Data-Dependent Hashing via Nonlinear Spectral Gaps

Alexandr Andoni
Columbia University
andoni@cs.columbia.edu

Assaf Naor
Princeton University
naor@math.princeton.edu

Aleksandar Nikolov
University of Toronto
anikolov@cs.toronto.edu

Ilya Razenshteyn
Microsoft Research Redmond
ilyaraz@microsoft.com

Erik Waingarten
Columbia University
eaw@cs.columbia.edu

April 16, 2018

Abstract

We establish a generic reduction from *nonlinear spectral gaps* of metric spaces to data-dependent Locality-Sensitive Hashing, yielding a new approach to the high-dimensional Approximate Near Neighbor Search problem (ANN) under various distance functions. Using this reduction, we obtain the following results:

- For *general* d-dimensional normed spaces and n-point datasets, we obtain a *cell-probe* ANN data structure with approximation $O(\frac{\log d}{\varepsilon^2})$, space $d^{O(1)}n^{1+\varepsilon}$, and $d^{O(1)}n^{\varepsilon}$ cell probes per query, for any $\varepsilon > 0$. No non-trivial approximation was known before in this generality other than the $O(\sqrt{d})$ bound which follows from embedding a general norm into ℓ_2.
- For ℓ_p and Schatten-p norms, we improve the data structure further, to obtain approximation $O(p)$ and sublinear query *time*. For ℓ_p, this improves upon the previous best approximation $2^{O(p)}$ (which required polynomial as opposed to near-linear in n space). For the Schatten-p norm, no non-trivial ANN data structure was known before this work.

Previous approaches to the ANN problem either exploit the low dimensionality of a metric, requiring space exponential in the dimension, or circumvent the curse of dimensionality by embedding a metric into a "tractable" space, such as ℓ_1. Our new generic reduction proceeds differently from both of these approaches using a novel partitioning method.

Contents

1 Introduction **3**
- 1.1 ANN for general distance functions . . . 3
- 1.2 Main results . . . 4
- 1.3 Techniques . . . 5
- 1.4 Related work . . . 8
- 1.5 Lower bounds . . . 9
- 1.6 Open problems . . . 10
- 1.7 Organization of the paper . . . 10

2 Preliminaries **11**

3 Partitioning general metrics **12**
- 3.1 Cutting modulus of a metric space . . . 12
- 3.2 Partitioning theorems . . . 13
  - 3.2.1 Partitioning with the (R, ε)-ball-or-cut property . . . 14
  - 3.2.2 Inner multiplicative weights update . . . 15
  - 3.2.3 Outer multiplicative weights update: proof of Theorem 3.6 . . . 17

4 Cell-probe data structure for general metrics **20**

5 Discretizing the space **23**

6 Bounding the cutting modulus of a normed space **25**

7 Algorithm for ℓ_p **29**
- 7.1 Rayleigh quotient inequality for ℓ_p spaces and proof of Lemma 7.4 . . . 30
- 7.2 Proof of Lemma 7.5 . . . 33

8 Algorithm for Schatten-p **34**
- 8.1 Rayleigh quotient inequality for S_p, $p > 2$ . . . 36
- 8.2 Proof of Lemma 8.3 . . . 42
- 8.3 The case of $1 \le p \le 2$ . . . 44

9 Lower bounds **46**
- 9.1 General norms do not admit succinct collections . . . 46
- 9.2 Lower bound for random partitions . . . 48

10 Acknowledgments **51**

1 Introduction

The *c-Approximate Near Neighbor Search* (c-ANN) problem is defined as follows. Given an n-point dataset $P \subset X$ lying in a metric space (X, d_X), we want to preprocess P to answer *approximate near neighbor queries* quickly. Namely, given a query point $q \in X$ such that there is a data point $p^* \in P$ with $d_X(q, p^*) \leq r$, the algorithm should return a data point $\widehat{p} \in X$ with $d_X(q, \widehat{p}) \leq cr$. We refer to $c > 1$ as the *approximation* and $r > 0$ as the *distance scale*; both parameters are known during the preprocessing. The main quantities to optimize are the *space* the data structure occupies and the *time* it takes to answer a query. In addition to being an indispensable tool for data analysis, ANN data structures have spawned two decades of theoretical developments (see, e.g., the surveys [AI17, AIR18] and the thesis [Raz17] for an overview).

1.1 ANN for general distance functions

The best-studied metrics in the context of ANN are the Hamming/Manhattan (ℓ_1) and the Euclidean (ℓ_2) distances. Both ℓ_1 and ℓ_2 are very common in applications and admit efficient algorithms based on *hashing*: in particular, Locality-Sensitive Hashing (LSH) [IM98, AI06] and its data-dependent counterparts [AINR14, AR15, ALRW17]. Hashing-based algorithms for ANN over ℓ_1 and ℓ_2 have now been the subject of a long line of work, leading to a comprehensive understanding of the respective time–space trade-offs.

Beyond ℓ_1 and ℓ_2, the ANN landscape is much more mysterious despite having received significant attention (see Section 1.4 for an overview). In summary, we are still very far from having a general recipe for ANN data structures for *general* metrics with a non-trivial approximation. This state of affairs motivates the following broad question.

Problem 1.1. *For a given approximation $c > 1$, which metric spaces allow efficient ANN algorithms?*

An algorithm for general metrics is highly desirable both in theory and in practice. From the theoretical perspective, we are interested in a theory of ANN algorithms for a wide class of distance functions. Such a theory would yield data structures (or impossibility results) for a variety of important distances for which we still do not know efficient ANN algorithms (e.g. the Earth Mover's Distance (EMD), the edit distance, generalized versions of the Hamming distance[1], etc). Perhaps even more tantalizing is the question of understanding what geometric properties of a metric space govern the hardness of ANN. In addition to the theoretical interest, in practice, one often needs to tune the distance function to the specifics of the application, and hence generic ANN algorithms are also preferred.

In this paper, we focus on the following important case of Problem 1.1, which was first raised in 2010 [And10].

Problem 1.2. *Solve Problem 1.1 for d-dimensional* normed *spaces.*

[1] E.g., a metric of interest in applications is (X^d, ρ_{X^d}), where X is a metric itself, with the distance between vectors $x, y \in X^d$ defined as $\rho_{X^d}(x, y) = \sum_{i=1}^{d} d_X(x_i, y_i)$.

Most metrics arising in applications are actually norms (e.g., the ℓ_p distances, matrix norms, the Earth Mover's Distance, etc.). Besides that, norms are geometrically nicer than general metrics, so there is more hope for a coherent theory (e.g., for the problems of *sketching* and *streaming* norms, see the general results of [AKR15, BBC+17], for ANN over general *symmetric* norms, see a recent result [ANN+17]).

1.2 Main results

In this paper, we make progress towards resolving Problem 1.2. Our main contribution is a data structure for the $O(\log d)$-ANN problem over a general d-dimensional norm in the *cell-probe* model introduced by Yao [Yao81]. Prior to this work, the only other ANN data structure for general norms achieved approximation $O(\sqrt{d})$ (see Section 1.4).

Theorem 1.3. *Let $0 < \varepsilon < 1$. Suppose that $(\mathbb{R}^d, \|\cdot\|)$ is a d-dimensional normed space. Then there exists a randomized data structure for $O\left(\frac{\log d}{\varepsilon^2}\right)$-ANN over $\|\cdot\|$ with the following parameters:*

- *The space used by the data structure is $n^{1+\varepsilon} \cdot d^{O(1)}$;*
- *The query procedure probes $n^{\varepsilon} \cdot d^{O(1)}$* words in memory*, where words consist of $O(\log n)$ bits*[2].

Let us emphasize that we do not claim any *time* bound on the query procedure. We only restrict the number of memory locations the data structure is allowed to probe (see Section 4 for a further discussion of the model). Nonetheless, we conjecture that one can in fact obtain a data structure for $O(\log d)$-ANN with sublinear *time* query complexity (as opposed to cell probe complexity only), provided a suitable oracle access to the norm.

Irrespective of the conjecture, our theorem can be thought of as a barrier for proving impossibility of efficient ANN data structures with approximation $O(\log d)$ for general norms. This is because *all* known unconditional data structure lower bounds proceed by proving a cell-probe lower bound [Mil99]. Thus, a potential strong lower bound for the ANN problem would require a completely new approach to data structure lower bounds.

The main tool behind Theorem 1.3 is a new random partition for sets of points in a general normed space, and is of independent interest. In particular, we show how to convert an estimate on the nonlinear spectral gap of a metric space into a data-dependent Locality-Sensitive Hashing (LSH) family (see Section 1.3 for an overview).

Finally, our technique also gives a natural approach to designing data structures for *specific* metric spaces with better parameters, including *sublinear time*. Indeed, we instantiate our technique with the ℓ_p and Schatten-p norms, for which, with additional work, we obtain data structures with better approximations and sublinear time. For the ℓ_p norms, we obtain approximation $c = O(p)$, which improves exponentially over the approximation factor of $2^{O(p)}$ from [NR06, BG15] (see Section 1.4).

Theorem 1.4. *Let $0 < \varepsilon < 1$ and $2 < p \leq \infty$. There exists a randomized data structure for $O(p/\varepsilon)$-ANN over the ℓ_p norm with the following parameters:*

[2] We assume that all the coordinates of the dataset and query points as well as r can be stored in $O(\log n)$ bits.

- *The space used by the data structure is* $n^{1+\varepsilon} \cdot d^{O(1)}$*;*
- *The query procedure takes* time $n^{\varepsilon} \cdot d^{O(1)}$*.*

Generalizing the theorem from above to Schatten-p norms, we obtain the first ANN data structures with a non-trivial approximation factor. Recall that the Schatten-p norm $\|\cdot\|_{S_p}$ of a matrix is the ℓ_p norm of the vector of its singular values[3]. The challenge of designing ANN under Schatten norms was posed in [And10].

We state our Schatten-p results for regimes $1 \le p \le 2$ and $2 < p < \infty$ separately.

Theorem 1.5. *Let* $0 < \varepsilon < 1$ *and* $1 \le p \le 2$*. There exists a randomized data structure for c-ANN over the Schatten-p norm, where* $c = O\left(\frac{1}{\varepsilon^{2/p}}\right)$ *with the following parameters:*

- *The space used by the data structure is* $n^{1+\varepsilon} \cdot d^{O(1)}$*;*
- *The query procedure takes time* $n^{\varepsilon} \cdot d^{O(1)}$*.*

We now state the data structure for Schatten-p norms with $p > 2$. Compared to the ℓ_p algorithm from Theorem 1.4, the result for Schatten-p has worse dependence on the dimension for the space and query time. We note that for $p > \log d$, the norm $\|x\|_{S_p}$ is a constant factor from $\|x\|_{S_{\log d}}$; thus, it suffices to consider the cases when $2 < p \le \log d$.

