range space $(\mathbb{W}_m, \mathcal{R}_H^r)$, where $\mathcal{R}_H^r$ denotes the set of all balls, of radius r, under the Hausdorff distance.

[3] The proofs in this section are written for polygonal curves, but they readily extend to (not-necessarily connected) sets of line segments in $\mathbb{R}^2$ of size m.

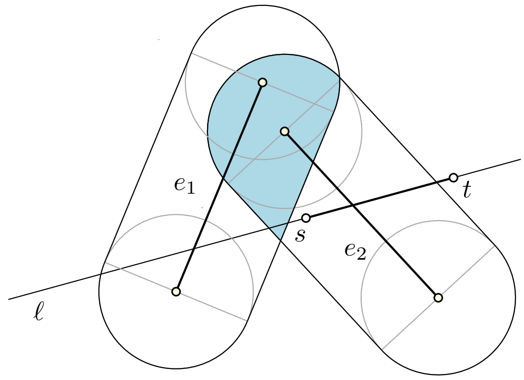

Figure 8.1: Illustration of the predicate P_7: The predicate evaluates to true if and only if the triple intersection of the line ℓ supporting $\overline{q_i q_{i+1}}$ with the two stadiums centered at $\overline{s_j s_{j+1}}$ and $\overline{s_t s_{t+1}}$ is non-empty. Note that $\overline{q_i q_{i+1}}$ may lie outside of the intersection.

8.6.1 Representation in terms of predicates

According to Alt, Behrends and Blömer [6], the directed Hausdorff distance $\mathrm{d}_{\overrightarrow{H}}(A, B)$ of two pairwise disjoint sets of line segments A and B is assumed either at some vertex of A or at some intersection point of A with a Voronoi-edge of B. We can re-use part of the predicates from the previous section for encoding the first type of event where the distance is assumed at a vertex of A. We need to derive a new set of predicates for the second type of event. In particular we need a predicate for testing if a line supporting an edge intersects the intersection of two stadiums (see Figure 8.1).

Consider any two polygonal curves $s \in \mathbb{W}_m$ and $q \in \mathbb{W}_k$. In order to encode the intersection of polygonal curves with metric balls under the Hausdorff metric, we will make use of the following predicates:

P_3 *(Vertex-edge (horizontal))* As defined in Section 8.4.

P_4 *(Vertex-edge (vertical))* As defined in Section 8.4.

P_7 *(Stadium-stadium-line (horizontal))* given one edge of q, $\overline{q_i, q_{i+1}}$, and two edges of s, $\overline{s_j, s_{j+1}}$, $\overline{s_t, s_{t+1}}$, this predicate is equal to $\ell(\overline{q_i, q_{i+1}}) \cap B_r(s_j, s_{j+1}) \cap B_r(s_t, s_{t+1}) \neq \emptyset$.

P_8 *(Stadium-stadium-line (vertical))* given one edge of s, $\overline{s_i, s_{i+1}}$, and two edges of q, $\overline{q_j, q_{j+1}}$, $\overline{q_t, q_{t+1}}$, this predicate is equal to $\ell(\overline{s_i, s_{i+1}}) \cap B_r(q_j, q_{j+1}) \cap B_r(q_t, q_{t+1}) \neq \emptyset$.

As in the proofs of Theorems 99 and 106, we argue that the truth values for the first predicate over all possible inputs, are uniquely defined by the set $P_3^r(q, s)$. Similarly, the truth values for predicate P_4 are uniquely defined by the set $P_4^r(q, s)$. Now predicate P_7 (resp. P_8) breaks to three simple predicates:

(j) given an edge $\overline{q_i q_{i+1}}$, an edge $\overline{s_j s_{j+1}}$, and a point s_t, determine whether $\ell(\overline{q_i q_{i+1}}) \cap R_r(\overline{s_j s_{j+1}}) \cap D_r(s_t) \neq \emptyset$,

(k) given an edge $\overline{q_i q_{i+1}}$, and edges $\overline{s_j s_{j+1}}$, $\overline{s_t s_{t+1}}$, determine whether $\ell(\overline{q_i, q_{i+1}}) \cap R_r(\overline{s_j s_{j+1}}) \cap R_r(\overline{s_t s_{t+1}}) \neq \emptyset$.

(l) given an edge $\overline{q_i q_{i+1}}$, and points s_j, s_t, determine whether $\ell(\overline{q_i q_{i+1}}) \cap D_r(s_j) \cap D_r(s_t) \neq \emptyset$,

Lemma 107. *For any two polygonal curves s, q, given the truth values of the predicates P_3, P_7 one can determine whether $\mathsf{d}_{\overrightarrow{H}}(q, s) \leq r$. Similarly, given the truth values of the predicates P_4, P_8 one can determine whether $\mathsf{d}_{\overrightarrow{H}}(s, q) \leq r$.*

Proof. We first assume for the sake of simplicity that q is a line segment in the plane with endpoints q_1 and q_2. We claim that $\mathsf{d}_{\overrightarrow{H}}(q, s) \leq r$ if and only if there exists a sequence of edges $\overline{s_{j_1} s_{(j_1+1)}}, \overline{s_{j_2} s_{(j_2+1)}}, \ldots, \overline{s_{j_v} s_{(j_v+1)}}$ for some integer value v, such that the predicates $P_3(q_1, \overline{s_{j_1} s_{(j_1+1)}})$, $P_3(q_2, \overline{s_{j_v} s_{(j_v+1)}})$ both evaluate to true and the conjugate

$$\bigwedge_{t=1}^{v-1} P_7(\overline{q_1, q_2}, \overline{s_{j_t} s_{(j_t+1)}}, \overline{s_{j_{t+1}} s_{(j_{t+1}+1)}})$$

evaluates to true.

