Dimensionality reduction for approximate near neighbor search in the Manhattan metric

Vasileios Margonis
R.N. A0005

Examination committee:
Ioannis Emiris, Department of Informatics & Telecommunications, NKUA.
Aris Pagourtzis, School of Electrical & Computer Engineering, NTUA.
Dimitris Fotakis, School of Electrical & Computer Engineering, NTUA.

Supervisor:
Ioannis Emiris, Professor, Department of Informatics & Telecommunications, NKUA.

April, 2019

Η παρούσα Διπλωματική Εργασία
εκπονήθηκε στα πλαίσια των σπουδών
για την απόκτηση του
Μεταπτυχιακού Διπλώματος Ειδίκευσης
«Αλγόριθμοι, Λογική και Διακριτά Μαθηματικά»
που απονέμει το
Τμήμα Πληροφορικής και Τηλεπικοινωνιών
του
Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την 11η Απριλίου 2019 από Εξεταστική Επιτροπή
αποτελούμενη από τους:

Ονοματεπώνυμο	**Βαθμίδα**
1. Ιωάννης Εμίρης	Καθηγητής
2. Άρης Παγουρτζής	Αναπλ. καθηγητής
3. Δημήτρης Φωτάκης	Αναπλ. καθηγητής

ABSTRACT

The approximate nearest neighbor problem is one of the fundamental problems in computational geometry and has received much attention during the past decades. Efficient and practical algorithms are known for data sets of low dimension. However, modern, high-dimensional data cannot be handled by these algorithms, because of the so called "curse of dimensionality". A new theory for approximate nearest neighbors in high dimensions emerged with an influential paper by Indyk and Motwani, in 1998, yielding algorithms that depend polynomially on the dimension.

Nevertheless, is has been realized that designing efficient ANN data structures is closely related with dimension-reducing embeddings. One popular dimension reduction technique is randomized projections. Starting with the celebrated Johnson-Lindenstrauss Lemma, such projections have been studied in depth for the Euclidean (ℓ_2) metric and, much less, for the Manhattan (ℓ_1) metric. In 2007, Indyk and Naor, in the context of approximate nearest neighbors, introduced the notion of nearest neighbor-preserving embeddings. These are randomized embeddings between two metric spaces with guaranteed bounded distortion only for the distances between a query point and a point set. Such embeddings are known to exist for both ℓ_2 and ℓ_1 metrics, as well as for doubling subsets of ℓ_2.

In this thesis, we consider the approximate nearest neighbor problem in doubling subsets of ℓ_1. We exploit the decision-with-witness version, called approximate *near* neighbor, which incurs a roughly logarithmic overhead, and we propose a dimension reducing, *near* neighbor-preserving embedding for doubling subsets of ℓ_1. Our approach is to represent the point set with a carefully chosen covering set, and then apply a random linear projection to that covering set, using a matrix of Cauchy random variables. We study two cases of covering sets: approximate nets and randomly shifted grids, and we discuss the differences between them in terms of computing time, target dimension, as well as their algorithmic implications.

Keywords and phrases Approximate nearest neighbor, dimensionality reduction, Manhattan metric, randomized embedding

ΣΎΝΟΨΗ

Οι τυχαίες προβολές αποτελούν μια απο τις πιο διαδεδομένες μεθόδους για το χειρισμό δεδομένων μεγάλης διάστασης. Ξεκινώντας από το περίφημο Johnson-Lindenstrauss Lemma, τέτοιου είδους προβολές έχουν μελετηθεί αρκετά για την Ευκλείδια (ℓ_2) μετρική, και πολύ λιγότερο για τη μετρική Μανχάταν (ℓ_1). Σε αυτή την εργασία εστιάζουμε στο πρόβλημα του προσεγγιστικού κοντινότερου γείτονα στη μετρική Μανχάταν, εκμεταλλεύοντας την αποφαντική εκδοχή του προβλήματος, που λέγεται προσεγγιστικός *κοντινός* γείτονας και επιβάλει ένα (περίπου) λογαριθμικό κόστος.

Το 2007, οι Indyk και Naor εισήγαγαν την έννοια των εμβυθίσεων που διατηρούν τον κοντινότερο γείτονα (nearest neighbor-preserving embeddings). Οι εμβυθίσεις αυτές είναι τυχαιοκρατικές και εγγυόνται για την αλλοίωση μόνο n αποστάσεων (μεταξύ ενός σημείου-query και n σημείων), αντί για όλα τις δυνατές $O(n^2)$. Τέτοιου είδους εμβυθίσεις υπάρχουν για τις μετρικές ℓ_2 και ℓ_1, καθώς και για διπλασιάζοντα (doubling) υποσύνολα της ℓ_2.

Σε αυτή την εργασία παρουσιάζουμε μια συνάρτηση εμβύθισης για την μείωση διάστασης, η οποία διατηρεί τον *κοντινό* γείτονα (*near* neighbor-preserving) για διπλασιάζοντα υποσύνολα της ℓ_1. Η τεχνική που εφαρμόζουμε είναι να προβάλουμε τυχαία όχι τα ίδια τα σημεία, αλλά ένα σύνολο αντιπροσώπων τους. Μελετούμε δύο είδη αντιπροσώπων, τα approximate nets και τα randomly shifted grids, και τα συγκρίνουμε ως προς την νέα διάσταση και το χρόνο υπολογισμού της συνάρτησης εμβύθισης.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor, Prof. Ioannis Emiris, for his guidance, support, and excellent advice.

Of course, I want to express my deep gratitude to Ioannis Psarros, for our productive and stimulating cooperation throughout the whole process of researching this thesis.

I also want to thank Prof. Aris Pagourtzis and Prof. Dimitris Fotakis, for their participation in the examination committee.

Finally, I want to thank my family and my friends, for their unconditional and continued support.

CONTENTS

CHAPTER 1

INTRODUCTION

The *nearest neighbor* problem is defined as follows: Given a set P of n points in a metric space (X, d_X), build a data structure that, given any query point $q \in X$, returns its "nearest neighbor" $\arg\min_{p \in P} d_X(q, p)$. A particularly interesting case is that of geometric spaces, where $X = \mathbb{R}^d$ and d_X is induced by some norm. The most popular metrics are the Euclidean (ℓ_2) and the Manhattan (ℓ_1). This problem, and its approximate versions, are some of the central problems in computational geometry, and have a wide range of applications in machine learning, computer vision, data compression and other fields [SDI06, Dub10, MO15].

A common relaxation is the *c-approximate nearest neighbor* problem, where the data structure is allowed to report any $p' \in P$ within distance $c \cdot \min_{p \in P} d_X(q, p)$; for some approximation factor $c \geq 1$. By known reductions [IM98], one may focus on the *decision-with witness* version, which incurs a polylogarithmic overhead:

Definition 1.1 (Approximate Near Neighbor). Let (X, d_X) be a metric space. Given $P \subseteq X$ and reals $R > 0$, $c \geq 1$, build a data structure $\mathcal{S}$ that, given a query point $q \in X$, performs as follows:

- If the nearest neighbor of q lies in distance at most R, then $\mathcal{S}$ is allowed to report any point $p^* \in P$ such that $d_X(q, p^*) \leq cR$.
- If all points lie at distance more than cR from q, then $\mathcal{S}$ should return $\perp$.

$\mathcal{S}$ is allowed to return either a point at distance $\leq cR$ or $\perp$.

From now on, we shall refer to this problem as c-ANN, or simply ANN. Typically, the performance of an ANN data structure is measured by three quantities: 1) *preprocessing* – time to build it, 2) *space* – amount of memory it occupies and, 3) *query time* – time it takes to return an answer, given a query.

Depending on the relation between the dimension d and the number of data points n, two main regimes have emerged: low- and high-dimensional. The low-dimensional regime corresponds to $d = o(\log n)$; (hence algorithms can afford to be exponential in the dimension) and the high-dimensional regime corresponds to $d = \omega(\log n)$.

1.1 Previous work

In the low-dimensional regime, efficient $(1+\varepsilon)$-ANN algorithms are known for the Euclidean space. One notable data-structure is the Balanced Box-Decomposition (BBD) tree, introduced in [Ary+98]. BBD trees achieve query time $O(c \log n)$, for $c \leq d/2\lceil 1 + 6d/\varepsilon \rceil^d$, space $O(dn)$ and preprocessing time $O(dn \log n)$, and can also be used to retrieve the $k \leq 1$ approximate nearest neighbors, with an extra $O(d \log n)$ cost per neighbor. They are very practical as well.

Another popular data structure for ℓ_2, is the Approximate Voronoi Diagrams (AVD) [AMM09], where a tradeoff is established between space requirement and query time. More specifically, for a tradeoff parameter $2 < \gamma < 1/\varepsilon$, the query time is $O(\log(n\gamma) + 1/(\varepsilon\gamma)^{\frac{d-1}{2}})$ and space is $O(n\gamma^{d-1} \log(1/\varepsilon))$. This data structure maintains a hierarchical subdivision of space into cells, each storing a number of representatives, such that for any query lying in some 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 [AFM18].

The bucketing method of [HIM12] admits a data structure that supports fast queries for any ℓ_p, $p \in [1, 2]$, with space and preprocessing time $O(1/\varepsilon)^d \times O(n)$ and query time $O(d)$. This is done by imposing a grid on the data set, and then storing grid points which lie close to data points in a hash function.

All the aforementioned methods however, are based on discretization of the input space, and therefore depend exponentially in d, making them unfit for high-dimensional data.

An important method conceived for high dimensional data is Locality-Sensitive Hashing (LSH), introduced in [IM98]. It relies on the existence of locality sensitive hash functions for the input space, which are more likely to map similar objects to the same bucket [1]. In general, LSH requires roughly $O(dn^{1+\rho})$ space and $O(dn^{\rho})$ query time for some parameter $\rho \in (0, 1)$. Upper bounds of ρ have been established for c-ANN in ℓ_2 and ℓ_1 norms; $\rho = 1/c^2 + o(1)$ for ℓ_2 and $\rho = 1/c + o(1)$ for ℓ_1 [AI08, IM98], as well as matching lower bounds [MNP07, OWZ14]. LSH schemes based on p-stable distributions also exist for $\ell_p, p \in (0, 2]$ [Dat+04], with $\rho \leq (1+\gamma) \cdot \max(1/c^p, 1/c)$, for any $\gamma > 0$.

Better bounds on ρ can be obtained via *data-dependent LSH*: random space partitions which depend on the data set. Namely, we get $\rho = 1/(2c - 1) + o(1)$ for ℓ_1 and $\rho = 1/(2c^2 - 1) + o(1)$ for ℓ_2 [And+14, AR15]. These bounds are also known to be tight in the data-dependent LSH framework [AR16]. Moreover, upper and lower bounds for time-space tradeoffs have also been studied for both data-dependent and data-independent LSH [And+17b].

Spaces which are considered to be harder in this context, like ℓ_∞, can also be treated [Ind01, Cha17], and are very interesting since they can be used as host spaces for various symmetric norms [And+17a] (e.g., top-k and Orlicz norms)

It has become apparent that designing efficient ANN algorithms, at least for high-dimensional data, is closely related to the task of designing *low-distortion* embeddings.