Theorem 1.6. *Let* $0 < \varepsilon < 1$ *and* $2 < p \le \infty$*. There exists a randomized data structure for c-ANN over the Schatten-p norm, where* $c = O(p/\varepsilon)$ *with the following parameters:*

- *The space used by the data structure is* $n^{1+\varepsilon} \cdot d^{O(p)}$*;*
- *The query procedure takes time* $n^{\varepsilon} \cdot d^{O(p)}$*.*

See Section 1.4 for a more detailed exposition of how Theorem 1.5 and Theorem 1.6 relate to previously known results.

Let us note that the preprocessing procedures in all the new data structures are inefficient. Improving the preprocessing time is left as an interesting open problem.

1.3 Techniques

Nonlinear spectral gaps. At the conceptual level, the main contribution of the paper is a reduction from bounds on the *nonlinear spectral gap* to a data-dependent Locality-Sensitive Hashing (LSH) family for a *general metric space*. Let $A = (a_{ij}) \in \mathbb{R}^{m \times m}$ be a symmetric doubly stochastic $m \times m$ matrix. Then, for a metric space (X, d_X) and $q \ge 1$ the nonlinear spectral gap $\gamma(A, d_X^q)$ is the smallest number for which the following holds. For every set of points $x_1, x_2, \ldots, x_m \in X$,

$$\frac{1}{m} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} d_X(x_i, x_j)^q \le \gamma(A, d_X^q) \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} \cdot d_X(x_i, x_j)^q.$$

[3]The Schatten-1 norm is known under the names of the nuclear or trace norm, the Schatten-2 norm is simply the Frobenius norm, and Schatten-∞ is known as the spectral or the operator norm.

For the ℓ_2 norm, $\gamma(A, \|\cdot\|_2^2) = \frac{1}{1-\lambda_2(A)}$, where $\lambda_2(A)$ is the second largest eigenvalue of A; i.e. in this case $\gamma(A, \|\cdot\|_2^2)$ is the inverse of the usual spectral gap of A. A systematic study of nonlinear spectral gaps of metric spaces was initiated in [MN14]. Similar inequalities can be found in earlier works (see, e.g., the introduction of [MN15] for a thorough literature review); we single out the reference [Mat97] which is instrumental for our results. In the above-mentioned works, bounds on the nonlinear spectral gap were primarily used to show strong non-embeddability results.

We use the following recent result from [Nao17] in order to build a cell-probe data structure for ANN over a general *norm*, as claimed in Theorem 1.3.

Theorem 1.7 ([Nao17])**.** *For every norm $\|\cdot\|$ defined on $\mathbb{R}^d$, one has:*

$$\gamma(A, \|\cdot\|^2) = O\left(\frac{\log^2 d}{(1-\lambda_2(A))^2}\right).$$

We present a simplified proof of this theorem with a slight generalization to a weighted setting (which we need for the actual reduction) in Section 6.

At a high level, a strong enough upper bound on $\gamma(\cdot, d_X^q)$ in terms of $\gamma(\cdot, \|\cdot\|_2^2)$ gives a cell-probe data structure for ANN over a given metric space (X, d_X) using the reduction given in this paper. For the *time-efficient* data structures over ℓ_p and Schatten-p spaces (Theorem 1.4 and Theorem 1.6), we need the nonlinear spectral gap inequality in a strong *Rayleigh quotient* form. For the ℓ_p norms, such a stronger inequality was shown by Matoušek [Mat97]. We adapt Matoušek's inequality to the weighted setting in Section 7. For Schatten-p, the corresponding inequality is stated and proved in Section 8. The new inequality is an extension of the Matoušek's inequality to the matrix setting using estimates from [Ric15]. An additional twist compared to [Mat97] is the need for a fixed-point statement similar to the Brouwer's theorem.

Data-dependent LSH. We now briefly describe how to utilize Theorem 1.7 to obtain a data-dependent LSH family for a general norm. Informally, for a given dataset, we would like to design a random partition of $\mathbb{R}^d$ that separates a query point from *far data points* often, while not separating a query point from *close data points* too often. With such a random partition, we can build the data structure as simply a collection of random decision trees. In each node, we sample a partition from the family, split the dataset among child nodes accordingly, and recurse on each child node. This connection has already been used in [Ind01, AR15, ALRW17] (however, let us note that in [Ind01] space partitions are used in a fundamentally different way; see the discussion in Section 1.5).

The construction of the data-dependent LSH incorporates three main ideas.

- We use the multiplicative weights update algorithm (MWU) [AHK12] to reduce the problem of constructing a random partition to the problem of finding a deterministic partition that works on average with respect to a given distribution over points. This step is non-trivial since the resulting random partition must depend on the dataset fairly weakly so that a sample from it can be stored in $\mathrm{poly}(d)$ space. We end up using two levels of MWU, where the "outer" part is responsible for "guessing" the dataset iteratively, while the "inner" part finds a required

random partition for a current guess.

- The problem of finding a deterministic partition can be seen as finding a sparse cut in an undirected graph embedded in $(\mathbb{R}^d, \|\cdot\|)$ so that the following conditions hold. First, we assume that the distance between the endpoints of every edge is at most 1. Additionally, suppose that the distance between a typical pair of vertices is $\gg \log d$. It suffices to prove that this graph cannot be a spectral expander, since we may then employ Cheeger's inequality [Che69, Chu96] to obtain a sparse cut.

- Finally, Theorem 1.7 directly implies that expanders do not embed into $(\mathbb{R}^d, \|\cdot\|)$, so the above graph cannot be a spectral expander. In fact, if $A \in \mathbb{R}^{m\times m}$ is a normalized adjacency matrix of a graph and $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$ are the points the vertices are mapped to, then the following holds. Since every edge has length at most 1,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}\cdot\|x_i - x_j\|^2 \le \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} = m. \tag{1}$$

  Since a typical pair of vertices is at distance $\gg \log d$ apart,

$$\frac{1}{m}\cdot\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i - x_j\|^2 \gg m\cdot\log^2 d. \tag{2}$$

  Combining (1) and (2) with Theorem 1.7, we get that $1-\lambda_2(A) \ll 1$, which implies that the graph is not an expander.

Algorithmically, we construct a randomized space partition by combining the two-level MWU algorithm together with a spectral partitioning procedure. The new data-dependent LSH construction gives a generic approach to ANN, which departs substantially from the commonly-used embeddings technique.

Partitions of normed spaces. As mentioned briefly, Theorems 1.3, 1.4, 1.5, and 1.6 follow from new partitioning results for sets of points lying in normed spaces. The specific partitioning results are given in Sections 6, 7 and 8, respectively. Let us now state the partitioning results for ℓ_p spaces and for general normed spaces.

A *box* in $\mathbb{R}^d$ is an intersection of sets of the form $\{x \in \mathbb{R}^d \mid x_k \le u\}$ or $\{x \in \mathbb{R}^d \mid x_k \ge u\}$, where $1 \le k \le d$ and $u \in \mathbb{R}$. In Section 7, we obtain the following partitioning result for ℓ_p spaces.

Theorem 1.8. *Let $0 < \varepsilon < 1$, $2 < p < \infty$ and $R > 0$. Consider any dataset $P \subset \mathbb{R}^d$ of n points lying in $B_p(0,R) = \{x \in \mathbb{R}^d \mid \|x\|_p \le R\}$. Either there is an ℓ_p-ball of radius $O(p/\varepsilon)$ containing $\Omega(n)$ points from P, or there exists a distribution $\mathcal{D}$ over* boxes *such that:*

1. *For every $u, v \in B_p(0,R)$ with $\|u-v\|_p \le 1$, a random box $S \sim \mathcal{D}$ separates u and v with probability at most ε.*

2. *For every box S from the support of $\mathcal{D}$, the number of points in P lying in S is between $\Omega(n)$ and $(1-\Omega(1))\cdot n$.*

Now let us state the partitioning result for general normed spaces, proved in Section 6.

Theorem 1.9. *Let $0<\varepsilon<1$, $X=(\mathbb{R}^d,\|\cdot\|_X)$ be a normed space and $0<R\le 2^{\mathrm{poly}(d)}$. There exists a collection $\mathcal{C}$ of measurable subsets of $B_X(0,R)=\{x\in\mathbb{R}^d \mid \|x\|_X\le R\}$ with $\log|\mathcal{C}|\le \mathrm{poly}(d)$ such that the following holds. Consider any dataset $P\subset\mathbb{R}^d$ of n points lying in $B_X(0,R)$. Either there is an X-ball of radius $O\left(\frac{\log d}{\varepsilon^2}\right)$ containing $\Omega(n)$ points from P, or there exists a distribution $\mathcal{D}$ over the elements of $\mathcal{C}$ such that:*

1. *For every $u,v\in B_X(0,R)$ with $\|u-v\|_X\le 1$, a random set $S\sim\mathcal{D}$ separates u and v with probability at most ε.*

2. *For every set S from the support of $\mathcal{D}$, the number of points in P lying in S is between $\Omega(n)$ and $(1-\Omega(1))\cdot n$.*

1.4 Related work

Prior to our work, the quest for efficient ANN data structures in high-dimensional spaces beyond ℓ_1 and ℓ_2 has proceeded via embeddings. The idea is to embed the original space into an algorithmically tractable target space, for which one then builds a data structure. The common targets are ℓ_1 and ℓ_2 which can be handled with $O(1)$-approximation by [ALRW17], ℓ_∞ which can be handled with $O(\log\log d)$-approximation with [Ind01], and ℓ_p-direct sums of these spaces, which can be handled with approximation $\mathrm{poly}(\log\log n)$ by [Ind02, Ind04, AIK09, And09]. This approach gives the best known ANN data structure for a general norm with approximation $O(\sqrt{d})$ [Joh48, Bal97]. It has also been successful for a $\mathrm{poly}(\log\log d)$-approximation for the Ulam metric [AIK09], a $O(\log d)$-approximation for EMD [Cha02, IT03], a $2^{\widetilde{O}(\sqrt{\log d})}$-approximation for edit distance [OR07], and a $\mathrm{poly}(\log d)$-approximation for Frechét distance [Ind02].