Assume such a sequence of edges exists. In this case, there exists a sequence of points $p_1, \ldots, p_v$ on the line supporting q, with $p_1 = q_1$, $p_v = q_2$ and such that $p_i \in B_r(s_{j_i}, s_{j_{i+1}})$ (for $1 \leq i < v$) and such that $p_i \in B_r(s_{j_{i-1}}, s_{j_i})$ (for $1 < i \leq v$). Since each stadium is a convex set, it follows that each line segment connecting two consecutive points of this sequence p_i, p_{i+1} is contained in one of the stadiums. Moreover, the curve that is formed by these edges is continuous and contained inside a line and as such the points on the curve form a convex set U. Since q_1 and q_2 are contained in U, it follows that q is contained inside the union of the stadiums and thus $\mathsf{d}_{\overrightarrow{H}}(q, s) \leq r$.

Now, in order to prove the other direction, let us assume that $\mathsf{d}_{\overrightarrow{H}}(q, s) \leq r$. We invoke the observation in [6], restricted in the case of polygonal curves, stating that the directed Hausdorff distance $\mathsf{d}_{\overrightarrow{H}}(q, s)$ is assumed either at some vertex of q or at some intersection point of q with a Voronoi-edge of the Voronoi-diagram of a set of pairwise disjoint line segments representing s. To this end, we split each edge of s that intersects another edge of s at the intersection point in order to obtain a set of pairwise disjoint line segments E' which represent s. The sequence of Voronoi cells of the Voronoi-diagram of E' that are intersected by q, induce a sequence of edges of s with the desired properties. Indeed, the matching induced by the Voronoi diagram is optimal, therefore the corresponding predicates evaluate to true.

In general, for any polygonal curve $q \in \mathbb{W}_k$ with vertices $q_1, \ldots, q_k$, we have that

$$\mathsf{d}_{\overrightarrow{H}}(q, s) \leq r \iff \bigwedge_{i=1}^{k-1} \left[\mathsf{d}_{\overrightarrow{H}}(\overline{q_i q_{i+1}}, s) \leq r\right].$$

Thus, we can apply the arguments above to each edge of q individually. Similarly, we can prove that given the truth values of the predicates P_4, P_8 one can determine whether $\mathsf{d}_{\overrightarrow{H}}(s, q) \leq r$. $\square$

8.6.2 Representation as a range space

We will make use of the following sets, defined in Sections 8.4 and 8.5:

$$P_3^r(q,s), P_4(q,s), P_{d\wedge e}(q,s), P'_{d\wedge e}(q,s), P_h(q,s), P'_h(q,s), P_i(q,s), P'_i(q,s).$$

In addition, we define the following new sets:

- $P_j(q,s) = \{h^+(\overline{q_iq_{i+1}}) \cap V_{RC}(s) \mid \overline{q_iq_{i+1}} \in E(q)\} \cup \{\ell(\overline{q_iq_{i+1}}) \cap V_{RC}(s) \mid \overline{q_iq_{i+1}} \in E(q)\}$, where $h^+(\overline{q_iq_{i+1}})$ denotes the right-side halfspace supported by the directed edge $\overline{q_iq_{i+1}}$ and
$$V_{RC}(s) = \bigcup_{\substack{e\in E(s)\\ p\in V(s)}} \partial R_r(e) \cap C_r(p),$$
- $P'_j(q,s) = \{h^+(\overline{s_is_{i+1}}) \cap V_{RC}(q) \mid \overline{s_is_{i+1}} \in E(s)\} \cup \{\ell(\overline{s_is_{i+1}}) \cap V_{RC}(q) \mid \overline{s_is_{i+1}} \in E(s)\}$, where $h^+(\overline{s_is_{i+1}})$ denotes the right-side halfspace supported by the directed edge $\overline{s_i, s_{i+1}}$ and
$$V_{RC}(q) = \bigcup_{\substack{e\in E(q)\\ p\in V(q)}} \partial R_r(e) \cap C_r(p),$$
- $P_k(q,s) = \{h^+(\overline{q_iq_{i+1}}) \cap V_{RR}(s) \mid \overline{q_iq_{i+1}} \in E(q)\} \cup \{\ell(\overline{q_iq_{i+1}}) \cap V_{RC}(s) \mid \overline{q_iq_{i+1}} \in E(q)\}$, where $h^+(\overline{q_iq_{i+1}})$ denotes the right-side halfspace supported by the directed edge $\overline{q_iq_{i+1}}$ and
$$V_{RR}(s) = \bigcup_{\substack{e_1,e_2\in E(s)\\ e_1\neq e_2}} \partial R_r(e_1) \cap \partial R_r(e_2),$$
- $P'_k(q,s) = \{h^+(\overline{s_is_{i+1}}) \cap V_{RR}(q) \mid \overline{s_is_{i+1}} \in E(s)\} \cup \{\ell(\overline{s_is_{i+1}}) \cap V_{RC}(q) \mid \overline{s_is_{i+1}} \in E(s)\}$, where $h^+(\overline{s_is_{i+1}})$ denotes the right-side halfspace supported by the directed edge $\overline{s_i, s_{i+1}}$ and
$$V_{RR}(q) = \bigcup_{\substack{e_1,e_2\in E(q)\\ e_1\neq e_2}} \partial R_r(e_1) \cap \partial R_r(e_2),$$

where $R_r(\overline{st}) = \text{conv}\{s-u, s+u, t+u, t-u\}$ and $u \in \mathbb{R}^2$ s.t. $\langle t-s, u\rangle = 0$ and $\|u\|_2 = r$.