Definition 1.2. A *bi-Lipschitz embedding* between two metric spaces (X, d_X) and (Y, d_Y) is a mapping $f : X \to Y$ such that for some scaling factor $C > 1$, and *distortion* $D \geq 1$, for every $p, q \in X$

$$C \cdot d_X(p, q) \leq d_Y(f(p), f(q)) \leq D \cdot C \cdot d_X(p, q).$$

[1] See also Section 2.3

Of particular importance, are low-distortion embeddings that map $\mathbb{R}^d$ to $\mathbb{R}^k$, where k is much smaller than d. The idea is to apply such an embedding as a preprocessing step, and then solve the ANN problem on the new space of lower dimension.

Such dimension-reducing embeddings do exist for the ℓ_2 norm if we allow randomization, as first shown in the influential paper by Johnson and Lindenstrauss:

Lemma 1.3 ([JL84]). Fix dimension $d > 1$ and "target" dimension $k < d$. Let A be a $k \times d$ random matrix where the A_{ij}'s are independent standard normal random variables, and define $f : \mathbb{R}^d \to \mathbb{R}^k$ as $f(x) = \frac{1}{\sqrt{k}} Ax$. Then, for any $\varepsilon \in (0, 1)$ and any $x, y \in \mathbb{R}^d$,

$$\mathbf{Pr}\left[(1-\varepsilon)\|x-y\|_2 \leq \|f(x)-f(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2\right] \geq 1 - \mathrm{e}^{-C\varepsilon^2 k},$$

where $C > 0$ is a constant (independent of d, k, ε).

Instead of a Gaussian matrix, we can even use a matrix whose entries are independent random variables with uniformly distributed values in $\{-1, +1\}$, and get the same guarantees [Ach03].

Applying this lemma for $k = O(\log n/\varepsilon^2)$ to a set $P \subset \ell_2^d$ of n points, shows that the map f has a $(1+\varepsilon)$ distortion on P, with probability at least $2/3$. Combining such a projection, (or a relevant variant) with known data structures for ℓ_2, yields better space and query bounds for ANN in high dimensions [AC09, AEP18]. This embedding has also the property of being *oblivious* to P. That is, it is well-defined over the whole space $\mathbb{R}^d$ and not just the point set P. This property is crucial because in general, a query point does not belong to the data set P.

More recently, it has been realized that the approximate nearest neighbor problem requires embedding properties that are somewhat different from definition 1.2. Apart from the obliviousness which we already mentioned, the main difference is that the embedding does not need to preserve *all* inter-point distances. This idea is captured by the following definition, introduced by Indyk and Naor in [IN07]:

Definition 1.4 (Nearest-neighbor-preserving embedding). Let (Y, d_Y); (Z, d_Z) be metric spaces and $X \subseteq Y$. We say that a distribution over mappings $f : Y \to Z$ with distortion $D \geq 1$ and probability of correctness $\mathcal{P} \in [0, 1]$, is a *D-NN-preserving embedding*, if for every $\alpha \geq 1$ and any $q \in Y$ the following holds with probability at least $\mathcal{P}$: if $x \in X$ is such that $f(x)$ is a α-approximate nearest neighbor of $f(q)$ in Z, then x is a $(D \cdot \alpha)$-approximate nearest neighbor of q in Y.

This notion is the appropriate generalization of oblivious embeddings à la Johnson and Lindenstrauss: We want f to be defined on the entire space of possible query points Y , and we require much less than a bi-Lipschitz condition. Clearly, the Johnson-Lindenstrauss lemma is an example of a NN-preserving embedding.

An analog of the Johnson-Lindenstrauss lemma for the ℓ_1 norm is impossible due to known lower bounds [BC05, LMN05]: there exists a family of data sets of n points in $(\mathbb{R}^d, \ell_1)$ such that any embedding to $(\mathbb{R}^k, \ell_1)$ with distortion D requires $k = n^{\Omega(1/D^2)}$ dimensions. The lower bound also holds for doubling subsets of ℓ_1 with doubling constant at least 6.

However, the following theorem by Indyk shows that NN-preserving embeddings allow us to overcome the impossibility results for the stronger notion of bi-Lipschitz embeddings, while being sufficient for the purpose of the nearest neighbor problem.

Theorem 1.5 ([Ind06]). For any $\varepsilon \leq 1/2$, $\delta > 0$, $\varepsilon > \gamma > 0$ there is a probability space over linear mappings $f : \mathbb{R}^d \to \mathbb{R}^k$, where $k = (\ln(1/\delta))^{1/(\varepsilon-\gamma)}/\zeta(\gamma)$, for a function $\zeta(\gamma) > 0$ depending only on γ, such that for any pair of points $p, q \in \ell_1^d$

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

Note that the mapping is defined as $f(u) = Au/T$, where A is a $k \times d$ 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 $T = \Theta(k \log(k/\varepsilon))$ (see Lemma 5 in [Ind06]).

The embedding is randomized but asymmetric: for any pair of points, the probability that the distance between the pair gets contracted is *very small*, while the probability of the distance being expanded by $(1+\varepsilon)$ is *constant*. This is all we want for a NN-preserving embedding: We don't want some far neighbor of q in the space $\mathbb{R}^d$ to become nearest neighbor in the space $\mathbb{R}^k$. Thus, for $\delta = 1/n$, we get constant probability of contraction, after a union bound on all the far points. On the other hand, we want the nearest neighbor of q to remain nearest after the embedding, so the constant probability of expansion in enough so that the whole embedding is correct with constant probability.

Significant amount of work has also been done for pointsets of low doubling dimension, a notion of dimension that measures the "volume growth" of X. [2]

Definition 1.6. Let (X, d_X) be a metric space and $B_X(x, r)$ the ball of radius $r > 0$ centered at $x \in X$. The *doubling constant* of X, denoted λ_X, is the smallest integer $\lambda \geq 1$ such that for any $p \in X$ and $r > 0$, the ball $B_X(p, r)$ can be covered by at most λ_X balls of radius $r/2$, centered at points in X. The *doubling dimension* of X, denoted $\dim(X)$, is defined to be $\log \lambda_X$.

For any finite metric space X of doubling dimension $\dim(X)$, there exists a data structure [CG06, HM06] with expected preprocessing time $O(2^{\dim(X)} n \log n)$, space $O(2^{\dim(X)} n)$ and query time $O(2^{\dim(X)} \log n + \varepsilon^{-O(\dim(X))})$. A notable series of results [Cla99, KR02, KL04, BKL06] concerned arbitrary metrics of bounded *expansion rate* (a notion similar to doubling dimension) and provided efficient data structures for ANN, which can be extended to metric spaces where $\dim(X) = O(1)$.

Indyk and Naor showed that for doubling subsets of ℓ_2, the Johnson-Lindenstrauss embedding is $(1+\varepsilon)$-NN-preserving:

Theorem 1.7 ([IN07]). Let $X \subseteq \ell_2^d$, $\varepsilon \in (0,1)$, $\delta \in (0, 1/2)$ and f be the Johnson-Lindenstrauss projection. Then, there exists $k = O\left(\log \lambda_X \cdot \frac{\log(2/\varepsilon)}{\varepsilon^2} \cdot \log(1/\delta)\right)$ such that for any $q \in X$ with nearest neighbor p^*, with probability at least $1 - \delta$,

$$\|f(p^*) - f(q)\|_2 \leq (1+\varepsilon)\|p^* - q\|_2,$$
$$\forall x : \|x - q\|_2 > (1+2\varepsilon)\|p^* - q\|_2 \implies \|f(x) - f(q)\|_2 > (1+\varepsilon)\|p^* - q\|_2.$$

Randomized embeddings have also been recently used for doubling subsets of ℓ_p, $2 < p < \infty$, yielding c-ANN data structures, where c depends on the doubling constant and the dimension of the data set [BG19]. Dimension reduction techniques for

[2] See also Section 2.4

doubling subsets of ℓ_p, $p \in [1, 2]$, also exist [BG16], but they rely on partition algorithms which require the whole pointset to be known in advance. Moreover, the guarantees concern only distances up to some scale $s > 1$, on which the target dimension depends. Hence, it is not clear whether these techniques can be used in the ANN context.

The next table sums up the randomized embeddings we mentioned.

Norm	Ref.	Target dimension (k)	Guarantee
ℓ_2	Lem 1.3	$O\left(\log n/\varepsilon^2\right)$	$(1+\varepsilon)$-bi-Lipschitz
	Thm 1.7	$O\left(\log \lambda_X \cdot \frac{\log(2/\varepsilon)}{\varepsilon^2} \cdot \log(1/\delta)\right)$	$(1+\varepsilon)$-NN-preserving
ℓ_1	Thm 1.5	$(\ln n)^{1/(\varepsilon-\gamma)}/\zeta(\gamma)$	$(1+\varepsilon)$-NN-preserving

1.2 Contribution

In this thesis, 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 of Definition 1.4.

Definition 1.8 (Near neighbor-preserving embedding). Let $(Y, d_Y), (Z, d_Z)$ be metric spaces and $X \subseteq Y$. A distribution over mappings $f : Y \to Z$ is a *near-neighbor preserving embedding* with range R, distortion $D \geq 1$ and probability of correctness $\mathcal{P} \in [0, 1]$ if, $\forall \alpha \geq 1$ and $\forall q \in Y$, with probability at least $\mathcal{P}$, when $x \in X$ is such that $d_Z(f(x), f(q)) \leq \alpha \cdot R$, then $d_Y(x, q) \leq D \cdot \alpha \cdot R$.

Considering a pointset $P \subset \ell_1^d$ of cardinality n, our approach is to represent P with an ε-covering set, and then apply a random linear projection to that set, using Cauchy variables as in Theorem 1.5. We study two cases of covering sets: c-approximate r-nets and randomly shifted grids. The two main results concern ℓ_1^k as the target space, where k depends on the doubling dimension of P. More specifically:

1. In Theorem 3.6, we prove that for every $\varepsilon \in (0, 1/2)$ and $c \geq 1$, there is a randomized mapping $h : \ell_1^d \to \ell_1^k$ that can be computed in time $\tilde{O}(dn^{1+1/\Omega(c)})$ and is *near* neighbor-preserving for P with distortion $1+6\varepsilon$ and probability of correctness $\Omega(\varepsilon)$, where

$$k = (\log \lambda_P \cdot \log(\mathbf{c}/\varepsilon))^{\Theta(1/\varepsilon)}/\zeta(\varepsilon).$$

   Although the mapping h depends on the pointset, the parameter c is user-defined and therefore provides a trade-off between preprocessing time and target dimension.

2. In Theorem 3.10, we show that for every $\varepsilon \in (0, 1/2)$, there is a randomized mapping $h' : \ell_1^d \to \ell_1^k$ that can be computed in time $O(dkn)$ and is *near* neighbor-preserving for P with distortion $1+6\varepsilon$ and probability of correctness $\Omega(\varepsilon)$, where

$$k = (\log \lambda_P \cdot \log(\mathbf{d}/\varepsilon))^{\Theta(1/\varepsilon)}/\zeta(\varepsilon).$$

   In this case, the function h' is oblivious to P and well-defined over the whole space, but the target dimension depends on d.