In a similar vein, the recent work [ANN+17] gives an ANN data structure for general *symmetric* norms with $\mathrm{poly}(\log\log n)$-approximation. It proceeds via a linear embedding of a d-dimensional symmetric norm into a $d^{O(1)}$-dimensional tractable universal space. However, the same paper shows that this approach fails for general norms.

For ANN under ℓ_p norms, constant factor approximations were known for $1\le p\le 2$ for near-linear space and sub-linear time [Ngu14]. The case when $p\ge 2$ is less clear. Prior to this work, the best algorithm for ℓ_p norms of [NR06, BG15] achieved approximation $2^{O(p)}$ with polynomial space (as opposed to near-linear space) and poly-logarithmic query time. For large p, there is a better algorithm with approximation $O(\log\log d)$ [AIK09, And09].

For ANN under Schatten-p norm, the previous best algorithm has polynomial in d approximation and follows from the relation between Schatten-p and ℓ_2 norms. An approximation $2^{O(p)}$ using polynomial space follows *implicitly* from a combination of the results from [NR06, BG15] with the estimate from [Ric15]. The related questions of *streaming*, *sketching* and *dimension reduction* of

Schatten-p norms have been actively studied over the past few years [LNW14, AKR15, LW16a, LW16b, LW17, NPS18].

For metrics with low intrinsic dimension, efficient ANN algorithms are known for *any metric space* [Cla99, KR02, KL04, BKL06]. These results depend *exponentially* on the intrinsic dimension, and therefore the latter is assumed to be low. This is in contrast to this paper, where we do not make such assumptions, and focus on the *high-dimensional* regime (when $\omega(\log n) \leq d \leq n^{o(1)}$), where we cannot afford to have an exponential dependence on the dimension.

1.5 Lower bounds

We complement our new algorithms with two impossibility results.

Limitation of efficient cuts. The reason that Theorem 1.3 is restricted to the cell-probe model is due to the inability to bound the time complexity of evaluating the random space partitions from Theorem 3.6 when working with general norms (even though we bound their space complexity). In constrast, for ℓ_p and Schatten-p norms, we manage to bound the time complexity and obtain *time-efficient* data structures. To explain this disparity, consider the following general scenario.

Let $G = (V, E)$ be a large graph embedded into an arbitrary normed space $(\mathbb{R}^d, \|\cdot\|)$ with edges between points at distance at most 1, and typical pair of vertices being well-separated. Following the discussion in Section 1.3, the graph G must have a sparse cut; however, the cut may not be induced by a "geometrically nice" subset of $\mathbb{R}^d$. During the algorithm from the proof of Theorem 1.3, graphs will have $d^{\Omega(d)}$ vertices, so we cannot afford to store the cut explicitly. Therefore, the query procedure re-computes the cuts on the fly. In order to achieve a time-efficient data structure for general norms, one would need to find geometrically nice cuts which can be evaluated efficiently.

For ℓ_p norms, we always find a sparse cut that is realized by a *coordinate cut* (that is, $\{v \in V \mid f(v)_k \leq u\}$ for some $1 \leq k \leq d$ and $u \in \mathbb{R}$). In our reduction we need to take intersections of cuts, which, in the case of coordinate cuts, are boxes, which are the main objects of Theorem 1.8. Thus, we store the boxes by storing the $2d$ values (lower and upper limits for each coordinate), and then we can easily evaluate on which side of a cut a given point lies. For Schatten-p norms, the argument is more delicate, but we are also able to store and compute cuts in an efficient manner.

In Section 9, we show that it is not enough to consider a fixed family of cuts with small *description complexity* for general norms; these include coordinate cuts and hyperplane cuts. More generally, Theorem 9.1 says that families of cuts used must be tailored to the particular normed space. We use a *random* norm construction similar to the one used by Gluskin in [Glu81] to prove Theorem 9.1. We note that this lower bound does not rule out *ball* cuts or other families of cuts that depend on the particular norm.

Optimality of data-dependent LSH. We show that for ℓ_p spaces, any *data-dependent LSH family* with sufficiently good parameters requires approximation $\Omega(\min\{p, \log d\})$,[4] thus our construction is

[4]Note that when $p > \log d$, ℓ_p is $O(1)$-close to $\ell_{\log d}$, so an $\Omega(p)$ lower bound when $1 \leq p \leq \log d$ covers all interesting values of p.

optimal within the data-dependent LSH framework. To show this, we embed a large expander into ℓ_p using a result from [Mat97]. We apply a similar argument to [AR16] to the embedded expander to show the desired lower bound. Thus, at least in some cases, embeddability of expanders captures the complexity of LSH *precisely*.

This result should be contrasted with the $O(\log \log d)$-ANN data structure for ℓ_∞ from [Ind01]. It also proceeds by certain[5] space partitions; the difference is that a dataset point is duplicated when inside some parts. This duplication allows the result of [Ind01] to overcome the above-mentioned $\Omega(\log d)$ lower bound.

1.6 Open problems

We state several natural open problems which seem approachable in light of the techniques developed in this paper.

- Can we get a *time-efficient* $O(\log d)$-ANN data structure for general norms? As mentioned in Section 1.5, randomized partitions from a family of "geometrically nice" cuts must be tailored to the norm of interest.
- Can we improve the approximation for general norms to $O(\log \log d)$ (even in the cell-probe model)? To accomplish this, we need to step out of the data-dependent LSH framework (see Section 1.5) to resemble the techniques from [Ind01]. A related (perhaps easier) question is to obtain an $O(\log p)$-ANN data structure over the ℓ_p or Schatten-p norm.
- Can we make the preprocessing time *polynomial* in n and d, even for the ℓ_p case?
- For the *edit distance* defined on $\{0,1\}^d$, can we obtain a $(\log d)^{O(1)}$-ANN data structure by bounding the nonlinear spectral gap? The best known ANN data structure proceeds by embedding the metric into ℓ_1 with distortion $2^{\widetilde{O}(\sqrt{\log d})}$ [OR07].
- For the Earth Mover's Distance on $[d]^2$, can we obtain a $o(\log d)$-ANN data structure by bounding the nonlinear spectral gap? The best known ANN data structure (aside from the cell-probe data structure from Theorem 1.3) proceeds by embedding into ℓ_1 with distortion $O(\log d)$ [Cha02, IT03, NS07].

1.7 Organization of the paper

In Section 3, we show how to construct a data-dependent LSH family for a general *finite* metric space assuming a good enough bound on the spectral gap. We state this result in terms of a *cutting modulus* of a metric space, a quantity we introduce in Section 3.1. In Section 4, we show how to use this LSH family to construct a cell-probe ANN data structure for a finite metric. In order to handle general normed spaces defined over $\mathbb{R}^d$ (and not just finite metrics), we discretize the ambient space; the corresponding argument is standard and appears in Section 5. In Section 6, we show a minor

[5] Deterministic.

generalization of Theorem 1.7, which bounds the spectral gap of a general norm. This allows us to give an upper bound on the cutting modulus of a normed space.

Using the results from Sections 4 and 5, we obtain a cell-probe data structure for $O(\log d)$-ANN, as claimed in Theorem 1.3. In Section 7, we address the case of ℓ_p norms and prove Theorem 1.4. In Section 8, we show a new spectral gap inequality for Schatten-p norms which implies Theorems 1.5 and 1.6.

Finally, in Section 9, we show the two impossibility results discussed in Section 1.5.

2 Preliminaries

We write χ_E as the indicator variable of event E. For any $m > 0$, we denote by $\Delta(m) \subset \mathbb{R}^{m\times m}$ the space of symmetric matrices $G = (g_{ij})$ with non-negative entries such that $\sum_{i=1}^m \sum_{j=1}^m g_{ij} = 1$. For $G \in \Delta(m)$, we denote the row sums as $\rho_G(i) = \sum_{j=1}^m g_{ij}$. The Laplacian of G is given by the $m \times m$ matrix

$$L_G = D - G,$$

and the normalized Laplacian of G is given by the $m \times m$ matrix

$$\mathcal{L}_G = I_m - D^{-1/2} G D^{-1/2},$$

where $D = \mathrm{diag}(\rho_G(1), \rho_G(2), \ldots, \rho_G(m))$ and I_m is the $m \times m$ identity matrix. We denote $0 = \lambda_1(\mathcal{L}_G) \leq \lambda_2(\mathcal{L}_G) \leq \ldots \leq \lambda_m(\mathcal{L}_G)$ the eigenvalues of the normalized Laplacian of G, and $\nu_1(\mathcal{L}_G), \ldots, \nu_m(\mathcal{L}_G) \in \mathbb{R}^m$ be the corresponding eigenvectors. For a subset $S \subseteq [m]$, we write $\overline{S} = [m] \setminus S$ and $\rho_G(S) = \sum_{i\in S} \rho_G(i)$. We will frequently refer to sequences of m points in X, $\boldsymbol{x} = (x_1, \ldots, x_m) \in X^m$. We will associate a subset $S \subset [m]$ with the corresponding subset of points $S_{\boldsymbol{x}} \subset X$ with $S_{\boldsymbol{x}} = \{x_i : i \in S\}$; and we often drop the subscript and refer to $S_{\boldsymbol{x}}$ as S when the sequence $\boldsymbol{x}$ is clear. In addition, for $S \subset X$, we write $S\colon X \to \{0, 1\}$ for the map $S(x) = \chi_{\{x\in S\}}$. For some finite subset $P \subset X$ and $x \in X$, we let $S(x, P) = \{p \in P : S(x) = S(p)\}$.

For a fixed matrix $G \in \Delta(m)$ and $S \subset [m]$, the conductance of S with matrix G is given by:

$$\Phi_G(S) = \frac{\sum_{\substack{i\in S \\ j\notin S}} g_{ij}}{\min\left\{\rho_G(S), \rho_G(\overline{S})\right\}}.$$

Definition 2.1. *For any $G \in \Delta(m)$, any metric space (X, d_X), and any $\boldsymbol{x} = (x_1, \ldots, x_m) \in X^m$, we define the Rayleigh quotient of $\boldsymbol{x}$ and G with respect to d_X^p by*

$$\mathsf{R}(\boldsymbol{x}, G, d_X^p) = \frac{\sum_{i=1}^m \sum_{j=1}^m g_{ij} d_X(x_i, x_j)^p}{\sum_{i=1}^m \sum_{j=1}^m \rho_G(i)\rho_G(j) d_X(x_i, x_j)^p}.$$

Via a straight-forward calculation, we have that when the metric space is $\mathbb{R}$ with $d_X(x_i, x_j) =$

$|x_i - x_j|$, if $x \in \mathbb{R}^m$ and $\sum_{i=1}^m \rho_G(i) x_i = 0$,

$$\mathsf{R}(x, G, |\cdot|^2) = \frac{\sum_{i=1}^m \sum_{j=1}^m g_{ij} |x_i - x_j|^2}{\sum_{i=1}^m \sum_{j=1}^m \rho_G(i)\rho_G(j)|x_i - x_j|^2} = \frac{x^T L_G x}{x^T d x}.$$

I.e. in this case $\mathsf{R}(x, G, |\cdot|^2)$ is the Rayleigh quotient $\frac{y^T \mathcal{L}_G y}{y^T y}$ for $y = D^{1/2}x$. Using this observation, we may state Cheeger's inequality with respect to $\mathsf{R}(x, G, |\cdot|^2)$.