Lemma 108. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (j) over all possible inputs $\overline{q_iq_{i+1}} \in E(q)$, $\overline{s_js_{j+1}} \in E(s)$, $s_t \in V(s)$ are uniquely defined by the sets $P_{d\wedge e}(q,s)$, $P_j(q,s)$.*

Proof. Let a, b be the two intersection points. The line $\ell(\overline{q_i, q_{i+1}})$ passes between a and b iff one of the supporting halfspaces contains only one of them. If the line passes between the two intersection points of $\partial R_r(\overline{s_js_{j+1}}) \cap C_r(s_t)$, then the predicate returns true. Now if the line does not pass between the two intersection points, then $\ell(\overline{q_iq_{i+1}}) \cap R_r(\overline{s_js_{j+1}}) \cap D_r(s_t) \neq \emptyset$ iff $s_t \in L_r(\overline{q_iq_{i+1}})$ and $[a \in h^+(\overline{q_iq_{i+1}}) \wedge b \in h^+(\overline{q_iq_{i+1}})] \vee [a \notin h^+(\overline{q_iq_{i+1}}) \wedge b \notin h^+(\overline{q_iq_{i+1}})]$. If there is just one intersection point, it suffices to check whether $\ell(\overline{q_iq_{i+1}})$ intersects that point. $\square$

Lemma 109. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (j) over all possible inputs $\overline{s_i s_{i+1}} \in E(s)$, $\overline{q_j q_{j+1}} \in E(q)$, $q_t \in V(q)$ are uniquely defined by the sets $P'_{d \wedge e}(q,s)$, $P'_j(q,s)$.*

Proof. The statement follows by the same arguments which were used in the proof of Lemma 108. □

Lemma 110. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (k) over all possible inputs $\overline{q_i q_{i+1}} \in E(q)$, $\overline{s_j s_{j+1}} \in E(s)$, $\overline{s_t s_{t+1}} \in E(s)$ are uniquely defined by the set $P_k(q,s)$.*

Proof. Suppose that $|\partial R_r(\overline{s_j s_{j+1}}) \cap \partial R_r(\overline{s_t, s_{t+1}})| > 1$. The intersection $R_r(\overline{s_j s_{j+1}}) \cap R_r(\overline{s_t, s_{t+1}})$ defines a convex polygon and the line $\ell(\overline{q_i, q_{i+1}})$ intersects it iff there exist two points $a, b \in \partial R_r(\overline{s_j s_{j+1}}) \cap \partial R_r(\overline{s_t, s_{t+1}})$ which are separated by $h^+(\overline{q_i, q_{i+1}})$. If $|\partial R_r(\overline{s_j s_{j+1}}) \cap \partial R_r(\overline{s_t, s_{t+1}})| = 1$, then it suffices to check whether the line $\ell(\overline{q_i, q_{i+1}})$ intersects that point. □

Lemma 111. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (k) over all possible inputs $\overline{s_i s_{i+1}} \in E(s)$, $\overline{q_j q_{j+1}} \in E(q)$, $\overline{q_t q_{t+1}} \in E(q)$ are uniquely defined by the set $P'_k(q,s)$.*

Proof. The statement follows by the same arguments which were used in the proof of Lemma 110. □

We repeat the following lemma from [3].

Lemma 112 (Lemma 14, [3]). *If and only if $h \vee (d \wedge e \wedge i)$ evaluates to true, then the line ℓ intersects the lens formed by the two disks of radius r at a_1 and a_2.*

Lemma 113. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (l) over all possible inputs $\overline{q_i q_{i+1}} \in E(q)$, $s_j \in V(s)$, $s_t \in V(s)$ are uniquely defined by the sets $P_{d \wedge e}(q,s), P_h(q,s), P_i(q,s)$.*

Proof. Predicate (l) is equivalent to $h \vee (d \wedge e \wedge i)$, according to Lemma 112. □

Lemma 114. *Let s be a polygonal curve in $\mathbb{W}_m$ and q a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The truth values for predicate (l) over all possible inputs $\overline{s_i s_{i+1}} \in E(s)$, $q_j \in V(q)$, $q_t \in V(q)$ are uniquely defined by the sets $P'_{d \wedge e}(q,s), P'_h(q,s), P'_i(q,s)$.*

Proof. Predicate (l) is equivalent to $h \vee (d \wedge e \wedge i)$, according to Lemma 112. □

Lemma 115. *Let s be a polygonal curve in $\mathbb{W}_m$ and q be a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The following sets are sufficient in order to determine whether $\mathrm{d}_{\overrightarrow{H}}(q,s) \leq r$ or $\mathrm{d}_{\overrightarrow{H}}(q,s) > r$:*

$$P_3^r(q,s), P_{d \wedge e}(q,s), P_h(q,s), P_i(q,s), P_j(q,s), P_k(q,s).$$

Proof. Lemmas 107, 108, 110, 113 imply the statement. □

Lemma 116. *Let s be a polygonal curve in $\mathbb{W}_m$ and q be a polygonal curve in $\mathbb{W}_k$. Fix any $r \geq 0$. The following sets are sufficient in order to determine whether $\mathrm{d}_{\overrightarrow{H}}(s,q) \leq r$ or $\mathrm{d}_{\overrightarrow{H}}(s,q) > r$:*

$$P_4^r(q,s), P'_{d\wedge e}(q,s), P'_h(q,s), P'_i(q,s), P'_j(q,s), P'_k(q,s).$$

Proof. Lemmas 107, 109, 111, 114 imply the statement. □

8.6.3 VC dimension bounds

Theorem 117. *Let $\mathcal{R}_H^r$ be the set of all balls of radius r, under the directed Hausdorff distance from polygonal curves in $\mathbb{W}_k$. The VC dimension of $(\mathbb{W}_m, \mathcal{R}_H^r)$ is $O(k\log(km))$.*

Proof. Let $S \subset \mathbb{W}_m$ be a set of t polygonal curves and let $q \in \mathbb{W}_k$. By Lemma 115, the set $\{s \in S \mid \mathrm{d}_{\overrightarrow{H}}(q,s) \leq r\}$ is uniquely defined by the sets:

$$\bigcup_{s\in S} P_3^r(q,s), \bigcup_{s\in S} P_{d\wedge e}(q,s), \bigcup_{s\in S} P_h(q,s), \bigcup_{s\in S} P_i(q,s), \bigcup_{s\in S} P_j(q,s), \bigcup_{s\in S} P_k(q,s).$$