On the low-preprocessing-time extreme, one can embed the dataset in near-linear time, but the target dimension is polynomial in $\log \log n$. This is to be juxtaposed to the analogous result by Indyk [Ind06], which provides with target dimension polynomial in $\log n$, without any assumption on the doubling dimension of the dataset. On the other hand, one can obtain a preprocessing time of $dn^{1+\delta}$ for any constant $\delta > 0$, and target dimension which depends solely on the doubling dimension.

One key observation here is that given an r-net in a space X of bounded diameter Δ, we can directly employ Theorem 1.5: The number of net points can be upper bounded by a function of λ_X, r, and Δ, and hence the new dimension depends only on these parameters. Therefore, one can reduce ANN in ℓ_1^d to ANN in ℓ_1^k, where $k := k(\lambda_X, r)$. This thesis proves better bounds on the target dimension than the ones of Theorem 1.5, for doubling subsets of ℓ_1^d, without any assumption on the diameter of the dataset.

CHAPTER 2

PRELIMINARIES

In this chapter, we provide the necessary notation, definitions, as well as some tools that will come in handy in Chapter 3.

2.1 Metric spaces

Definition 2.1. Let X be a set and $d_X : X \times X \to \mathbb{R}^+$ a distance function. The pair (X, d_X) is called a *metric space* if for any $x, y, z \in X$, d_X satisfies

- $d_X(x, x) = 0$,
- $d_X(x, y) > 0$, iff $x \neq y$,
- $d_X(x, y) = d_X(y, x)$,
- $d_X(x, y) + d_X(y, z) \leq d_X(x, z)$.

Metrics can be defined on completely arbitrary sets, and specify distances for pairs of points. A norm is defined only on a vector space, and for each point it specifies its distance from the origin.

By definition, a *norm* on a real vector space Z is a mapping $\|\cdot\| : Z \to \mathbb{R}^+$ so that:

- $\|x\| = 0$ iff $x = \mathbf{0}$,
- $\|\alpha x\| = |\alpha| \cdot \|x\|$, for all $\alpha \in \mathbb{R}$,
- $\|x + y\| \leq \|x\| + \|y\|$ (sub-additivity).

Every norm $\|x\|$ on Z defines a metric, in which the distance of points x, y equals $\|x - y\|$. However, not all metrics derive from norms.

The *unit ball* $\{x \in Z : \|x\| \leq 1\}$ of any norm is a closed convex set K, that is symmetric around the origin $\mathbf{0}$ and contains $\mathbf{0}$ in the interior. Conversely, any $K \subset Z$ with these properties is the unit ball of a uniquely determined norm. Therefore, norms and symmetric convex sets can be considered as different views of the same objects.

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

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} . \tag{2.1}$$

We denote by ℓ_p^d the normed space $(\mathbb{R}^d, \|\cdot\|_p)$.

The most popular norms in this family are the *Euclidean* (ℓ_2), the *Manhattan* (ℓ_1), and the *maximum norm* (ℓ_∞). The ℓ_∞ norm is given by $\|x\|_\infty = \max_i |x_i|$ which is the limit of (2.1) as $p \to \infty$. To get some intuition about these norms, take a look at their unit balls in the plane, in Figure 2.1.

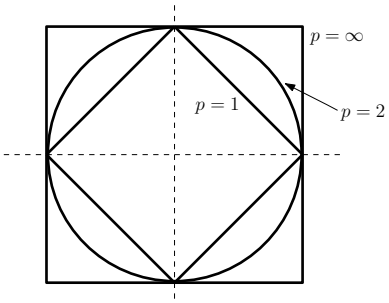

Figure 2.1: Unit balls of ℓ_1, ℓ_2 and ℓ_∞ norms.

Notice that all three are indeed closed convex bodies. For $p = 2$, we have the ordinary disk. As p decreases to 1, the unit ball shrinks towards the rhombus. For $p \geq 2$, the unit ball expands towards the square, as $p \to \infty$. Note that, only the unit balls of ℓ_1 and ℓ_∞ have sharp corners – for any $p > 1$ the unit ball is differentiable everywhere.

For $p \in (0, 1)$, the mapping (2.1) still defines a metric on $\mathbb{R}^d$, which may be useful for some applications, but it no longer defines a norm – the sub-additive property no longer holds. Consequently, the unit ball is not a convex set, as Figure 2.2 demonstrates.

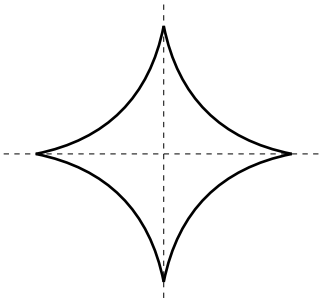

Figure 2.2: The unit ball for $p = 2/3$.

The next claim provides a connection between ℓ_p metrics.

Claim 2.2. For any vector $x \in \mathbb{R}^d$ and $p > q > 0$

$$\|x\|_p \leq \|x\|_q \leq d^{1/q - 1/p} \|x\|_p .$$

Proof. Appendix A.

2.2 Concentration bounds and stable distributions

A simple, yet fundamental and tremendously useful inequality in probability theory, is the union bound, also known as Boole's inequality: For any events $A_1 \ldots, A_n$ we have

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}[A_i].$$

The equality holds only if the events are pairwise mutually disjoint.

Concentration inequalities

Concentration inequalities provide bounds on how a random variable deviates from some value, typically its expected value. The laws of large numbers of classical probability theory states that sums of independent random variables are, under very mild conditions, close to their expectation with a large probability. Such sums are the most basic examples of random variables concentrated around their mean.

Markov's inequality gives an upper bound for the probability that a non-negative random variable is greater than or equal to some positive constant.

Theorem 2.3 (Markov's inequality). Let X be a real-valued non-negative random variable. Then for all $\alpha > 0$

$$\mathbf{Pr}[X \geq \alpha] \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Setting $\alpha = \tilde{\alpha}\,\mathbb{E}[X]$, for some $\tilde{\alpha} > 0$, one can rewrite

$$\mathbf{Pr}[X \geq \tilde{\alpha}\,\mathbb{E}[X]] \leq \frac{1}{\tilde{\alpha}}.$$

Another popular concentration inequality is the Chernoff bound, which gives exponentially decreasing bounds on tail distributions of sums of independent random variables. Although it is a sharper bound than Markov's inequality, the Chernoff bound requires that the variables be independent – a condition that is not required for Markov's inequality.

The generic Chernoff bound is derived using the *moment generating function*:

Definition 2.4. The moment-generating function of a random variable X is

$$M_X(t) = \mathbb{E}[\mathrm{e}^{tX}],\ t \in \mathbb{R},$$

whenever this expectation exists.

Theorem 2.5 (Generic Chernoff bound). Let S be the sum of n *independent* random variables, $X_1, \ldots, X_n$ and $\alpha \in \mathbb{R}$. Then, for every $t > 0$

$$\mathbf{Pr}[S \geq \alpha] \leq \mathrm{e}^{-t\alpha} \cdot \prod_{i=1}^{n} \mathbb{E}[\mathrm{e}^{tX_i}],$$

$$\mathbf{Pr}[S \leq \alpha] \leq \mathrm{e}^{t\alpha} \cdot \prod_{i=1}^{n} \mathbb{E}[\mathrm{e}^{-tX_i}].$$

Proof. By Markov's inequality, and since X_i's are independent:

$$\mathbf{Pr}[S \geq \alpha] = \mathbf{Pr}[\mathrm{e}^{tS} \geq \mathrm{e}^{t\alpha}] \leq \frac{\mathbb{E}[\mathrm{e}^{tS}]}{\mathrm{e}^{t\alpha}} = \mathrm{e}^{-t\alpha} \cdot \mathbb{E}\left[\prod_{i=1}^{n} \mathrm{e}^{tX_i}\right] = \mathrm{e}^{-t\alpha} \cdot \prod_{i=1}^{n} \mathbb{E}[\mathrm{e}^{tX_i}].$$

The second inequality is proved similarly. ■

Notice that with the Chernoff bound, one can derive a concentration inequality for a sum of independent random variables, by using a bound on the moment generating function.

Stable distributions

Stable distributions [Zol86] are defined as limits of normalized sums of independent identically distributed variables (an alternate definition follows). The most known example of a stable distribution is Gaussian distribution. However, the class is much wider; for example, it includes heavy-tailed distributions.

Definition 2.6. A distribution $\mathcal{D}$ over $\mathbb{R}$ is called p-stable, if there exists $p > 0$ such that for any d real numbers $v_1, \ldots, v_d$ and i.i.d. variables $X_1, \ldots, X_d$ with distribution $\mathcal{D}$, the random variable $\sum_i v_i X_i$ has the same distribution as $X \cdot (\sum_i |v_i|^p)^{1/p}$, where $X \sim \mathcal{D}$.

Stable distributions are known to exist for any $p \in (0, 2]$. In particular:

- The Gaussian distribution ($\mathcal{D}_G$), with density $g(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$, is 2-stable.
- The Cauchy distribution ($\mathcal{D}_C$), with density $c(x) = \frac{1}{\pi} \frac{1}{1+x^2}$, is 1-stable.

From a practical point of view, despite the lack of closed form density and distribution functions, it is known [CMS76] that one can generate a p-stable random variable X by taking

$$X = \frac{\sin(p\theta)}{\cos^{1/p}\theta} \left(\frac{\cos(\theta(1-p))}{-\ln r} \right)^{(1-p)/p},$$

where θ is uniform on $[-\pi/2, \pi/2]$ and r is uniform on $[0, 1]$. Stable distributions have found numerous applications in various fields [Nol18]. In computer science, they are used for sketching and dimension reduction.

The main property of p-stable distributions directly translates into a sketching technique. For example, let $0 < k < d$ and $f : \mathbb{R}^d \to \mathbb{R}^k$, such that for any $v \in \mathbb{R}^d$, $f(v) = A \cdot v$, where A is a random $k \times d$ matrix, with each entry being an i.i.d. variable from a p-stable distribution. Let $A_j, j \in [k]$ be the j-th row of A. Then, $f(v)$ is a tuple of k dot products:

$$f(v) = (\langle A_1, v\rangle, \ldots, \langle A_k, v\rangle).$$

By the p-stability property, $f(v)$ is distributed as

$$(Y_1 \cdot \|v\|_p, \ldots, Y_k \cdot \|v\|_p) = \|v\|_p \cdot (Y_1, \ldots, Y_k),$$

where the Y_j's are i.i.d. with p-stable distribution.

Now, one can use $f(v)$, which is termed as the sketch of v, to analyze $\|v\|_p$. In addition, the map f defines a randomized projection of $\mathbb{R}^d$ to $\mathbb{R}^k$.