Theorem 2.2 (Cheeger's Inequality, [Che69, Chu96], see also [Spi15])**.** *For $x \in \mathbb{R}^m$ with $\sum_{i=1}^m \rho_G(i) x_i = 0$, there exists $t \in \mathbb{R}$ for which the set $S_t = \{i \in [m] : x_i < t\}$ satisfies:*

$$\Phi_G(S_t) \leq \sqrt{\frac{\mathsf{R}(x, G, |\cdot|^2)}{2}}.$$

Letting $x = D^{-1/2}\nu_2(\mathcal{L}_G)$, there exists a subset $S \subset [m]$ which satisfies:

$$\Phi_G(S) \leq \sqrt{2 \cdot \lambda_2(\mathcal{L}_G)}.$$

Remark 2.3 (Oracle access to a norm)**.** *When working with a general normed space $(\mathbb{R}^d, \|\cdot\|_X)$, we assume oracle access to the function $\|\cdot\| \colon \mathbb{R}^d \to \mathbb{R}^{\geq 0}$. We also assume John's ellipsoid of $(\mathbb{R}^d, \|\cdot\|_X)$, i.e. the maximum volume centered ellipsoid in $\mathbb{R}^d$ contained in the unit ball of $\|\cdot\|_X$, is given by the d vectors in $\mathbb{R}^d$ specifying the ellipsoid.*

3 Partitioning general metrics

In this section, we give a general approach for constructing LSH schemes for general metric spaces. Section 3.1 defines the cutting modulus of a metric space. At a high level, the cutting modulus captures the following property of a metric space (X, d_X): for any probability distribution on pairs of close points in X, either X contains a small ball with most of the mass (with respect to the marginal distribution), or there is a balanced partition of X which separates a small fraction of neighboring pairs.

The cutting modulus determines the approximation of the data structure and is an interface between the data structure description and nonlinear spectral gaps. We describe the data structure with cutting modulus as a parameter of the metric space, and we bound the cutting modulus of various metric spaces with bounds on the non-linear spectral gap.

3.1 Cutting modulus of a metric space

We consider a metric space (X, d_X). The goal of this section is to define the cutting modulus of a metric space.

Definition 3.1. *Fix some $G \in \Delta(m)$. We say $\boldsymbol{x} = (x_1, \ldots, x_m) \in X^m$ has a* β-dense ball *of radius R if there exists a point $c \in X$ such that $\rho_G(\{i \in [m] : x_i \in B_X(c, R)\}) \geq \beta$.*

Definition 3.2. *Let $\mathfrak{S}$ be family of subsets of the metric space X. We say that $G \in \Delta(m)$ has the* (R, ε)-ball-or-cut *property with respect to $\mathfrak{S}$ if for every m points $\boldsymbol{x} = (x_1, \ldots, x_m) \in X^m$ where $d_X(x_i, x_j) \leq 1$ if $g_{ij} > 0$, one of the two properties hold:*

- *Either $\boldsymbol{x}$ has a $\frac{1}{2}$-dense ball of radius R, or*
- *There exists a subset $S \in \mathfrak{S}$ such that $S_{\boldsymbol{x}} = \{i : x_i \in S\}$ satisfies $\Phi_G(S_{\boldsymbol{x}}) \leq \varepsilon$.*

If $\mathfrak{S}$ contains all finite subsets of X, then we say that G has the (R, ε)-ball-or-cut *property.*

We may now formally define the notion of cutting modulus of a metric space.

Definition 3.3. *We say that the* ε-cutting modulus *of a metric space (X, d_X) with respect to a family $\mathfrak{S}$ of subsets of X, $\Xi_{\mathfrak{S}}(X, \varepsilon)$, is given by:*

$$\Xi_{\mathfrak{S}}(X, \varepsilon) = \inf\{R \in \mathbb{R} : \forall m \in \mathbb{N}, \forall G \in \Delta(m), \text{ matrix } G \text{ has } (R, \varepsilon)\text{-ball-or-cut property w.r.t. } \mathfrak{S}\}.$$

If $\mathfrak{S}$ contains all finite subsets of X, we denote $\Xi_{\mathfrak{S}}(X, \varepsilon)$ simply by $\Xi(X, \varepsilon)$.

At a high level, the ε-cutting modulus of a metric space will govern the approximation ratio one may achieve with space $\mathrm{poly}(d) \cdot n^{1+O(\varepsilon)}$ and query time $\mathrm{poly}(d) \cdot n^{O(\varepsilon)}$. In particular, suppose (X, d_X) has $\Xi(X, \varepsilon) = R$. Consider any sequence of points $x_1, \ldots, x_m \in X$, and form a graph by connecting points lying at distance at most 1. The graph defines a normalized adjacency matrix $G \in \Delta(m)$ which has the (R, ε)-ball-or-cut property. If there exists a dense ball, then we know that a constant fraction of the points lie close to each other (within distance $2R$). Otherwise, there is a sparse cut of the points which does not cut many edges of G. Roughly speaking, the data-dependent LSH will be built by recursively applying this procedure, and using the multiplicative weights update rule in order to handle any possible distribution over datasets and queries. For our cell-probe algorithms we will allow $\mathfrak{S}$ to contain all finite subsets of X. However, our efficient data structures will use a restricted family $\mathfrak{S}$ which allows us to quickly determine which side of a cut a point lies on.

3.2 Partitioning theorems

The goal of this section is to prove the main partitioning theorem. We consider a metric space (X, d_X) which consists of N points. Let $0 < \varepsilon < \varepsilon_0$ be a small positive parameter and $R = \Xi(X, \varepsilon)$.

We first define the notion of balanced collections of balls and cuts.

Definition 3.4. *Let $\mathcal{S}$ be a collection of subsets $S_1, \ldots, S_m \subseteq X$. We say $\mathcal{S}$ is* ε-sparse *if for every two points $x, y \in X$ with $d_X(x, y) \leq 1$, at most an ε-fraction of subsets from $\mathcal{S}$ split x and y, i.e.,*

$$\Pr_{i \sim [m]}[S_i(x) \neq S_i(y)] \leq \varepsilon.$$

Definition 3.5. *Consider a dataset $P \subseteq X$ of n points. Let $\mathcal{S}$ be a collection of subsets $S_1, \ldots, S_m \subseteq X$. We say that $\mathcal{S}$ is* γ-balanced under P *if for any $S \in \mathcal{S}$ we have*

$$(1-\gamma)n \leq |S \cap P| \leq \gamma n.$$

These two notions of sparsity and balancedness will measure the quality of the data-dependent LSH. Intuitively, the data-dependent LSH is constructed by recursively partitioning the space with a random subset from a particular collection. We want the collection to be balanced, to ensure the algorithm makes progress, and sparse, to maintain a low probability of error. Lastly, we want collections of subsets which can be written succinctly; such a condition will ensure the querying algorithm can utilize the data-dependent LSH. We ensure our collection can be written succinctly by requiring there are not too many of them, and that the collections do not have too many sets.

We may now state the main partitioning theorem for general metric spaces.

Theorem 3.6. *Let $R = \Xi(X, \varepsilon)$ for some $\varepsilon \in (0, \frac{1}{4})$, and fix any $n \in \mathbb{N}$. There exists a collection $\mathcal{C}$ of subsets of X with $\log|\mathcal{C}| = O(\log(N)\log(\log(N)/\varepsilon))$ such that for any dataset $P \subseteq X$ of n points,*

- *Either there exists a point $x_0 \in X$ with $|P \cap B_X(x_0, R)| \geq \frac{n}{50}$, or*
- *There exists a subcollection $\mathcal{S} \subseteq \mathcal{C}$ of subsets of X such that:*
  - *$\mathcal{S}$ is 50ε-sparse,*
  - *$\mathcal{S}$ is $\frac{49}{50}$-balanced under P.*

Theorem 3.6 suggests a very natural data-dependent LSH. At each step of the algorithm, either we have a dense ball, or we have a collection of subsets with a distribution which decreases the size of the dataset and does not split the query from its dataset point too often. Note that the set $\mathcal{C}$ does not depend on P. This means the querying algorithm will know $\mathcal{C}$, and needs to read $O(\log(N)\log(\log(N)/\varepsilon)/\varepsilon)$ many bits from the data-structure in order to specify any particular set $S \in \mathcal{C}$.

We now turn to proving Theorem 3.6. The proof is algorithmic and requires a few lemmas, which correspond to particular subroutines.

3.2.1 Partitioning with the (R, ε)-ball-or-cut property

Let $X = \{x_1, \ldots, x_N\}$ be the points of the metric space of size N. For the remainder of the section, let $G \in \Delta(N)$ be a fixed matrix with $g_{ij} > 0$ only if $d_X(x_i, x_j) \leq 1$. We will frequently interchange between subsets $S \subseteq X$ and $S \subseteq [N]$ by associating $x_i \in X$ with $i \in [N]$. In addition, we frequently write $\overline{S} = [N] \setminus S$. The goal of this section is to use the (R, ε)-ball-or-cut property to give a subroutine which when given a matrix $G \in \Delta(N)$, outputs a dense ball with respect to G, or a particular subset of vertices which cuts few edges with respect to G.

Lemma 3.7. *Let $R = \Xi(X, \varepsilon)$ for some $\varepsilon \in (0, \frac{1}{4})$. Then there either exists a $\frac{1}{4}$-dense ball of radius R with respect to G, or there exists a subset $S \subseteq X$ where*

$$\frac{1}{3} \leq \rho_G(S) \leq \frac{3}{4} \qquad \text{and} \qquad \sum_{i \in S, j \notin S} g_{ij} \leq 2\varepsilon.$$

Proof. We give an iterative procedure which begins with a set $S := \emptyset$, and at each step, either finds a dense ball of radius R, or adds some points to S while keeping $\rho_G(S) \leq \frac{3}{4}$ and $\sum_{i \in S, j \notin S} g_{ij} \leq 2\varepsilon$.