For any $s \in S$, recall that $V_{RC}(s)$ is the set of points which belong to all possible intersections formed by rectangles centered at edges in $E(s)$ and circles of radius r centered at points in $V(s)$. Formally,

$$V_{RC}(s) = \bigcup_{\substack{e\in E(s)\\ p\in V(s)}} R_r(e) \cap C_r(p),$$

where $R_r(\overline{st}) = \mathrm{conv}\{s-u, s+u, t+u, t-u\}$ and $u \in \mathbb{R}^2$ s.t. $\langle t-s, u\rangle = 0$ and $\|u\|_2 = r$. Let $V_{RC}(S) = \bigcup_{s\in S} V_{RC}(s)$. Notice that $|V_{RC}(S)| = tm^{O(1)}$. We need to bound the number of different sets

$$\{h^+(q_i, q_{i+1}) \cap V_{RC}(S) \mid \overline{q_i q_{i+1}} \in E(q)\}$$

over all possible $q \in \mathbb{W}_k$, where $h^+(q_i, q_{i+1})$ defines either one of the two halfspaces defined by points q_i, q_{i+1}. The shattering dimension of the range space of points and halfspaces is $O(1)$, hence we get an upper bound of $(tm)^{O(k)}$.

Now, for any $s \in S$, recall that $V_{RR}(s)$ is the set of points which belong to all possible intersections formed by two rectangles centered at different edges in $E(s)$. Formally,

$$V_{RR}(s) = \bigcup_{\substack{e_1,e_2\in E(s)\\ e_1\neq e_2}} R_r(e_1) \cap R_r(e_2).$$

Similarly, we get an upper bound of $(tm)^{O(k)}$ on the number of different sets

$$\{h^+(q_i, q_{i+1}) \cap V_{RR}(S) \mid \overline{q_i q_{i+1}} \in E(q)\}$$

over all possible $q \in \mathbb{W}_k$. It remains to reclaim, as we did in the proof of Theorem 106, that the number of all possible sets $\bigcup_{s\in S} P_3^r(q,s)$, $\bigcup_{s\in S} P_{d\wedge e}(q,s)$, $\bigcup_{s\in S} P_h(q,s)$, $\bigcup_{s\in S} P_i(q,s)$ is bounded by $(tm)^{O(k)}$. Hence, the VC dimension of this range space is $O(k\log(km))$.

□

Theorem 118. *Let $\mathcal{R}_H^r$ be the set of all balls of radius r, under the directed Hausdorff distance to polygonal curves in $\mathbb{W}_k$. The VC dimension of $(\mathbb{W}_m, \mathcal{R}_H^r)$ is $O(k^2 \log(km))$.*

Proof. We able to follow the same analysis as in the proof of Theorem 117. However, notice that $|V_{RC}(q)| = O(k^2)$, and $|V_{RR}(q)| = O(k^2)$. Due to Lemma 116, we can employ similar arguments to the ones we used in the proof of Theorem 117, now for the dual range space of the points-halfspaces range space, and for the sets $\bigcup_{s \in S} P_4^r(q,s)$, $\bigcup_{s \in S} P'_{d \wedge e}(q,s)$, $\bigcup_{s \in S} P'_h(q,s)$, $\bigcup_{s \in S} P'_i(q,s)$ imply that the VC dimension of this range space is $O(k^2 \log(km))$. □

Theorem 119. *Let $\mathcal{R}_H^r$ be the set of all balls of radius r, under the symmetric Hausdorff distance in $\mathbb{W}_k$. The VC dimension of $(\mathbb{W}_m, \mathcal{R}_H^r)$ is $O(k^2 \log(km))$.*

Proof. Lemmas 115 and 115 imply that the set $\{s \in S \mid \mathsf{d}_H(q,s) \leq r\}$ is uniquely defined by the sets:

$$\bigcup_{s \in S} P_3^r(q,s), \bigcup_{s \in S} P_{d \wedge e}(q,s), \bigcup_{s \in S} P_h(q,s), \bigcup_{s \in S} P_i(q,s), \bigcup_{s \in S} P_j(q,s), \bigcup_{s \in S} P_k(q,s),$$

and

$$\bigcup_{s \in S} P_4^r(q,s), \bigcup_{s \in S} P'_{d \wedge e}(q,s), \bigcup_{s \in S} P'_h(q,s), \bigcup_{s \in S} P'_i(q,s), \bigcup_{s \in S} P'_j(q,s), \bigcup_{s \in S} P'_k(q,s).$$

Now bounding the number of all possible such sets, as we did in the proofs of Theorems 117 and 118, implies the statement. □

8.7 The discrete case in higher dimensions

In the following sections we focus on Euclidean spaces of higher dimension ($d > 2$) being the ambient space of the curves of the ground set. In this section we discuss our bounds in the discrete setting. Let $\mathbb{X}_m^d = \left(\mathbb{R}^d\right)^m$; we treat the elements of this set as ordered sets of points in $\mathbb{R}^d$ of size m.

Theorem 120. *Let $(\mathbb{X}_m^d, \mathcal{R}_{H,k})$ be the range space with $\mathcal{R}_{H,k}$ the set of all balls under the Hausdorff distance centered at sets in $\mathbb{X}_k^d$. The VC dimension is $O\left(kd \log(kdm)\right)$.*

Proof. The proof is similar to the one from Theorem 93. We are able to extend it to higher dimensions by making use of known bounds for balls in any dimension instead of just disks. Let $\{S_1, \ldots, S_t\} \subseteq \mathbb{X}_m$ and $S = \bigcup_i S_i$; we define S so that it ignores the ordering with each S_i and is a single set of size tm. Any intersection of a Hausdorff ball with $\{S_1, \ldots, S_t\}$ is uniquely defined by a set $\{D_1^d \cap S, \ldots, D_k^d \cap S\}$, where $D_1^d, \ldots, D_k^d$ are balls in $\mathbb{R}^d$.