2.3 Locality-Sensitive Hashing

The main idea behind LSH, as mentioned in Section 1.1, is *random space partitions*, which have the property that a pair of close points (at distance at most r) is more likely to belong to the same part than a pair of far points (at distance more than cr). Given such a partition, the data structure splits the dataset P accordingly, and, given a query, retrieves all the data points which belong to the same part as the query.

Definition 2.7 (Locality-Sensitive Hashing). Fix a metric space (X, d_X), $r > 0$, approximation $c > 1$, and a set U. Then a distribution $\mathcal{H}$ over maps $h : X \to U$ is (r, cr, p_1, p_2)-sensitive if for any $x, y \in X$

$$d_X(x, y) \leq r \implies \mathop{\mathbf{Pr}}_{h \sim \mathcal{H}}[h(x) = h(y)] \geq p_1,$$
$$d_X(x, y) \geq cr \implies \mathop{\mathbf{Pr}}_{h \sim \mathcal{H}}[h(x) = h(y)] \leq p_2.$$

The distribution $\mathcal{H}$ is called LSH family and has quality $\rho = \rho(\mathcal{H}) = \frac{\log(1/p_1)}{\log(1/p_2)}$.

An LSH family is meaningful when $p_1 > p_2$, and therefore $\rho < 1$. Notice that LSH mappings are oblivious: they are well-defined over the whole metric space. Hence, data structures based on LSH can naturally work in an online setting with insertions and deletions.

Theorem 2.8 ([IM98]). Let $P \subset (X, d_X)$, $|P| = n$, and let $\mathcal{H}$ be an (r, cr, p_1, p_2)-sensitive LSH family for (X, d_X) with quality ρ. Then there exists a fully dynamic data structure for (c, r)-ANN with space and preprocessing time $O(n^{1+\rho} + dn)$ and query time $O(dn^{\rho})$. The data structure is correct with constant probability.

Theorem 2.9 ([IM98]). There exist an (r, cr, p_1, p_2)-sensitive LSH family for the Hamming space with quality $\rho = 1/c$.

The distribution $\mathcal{H}$ is simply defined as projections on random coordinates: $\mathcal{H} = \{h_i : h_i(x) = x_i, i = 1, \ldots, d\}$. This family is $(r, cr, 1 - r/d, 1 - cr/d)$-sensitive and therefore, $\rho \leq 1/c$. Note that this LSH family can be extended to the Manhattan (ℓ_1).

Theorem 2.10 ([AI08]). There exist an (r, cr, p_1, p_2)-sensitive LSH family for the Euclidean metric with quality $\rho = 1/c^2 + o(1)$.

This LSH family partitions the space into Euclidean balls. It proceeds in two steps: first it performs a random dimension reduction to dimension t, where t is a parameter, and then partitions $\mathbb{R}^t$ into balls.

2.4 Doubling sets and covering nets

Doubling dimension

Let's restate the definition in a more rigorous way:

Definition. Let (X, d_X) be a metric space and $B_X(x, r) = \{y \in X : d_X(x, y) \leq r\}$. The *doubling constant* of X, denoted λ_X, is the smallest integer $\lambda \geq 1$ such that for any $p \in X$ and $r > 0$, there exist a set $S \subseteq X$ of cardinality at most λ_X, such that

$$B_X(p, r) \subseteq \bigcup_{s \in S} B_X(s, r/2).$$

The *doubling dimension* of X is defined as $\dim(X) = \log \lambda_X$.

The following properties demonstrate that $\dim(X)$ is a robust and meaningful notion.

1. If $|X| = n$, then $\dim(X) \leq \log n$.
2. For $X = \mathbb{R}^d$ equipped with any norm, $\dim(X) = \Theta(d)$.
3. If $S \subseteq X$, then $\dim(S) \leq \dim(X)$.
4. $\dim(X_1 \cup \ldots \cup X_m) \leq \max_i\{\dim(X_i)\} + \log m$.

Motivation. Doubling metrics often occur naturally in practical applications, where the data set P is contained in the union of low-dimensional manifolds lying in some very high-dimensional space $\mathbb{R}^d$, and the distance function is some norm of $\mathbb{R}^d$. In such cases, algorithms that exploit the doubling dimension of the data set might be superior to algorithms which consider only the structure of the high dimensional host space.

Exact and approximate r-nets

Nets play an important role in the study of embeddings, as well as in designing efficient data structures for doubling spaces. The formal definition is followed by an illustration.

Definition 2.11 (r-net). Let (X, d_X) be a metric space and $r > 0$. A subset $\mathcal{N} \subseteq X$ is called an r-net if it satisfies the following properties:

- r-*Packing*: For every $s, s' \in \mathcal{N}$, $d_X(s, s') > r$.
- r-*Covering*: For every $x \in X$, there exists $s \in \mathcal{N}$ such that $d_X(x, s) \leq r$.

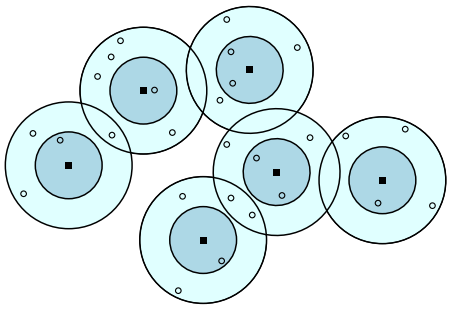

Figure 2.3: An r-net of a pointset $P \subset \mathbb{R}^2$. Net points are depicted as squares. The small disks have radius $r/2$ and the large disks have radius r. Observe that by the packing property, the small disks are disjoint, and by the covering property, the large disks cover all points of P.

Nets are "sparse" representative sets that try to capture the geometry of a metric space. The following lemma demonstrates why doubling spaces and r-nets are closely related.

Lemma 2.12. Let (X, d_X) be a metric space of bounded diameter Δ and doubling constant λ_X. Let also $\mathcal{N} \subseteq X$ be an r-net of X. Then

$$|\mathcal{N}| \leq \lambda_X^{\lceil \log (2\Delta/r) \rceil}.$$

Proof. For some $s \in \mathcal{N}$ and $R > 0$, let $B_{\mathcal{N}}(s, R) := \{s' \in \mathcal{N} : d_V(s, s') \leq R\}$. Since $\mathcal{N}$ is a subset of X, by the definition of λ_X, there exists a set $S_1 \subseteq \mathcal{N}$ of cardinality at most λ_X, such that for every $s \in \mathcal{N}$,

$$\mathcal{N} = B_{\mathcal{N}}(s, \Delta) \subseteq \bigcup_{s' \in S_1} B_{\mathcal{N}}(s', \Delta/2).$$

Applying the definition k times in a recursive fashion, yields a set $S_k \subseteq \mathcal{N}$ of cardinality at most λ_X^k, such that for every $s \in \mathcal{N}$,

$$\mathcal{N} \subseteq \bigcup_{s' \in S_k} B_{\mathcal{N}}(s', \Delta/2^k).$$

Therefore,

$$|\mathcal{N}| \leq \sum_{s' \in S_k} |B_{\mathcal{N}}(s', \Delta/2^k)|.$$

By the packing property of the r-net, for every $s \in \mathcal{N}$, $B_{\mathcal{N}}(s, r/2) = \{s\}$. Hence, for $k = \lceil \log(2\Delta/r) \rceil$,

$$|\mathcal{N}| \leq \sum_{s' \in S_k} |B_{\mathcal{N}}(s', r/2)| \leq \lambda_X^{\lceil \log(2\Delta/r) \rceil}.$$ ■

For set P of n points in some metric space (X, d_X) and $r > 0$, an r-net can be computed in $O(n^2)$ time by a greedy algorithm: Pick an arbitrary ordering of the points, $p_1, \ldots, p_n$. Set $\mathcal{N} = \{p_1\}$ and cover all points that are at distance at most r from p_1. Continue with p_2; if it is covered, proceed to p_3, else, add p_2 to $\mathcal{N}$ and cover all points within distance r from it. Continue until all points are processed or covered, and return $\mathcal{N}$.

In [EHS15], the notion of r-nets was extended to c-approximate r-nets, where the covering property is relaxed. Notice that Lemma 2.12 applies to c-approximate r-nets as well, since only the covering property has changed.

Definition 2.13 (c-approximate r-net). For $c \geq 1$, $r > 0$ and metric space (X, d_X), a c-approximate r-net of X is a subset $\mathcal{N} \subseteq X$ that satisfies

- *r-Packing*: For every $s, s' \in \mathcal{N}$, $d_X(s, s') > r$.
- *cr-Covering*: For every $x \in X$, there exists $s \in \mathcal{N}$ such that $d_X(x, s) \leq cr$.

Theorem 2.14 ([EHS15]). Let $P \subset \ell_2^d$ such that $|P| = n$. Then, for any $r > 0$, $\varepsilon > 0$, one can compute a $(1 + \varepsilon)$-approximate r-net of P in expected running time $O(\varepsilon^{-2} n^{1+1/(1+\varepsilon)^2 + o(1)})$. The result is correct with high probability.

The algorithm consists of a Johnson-Lindenstrauss projection as a preprocessing step, and a combination of the greedy algorithm with the LSH family of Theorem 2.10. We use this technique in Section 3.2 to compute approximate nets for the ℓ_1 norm.

2.5 Randomly shifted grids

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.

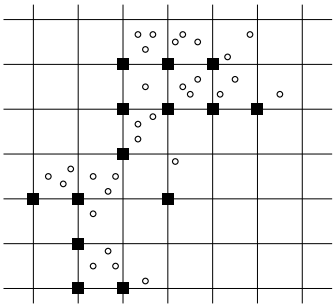

Figure 2.4: A pointset P (circles) and $g_w(P)$ (squares). Each point of P is mapped onto the bottom left point of the cell in which it lies.

The next claim concerns the expected number of grid points that are contained in some closed interval.

Claim 2.15. Let $[a, b] \subset \mathbb{R}$ be an interval of length $L > 0$ and $w > 0$. Then

$$\mathbb{E}[|[a, b] \cap h_{w,t}(\mathbb{R})|] = L/w.$$

Proof. Let $M = |[a, b] \cap h_{w,t}(\mathbb{R})|$ and assume wlog that $[a, b] = [0, L]$. Think of the case where $t = 0$, and then observe how M changes as $t \to w$.

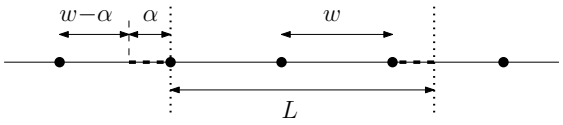

Figure 2.5

Increasing t corresponds to moving the interval to the left by t. Therefore, there exists a threshold $\alpha > 0$, such that

$$M = \begin{cases} \lfloor L/w \rfloor + 1, & \text{if } t \in [0, \alpha] \\ \lfloor L/w \rfloor, & \text{if } t \in (\alpha, w). \end{cases}$$

As we can see in figure 2.5, $\alpha = L - \lfloor L/w \rfloor w$. Moreover, since we choose t uniformly in $[0, w)$,

$$\mathbb{E}[M] = \frac{\alpha}{w}\left(\left\lfloor \frac{L}{w} \right\rfloor + 1\right) + \frac{(w-\alpha)}{w}\left\lfloor \frac{L}{w} \right\rfloor = \left\lfloor \frac{L}{w} \right\rfloor + \frac{\alpha}{w} = \frac{L}{w}. \qquad \blacksquare$$

Let $B_1(x, r)$ denote the ℓ_1-ball of radius $r > 0$ around a point $x \in \mathbb{R}^d$. The next lemma, provides a bound on the expected number of grid points that the ball contains.