At the beginning of an iteration, assume $\rho_G(S) < \frac{1}{3}$. We repeat the following procedure:

1. Consider the matrix $\widetilde{G} \in \Delta(|\overline{S}|)$ obtained by restricting G on the rows and columns corresponding to $\overline{S}$ and scaling the entries so they sum to 1. Note that we still have $g_{ij} > 0$ only if $d_X(x_i, x_j) \leq 1$.

2. The matrix $\widetilde{G}$ has the (R, ε)-ball-or-cut property, so either there exists a $\frac{1}{2}$-dense ball of radius R in $\overline{S}$ with respect to $\widetilde{G}$, or there exists a subset $\widetilde{S} \subset \overline{S}$ with $\Phi_{\widetilde{G}}(\widetilde{S}) \leq \varepsilon$.

   (a) Suppose $\overline{S}$ has a $\frac{1}{2}$-dense ball of radius R with respect to $\widetilde{G}$. Then, that ball is $\frac{1}{4}$-dense with respect to G, since $\widetilde{G}$ was rescaled by at least $1 - \rho_G(S) - 2\varepsilon\rho_G(S) \geq \frac{1}{2}$.

   (b) Suppose $\widetilde{S} \subset \overline{S}$ is a subset with $\Phi_{\widetilde{G}}(\widetilde{S}) \leq \varepsilon$, and assume, without loss of generality, that $\widetilde{S}$ has $0 < \rho_{\widetilde{G}}(\widetilde{S}) \leq \frac{1}{2}$, since otherwise, we can switch $\widetilde{S}$ and $\overline{S} \setminus \widetilde{S}$. Then, let $S \leftarrow S \cup \widetilde{S}$.

The quantity $\rho_G(S)$ is monotonically increasing with the iterations, and the procedure terminates when $\rho_G(S) \geq \frac{1}{3}$. Thus, we just need to show that, as long as we do not return a $\frac{1}{4}$-dense ball with respect to G, we always have $\rho_G(S) \leq \frac{3}{4}$ and $\Phi_G(S) \leq 2\varepsilon$.

Consider the final iteration of the algorithm before S is returned; we have that $S \subset [N]$ satisfies $\rho_G(S) < \frac{1}{3}$ and $\rho_{\widetilde{G}}(\widetilde{S}) \leq \frac{1}{2}$. Additionally, assume $\Phi_G(S) \leq 2\varepsilon$ and $\Phi_{\widetilde{G}}(\widetilde{S}) \leq \varepsilon$. Then,

$$\sum_{\substack{i \in S \cup \widetilde{S} \\ j \notin S \cup \widetilde{S}}} g_{ij} \leq \sum_{\substack{i \in S \\ j \notin S}} g_{ij} + \sum_{\substack{i \in \widetilde{S} \\ j \notin S \cup \widetilde{S}}} g_{ij} \leq 2\varepsilon \cdot \rho_G(S) + \varepsilon \cdot \rho_{\widetilde{G}}(\widetilde{S}) \leq 2\varepsilon \left(\rho_G(S) + \rho_G(\overline{S})\right) = 2\varepsilon \cdot \rho_G(S \cup \widetilde{S}),$$

where we used the fact that $\rho_{\widetilde{G}}(\widetilde{S}) \leq 2\rho_G(\widetilde{S})$, because the matrix $\widetilde{G}$ was normalized by a factor of at least $\frac{1}{2}$. Therefore, we have $\Phi_G(S \cup \widetilde{S}) \leq 2\varepsilon$. Finally, note that:

$$\rho_G(S \cup \widetilde{S}) \leq \rho_G(S) + \rho_G(\widetilde{S}) \leq \rho_G(S) + \frac{1}{2}\left(1 - \rho_G(S) - \Phi_G(S)\right) + \Phi_G(S) \leq \frac{2}{3} + \frac{\varepsilon}{3} \leq \frac{3}{4}.$$

□

3.2.2 Inner multiplicative weights update

The goal of this subsection is to use the partitioning procedure from Lemma 3.7 in order to either find a dense ball (with respect to a given distribution over X), or build a sparse collection of subsets.

For the rest of the section, we let E be the set of unordered pairs of close points in X (at distance at most 1).

Lemma 3.8. *Let $R = \Xi(X, \varepsilon)$ for some $\varepsilon \in (0, \frac{1}{4})$, and let ν be a probability measure over points in X. Then, either there exists a ball B of radius R such that $\nu(B) \geq \frac{1}{6}$, or there exists a collection $\mathcal{S}$ of $O\left(\frac{\log N}{\varepsilon}\right)$ subsets $S \subseteq X$ such that:*

- *$\mathcal{S}$ is 50ε-sparse, and*
- *Every $S \in \mathcal{S}$ satisfies $\frac{1}{4} \leq \nu(S) \leq \frac{5}{6}$.*

Proof. We prove the lemma by giving an algorithm which produces the collection $\mathcal{S}$ via the multiplicative weights update algorithm. More specifically, we give an iterative procedure where for $t = 0, \ldots, O\left(\frac{\log N}{\varepsilon}\right)$, maintains at most N^2 weights, $w_t \colon E \to \mathbb{R}^{\geq 0}$. At each step, the procedure produces a matrix $G \in \Delta(N)$, checks the conditions of Lemma 3.7, and either outputs a dense ball or updates the weights w_{t+1}. Fix $\delta = \frac{1}{10}$. The procedure does the following:

1. For $t = 0, \ldots, T = \left\lceil \frac{\log_2 N}{\varepsilon} \right\rceil$, maintain weights $w_t \colon E \to \mathbb{R}^{\geq 0}$, where initially, $w_0(x, y) = 1$ for all $(x, y) \in E$, and $\Psi_t = \sum_{(x,y)\in E} w_t(x, y)$. Start with $\mathcal{S} = \emptyset$.
2. Let $G^{(t)} \in \Delta(N)$ be given by:

$$g_{ij}^{(t)} = \begin{cases} \delta \cdot \frac{w_t(x_i, x_j)}{2\Psi_t} & i \neq j, (x_i, x_j) \in E \\ 0 & i \neq j, (x_i, x_j) \notin E \\ (1-\delta)\nu(x_i) & i = j \end{cases},$$

   and consider the possible outcomes of Lemma 3.7 with matrix $G^{(t)}$:

   (a) If there exists a $\frac{1}{4}$-dense ball B of radius R with respect to $G^{(t)}$, then $\frac{1}{4} \leq \rho_{G^{(t)}}(B) = \sum_{i\in B}(1-\delta)\nu(i) + \sum_{i\in B}\sum_{j\neq i} \delta \frac{w_t(x_i,x_j)}{2\Psi_t} \leq (1-\delta)\nu(B) + \frac{\delta}{2}$. Return B, since $\nu(B) \geq \frac{1}{6}$.

   (b) If there exists a subset $S^{(t)} \subset X$ with $\frac{1}{3} \leq \rho_{G^{(t)}}(S^{(t)}) \leq \frac{3}{4}$ and $\sum_{i\in S^{(t)}, j\notin S^{(t)}} g_{ij}^{(t)} \leq 2\varepsilon$, then let $\mathcal{S} \leftarrow \mathcal{S} \cup \{S^{(t)}\}$ and for all $(x, y) \in E$, we let:

$$w_{t+1}(x, y) = w_t(x, y)\left(1 + \chi_{\{S^{(t)}(x) \neq S^{(t)}(y)\}}\right).$$

3. After T iterations, if the procedure has not returned a ball, return $\mathcal{S}$.

It remains to show that if the procedure does not return a ball B, then the collection $\mathcal{S}$ is 50ε-sparse, and every $S^{(t)} \in \mathcal{S}$ satisfies $\frac{1}{4} \leq \nu(S^{(t)}) \leq \frac{5}{6}$. Note that $|\mathcal{S}| = O\left(\frac{\log N}{\varepsilon}\right)$ since $T = O\left(\frac{\log N}{\varepsilon}\right)$. In order to show that $\frac{1}{4} \leq \nu(S) \leq \frac{5}{6}$ for all $S^{(t)} \in \mathcal{S}$, note that, similarly to the case with B,

$$\frac{1}{3} \leq \rho_{G^{(t)}}(S^{(t)}) \leq \frac{\delta}{2} + (1-\delta)\nu(S^{(t)}) \qquad \text{and} \qquad (1-\delta)\nu(S^{(t)}) \leq \rho_{G^{(t)}}(S^{(t)}) \leq \frac{3}{4},$$

where the claim follows since $\delta = \frac{1}{10}$. We now turn to showing that $\mathcal{S}$ is 50ε-sparse. On the one hand, we have:

$$\begin{aligned}\Psi_{t+1} &= \sum_{(x,y)\in E} w_{t+1}(x,y) = \sum_{(x,y)\in E} w_t(x,y)\left(1+\chi_{\{S^{(t)}(x)\neq S^{(t)}(y)\}}\right) \\ &\leq \Psi_t + \Psi_t \cdot \frac{2}{\delta} \sum_{i\in S^{(t)}, j\notin S^{(t)}} \delta \cdot \frac{w_t(x_i,x_j)}{2\Psi_t} \leq \Psi_t\left(1+\frac{4\varepsilon}{\delta}\right), \end{aligned} \tag{3}$$

since $\delta \cdot \frac{w_t(x_i,x_j)}{2\Psi_t} = g_{ij}$ for every close pair (x_i, x_j), and $\sum_{i\in S^{(t)}, j\notin S^{(t)}} g_{ij} \leq 2\varepsilon$. Thus,

$$\Psi_{T+1} \leq \Psi_0\left(1+\frac{4\varepsilon}{\delta}\right)^T \leq N^2\left(1+\frac{4\varepsilon}{\delta}\right)^T.$$

On the other hand, for each pair $(x,y) \in E$,

$$\Psi_{T+1} \geq 2^{p(x,y)\cdot T}, \tag{4}$$

where $p(x,y) = \Pr_{t\in[T]}[S^{(t)}(x) \neq S^{(t)}(y)]$. Combining (3) and (4), and taking logarithms, we have:

$$p(x,y) \leq \frac{2\log_2 N}{T} + \log_2\left(1+\frac{4\varepsilon}{\delta}\right) \leq \frac{2\log_2 N}{T} + \frac{4\varepsilon}{\delta} \leq 2\varepsilon + 40\varepsilon \leq 50\varepsilon.$$