Consider the range space $(\mathbb{R}^d, \mathcal{D}_d)$, where $\mathcal{D}_d$ is the set of balls in $\mathbb{R}^d$. We know that the VC dimension is $d+1$. Hence, since the shattering dimension is upper bounded by the VC dimension,

$$\max_{S \subseteq \mathbb{R}^2, |S| = tm} |\mathcal{D}_{|S}| = O((tm)^{d+1}).$$

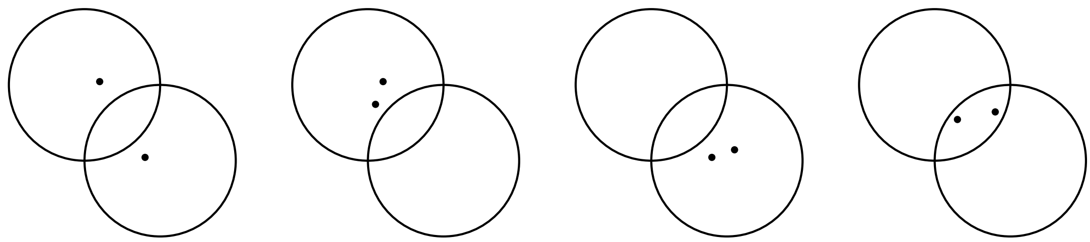

Figure 8.2: The lower bound for $(\mathbb{X}_1, \mathcal{R}_{dF,2})$. The two disks correspond to the two polygonal curves of the ground set. The set of these two polygonal curves is shattered by $\mathcal{R}_{dF,2}$.

This implies that $|\{\{D_1^d \cap S, \ldots, D_k^d \cap S\} \mid D_1^d, \ldots, D_k^d$ are balls in $\mathbb{R}^d\}| \leq O((tm)^{(d+1)k})$, and hence,

$$2^t \leq 2^{O(dk \log(tm))} \implies t = O\left(dk \log(dkm)\right). \qquad \square$$

Theorem 121. *Let $(\mathbb{X}_m, \mathcal{R}_{dF,k})$ be the range space with $\mathcal{R}_{dF,k}$ the set of all balls under the Discrete Fréchet distance centered at polygonal curves in $\mathbb{X}_k$. The VC dimension is $O\left(kd \log(kdm)\right)$.*

Proof. Similar to the proof of Theorem 120. The only difference is that, as with the proof of Theorem 94, we need to bound the number of sequences $D_1^d \cap S, \ldots, D_k^d \cap S$, which is also $O((tm)^{d+1})$. $\square$

8.8 Lower bounds

We now state the lower bounds. We denote by $\mathcal{R}_{dF,k}$ be the set of all balls, under the Discrete Fréchet distance, centered at polygonal curves in $\mathbb{X}_k$. We also denote by $\mathcal{R}_{wF,k}$, $\mathcal{R}_{F,k}$, $\mathcal{R}_{H,k}$, the sets of all balls, under the Weak Fréchet distance, under the Fréchet distance and under the Hausdorff distance respectively, where balls are centered at polygonal curves in $\mathbb{W}_k$.

We start with a weaker result about Discrete Fréchet distance, that will be easier to extend to continuous metrics.

Lemma 122. *Let $\mathcal{R}_{dF,k}$ be the set of all balls, under the Discrete Fréchet distance, centered at polygonal curves in $\mathbb{X}_k$. The VC-dimension of the range space $(\mathbb{X}_m, \mathcal{R}_{dF,k})$ is $\geq k$.*

Proof. We will show that there exists a configuration S of k polygonal curves of complexity $m = 1$, i.e. points in $\mathbb{R}^2$, which are shattered by Discrete Fréchet balls centered at polygonal curves of complexity k. Consider k disks $D_1, \ldots, D_k$ centered at the k polygonal curves of S and let $p_1, \ldots, p_k$ be the vertices of the polygonal curve which is the center of the Discrete Fréchet ball. Any intersection between a Discrete Fréchet ball and the set of polygonal curves is defined by the disks which are commonly stabbed by all points $p_1, \ldots, p_k$.

First, we will show that there exists a configuration of disks $D_1, \ldots, D_k$ such that:

$$\text{area}\left(\bigcap_{i=1}^{k} D_i\right) > 0,$$

$$\bigcap_{\substack{i \neq j \\ i=1,\ldots,k}} D_i \neq \bigcap_{i=1}^{k} D_i \qquad \forall j \in [k]$$

$$\text{area}\left(\bigcap_{\substack{i \neq j \\ i=1,\ldots,k}} D_i\right) > 0 \qquad \forall j \in [k].$$

We can easily prove this by induction: two disks can be placed in a way that area$(D_1 \cap D_2) > 0$, $D_1 \neq D_2$. Now consider t disks $D_1, \ldots, D_t$ which satisfy the induction hypothesis. Since area $\left(\bigcap_{i=1}^{t} D_i\right) > 0$, we can simply place a disk D_{t+1} such that its boundary ∂D_{t+1} halves area $\left(\bigcap_{i=1}^{t} D_i\right)$.

Then, the set S of polygonal curves which consists of the k centers of the disks $D_1, \ldots, D_k$ is shattered as follows: each point p_j either stabs $\bigcap_{i=1}^{k} D_i$ or it stabs $\left(\bigcap_{i \neq j, i \in [k]} D_i\right) \setminus D_j$ and hence the corresponding polygonal curve either belongs to the intersection of the set of polygonal curves with the Discrete Fréchet ball or not. The simple case $k = 2$ is depicted in Figure 8.2. □

However, we can strengthen this bound for this distance measure.