Lemma 2.16. For any $x \in \mathbb{R}^d$ and $r > 0$, we have

$$\mathbb{E}[|B_1(x,r) \cap g_w(\mathbb{R}^d)|] \leq (1 + 2r/w)^d.$$

Proof. Let $M := |B_1(x,r) \cap g_w(\mathbb{R}^d)|$ and $\mathcal{C}$ be the total number of grid cells intersecting $B_1(x,r)$. Each grid point corresponds to a grid cell, hence M is bounded by $\mathcal{C}$. Now, let $\mathcal{C}_i$ denote the number of grid cells that $B_1(x,r)$ intersects in the axis i. Obviously, $\mathcal{C} \leq \prod_{i\in[d]} \mathcal{C}_i$ (see Figure 2.6). Therefore,

$$\mathbb{E}[M] \leq \mathbb{E}[\mathcal{C}] \leq \mathbb{E}\left[\prod_{i=1}^{d} \mathcal{C}_i\right].$$

By Claim 2.15, we have that $\mathbb{E}[\mathcal{C}_i] = 1 + 2r/w$. Recall that the random shift t_i is independent per axis and so the $\mathcal{C}_i$'s are also independent. Consequently,

$$\mathbb{E}[M] \leq \mathbb{E}\left[\prod_{i=1}^{d} \mathcal{C}_i\right] = \prod_{i=1}^{d} \mathbb{E}[\mathcal{C}_i] = (1 + 2r/w)^d. \qquad \blacksquare$$

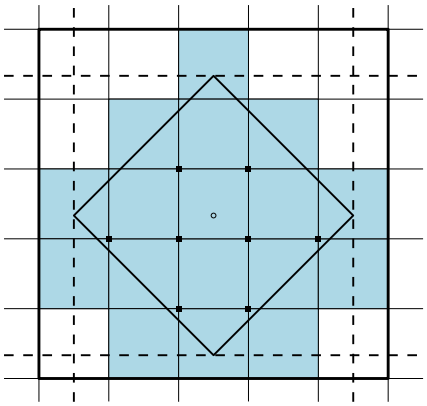

Figure 2.6: Grid cells intersecting an ℓ_1 ball.

Notice that the bound of Lemma 2.16 also holds for any ℓ_p ball of radius r, as the arguments in the proof apply for the ℓ_∞ ball as well.

CHAPTER 3

RANDOMIZED EMBEDDINGS FOR DOUBLING SUBSETS OF ℓ_1

3.1 A concentration bound for sums of Cauchy variables

In this section, we present a concentration inequality for sums of Cauchy variables, which serves as one of main tools for the analysis.

Recall the Cauchy distribution, denoted $\mathcal{D}_C$, with density

$$c(x) = \frac{1}{\pi}\frac{1}{1+x^2}.$$

Unlike the Gaussian distribution, the Cauchy distribution has no mean, variance, and moment generating function. However, for $0 < q < 1$, the mean of the qth power of the absolute value of a Cauchy random variable can be defined. More specifically, for some $X \sim \mathcal{D}_C$ we have

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

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

Lemma 3.1. Let $X \sim \mathcal{D}_C$. Then for any $\beta > 1$:

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

Proof. For any constant $\beta > 0$, we have [1]

$$\int_0^1 \mathrm{e}^{-\beta x^{1/2}}\,\mathrm{d}x = \frac{2}{\beta^2}\left(1 - \frac{\beta+1}{\mathrm{e}^\beta}\right). \tag{3.1}$$

Moreover,

$$\mathrm{e}^{-\beta x^{1/2}} \leq \mathrm{e}^{-\beta}, \quad \forall x \geq 1, \beta > 0,$$
$$\frac{1}{1+x} \leq 1, \qquad \forall x \in [0,1].$$

[1] See Appendix B for the proof of (3.1).

Hence, for any $\beta > 1$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\beta|X|^{1/2}\right)\right] &= \int_{-\infty}^{\infty} \mathrm{e}^{-\beta|x|^{1/2}} \cdot c(x)\,\mathrm{d}x \\
&= \frac{2}{\pi}\int_{0}^{\infty} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\,\mathrm{d}x \\
&= \frac{2}{\pi}\int_{0}^{1} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\,\mathrm{d}x + \frac{2}{\pi}\int_{1}^{\infty} \mathrm{e}^{-\beta x^{1/2}} \cdot \frac{1}{1+x^2}\,\mathrm{d}x \\
&\le \frac{2}{\pi}\int_{0}^{1} \mathrm{e}^{-\beta x^{1/2}}\,\mathrm{d}x + \frac{2}{\pi}\int_{1}^{\infty} \mathrm{e}^{-\beta} \cdot \frac{1}{1+x^2}\,\mathrm{d}x \\
&= \frac{2}{\pi}\cdot\frac{2}{\beta^2}\left(1 - \frac{\beta+1}{\mathrm{e}^{\beta}}\right) + \frac{1}{2\mathrm{e}^{\beta}} \\
&\le \frac{4}{\pi\beta^2} + \frac{1}{2\mathrm{e}^{\beta}} \\
&\le \frac{3}{2\beta^2} + \frac{1}{2\mathrm{e}^{\beta}} \\
&\le \frac{2}{\beta}.
\end{aligned}$$

■

Now, consider a collection of k i.i.d. Cauchy variables, $X_1, \ldots, X_k$, and let $S := \sum_{j=1}^{k} |X_j|$. The non-existence of mean and moment generating function of the Cauchy distribution makes it difficult to prove concentration bounds for S. Therefore, we study the sum $\tilde{S} := \sum_{j=1}^{k} |X_j|^{1/2}$ instead:

Lemma 3.2. For any $t > 0$,

$$\mathbf{Pr}[S \le t] \le \mathbf{Pr}[\tilde{S} \le \sqrt{tk}].$$

Proof. Set $x = (X_1, \ldots, X_k)$ and observe that $S = \|x\|_1$ and $\tilde{S} = \|x\|_{1/2}^{1/2}$. Hence, by Claim 2.2 for $p = 1$ and $q = 1/2$,

$$S \le \tilde{S}^2 \le k \cdot S.$$

Recall that S and $\tilde{S}$ are random variables. Therefore, for any $t > 0$,

$$(S \le t \implies \tilde{S} \le \sqrt{tk}) \implies \mathbf{Pr}[S \le t] \le \mathbf{Pr}[\tilde{S} \le \sqrt{tk}].$$ ■

We use the bound on the moment-generating function, to prove a Chernoff-type concentration bound for $\tilde{S}$, which by Lemma 3.2 translates directly into a concentration bound for S.

Lemma 3.3. For every $D > 1$,

$$\mathbf{Pr}\left[\tilde{S} \le \frac{\mathbb{E}[\tilde{S}]}{D}\right] \le \left(\frac{10}{D}\right)^{k}.$$

Proof. Since X_j's are independent, $\mathbb{E}[\tilde{S}] = \sqrt{2}k$. Then, by Lemma 3.1 and Markov's

inequality, for any $\beta > 1$,

$$\begin{aligned}
\mathbf{Pr}\left[\tilde{S} \leq \frac{\mathbb{E}[\tilde{S}]}{D}\right] &= \mathbf{Pr}\left[-\beta\tilde{S} \geq -\beta \cdot \frac{\mathbb{E}[\tilde{S}]}{D}\right] \\
&= \mathbf{Pr}\left[\exp(-\beta\tilde{S}) \geq \exp\left(-\beta \cdot \frac{\mathbb{E}[\tilde{S}]}{D}\right)\right] \\
&\leq \frac{\mathbb{E}[\exp(-\beta\tilde{S})]}{\exp(-\beta\,\mathbb{E}[\tilde{S}]/D)} \\
&= \frac{\mathbb{E}[\exp(-\beta|X_j|^{1/2})]^k}{\exp(-\beta\sqrt{2k}/D)} \\
&\leq \left(\frac{2}{\beta}\right)^k \cdot \mathrm{e}^{\sqrt{2}\beta k/D}.
\end{aligned}$$

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

3.2 Computing approximate nets in ℓ_1

One of our two embeddings requires the computation of an approximate r-net. We follow the same idea as in Theorem 2.14, combining the greedy algorithm with an LSH family for the Hamming space.

Theorem 3.4. Let $P \subset \ell_1^d$ such that $|P| = n$. 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 to one point of the net, which covers it.

Proof. 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 greater than $1 - \|p - q\|_1/2$. 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

$$G(x) = (g(x_1), \ldots, g(x_d)),$$

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 which are obtained by the above procedure. Notice that,

$$\|f(p) - f(q)\|_1 \in \left(\frac{2}{\delta}\right) \cdot \|p - q\|_1 \pm 1.$$

Hence,

$$\begin{aligned}
\|p - q\|_1 \leq 1 &\implies \|f(p) - f(q)\|_1 \leq \left(\frac{2}{\delta}\right) + 1, \\
\|p - q\|_1 \geq c &\implies \|f(p) - f(q)\|_1 \geq \left(\frac{2}{\delta}\right) \cdot c - 1.
\end{aligned}$$

Now, we employ the LSH family of [HIM12], 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 $c' = \Omega(c)$.

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 at least $p_1/2$. Any two points $p, q \in P$, such that $\|p - q\|_1 \geq c$, fall into the same bucket with probability at most p_2.

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 $p_1, \ldots, p_n$ of the points, and compute the approximate net in a greedy fashion. We start with 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 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. During 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 $\leq (1 - p_1/2)^k \leq n^{-10}$. Hence, 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 $\leq k \cdot n^2 \cdot p_2 = \tilde{O}(n^{1+1/c'})$. 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. ■

3.3 Dimension reduction via approximate nets

In this section we describe the dimension reduction mapping for the ℓ_1 norm 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 with $B_1(x, r)$ the ℓ_1-ball of radius r around x. The embedding is non-linear and is carried out in two steps: Given P, $\varepsilon > 0$ and $c \geq 1$,

1. we compute a c-approximate (ε/c)-net of P with the algorithm of Theorem 3.4. Let $\mathcal{N}$ be the output of the algorithm. In every iteration, a subset $P_s \subseteq P$ is covered by some point $s \in \mathcal{N}$. Let $g : P \to \mathcal{N}$ be the function that maps every point of P_s to s.
2. Then, for every $s \in \mathcal{N}$ and any query point $q \in \ell_1^d$, we apply the linear map of Theorem 1.5. That is, $f(s) = As/T$, where A is a $k \times d$ matrix with each element being an i.i.d. Cauchy random variable. Recall that T is a scaling factor such that $T = \Theta(k \log(k/\varepsilon))$.