□

3.2.3 Outer multiplicative weights update: proof of Theorem 3.6

The goal of this subsection is to prove Theorem 3.6. Similarly to Lemma 3.8, we use the multiplicative weights update rule to design an algorithm which incorporates (limited) information about the dataset P; in each update round, we call Lemma 3.8. We analyze this outer multiplicative weights update process using KL-divergence as a potential function. In particular, we use the following lemma, which is well known (see Theorem 2.4. in [AHK12]), and has been used, for example, in the literature on differential privacy (Lemma IV.1. in [HR10]). We give the short proof here for completeness. Below, KL divergence will be defined with respect to the natural logarithm, i.e. for two measures μ and ν on X we have

$$D_{KL}(\mu\|\nu) = \sum_{x\in X} \mu(x)\ln\frac{\mu(x)}{\nu(x)}.$$

Lemma 3.9. *Let μ and ν be probability measures over X. For a subset $S \subseteq X$, let $\sigma = \text{sign}(\mu(S) - \nu(S))$, and define a new probability measure ν' over X by*

$$\nu'(x) = \frac{\nu(x)e^{\eta\sigma S(x)}}{\sum_{y\in X}\nu(y)e^{\eta\sigma S(y)}}.$$

Then,

$$D_{KL}(\mu\|\nu') - D_{KL}(\mu\|\nu) \leq -\eta|\mu(S) - \nu(S)| + \eta^2.$$

Proof. By the definition of KL-divergence we have

$$\begin{aligned}
D_{KL}(\mu\|\nu') - D_{KL}(\mu\|\nu) &= \sum_{x\in X} \mu(x) \ln \frac{\nu(x)}{\nu'(x)} \\
&= \sum_{x\in X} \mu(x) \ln \frac{\sum_{y\in X} \nu(y) e^{\eta\sigma S(y)}}{e^{\eta\sigma S(x)}} \\
&= -\eta\sigma\mu(S) + \ln \sum_{y\in X} \nu(y) e^{\eta\sigma S(y)} \\
&\leq -\eta\sigma\mu(S) + \ln \sum_{y\in X} \nu(y)(1 + \eta\sigma S(y) + \eta^2 S(y)) \\
&= -\eta\sigma\mu(S) + \ln(1 + \eta\sigma\nu(S) + \eta^2\nu(S)) \\
&\leq -\eta\sigma(\mu(S) - \nu(S)) + \eta^2 \\
&= -\eta|\mu(S) - \nu(S)| + \eta^2.
\end{aligned}$$

The first inequality above follows from $e^z \leq 1 + z + z^2$ for all $|z| \leq 1$. The second inequality follows from $\ln(1 + z) \leq z$. □

In particular, notice that Lemma 3.9 implies that if $|\mu(S) - \nu(S)| > \alpha$, and we set $\eta = \frac{\alpha}{2}$, then the KL-divergence dicreases by at least $\frac{\alpha^2}{4}$.

Proof of Theorem 3.6. Similarly to Lemma 3.8, we give an iterative procedure where at each time step $t = 0, \ldots, T = O(\log N)$, we maintain N weights, $w_t \colon X \to \mathbb{R}^{\geq 0}$. At each step, the procedure produces a probability measure ν supported on points in X and uses Lemma 3.8 to get a collection of subsets of X. The procedure is defined as follows:

1. For $t = 0, \ldots, T = 400 \ln N$, maintain weights $w_t \colon X \to \mathbb{R}^{\geq 0}$, where initially, $w_0(x) = 1$ for all $x \in X$.
2. Let $\nu^{(t)}$ be the probability measure supported on X given by $\nu^{(t)}(x) = \frac{w_t(x)}{\sum_{y\in X} w_t(x)}$. Consider the possible outcomes of Lemma 3.8 with measure $\nu^{(t)}$:
   (a) If there exists a ball $B^{(t)}$ of radius R such that $\nu^{(t)}(B) \geq \frac{1}{6}$ and $|P \cap B| \geq \frac{n}{50}$, then return $B = B^{(t)}$.
   (b) If there exists a ball $B^{(t)}$ of radius R such that $\nu^{(t)}(B) \geq \frac{1}{6}$ but $|P \cap B| < \frac{n}{50}$, then set
   $$w_{t+1}(x) = w_t(x) e^{-B^{(t)}(x)/20},$$
   and continue with the next iteration.
   (c) If there exists a collection $\mathcal{S}^{(t)}$ of subsets of X satisfying the conditions of Lemma 3.8, and $\frac{n}{25} \leq |S \cap P| \leq \frac{24n}{25}$ for all $S \in \mathcal{S}^{(t)}$, then return $\mathcal{S} = \mathcal{S}^{(t)}$.

(d) If there exists a collection $\mathcal{S}^{(t)}$ of subsets of X satisfying the conditions of Lemma 3.8, and for some $S \in \mathcal{S}^{(t)}$ we have $|S \cap P| < \frac{n}{25}$ or $|S \cap P| > \frac{24n}{25}$, then set $\sigma = \operatorname{sign}\left(\frac{|S\cap P|}{n} - \nu^{(t)}(S)\right)$, update the weights as

$$w_{t+1}(x) = w_t(x)e^{\sigma S(x)/20},$$

and continue with the next iteration.

Note that the procedure returns $B = B^{(t)}$ only if it is a ball of radius R that contains at least $\frac{n}{50}$ points, and it returns the collection $\mathcal{S} = \mathcal{S}^{(t)}$ only if it is 50ε-sparse and $\frac{49}{50}$-balanced. So, if the procedure returns B or $\mathcal{S}$, then we know it satisfies the condition of the theorem. Therefore, we just need to show that the procedure will return B or $\mathcal{S}$ in the first T iterations, and that $\mathcal{S}$ is a subcollection of a sufficiently small collection $\mathcal{C}$. Since in each iteration we either return B or $\mathcal{S}$, or we update w_t, for the first claim it is enough to show that w_t is updated fewer than T times. We do so using KL-divergence as a potential function.

Let μ be the empirical distribution induced by the dataset, i.e. $\mu(x) = \frac{1}{n}$ for every $x \in P$ and $\mu(x) = 0$ for every $x \in X \setminus P$. At step 0, we have

$$D_{KL}(\mu\|\nu^{(0)}) = \ln N - H(\mu) \leq \ln N, \tag{5}$$

where $H(\mu)$ is the Shannon entropy of μ, which is always non-negative. If we update w_t because there exists a ball $B^{(t)}$ with $\nu^{(t)}(B^{(t)}) \geq \frac{1}{6}$ but $\mu(B^{(t)}) = \frac{|P\cap B|}{n} < \frac{1}{50}$, then we have $|\mu(B^{(t)}) - \nu^{(t)}(B^{(t)})| > \frac{1}{6} - \frac{1}{50} > \frac{1}{10}$, so, by Lemma 3.9

$$D_{KL}(\mu\|\nu^{(t+1)}) < D_{KL}(\mu\|\nu^{(t)}) - \frac{1}{400}. \tag{6}$$

Similarly, if we update w_t because there exists a set $S \in \mathcal{S}^{(t)}$ with $\mu(S) = \frac{|S\cap P|}{n} < \frac{1}{25}$ or $\mu(S) > \frac{24}{25}$, then, by Lemma 3.8 we know that $\frac{1}{4} \leq \nu^{(t)} \leq \frac{5}{6}$, and, therefore,

$$|\mu(B) - \nu^{(t)}(B)| > \frac{24}{25} - \frac{5}{6} > \frac{1}{10}.$$

So, by Lemma 3.9, the inequality (6) holds in this case, too. By (5) and (6), and because KL-divergence is always non-negative, we have that w_t can be updated at most $400 \ln N \leq T$ times. Therefore, after one of the T iterations the procedure will return either a ball B or a collection $\mathcal{S}$ satisfying the conditions of the theorem.

To finish the proof, we need to argue that $\mathcal{S}$ is a subcollection of a small collection $\mathcal{C}$ of subsets of X. Let M be the number of distinct collections $\mathcal{S}$ that the iterative procedure can return. Lemma 3.8 guarantees that, for any such collection, $|\mathcal{S}| = O(\log N/\varepsilon)$, and if we define $\mathcal{C}$ to be the union of all possible $\mathcal{S}$, then we have the bound $|\mathcal{C}| = O(M \log(N/\varepsilon))$. To bound M, observe that $\mathcal{S}$ depends on the dataset P only to determine, for each $t = 1, \ldots, T$, whether the procedure has returned B or $\mathcal{S}$, or, otherwise, to determine the identity of a set $S \in \mathcal{S}^{(t)}$ such that $\mu(S) = \frac{|S\cap P|}{n} < \frac{1}{25}$

or $\mu(S) > \frac{24}{25}$, and the sign of $\mu(S) - \nu^{(t)}(S)$. Since $|\mathcal{S}^{(t)}| = O(\log N/\varepsilon)$, any set $S \in \mathcal{S}^{(t)}$ can be specified in $O(\log(\log(N)/\varepsilon))$ bits. Overall, $\mathcal{S}$ depends only on $O(T(1 + \log(\log(N)/\varepsilon))) = O(\log N \log(\log(N)/\varepsilon))$ bits from P, which gives the desired bound on M, and, therefore, on $\log |\mathcal{C}|$. □

4 Cell-probe data structure for general metrics

Here, we describe a cell-probe data structure solving c-ANN for (X, d_X), where $|X| = N$. Along the way, we use Theorem 3.6 as the main tool.

We first define the cell-probe model (as used in Theorem 1.3). Given a dataset, the cell-probe algorithm is allowed unbounded preprocessing time and eventually stores some memory as a sequence of *cells* of $O(\log n)$ bits each. Then, given a query point, a cell-probe algorithm is allowed to probe some cells (possibly adaptively) to read the contents of a cell. The algorithm performs unbounded auxiliary computations and uses unbounded auxiliary memory. The complexity of a cell-probe algorithm is measured by the number of cells, or the space, the data structure uses, and the number of probes the algorithm makes during a query. We will assume that $\log \log N = O(\log n)$ and that any point in X can be specified using $O(\log N)$ cells.

The main theorem in this section is:

Theorem 4.1. *For any metric space X of size N, and $\alpha \in (0, \frac{1}{4})$, there exists a cell-probe data structure for $(2 \cdot \Xi(X, \Theta(\alpha)) + 1)$-ANN that uses $O(n^{1+\alpha} \cdot \log N)$ words of space and $O(n^{\alpha} \cdot \log n \cdot \log N)$ cell probes per query.*

While we do not measure time complexity in this section, we note the cell-probe algorithm described may be implemented with preprocessing time and query time which depend exponentially on the dimension.