Lemma 123. *Let $\mathcal{R}_{dF,k}$ be the set of all balls, under the Discrete Fréchet distance, centered at polygonal curves in $\mathbb{X}_k$. The VC-dimension of the range space $(\mathbb{X}_m, \mathcal{R}_{dF,k})$ is $\Omega(k \log k)$.*

Proof. We will show that there exists a configuration S of $\kappa = \Omega(k \log k)$ polygonal curves of complexity $m = 1$, i.e. points in $\mathbb{R}^2$, which are shattered by Discrete Fréchet balls centered at polygonal curves of complexity k. Consider k disks $D_1, \ldots, D_\kappa$ centered at the κ polygonal curves of S and let $p_1, \ldots, p_k$ be the vertices of the polygonal curve which is the center of the Discrete Fréchet ball. Any intersection between a Discrete Fréchet ball and the set of polygonal curves is defined by the disks which are commonly stabbed by all points $p_1, \ldots, p_k$.

We now show this result by reducing to a recent lower bound of Csikos *et al.* [31] which gave an $\Omega(k \log k)$ lower bound for a related range space. This is defined on a ground set $P \subset \mathbb{R}^2$ with ranges $\mathcal{R}_k$ defined so each range $R \in \mathcal{R}_k$ is the intersection of k halfspaces. The first step is to observe that we can set r sufficiently large so that with respect to all $p_1, \ldots, p_k$ we consider each disk D_j has the same inclusion properties as some halfspace H_j. That is, we now need to show a set of κ halfspaces can be shattered by a set of k points, where a ground set object H_j is contained in the range defined by those k points if it includes all of them.

The second step is to observe that the standard point-line duality transforms this problem into the one considered by Csikos *et al.*. Under this transform a dual point q_j (corresponding to primal halfspace H_j) is contained in a dual halfspace h_i (corresponding to primal point p_i). Thus the primal halfspace H_j is contained in the range defined by the k points $p_1, \ldots, p_k$ if and only if its dual representation, the point q_j, is contained in all of the halfspaces $h_1, \ldots, h_k$ which are the dual representations of the points $p_1, \ldots, p_k$.

Finally, the lower bound by Csikos *et al.* [31] shows that there exist a set of $\kappa = \Omega(k \log k)$ points q_j which can be shattered by such ranges. $\square$

Lemma 124. *Let $\mathcal{R}_{dF}$ be the set of all balls, under the Discrete Fréchet distance, centered at polygonal curves in $\mathbb{X}_k$. The VC dimension of the range space $(\mathbb{X}_m, \mathcal{R}_{dF})$ is $\Omega(\log m)$.*

Proof. Theorem 122 and [50, Lemma 5.18], which bounds the VC dimension of the dual range space as a function of the VC dimension of the primal space, imply the theorem. $\square$

The following constructions also works directly for the discrete case of the Hausdorff distance. We conjecture that they can also be extended for the weak Fréchet, Fréchet, and Hausdorff for continuous curves, but do not have a complete proof. We can however extend the weaker bound in Theorem 122. We denote by $\mathcal{R}_{wF,k}$, $\mathcal{R}_{F,k}$, $\mathcal{R}_{H,k}$, the sets of all balls, under the Weak Fréchet distance, under the Fréchet distance and under the Hausdorff distance respectively, where balls are centered at polygonal curves in $\mathbb{W}_k$.

Lemma 125. *The VC-dimension of the range spaces $(\mathbb{W}_m, \mathcal{R}_{wF,3k})$, $(\mathbb{W}_m, \mathcal{R}_{F,3k})$, and $(\mathbb{W}_m, \mathcal{R}_{H,3k})$ is $\geq k$.*

Proof. Consider the case $m = 1$, that is X consisting of all polygonal curves with 1 vertex. We place k polygonal curves as in the proof of Thm. 122. Now, consider the corresponding disks $D_1, \ldots, D_k$. The continuous Fréchet balls of complexity $3k$ shatter X as follows: let $3k$ points $p_1, \ldots, p_k$, $q_1, \ldots, q_k$, $p'_1, \ldots, p'_k$ s.t. for any $j \in [k]$, $p_j, p'_j \in \left(\bigcap_{i=1}^k D_i\right) \cap \partial D_j$. For each $i \in [k]$, we have a segment $\overline{p_i q_i}$, a segment $\overline{q_i p'_i}$ and for any $i \in [k-1]$, we have segments $\overline{p'_i p_{i+1}}$. Then, either $q_j \in \bigcap_{i=1}^k D_i$ or $q_j \in \left(\bigcap_{i \neq j, i \in [k]} D_i\right) \setminus D_j$ which determines whether the continuous Fréchet ball covers the jth polygonal curve. Notice that if $q_j \in \bigcap_{i=1}^k D_i$ then the segments $\overline{p_j q_j}$, $\overline{q_j p'_j}$ lie inside $\bigcap_{i=1}^k D_i$ due to convexity. Similarly, if $q_j \in \left(\bigcap_{i \neq j, i \in [k]} D_i\right) \setminus D_j$ then the segments $\overline{p_j q_j}$, $\overline{q_j p'_j}$ lie inside $\bigcap_{i \neq j, i \in [k]}^k D_i$. $\square$

Lemma 126. *Let $\mathcal{R}_F$ be the set of all balls, under the Fréchet distance, centered at polygonal curves in $\mathbb{W}_k$. The VC dimension of the range space $(\mathbb{W}_m, \mathcal{R}_{wF,k})$, $(\mathbb{W}_m, \mathcal{R}_{F,k})$, $(\mathbb{W}_m, \mathcal{R}_{H,k})$ is $\Omega(\log m)$.*

Proof. Theorem 125 and [50, Lemma 5.18], which bounds the VC dimension of the dual range space as a function of the VC dimension of the primal space, imply the theorem. $\square$