We define the embedding to be $h = f \circ g$. We apply h for every point in P, and only f for any query q. It is clear from the properties of the net that g incurs an additive error of ε on the distances between q and P, so it suffices to study the distortion of f. By the 1-stability property of the Cauchy distribution, the j-th coordinate of $f(s)$ is distributed as $\|s\|_1 Y_j$, for some i.i.d. $Y_j \sim \mathcal{D}_C$. Hence, $\|f(s)\|_1 = \|s\|_1 \cdot S$ where $S := \sum_{j \in [k]} |Y_j|$.

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 can directly apply Theorem 1.5, as it has bounded diameter.

Handling far points. The next lemma guarantees the low distortion for the points of the latter set, i.e. 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 utilize the tools of section 3.

Lemma 3.5. Fix a query point $q \in \ell_1^d$. For any $\varepsilon \in (0, 1/2)$, $c \geq 1$, $\delta \in (0, 1)$, there exists $D_0 = O(\log(k/\varepsilon))$ such that for $k = \Theta\left(\log^2 \lambda_P \cdot \log(c/\varepsilon) + \log(1/\delta)\right)$, 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. Let $D_0 > 1$ and assume wlog that the query point lies at the origin ($q = \mathbf{0}$). We define the following subsets of $\mathcal{N}$:

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

Notice that $N_i \subseteq B_1(q, D_{i+1}) \cap \mathcal{N}$. Then, by Lemma 2.12 for $r = \varepsilon/c$, $|N_i|$ is at most $\lambda_P^{\lceil \log(4cD_{i+1}/\varepsilon) \rceil} \leq \lambda_P^{4\log(cD_{i+1}/\varepsilon)}$. Therefore, by the union bound, and Lemma 3.2

$$\begin{aligned}
\mathbf{Pr}\left[\exists i \exists s \in N_i : \|f(s)\|_1 \leq \frac{4\|s\|_1}{D_i}\right] &= \mathbf{Pr}\left[\exists i \exists s \in N_i : S \leq \frac{4T}{D_i}\right] \\
&\leq \sum_{i=0}^{\infty} |N_i| \mathbf{Pr}\left[\tilde{S} \leq \frac{\sqrt{4kT}}{\sqrt{D_i}}\right] \\
&= \sum_{i=0}^{\infty} |N_i| \mathbf{Pr}\left[\tilde{S} \leq \mathbb{E}[\tilde{S}] \cdot \sqrt{\frac{2T}{k 2^{2i} D_0}}\right]
\end{aligned}$$

For $D_0 = \lceil 800T/k \rceil = \Theta(\log(k/\varepsilon))$ and $k > 4 \cdot \log \lambda_P \cdot \log(cD_0/\varepsilon) + 2\log(2\lambda_P/\delta)$, by Lemma 3.3:

$$\begin{aligned}
\sum_{i=0}^{\infty} |N_i| \mathbf{Pr}\left[\tilde{S} \leq \frac{\mathbb{E}[\tilde{S}]}{10 \cdot 2^{i+1}}\right] &\leq \sum_{i=0}^{\infty} \lambda_P^{4\log(cD_{i+1}/\varepsilon)} \left(\frac{1}{2^{i+1}}\right)^k \\
&= \sum_{i=0}^{\infty} \frac{2^{\log(\lambda_P)(4\log(cD_0/\varepsilon)+2i+2)}}{2^{k(i+1)}} \\
&\leq \sum_{i=0}^{\infty} \frac{2^{\log(\lambda_P)\cdot 4\log(cD_0/\varepsilon)} \cdot 2^{2\log(\lambda_P)(i+1)}}{2^{(4\cdot\log\lambda_P\cdot\log(cD_0/\varepsilon))(i+1)} \cdot 2^{2\log(2\lambda_P/\delta))(i+1)}} \\
&\leq \sum_{i=0}^{\infty} 2^{-2\log(2/\delta))(i+1)} \\
&= \sum_{i=0}^{\infty} \left(\frac{\delta^2}{4}\right)^i - 1 \\
&= \frac{\delta^2}{4 - \delta^2} \\
&\leq \delta.
\end{aligned}$$

Finally, for some large enough constant C, we demand

$$\begin{aligned} 4 \cdot \log \lambda_P \cdot \log(cD_0/\varepsilon) + 2\log(2\lambda_P/\delta) &< 4(\log \lambda_P \cdot \log(cD_0/\varepsilon) + \log(\lambda_P/\delta)) \\ &< 8(\log \lambda_P \cdot \log(cD_0/\varepsilon) + \log(1/\delta)) \\ &< C(\log \lambda_P \cdot \log(c \log k/\varepsilon) + \log(1/\delta)) \\ &< k. \end{aligned}$$

which is satisfied for $k = \Theta\left(\log^2 \lambda_P \cdot \log(c/\varepsilon) + \log(1/\delta)\right)$. ■

Now that we handled the (very) far neighbors, we can prove that the mapping satisfies the desirable conditions for a near neighbor-preserving embedding.

Theorem 3.6. Let $P \subset \ell_1^d$ such that $|P| = n$. For any $\varepsilon < 1/2$ and $c \geq 1$, there is a non-linear randomized embedding $h = f \circ g : \ell_1^d \to \ell_1^k$ with target dimension $k = (\log \lambda_P \cdot \log(c/\varepsilon))^{\Theta(1/\varepsilon)}/\zeta(\varepsilon)$, for a function $\zeta(\varepsilon) > 0$ 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 $\Omega(\varepsilon)$,

$$\begin{aligned} &\|h(p^*) - f(q)\|_1 \leq 1 + 3\varepsilon, \\ &\forall p \in P : \|p - q\|_1 > 1 + 9\varepsilon \implies \|h(p) - f(q)\|_1 > 1 + 3\varepsilon. \end{aligned}$$

The 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 $O(dk)$.

Proof. Let $D_0 = \Theta(\log(k/\varepsilon))$ and assume for simplicity (wlog) that $q = \mathbf{0}$. Then, by Lemma 3.5 for $k = \Theta\left(\log^2 \lambda_P \cdot \log(c/\varepsilon)\right)$, with probability at least $1 - \varepsilon/5$

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

By Theorem 1.5 for $\gamma = \varepsilon/10$ and $\delta = \varepsilon/(5\lambda_P^{8\log(cD_0/\varepsilon)})$, with probability at least $1 - \varepsilon/5$

$$\begin{aligned} \forall p \in P : \|p - q\|_1 \in (1 + 9\varepsilon, D_0 + \varepsilon) \implies \|h(p) - f(q)\|_1 &\geq (1 + 8\varepsilon)(1 - \varepsilon) \\ &> 1 + 3\varepsilon. \end{aligned}$$

Moreover,

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

The target dimension then, needs to satisfy

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

Hence, for $k = (\log \lambda_P \cdot \log(c/\varepsilon))^{\Theta(1/\varepsilon)}/\zeta(\varepsilon)$ we achieve total probability of success $\Omega(\varepsilon)$, which completes the proof. ■

3.4 Dimension reduction via randomly shifted grids

In this section, we present a simplified embedding which grids instead of nets. Recall the randomly shifted grid function

$$g_w(x) = \left(\left\lfloor \frac{x_1 - t_1}{w} \right\rfloor, ..., \left\lfloor \frac{x_d - t_d}{w} \right\rfloor \right),$$

where $t_1, \ldots, t_d$ independent uniform random variables in the interval $[0, w]$. induces a randomly shifted grid in $\mathbb{R}^d$.

Now, let $P \subset \ell_1^d$ be a set of n points with doubling constant λ_P, $q \in \ell_1^d$ a query point, and $\varepsilon > 0$. The embedding is similar as before: First, we compute $\mathcal{G} := g_{\varepsilon/d}(P)$, and then we randomly project $\mathcal{G}$ and q with f, as in Theorem 1.5.

By choosing side length $w = \varepsilon/d$, we have that the ℓ_1-diameter of each cell is ε, and therefore $\mathcal{G}$ is an ε-covering set of P. That is, every point $p \in P$ lies within distance ε from some point in $\mathcal{G}$.

Properties of $\mathcal{G}$.

We can bound the number of points of $\mathcal{G}$ that lie inside a bounded ball around the query, using the doubling constant:

Lemma 3.7. Let $R > 1$ and $P' := B_1(q, R) \cap P$. Then, for $w = \varepsilon/d$

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

Proof. By the doubling constant, there exists a set of balls of radius ε/d^2 centered at points in P', of cardinality at most $\lambda_P^{2\log(dR/\varepsilon)}$ which covers P'. By Lemma 2.16, the expected number of grid points contained in each ball is at most

$$\left(1 + 2(\varepsilon/d^2)/(\varepsilon/d)\right)^d = (1 + 2/d)^d \leq \mathrm{e}^2.$$

Hence, the proof 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 3.8. Let $\{D_i\}_{i \in \mathbb{N}}$ be a sequence of radii such that for any i: $D_{i+1} = 4 \cdot D_i$, and 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 $1/3$,

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

Proof. By Lemma 3.7, $\mathbb{E}[|A_i|] \leq 8\lambda_P^{2\log(2dD_{i+1}/\varepsilon)}$ for every $i \in \{-1, 0, \ldots\}$. Then, a union bound followed by Markov's inequality yields

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

In addition,

$$\mathbf{Pr}\left[|A_{-1}| \geq 4\, \mathbb{E}[|A_i|]\right] \leq 1/4.$$ ■

Embedding analysis

The analysis is the same as with approximate nets; first we handle the representatives the lie sufficiently far from the query, and we invoke Theorem 1.5 for the rest.

The next lemma is analogous to Lemma 3.5.