In the rest of this section we fix $R = \Xi(X, \varepsilon)$ for a parameter $\varepsilon = \Theta(\alpha)$, to be determined later.

Preprocessing. Next we describe how to build the data structure (for the pseudocode, see Figure 1). Let $P \subset X$ be a dataset of n points. The data structure is a collection of independently generated random decision trees. Each node v of a tree stores the following fields:

- v.type: the type of the node;
- $v.P$: a subset of the dataset points;
- v.center: a point in X;
- $v.S$: $O(\log(N) \log(\log(N)/\varepsilon))$ bits used to indicate a set S in the collection $\mathcal{C}$ guaranteed by Theorem 3.6, defining a cut node;
- v.left and v.right: pointers to child nodes.

function PROCESS(P, ℓ, v)
  if $\ell = t$ or $|P| \leq 100$ **then**
    v.type $\leftarrow$ "leaf."
    $v.P \leftarrow P$.
  else if $\exists x_0$ such that $|P \cap B_X(x_0, R)| \geq \frac{|P|}{50}$ **then**
    call PROCESSBALL(P, x_0, ℓ, v)
  else
    $\mathcal{S} \leftarrow$ MWU(P).
    v.mwu $\leftarrow$ mwu
    sample S uniformly from $\mathcal{S}$
    store bits necessary to identify $S \in \mathcal{C}$ in $v.S$
    PROCESSCUT(P, S, ℓ, v).

function MWU(P)
  $\mathcal{S} \subseteq \mathcal{C}$ obtained from Theorem 3.6 with P.
  return $\mathcal{S}$.

function PROCESSBALL(P, x, ℓ, v)
  v.type $\leftarrow$ "ball."
  v.center $\leftarrow x$.
  $v.P \leftarrow P \cap B_X(x, R)$.
  PROCESS$(P \setminus B_X(x_0, R), \ell + 1, v.\text{left})$.

function PROCESSCUT(P, S, ℓ, v)
  v.type $\leftarrow$ "cut."
  $P_l = P \cap S_i$, $P_r = P \setminus S$.
  PROCESS$(P_l, \ell + 1, v.\text{left})$.
  PROCESS$(P_r, \ell + 1, v.\text{right})$.
  $v.P \leftarrow \emptyset$.

Figure 1: Pseudocode for constructing the data-structure

We keep a counter ℓ, which denotes the current level of the tree we are processing. Initially, $\ell = 0$, and it is incremented on each recursive call. Once ℓ reaches some threshold t (to be specified shortly), we store a leaf node v and save the points of the dataset which reached v in $v.P$. Thus the depth of the tree is bounded by t a priori.

1. If there exists a point $x_0 \in X$ such that $|P \cap B_X(x_0, R)| \geq \frac{n}{50}$, we build a *ball node*. In this case, the ball node saves x_0 in v.center and $P \cap B_X(x_0, R)$ in $v.P$. We then recurse by building a data structure on $P \setminus B_X(x_0, R)$. (See PROCESSBALL in Figure 1).

2. Otherwise, the second condition of Theorem 3.6 holds, and the set $\mathcal{C}$ guaranteed by the theorem contains a subcollection $\mathcal{S} \subseteq \mathcal{C}$ of subsets of X which is 50ε sparse and $\frac{96}{100}$-balanced. We sample a uniformly random $S \in \mathcal{S}$, and we build a *cut node* v. We store the $O(\log(N) \log(\log(N)/\varepsilon))$ bits necessary to identify S in $v.S$, and recursively create two child nodes, holding the points $P \cap S$ and $P \setminus S$. (See PROCESSCUT in Figure 1).

The final data structure consists of $k = O(n^\alpha)$ independent trees, rooted at the nodes $v_1, \ldots, v_k$, where the i-th tree was built by a call to PROCESS$(P, 0, v_i)$.

Querying the Data Structure. We now specify how to query the data structure; the pseudocode is given in Figure 2. For each of the k trees in the data structure, we start the query procedure at the root of the tree, and proceed by cases, according to the type of node, as follows:

- *Leaf nodes*: If a query $q \in X$ queries a leaf node v, then the query scans $v.P$ and returns the first point which lies within distance $2R + 1$. If no such point is found, return $\perp$.

- *Ball nodes*: If a query $q \in X$ queries a ball node v, we test whether our query is close to the ball centered at v.center of radius R. In particular, if $d_X(q, v.\text{center}) \leq R + 1$ and $v.P \neq \emptyset$, we return an arbitrary $p \in v.P$. Otherwise, we recurse on the child node of v.

- *Cut nodes*: If a query $q \in X$ queries a cut node v, the querying algorithm runs the multiplicative weights algorithm, accessing the values stored in v.mwu. Once it determines the collection $\mathcal{S}$,

```
function Query(q, v)
    if v.type = "leaf" then
        for p ∈ v.P do
            return p if d_X(q, p) ≤ 2R + 1.
        return ⊥.
    if v.type = "ball" then
        p ← QueryBall(q, v).
        return p if p ≠ ⊥.
    if v.type = "cut" then
        p ← QueryCut(q, v).
        return p if p ≠ ⊥.

function QueryBall(q, v)
    x_0 ← v.center.
    if d_X(x_0, q) ≤ R + 1 then
        return any p ∈ v.P.
    return Query(q, v.left).

function QueryCut(q, v)
    Identify S ∈ C from v.S
    if q ∈ S then
        return Query(q, v.left).
    return Query(q, v.right).
```

Figure 2: Pseudocode for querying the data-structure

the querying algorithm checks the index of the set $S_i \in \mathcal{S}$, which is stored in $v.S$. If $q \in S_i$, then the querying algorithm recurses on the left child, otherwise, it recurses on the right child of v.

We collect some simple facts about the data structure which we use later in the analysis.

Claim 4.2. *The following statements are true:*

- *The sets $v.P$ for nodes v partition the dataset P.*
- *If $\textsc{Query}(q, v)$ returns a point $p \in P$, then $d_X(p, q) \leq 2R + 1$.*

Analysis.

It remains to set the parameters t and ε. We let $t = \left\lceil \frac{\log n}{\log(50/49)} \right\rceil$ and $\varepsilon = \left\lfloor \frac{\alpha \cdot \log(50/49)}{50} \right\rfloor$ in order to have $(1 - 50\varepsilon)^t \geq n^{-\alpha}$.

Consider a fixed dataset P, and let $q \in X$ be any query, which is promised to have a point $p \in P$ with $d_X(p, q) \leq 1$. If there are multiple such points for q, we fix one arbitrarily. Let v be a node of the data structure built by a call to $\textsc{Process}(P_v, \ell, v)$ for some $P_v \subset P$ and $\ell < t$. We let $U = C(v, q)$ be the random variable (over the random choice of $S_i \in \mathcal{S}$ if v is a cut node) which specifies the child node followed by $\textsc{Query}(q, v)$, and $\perp$ if $\textsc{Query}(q, v)$ does not recurse down a child. We also consider the random variable P_U consisting of the dataset involved in the call $\textsc{Process}(P_U, \ell + 1, U)$ which builds the node U when $U \neq \perp$.

We first claim that for any node v of the data structure, if $p \in P_v$, then,

$$\Pr[p \in P_U \mid U \neq \perp] \geq 1 - 50\varepsilon. \tag{7}$$

To see this, first consider the case in which $\textsc{Process}(P_v, \ell, v)$ calls $\textsc{ProcessBall}$, and let x be the center of the ball. If $U \neq \perp$, then $\textsc{QueryBall}(q, v)$ did not return any point and $d_X(x, q) > R + 1$, so $p \notin B_X(x, R)$. Then $p \in P_v \setminus B_X(x, R) = P_U$ with probability 1. For the remaining case, when $\textsc{Process}$ calls $\textsc{ProcessCut}$, we have:

$$\Pr_{S \sim \mathcal{S}}[p \in P_U] = \Pr_{S \sim \mathcal{S}}[S(p) = S(q)] \geq 1 - 50\varepsilon,$$

since $\mathcal{S}$ is guaranteed to be 50ε-sparse by Theorem 3.6.

By Claim 4.2, any point p' returned by $\textsc{Query}(q, v_i)$, where v_i is the root of one of the data structure trees, satisfies $d_X(p', q) \leq 2R + 1$. To prove correctness, it remains to argue that, with sufficiently high probability, at least one of the $\textsc{Query}(q, v_i)$ calls, for $i = 1, \ldots, k$, does in fact return a point. Fix some i between 1 and k, and define a random sequence $U_0, U_1, \ldots, U_s$ of nodes of the tree rooted at v_i by $U_0 = v_i$ and $U_\ell = C(U_{\ell-1}, q)$; U_s is the first node in this sequence for which $C(U_s, q) = \perp$. Notice that $s \leq t$. Clearly, $\textsc{Query}(q, v_i)$ will return a point if $p \in P_{U_s}$. By (7) and the choice of t, this happens with probability at least $(1 - 50\varepsilon)^s \geq (1 - 50\varepsilon)^t \geq n^{-\alpha}$. By picking the number k of trees in the data structure to be a sufficiently large multiple of n^α, we can guarantee that with large constant probability the data structure returns a point p' such that $d_X(p', q) \leq 2R + 1$.

To finish the analysis, we need to bound the number of cells stored by the data structure, and the number of cell probes made by the query procedure. Each of the points stored in the leaves of each tree form a partition of the point set P, so each tree has at most n internal nodes. Each internal node stores $O(\log(N))$ cells, and all the leaves together use $O(n \log N)$ cells of space ($O(\log N)$ per point in P). Therefore, the total space used by the data structure is $O(n^{1+\alpha} \log N)$ cells.

The query procedure probes $O(\log N)$ cells at each internal node of a tree. The number of cells probed at a leaf node v is proportional to $O(|v.P| \cdot \log N)$. We claim that $v.P$ is bounded by a constant. Suppose that u is a child of a node v, and also that v was created by a call to $\textsc{Process}(P_v, \ell, v)$ and u by a call to $\textsc{Process}(P_u, \ell+1, u)$. Then, by the guarantees of Theorem 3.6, $|P_u| \leq \frac{49}{50}|P_v|$, so the number of points that can reach a leaf of a tree is bounded by $n\left(\frac{49}{50}\right)^t$. By the choice of t, this number is bounded by a constant, as we claimed. Therefore, the total number of cells probed by the query procedure is $O(kt \log N) = O(n^\alpha \log n \log N)$. This completes the proof of Theorem 4.1.