Theorem 127. *The VC-dimension of the range spaces $(\mathbb{X}_m, \mathcal{R}_{dF,k})$, and $(\mathbb{X}_m, \mathcal{R}_{H,k})$ is $\Omega(\max(k \log k, \log m))$, and for $(\mathbb{W}_m, \mathcal{R}_{wF,k})$, $(\mathbb{W}_m, \mathcal{R}_{F,k})$, and $(\mathbb{W}_m, \mathcal{R}_{H,k})$ is $\Omega(\max(k, \log m))$.*

Proof. The statement essentially combines Lemmas 125, 123, 126, 122 and 124. □

ABBREVIATIONS - ACRONYMS

ΕΚΠΑ	Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
NKUA	National and Kapodistrian University of Athens
ΠΙΣ	Πανεπιστήμιο του Ιλλινόις στο Σικάγο
UIC	University of Illinois at Chicago
ΕΜΠ	Εθνικό Μετσόβιο Πολυτεχνείο
NTUA	National Technical University of Athens
JL	Johnson-Lindenstrauss
DFD	Discrete Fréchet Distance
DTW	Dynamic Time Warping
LSH	Locality Sensitive Hashing
VC	Vapnik–Chervonenkis

REFERENCES

[1] Ittai Abraham, Yair Bartal, T-H. Hubert Chan, Kedar Dhamdhere Dhamdhere, Anupam Gupta, Jon Kleinberg, Ofer Neiman, and Aleksandrs Slivkins. Metric embeddings with relaxed guarantees. In *Proc. of the 46th Annual IEEE Symp. on Foundations of Computer Science*, FOCS '05, pages 83–100, Washington, DC, USA, 2005. IEEE Computer Society.

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

[3] Peyman Afshani and Anne Driemel. On the complexity of range searching among curves. *CoRR*, arXiv:1707.04789v1, 2017.

[4] Peyman Afshani and Anne Driemel. On the complexity of range searching among curves. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 898–917, 2018.

[5] Josh Alman, Timothy M. Chan, and Ryan Williams. Polynomial representations of threshold functions and algorithmic applications. In *IEEE Symposium on Foundations of Computer Science (FOCS), New Brunswick, NJ, USA*, pages 467–476, 2016.

[6] Helmut Alt, Bernd Behrends, and Johannes Blömer. Approximate matching of polygonal shapes. *Annals of Mathematics and Artificial Intelligence*, 13(3):251–265, Sep 1995.

[7] Helmut Alt and Michael Godau. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 05:75–91, 1995.

[8] 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.

[9] Alexandr Andoni and Piotr Indyk. Efficient algorithms for substring near neighbor problem. In *Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA), Miami, Florida*, pages 1203–1212, 2006.

[10] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, 2008.

[11] 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.

[12] 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.

[13] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[14] Sunil Arya, Guilherme D. da Fonseca, and David M. Mount. Approximate polytope membership queries. In *Proc. 43rd Annual ACM Symp. Theory of Computing*, STOC'11, pages 579–586, 2011.

[15] Sunil Arya, Theocharis Malamatos, and David M. Mount. Space-time tradeoffs for approximate nearest neighbor searching. *J. ACM*, 57(1):1:1–1:54, 2009.

[16] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, 1998.

[17] Sunil Arya, David M. Mount, Antoine Vigneron, and Jian Xia. Space-time tradeoffs for proximity searching in doubling spaces. In *Algorithms - ESA 2008, 16th Annual European Symposium, Karlsruhe, Germany, September 15-17, 2008. Proceedings*, pages 112–123, 2008.

[18] Maria Astefanoaei, Paul Cesaretti, Panagiota Katsikouli, Mayank Goswami, and Rik Sarkar. Multi-resolution sketches and locality sensitive hashing for fast trajectory processing. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2018)*, volume 10, 2018.

[19] 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.

[20] Julian Baldus and Karl Bringmann. A fast implementation of near neighbors queries for Fréchet distance (GIS Cup). In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'17, pages 99:1–99:4, 2017.

[21] Yair Bartal and Lee-Ad Gottlieb. Approximate nearest neighbor search for ℓ_p-spaces ($2 < p < \infty$) via embeddings. *CoRR*, abs/1512.01775, 2015.

[22] Yair Bartal and Lee-Ad Gottlieb. Dimension reduction techniques for ℓ_p, (1<p<2), with applications. In *32nd International Symposium on Computational Geometry, SoCG 2016, June 14-18, 2016, Boston, MA, USA*, pages 16:1–16:15, 2016.

[23] Yair Bartal, Ben Recht, and Leonard J. Schulman. Dimensionality reduction: Beyond the johnson-lindenstrauss bound. In *Proc. of the 22nd Annual ACM-SIAM Symp. on Discrete Algorithms*, SODA '11, pages 868–887. SIAM, 2011.

[24] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *Proc. 23rd Intern. Conf. Machine Learning*, ICML'06, pages 97–104, 2006.

[25] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[26] Hervé Brönnimann and Michael T. Goodrich. Almost optimal set covers in finite VC-dimension. *Discrete & Computational Geometry*, 1995.

[27] Kevin Buchin, Yago Diez, Tom van Diggelen, and Wouter Meulemans. Efficient trajectory queries under the Fréchet distance (GIS Cup). In *Proc. 25th Int. Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 101:1–101:4, 2017.

[28] Moses Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. 34th Annual ACM Symposium on Theory of Computing, 2002, Montréal, Québec, Canada*, pages 380–388, 2002.

[29] Bernard Chazelle and Emo Welzl. Quasi-optimal range searching in spaces of finite VC-dimension. *Discrete and Computational Geometry*, 4:467–489, 1989.

[30] Richard Cole and Lee-Ad Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing*, STOC '06, pages 574–583, New York, NY, USA, 2006. ACM.

[31] Monika Csikos, Andrey Kupavskii, and Nabil H. Mustafa. Optimal bounds on the VC-dimension. *arXiv:1807.07924*, 2018.