Lemma 3.9. Fix a query point $q \in \ell_1^d$. For any $\varepsilon < 1/2$, $\delta \in (0,1)$, there exists $D_0 = O(\log(k/\varepsilon))$ such that for $k = \Theta\left(\log^2 \lambda_P \cdot \log(d/\varepsilon) + \log(1/\delta)\right)$, with probability at least $1-\delta$,

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

Proof. Let $D_0 > 1$ and assume wlog that the query point lies at the origin ($q = \mathbf{0}$). We define the following subsets of $\mathcal{N}$:

$$G_i = \{s \in \mathcal{G} \mid D_i \leq \|s\|_1 < D_{i+1}\},\ D_i = 2^{2i} D_0,\ i = 0,1,2,\ldots$$

Notice that $G_i \subseteq B_1(q, D_{i+1}) \cap \mathcal{G}$. Then, by Lemma 3.8, with constant probability,

$$\forall i : |G_i| \leq 4^{i+3} \lambda_P^{2\log(2dD_{i+1}/\varepsilon)}.$$

Therefore, by the union bound, and Lemma 3.2

$$\begin{aligned}
\mathbf{Pr}\left[\exists i \exists s \in G_i : \|f(s)\|_1 \leq \frac{4\|s\|_1}{D_i}\right] &= \mathbf{Pr}\left[\exists i \exists s \in G_i : S \leq \frac{4T}{D_i}\right] \\
&\leq \sum_{i=0}^{\infty} |G_i| \,\mathbf{Pr}\left[\tilde{S} \leq \frac{\sqrt{4kT}}{\sqrt{D_i}}\right] \\
&= \sum_{i=0}^{\infty} |G_i| \,\mathbf{Pr}\left[\tilde{S} \leq \mathbb{E}[\tilde{S}] \cdot \sqrt{\frac{2T}{k2^{2i}D_0}}\right]
\end{aligned}$$

For $k > 2\log\lambda_P \cdot \log(2dD_0/\varepsilon) + 4\log(2\lambda_P/\delta) + 6$ and $D_0 = \lceil 800T/k \rceil = \Theta(\log(k/\varepsilon))$, by Lemma 3.3:

$$\begin{aligned}
\sum_{i=0}^{\infty} |G_i| \,\mathbf{Pr}\left[\tilde{S} \leq \frac{\mathbb{E}[\tilde{S}]}{10 \cdot 2^{i+1}}\right] &\leq \sum_{i=0}^{\infty} 4^{i+3} \lambda_P^{2\log(2dD_{i+1}/\varepsilon)} \left(\frac{1}{2^{i+1}}\right)^k. \\
&= \sum_{i=0}^{\infty} \frac{2^{2\log(\lambda_P)(\log(2dD_0/\varepsilon)+2i+2)+2i+6}}{2^{k(i+1)}} \\
&\leq \sum_{i=0}^{\infty} \frac{2^{2\log(\lambda_P)\cdot\log(2dD_0/\varepsilon)} \cdot 2^{4\log(\lambda_P)(i+1)} \cdot 2^{2i+6}}{2^{[\log(\lambda_P)\cdot\log(2dD_0/\varepsilon)+4\log(2\lambda_P/\delta)+6](i+1)}} \\
&\leq \sum_{i=0}^{\infty} 2^{-2\log(4/\delta)(i+1)} \\
&= \sum_{i=0}^{\infty} \left(\frac{\delta^4}{16}\right)^i - 1 \\
&= \frac{\delta^4}{16-\delta^4} \\
&\leq \delta.
\end{aligned}$$

For some large enough constant C, we demand

$$\begin{aligned} 2\log\lambda_P \cdot \log(2dD_0/\varepsilon) + 4\log(2\lambda_P/\delta) + 6 &< 8[\log\lambda_P \cdot \log(dD_0/\varepsilon) + \log(\lambda_P/\delta)] \\ &< 16[\log\lambda_P \cdot \log(dD_0/\varepsilon) + \log(1/\delta)] \\ &< C[\log\lambda_P \cdot \log(d\log k/\varepsilon) + \log(1/\delta)] \\ &< k. \end{aligned}$$

which is satisfied for $k = \Theta\left(\log^2\lambda_P \cdot \log(d/\varepsilon) + \log(1/\delta)\right)$. ■

Finally, we prove that the embedding is $(1+6\varepsilon)$-Near neighbor-preserving, with probability of correctness $\Omega(\varepsilon)$. We follow the same reasoning as in the proof of Theorem 3.6.

Theorem 3.10. Let $P \subset \ell_1^d$ such that $|P| = n$. For any $\varepsilon < 1/2$, there is a non-linear randomized embedding $h' : \ell_1^d \to \ell_1^k$, where $k = (\log\lambda_P \cdot \log(d/\varepsilon))^{\Theta(1/\varepsilon)}/\zeta(\varepsilon)$, for a function $\zeta(\varepsilon) > 0$ depending only on ε, such that for any $q \in \ell_1^d$, if there exists $p^* \in P$ such that $\|p^* - q\|_1 \le 1$, then with probability $\Omega(\varepsilon)$,

$$\begin{aligned} &\|h'(p^*) - f(q)\|_1 \le 1 + 3\varepsilon, \\ &\forall p \in P : \|p - q\|_1 > 1 + 9\varepsilon \implies \|h'(p) - f(q)\|_1 > 1 + 3\varepsilon. \end{aligned}$$

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

Proof. Let $h' = f \circ g_{\varepsilon/d}$, where f is the randomized linear map defined in section 4. As before, we apply h' for every point in P, and only f for queries. Note that the randomly shifted grid incurs an additive error of ε in the distances between q and P.

Let $D_0 = \Theta(\log(k/\varepsilon))$ and assume for simplicity (wlog) that $q = \mathbf{0}$. Then, by Lemma 3.9, for $k = \Theta\left(\log^2\lambda_P \cdot \log(d/\varepsilon)\right)$, with probability at least $1 - \varepsilon/5$

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

Now, we are able to use Theorem 1.5 for points which are at distance at most $D_0 + \varepsilon$ from q, and the near neighbor. By Lemma 3.8, with constant probability, the number of grid points at distance $\le D_0 + \varepsilon$, is at most $32 \cdot \lambda_P^{4\log(dD_0/\varepsilon)}$. Hence, by Theorem 1.5 for $\gamma = \varepsilon/10$ and $\delta = \varepsilon/(160\lambda_P^{4\log(dD_0/\varepsilon)})$, with probability at least $1 - \varepsilon/5$

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

Moreover, with probability at least $\varepsilon/2$

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

As in Theorem 3.6, the target dimension needs to satisfy

$$k \ge \frac{\left(\ln\left(160\lambda_P^{4\log(dD_0/\varepsilon)}/\varepsilon\right)\right)^{2/\varepsilon}}{\zeta(\varepsilon)} = \frac{\left(\Theta(\log\log k \cdot \log\lambda_P + \log\lambda_P \cdot \ln(d/\varepsilon))\right)^{2/\varepsilon}}{\zeta(\varepsilon)}$$

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

CHAPTER 4

CONCLUSION

Discussion on results

The results in this thesis show that efficient dimensionality reduction for doubling subsets of ℓ_1 is possible, in the context of ANN searching. We presented two versions of a non-linear, randomized embedding. In each version, we represent the pointset with a covering set, and then randomly project this set and any query. The embedding is $(1+6\varepsilon)$-near neighbor-preserving, with probability of correctness $\Omega(\varepsilon)$. In the first version–Theorem 3.6–we used approximate nets as the representative set, while the second version–Theorem 3.10–considers randomly shifted grids instead. The following table compares target dimension and computation time.

Embedding	Target dimension	Computation time
Theorem 3.6	$k = (\log \lambda_P \cdot \log(\mathbf{c}/\varepsilon))^{\Theta(1/\varepsilon)} / \zeta(\varepsilon)$	$\tilde{O}(dn^{1+1/\Omega(c)})$
Theorem 3.10	$k = (\log \lambda_P \cdot \log(\mathbf{d}/\varepsilon))^{\Theta(1/\varepsilon)} / \zeta(\varepsilon)$	$O(dkn)$

Theorem 3.6 provides with a trade-off between the preprocessing time required and the target dimension, via the parameter $c > 1$. For example, for $c = O(1)$, target dimension depends only on $\log \lambda_P$ and ε, and preprocessing time is $\tilde{O}(dn^{1+\rho})$, for $\rho < 1$. Setting, $c = \log n$, reduces the preprocess to near linear, $\tilde{O}(dn^{1+o(1)})$, but k becomes polynomial in $\log \log n$.

On the other hand, Theorem 3.10 has the advantage of fast preprocessing; any point can be embedded in $O(dk)$ time. In addition, the embedding is oblivious to the pointset, as it is well defined on the grid, and may also be of practical interest since it is very easy to implement. However, the target dimension now depends on $\log d$ instead of $\log c$. For near-linear preprocessing, Theorem 3.10 is obviously stronger than Theorem 3.6 when $d = \text{polylog}(n)$, as we can achieve the same target dimension with exact linear preprocessing time.

Notice that any potential improvements to Theorem 1.5 could lead to improvements of Theorems 3.6 and 3.10. The target dimension in both theorems derives from a direct application of Theorem 1.5 to the representative points which lie inside a bounding ball centered at the query.

Algorithmic implications

Our embedding can be combined with the bucketing method of [HIM12] for the $(1+\varepsilon)$-ANN problem in ℓ_1^d, admitting better space bounds, while preserving the support of fast queries. This data structure requires preprocessing time and space $O(n)\times O(1/\varepsilon)^d$ and has query time $O(d)$. Space and preprocessing bounds can be improved to $n^{O(\log(1/\varepsilon)/\varepsilon^2)}$, with slightly slower $O(d\cdot\log n/\varepsilon^2)$ query time, for general subsets of ℓ_1. This is done via a JL-like dimension reduction, where the points are first mapped to the Hamming space, which is isometric to ℓ_2.

However, for subsets of ℓ_1 of fixed doubling dimension, one can combine the general bucketing data structure with the embedding of Theorem 3.6, improving upon the bounds obtained by the JL-like embedding, both in space and query time. The following table sums up the comparison.

	JL	Thm 3.6 ($c=\log n$)	Thm 3.6 ($c=O(1)$)
Preprocess	$n^{O(\log(\frac{1}{\varepsilon})/\varepsilon^2)}$	$\tilde{O}(dn^{1+o(1)})$	$\tilde{O}(dn^{1+1/c})$
Space	$n^{O(\log(\frac{1}{\varepsilon})/\varepsilon^2)}$	$n^{1+o(1)}$	$n^{1+o(1)}$
Query	$O(d\cdot\log n/\varepsilon^2)$	$O(d)\times(\log\log n)^{O(1/\varepsilon)}$	$O(d)\times(\log(\frac{1}{\varepsilon}))^{O(1/\varepsilon)}$

Open Questions

An immediate open question is whether better bounds on the target dimension can be proved for the embedding of Theorem 1.5. As already mentioned, a positive answer would imply better bounds for Theorems 3.6 and 3.10 as well. Moreover, an analysis of our embedding which does not rely on the employment of Theorem 1.5, might also be of interest.

Stable distributions have been used in designing LSH families and non-linear range embeddings for ℓ_p norms, with provable guarantees. An interesting open problem then is whether the linear mapping à la Johnson-Lindenstrauss can be proven to be NN-preserving for general (or even doubling) subsets of ℓ_p, $p\in(0,2]$, by substituting the Gaussian matrix with a matrix whose elements are i.i.d. p-stable variables.

APPENDICES

A Proof of Claim 2.2

Claim. For any vector $x \in \mathbb{R}^d$ and $p > q > 0$

$$\|x\|_p \leq \|x\|_q \leq d^{1/q-1/p} \|x\|_p .$$

Proof. We start with the left part of the inequality by showing that $\|x\|_p \leq \|x\|_q$. If $x = \mathbf{0}$, its obviously true. Otherwise, let $y_i = |x_i| / \|x\|_p \leq 1$, for every $i \in [d]$. Therefore,

$$\forall i \in [d] : y_i^p \leq y_i^q \implies 1 \leq \|y\|_q \implies \|x\|_p \leq \|x\|_q .$$

For the right part of the inequality, we invoke Hölder's Inequality: Let $r \in [1, \infty)$. Then for any $a, b \in \mathbb{R}^d$

$$\sum_{i=1}^{d} |a_i b_i| \leq \left(\sum_{i=1}^{d} |a_i|^r \right)^{1/r} \left(\sum_{i=1}^{d} |b_i|^{\frac{r}{r-1}} \right)^{1-1/r} .$$

Now, set $r = p/q > 1$, $a = (|x_1|^q, \ldots, |x_d|^q)$ and $b = \mathbf{1}$. Then,

$$\begin{aligned}
\sum_{i=1}^{d} |a_i b_i| &\leq \left(\sum_{i=1}^{d} |a_i|^r \right)^{1/r} \left(\sum_{i=1}^{d} |b_i|^{\frac{r}{r-1}} \right)^{1-1/r} , \\
\sum_{i=1}^{d} |x_i|^q &\leq \left(\sum_{i=1}^{d} |x_i|^p \right)^{q/p} d^{1-q/p}, \\
\left(\sum_{i=1}^{d} |x_i|^q \right)^{1/q} &\leq \left(\left(\sum_{i=1}^{d} |x_i|^p \right)^{q/p} d^{1-q/p} \right)^{1/q} .
\end{aligned}$$

Subsequently,

$$\|x\|_q \leq d^{1/q-1/p} \|x\|_p . \qquad \blacksquare$$

B Proof of Equality (3.1)

Claim. For any $t > 0$

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

Proof. We have

$$\int_0^1 \mathrm{e}^{-tx^{1/2}}\,\mathrm{d}x = \int_0^1 \frac{\mathrm{e}^{-tx^{1/2}}}{x^{1/2}} \cdot x^{1/2}\,\mathrm{d}x = \int_0^1 \left(\frac{-2\mathrm{e}^{-tx^{1/2}}}{t}\right)' \cdot x^{1/2}\,\mathrm{d}x.$$

Integration by parts yields

$$\begin{aligned}\int_0^1 \left(\frac{-2\mathrm{e}^{-tx^{1/2}}}{t}\right)' \cdot x^{1/2}\,\mathrm{d}x &= \left[\frac{-2\mathrm{e}^{-tx^{1/2}}}{t} \cdot x^{1/2}\right]_0^1 - \int_0^1 \frac{-2\mathrm{e}^{-tx^{1/2}}}{t} \cdot (x^{1/2})'\,\mathrm{d}x \\ &= \frac{-2\mathrm{e}^{-t}}{t} + \int_0^1 \frac{\mathrm{e}^{-tx^{1/2}}}{tx^{1/2}}\,\mathrm{d}x.\end{aligned}$$

Set $u = tx^{1/2}$, which implies

$$\frac{2}{t}\,\mathrm{d}u = \frac{1}{x^{1/2}}\,\mathrm{d}x.$$

Therefore,

$$\begin{aligned}\frac{-2\mathrm{e}^{-t}}{t} + \int_0^1 \frac{\mathrm{e}^{-tx^{1/2}}}{tx^{1/2}} &= \frac{-2\mathrm{e}^{-t}}{t} + \frac{2}{t^2}\int_0^t \mathrm{e}^{-u}\,\mathrm{d}u. \\ &= \frac{-2t\mathrm{e}^{-t}}{t^2} + \frac{2}{t^2}(1 - \mathrm{e}^{-t}) \\ &= \frac{2}{t^2}\left(1 - \frac{t+1}{\mathrm{e}^t}\right).\end{aligned}$$

■

BIBLIOGRAPHY

[AC09] N. AILON and B. CHAZELLE. "The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors". *SIAM Journal on Computing* 39.1 (2009), pp. 302–322.

[Ach03] D. ACHLIOPTAS. "Database-friendly random projections: Johnson Lindenstrauss with binary coins". *Journal of Computer and System Sciences* 66.4 (2003), pp. 671–687.

[AEP18] E. ANAGNOSTOPOULOS, I. Z. EMIRIS, and I. PSARROS. "Randomized Embeddings with Slack and High-Dimensional Approximate Nearest Neighbor". *ACM Transactions on Algorithms* 14.2 (2018), pp. 1–21.

[AFM18] S. ARYA, G. D. da FONSECA, and D. M. MOUNT. "Approximate Polytope Membership Queries". *SIAM Journal on Computing* 47.1 (2018), pp. 1–51.

[AI08] A. ANDONI and P. INDYK. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions". *Communications of the ACM* 51.1 (2008), p. 117.

[AMM09] S. ARYA, T. MALAMATOS, and D. M. MOUNT. "Space-time tradeoffs for approximate nearest neighbor searching". *Journal of the ACM* 57.1 (2009), pp. 1–54.

[And+14] A. ANDONI et al. "Beyond Locality-Sensitive Hashing". *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, Portland, Oregon, USA*. 2014, pp. 1018–1028.

[And+17a] A. ANDONI et al. "Approximate near neighbors for general symmetric norms". *Proceedings of the 49th Annual ACM Symposium on Theory of Computing, STOC, Montreal, QC, Canada*. 2017, pp. 902–913.

[And+17b] A. ANDONI et al. "Optimal Hashing-based Time-Space Trade-offs for Approximate Near Neighbors". *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, Barcelona, Spain*. 2017, pp. 47–66.

[AR15] A. ANDONI and I. P. RAZENSHTEYN. "Optimal Data-Dependent Hashing for Approximate Near Neighbors". *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, Portland, OR, USA*. 2015, pp. 793–801.

[AR16] A. ANDONI and I. P. RAZENSHTEYN. "Tight Lower Bounds for Data-Dependent Locality-Sensitive Hashing". *32nd International Symposium on Computational Geometry, SoCG, Boston, MA, USA*. 2016, 9:1–9:11.

[Ary+98] S. ARYA et al. "An optimal algorithm for approximate nearest neighbor searching fixed dimensions". *Journal of the ACM* 45.6 (1998), pp. 891–923.

[BC05] B. BRINKMAN and M. CHARIKAR. "On the impossibility of dimension reduction in l1". *Journal of the ACM* 52.5 (2005), pp. 766–788.

[BG16] Y. BARTAL and L. GOTTLIEB. "Dimension Reduction Techniques for l_p (1<p<2), with Applications". *32nd International Symposium on Computational Geometry, SoCG, Boston, MA, USA*. 2016, 16:1–16:15.

[BG19] Y. BARTAL and L. GOTTLIEB. "Approximate nearest neighbor search for l_p-spaces (2<p<∞)". *Theor. Comput. Sci.* 757 (2019), pp. 27–35.

[BKL06] A. BEYGELZIMER, S. KAKADE, and J. LANGFORD. "Cover trees for nearest neighbor". *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML), Pittsburgh, Pennsylvania, USA*. 2006, pp. 97–104.

[CG06] R. COLE and L. GOTTLIEB. "Searching dynamic point sets in spaces with bounded doubling dimension". *Proceedings of the 38th Annual ACM Symposium on Theory of Computing, Seattle, WA, USA*. 2006, pp. 574–583.

[Cha17] T. M. CHAN. "Orthogonal Range Searching in Moderate Dimensions: k-d Trees and Range Trees Strike Back". *33rd International Symposium on Computational Geometry, SoCG, Brisbane, Australia*. 2017, 27:1–27:15.

[Cla99] K. L. CLARKSON. "Nearest Neighbor Queries in Metric Spaces". *Discrete & Computational Geometry* 22.1 (1999), pp. 63–93.

[CMS76] J. M. CHAMBERS, C. L. MALLOWS, and B. W. STUCK. "A Method for Simulating Stable Random Variables". *Journal of the American Statistical Association* 71.354 (1976), pp. 340–344.

[Dat+04] M. DATAR et al. "Locality-sensitive hashing scheme based on p-stable distributions". *Proceedings of the 20th ACM Symposium on Computational Geometry, Brooklyn, New York, USA*. 2004, pp. 253–262.

[Dub10] M. DUBINER. "Bucketing Coding and Information Theory for the Statistical High-Dimensional Nearest-Neighbor Problem". *IEEE Transactions on Information Theory* 56.8 (2010), pp. 4166–4179.

[EHS15] D. EPPSTEIN, S. HAR-PELED, and A. SIDIROPOULOS. "Approximate Greedy Clustering and Distance Selection for Graph Metrics". *CoRR* abs/1507.01555 (2015). arXiv: 1507.01555.

[HIM12] S. HAR-PELED, P. INDYK, and R. MOTWANI. "Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality". *Theory of Computing* 8.1 (2012), pp. 321–350.

[HM06] S. HAR-PELED and M. MENDEL. "Fast Construction of Nets in Low-Dimensional Metrics and Their Applications". *SIAM Journal on Computing* 35.5 (2006), pp. 1148–1184.

[IM98] P. INDYK and R. MOTWANI. "Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality". *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA*. 1998, pp. 604–613.

[IN07] P. INDYK and A. NAOR. "Nearest-neighbor-preserving embeddings". *ACM Transactions on Algorithms* 3.3 (2007), 31–es.

[Ind01] P. INDYK. "On Approximate Nearest Neighbors under l∞ Norm". *Journal of Computer and System Sciences* 63.4 (2001), pp. 627–638.

[Ind06] P. INDYK. "Stable distributions, pseudorandom generators, embeddings, and data stream computation". *Journal of the ACM* 53.3 (2006), pp. 307–323.

[JL84] W. B. JOHNSON and J. LINDENSTRAUSS. "Extensions of Lipschitz mappings into a Hilbert space". *Proc. Conf. in modern analysis and probability (New Haven, Conn.)* Vol. 26. Contemporary Mathematics. American Mathematical Society, 1984, pp. 189–206.

[KL04] R. KRAUTHGAMER and J. R. LEE. "Navigating nets: simple algorithms for proximity search". *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA, New Orleans, Louisiana, USA*. 2004, pp. 798–807.

[KR02] D. R. KARGER and M. RUHL. "Finding nearest neighbors in growth-restricted metrics". *Proceedings on 34th Annual ACM Symposium on Theory of Computing, Montréal, Québec, Canada*. 2002, pp. 741–750.

[LMN05] J. R. LEE, M. MENDEL, and A. NAOR. "Metric structures in L_1: dimension, snowflakes, and average distortion". *Eur. J. Comb.* 26.8 (2005), pp. 1180–1190.

[MNP07] R. MOTWANI, A. NAOR, and R. PANIGRAHY. "Lower Bounds on Locality Sensitive Hashing". *SIAM J. Discrete Math.* 21.4 (2007), pp. 930–935.

[MO15] A. MAY and I. OZEROV. "On Computing Nearest Neighbors with Applications to Decoding of Binary Linear Codes". *Advances in Cryptology - EUROCRYPT - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, Bulgaria*. 2015, pp. 203–228.

[Nol18] J. P. NOLAN. *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhauser, 2018.

[OWZ14] R. O'DONNELL, Y. WU, and Y. ZHOU. "Optimal Lower Bounds for Locality-Sensitive Hashing (Except When q is Tiny)". *ACM Transactions on Computation Theory* 6.1 (2014), pp. 1–13.

[SDI06] G. SHAKHNAROVICH, T. DARRELL, and P. INDYK. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. The MIT Press, 2006.

[Zol86] V. M. ZOLOTAREV. *One-Dimensional Stable Distributions (Translations of Mathematical Monographs - Vol 65)*. American Mathematical Society, 1986.