[32] Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proc. 40th Annual ACM Symp. Theory of Computing*, STOC'08, pages 537–546, 2008.

[33] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. 20th Annual Symp. Computational Geometry*, SCG'04, pages 253–262, 2004.

[34] Mark De Berg, Atlas F Cook, and Joachim Gudmundsson. Fast Fréchet queries. 46(6):747–755, 2013.

[35] Mark de Berg and Ali D. Mehrabi. Straight-path queries in trajectory data. In *WALCOM: Algorithms and Computation - 9th Int. Workshop, WALCOM 2015, Dhaka, Bangladesh, February 26-28, 2015. Proceedings*, pages 101–112, 2015.

[36] 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.

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

[38] Anne Driemel and Francesco Silvestri. Locally-sensitive hashing of curves. In *33st International Symposium on Computational Geometry, SoCG 2017*, pages 37:1–37:16, 2017.

[39] Fabian Dütsch and Jan Vahrenhold. A filter-and-refinement- algorithm for range queries based on the Fréchet distance (GIS Cup). In *Proc. 25th Int. Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 100:1–100:4, 2017.

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

[41] 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.

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

[43] Arnold Filtser, Omrit Filtser, and Matthew J. Katz. Approximate nearest neighbor for curves - simple, efficient, and deterministic. *CoRR*, abs/1902.07562, 2019.

[44] Alexander Gilbers and Rolf Klein. A new upper bound for the VC-dimension of visibility regions. *Computational Geometry: Theory and Applications*, 74:61–74, 2014.

[45] Lee-Ad Gottlieb and Robert Krauthgamer. A nonlinear approach to dimension reduction. *Discrete & Computational Geometry*, 54(2):291–315, 2015.

[46] Joachim Gudmundsson and Michael Horton. Spatio-temporal analysis of team sports. *ACM Comput. Surv.*, 50(2):22:1–22:34, April 2017.

[47] Joachim Gudmundsson and Michiel Smid. Fast algorithms for approximate Fréchet matching queries in geometric trees. *Computational Geometry*, 48(6):479 – 494, 2015.

[48] Anupam Gupta, Robert Krauthgamer, and James R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proc. 44th Annual IEEE Symp. Foundations of Computer Science*, FOCS'03, pages 534–541, 2003.

[49] Sariel Har-Peled. Clustering motion. *Discrete & Computational Geometry*, 31(4):545–565, 2004.

[50] Sariel Har-Peled. *Geometric Approximation Algorithms*. American Mathematical Society, Boston, MA, USA, 2011.

[51] 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.

[52] 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.

[53] Sariel Har-Peled and Manor Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comput.*, 35(5):1148–1184, 2006.

[54] Sariel Har-Peled and Benjamin Raichel. Net and prune: A linear time algorithm for euclidean distance problems. *J. ACM*, 62(6):44, 2015.

[55] 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.

[56] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[57] 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.

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

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

[60] Sarang Joshi, Raj Varma Kommaraju, Jeff M. Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *ACM SoCG*, 2011.

[61] David R. Karger and Matthias Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proc. 34th Annual ACM Symp. Theory of Computing*, STOC'02, pages 741–750, 2002.

[62] Marek Karpinski and Angus Macintyre. Polynomial bounds for vc dimension of sigmoidal neural networks. In *STOC*, 1995.

[63] Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *Proc. 15th Annual ACM-SIAM Symp. Discrete Algorithms*, SODA'04, pages 798–807, 2004.

[64] Elmar Langetepe and Simone Lehmann. Exact VC-dimension for L1-visibility of points in simple polygons. *arXiv:1705.01723*, 2017.

[65] J. K. Laurila, Daniel Gatica-Perez, I. Aad, Blom J., Olivier Bornet, Trinh-Minh-Tri Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, 2012.

[66] J.R. Lee, M. Mendel, and A. Naor. Metric structures in L_1: dimension, snowflakes, and average distortion. *Eur. J. Comb.*, 26(8):1180–1190, 2005.

[67] Jiri Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, September 2008.

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

[69] S. Meiser. Point location in arrangements of hyperplanes. *Inf. Comput.* , 106(2):286–303, 1993.

[70] Michael Mitzenmacher and Eli Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[71] Ryan O'Donnell, Yi Wu, and Yuan Zhou. Optimal lower bounds for locality-sensitive hashing (except when q is tiny). *ACM Trans. Comput. Theory*, 6(1):5:1–5:13, March 2014.

[72] An Elementary Proof of a Theorem of Johnson and Lindenstrauss. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 2003.

[73] Rina Panigrahy. Entropy based nearest neighbor search in high dimensions. In *Proc. 17th Annual ACM-SIAM Symp. Discrete Algorithms*, SODA'06, pages 1186–1195, 2006.

[74] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

[75] Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1), 1972.

[76] Yifang Sun, Wei Wang, Jianbin Qin, Ying Zhang, and Xuemin Lin. Srs: Solving c-approximate nearest neighbor queries in high dimensional euclidean space with a tiny index. *Proc. VLDB Endow.*, 8(1):1–12, September 2014.

[77] G. Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *J. ACM*, 62(2):13, 2015.

[78] Pavel Valtr. Guarding galleries where no point sees a small area. *Israel Journal of Mathematics*, 104:1–16, 1998.

[79] Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[80] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[81] Santosh Vempala. Randomly-oriented k-d trees adapt to intrinsic dimension. In *Proc. Foundations of Software Technology & Theor. Computer Science*, pages 48–57, 2012.

[82] Martin Werner and Dev Oliver. ACM SIGSPATIAL GIS Cup 2017: Range queries under Fréchet distance. *SIGSPATIAL Special*, 10(1):24–27, June 2018.

[83] Feng Zheng and Thomas Kaiser. *Digital Signal Processing for RFID*. Wiley, 2016.