5 Discretizing the space

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ with unit ball K. Let $\mathcal{E}$ be the John Ellipsoid of K, i.e. the largest volume ellipsoid contained inside K. By John's theorem [Joh48],

$$\mathcal{E} \subset K \subset \sqrt{d} \cdot \mathcal{E}.$$

We let $\mathcal{C} \supset \mathcal{E}$ be the *smallest rotated box* (with side-length 2 in $\|\cdot\|$) containing $\mathcal{E}$. More formally, consider the affine transform $F \colon \mathbb{R}^d \to \mathbb{R}^d$ which maps B_2^d (the unit ball of $\|\cdot\|_2$) to $\mathcal{E}$. Then $\mathcal{C} = F(B_\infty^d)$. Note that the collection

$$\mathcal{H}_s = \{F(2s \cdot x) + s \cdot \mathcal{C} \subset \mathbb{R}^d : x \in \mathbb{Z}^d\},$$

partitions $\mathbb{R}^d$ into disjoint translated copies of $\mathcal{C}$ with side-length $2s$.

In this section, we reduce the problem of c-ANN for $\|\cdot\|$ over $\mathbb{R}^d$ to the problem of c-ANN for

$\|\cdot\|$ over a finite set of points. We first reduce to the case when the dataset and query are bounded by a high-dimensional box, then we will show how to discretize the boxes in order to reduce to a finite set of points.

Lemma 5.1. *Let A be a data structure solving c-ANN for $\|\cdot\|$ over $s\cdot\mathcal{C}$ where $s = O(d)$ with success probability $\frac{9}{10}$, query time $T(n)$ and space $S(n) = \Omega(dn)$. Then there exists a data structure A' solving c-ANN for $\|\cdot\|$ over $\mathbb{R}^d$ which solves the problem with probability $\frac{8}{10}$, query time $T(n) + O(d)$ and space $S(n) + O(dn)$.*

Proof. The data structure A', upon receiving the dataset P, proceeds in the following way:

- Partition the space by a randomly shifted $s\cdot\mathcal{C}$ where $s = 5d$ (with respect to $\|\cdot\|$). More formally, we sample $y \sim [0, 2s]^d$ and consider the collection:

$$\mathcal{H}_{s,y} = \{F(y) + H \subset \mathbb{R}^d : H \in \mathcal{H}_s\}.$$

- For each $H \in \mathcal{H}_{s,y}$, we take the dataset $P \cap H$ falling inside this location, translate the dataset by the center of H and invoke the data structure A on the translated points of $P \cap H$.

On a query q, we identify the location $q \in H \in \mathcal{H}_{s,y}$. We translate the query by the center of H, and query the corresponding data structure holding $P \cap H$.

We say that two points $p, q \in \mathbb{R}^d$ are split if they lie in different cells of the partition $\mathcal{H}_{s,y}$. For any p and q with $\|p - q\| \le 1$, we have

$$\Pr[p \text{ and } q \text{ split}] \le \frac{d\cdot\|p-q\|}{2s} \le \frac{1}{10}$$

where we used the fact that after the affine transform F which maps $e_1, \ldots, e_d$ to the major axes of $\mathcal{E}$, we have the probability that we split points p and q is at most

$$\begin{aligned}\frac{1}{2s}\sum_{i=1}^{d} |(F^{-1}(p-q))_i| = \frac{1}{s}\left\|F^{-1}(p) - F^{-1}(q)\right\|_1 &\le \frac{\sqrt{d}}{2s}\left\|F^{-1}(p) - F^{-1}(q)\right\|_2 \\ &= \frac{\sqrt{d}}{2s}\|p-q\|_{\mathcal{E}} \le \frac{\|p-q\|}{2s}.\end{aligned}$$

Thus, with probability $\frac{9}{10}$, the query point and the dataset point fall in the same grid location. The query time of $T(n) + O(d)$ is immediate, and the space $S(n) + O(dn)$ follows from the fact that we must store a hash of the non-empty values of $P \cap H$ where $H \in \mathcal{H}_{s,y}$, as well as y, as well as the fact that $S(n) = \Omega(n)$. □

We now proceed to the second step where we reduce to the case the dataset and query lie within a fixed set of points. We let X be a greedily constructed γ-net of $s\cdot\mathcal{C}$ (where distances are measured with respect to $\|\cdot\|$). Let $(X, \|\cdot\|)$ be the metric space obtained by restricting the norm to T.

A standard volume argument gives the following fact.

Fact 5.2. *We have that* $|X| \leq \exp\left(O(d\log(d/\gamma))\right)$.

Since X is a γ-net, we may identify points with their closest neighbor in X. The following lemma is immediate, and finishes the reduction.

Lemma 5.3. *Let A be a data structure solving c-ANN for $(X, \|\cdot\|)$ with success probability $\frac{9}{10}$, time $T(n)$ and space $S(n)$. There exists a data structure A' solving $c \cdot (\frac{1+2\gamma}{1-2\gamma})$-ANN for $\|\cdot\|$ over $s \cdot \mathcal{C}$ with success probability $\frac{9}{10}$ in time $T(n) + O(d)$ and space $S(n)$.*

6 Bounding the cutting modulus of a normed space

For $G = (g_{ij}) \in \Delta(m)$, we denote $D = \mathrm{diag}(\rho_G(1), \rho_G(2), \ldots, \rho_G(m))$. We set $A = (a_{ij}) = D^{-1/2}GD^{-1/2}$, so that $a_{ij} = \frac{g_{ij}}{\sqrt{\rho_G(i)\rho_G(j)}}$ and $\mathcal{L}_G = I - A$. For a metric space $(\mathcal{M}, d_{\mathcal{M}})$ and $q > 0$, we define (the inverse of) the nonlinear spectral gap $\gamma(G, d_{\mathcal{M}}^q)$ to be the infimum over $\gamma > 0$ such that for every $u_1, u_2, \ldots, u_m \in \mathcal{M}$, one has:

$$\sum_{i,j=1}^{m} \rho_G(i)\rho_G(j) \cdot d_{\mathcal{M}}(u_i, u_j)^q \leq \gamma \sum_{i,j=1}^{m} g_{ij} \cdot d_{\mathcal{M}}(u_i, u_j)^q.$$

Note that this definition agrees with the one from the Introduction if G is (a multiple of) a doubly-stochastic matrix.

In this section, we show that for every d-dimensional normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ and every $0 < \varepsilon < 1/2$, one has $\Xi(X, \varepsilon) = O\left(\frac{\log d}{\varepsilon^2}\right)$. This bound easily follows (see Theorem 6.6) from a slight extension of Theorem 1.7 to the case when A is not necessarily doubly stochastic. This extension can be obtained by examining the proof from [Nao17], but instead we present a new, shorter and more elementary argument, which constitutes the bulk of the present section (for a slightly different exposition of the same argument, see [Nao18]).

Recall that for normed spaces $X = (\mathbb{R}^d, \|\cdot\|_X)$ and $Y = (\mathbb{R}^d, \|\cdot\|_Y)$ and a linear map $T\colon \mathbb{R}^d \to \mathbb{R}^d$, the operator norm $\|T\|_{X\to Y}$ is defined as follows: $\|T\|_{X\to Y} = \sup_{\|x\|_X=1} \|Tx\|_Y$. The Banach–Mazur distance $d_{\mathrm{BM}}(X, Y)$ between X and Y is defined as follows: $d_{\mathrm{BM}}(X, Y) = \inf_{T\colon \mathbb{R}^d \to \mathbb{R}^d} \|T\|_{X\to Y} \cdot \|T^{-1}\|_{Y\to X}$. By John's theorem, one always has: $d_{\mathrm{BM}}(X, \ell_2^d) \leq \sqrt{d}$.

Theorem 6.1. *For every normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ and every $G = (g_{ij}) \in \Delta(m)$, one has $\gamma(G, \|\cdot\|_X^2) = O\left(\left(\frac{1+\log \mathsf{d}}{\lambda_2(\mathcal{L}_G)}\right)^2\right)$, where $\mathsf{d} = d_{\mathrm{BM}}(X, \ell_2^d) \leq \sqrt{d}$. In particular, one always has: $\gamma(G, \|\cdot\|_X^2) = O\left(\left(\frac{1+\log d}{\lambda_2(\mathcal{L}_G)}\right)^2\right)$.*

Let $V \subset (\mathbb{R}^d)^m$ be the following codimension-1 subspace:

$$V = \left\{ (v_1, v_2, \ldots, v_m) \in (\mathbb{R}^d)^m \;\middle|\; \sum_{i=1}^{m} \sqrt{\rho_G(i)} \cdot v_i = 0 \right\}.$$

We denote by $V_X = (V, \|\cdot\|_{V_X})$ the normed space where for $\boldsymbol{v} = (v_1, v_2, \ldots, v_m) \in V$, the norm is given by $\|\boldsymbol{v}\|_{V_X} = \sqrt{\sum_{i=1}^m \|v_i\|_X^2}$. Denote by $\mathcal{A}\colon V \to V$ the following linear map: $(\mathcal{A}\boldsymbol{v})_i =$

$\sum_{j=1}^m a_{ij} v_j = \sum_{j=1}^m \frac{g_{ij} v_j}{\sqrt{\rho_G(i)\rho_G(j)}}$. In words, $\mathcal{A}$ acts on a tuple of d-dimensional vectors the same way as $A = D^{-1/2} G D^{-1/2}$ acts on a tuple of scalars. It is immediate to check that the image of $\mathcal{A}$ indeed lies in V; this follows from the fact that $\left(\sqrt{\rho_G(1)}, \sqrt{\rho_G(2)}, \ldots, \sqrt{\rho_G(m)}\right)$ is an eigenvector of A. Let $\mathcal{I} \colon V \to V$ be the identity map.

We show that Theorem 6.1 readily follows from the following lemma.

Lemma 6.2. *One has:* $\left\|(\mathcal{I} - \mathcal{A})^{-1}\right\|_{V_X \to V_X} = O\left(\frac{1+\log \mathsf{d}}{\lambda_2(\mathcal{L}_G)}\right)$.

Proof of the implication "Lemma 6.2 $\Rightarrow$ Theorem 6.1". Indeed, an immediate reformulation of Lemma 6.2 is that for every $v_1, v_2, \ldots, v_m \in \mathbb{R}^d$ such that $\sum_{i=1}^m \sqrt{\rho_G(i)} \cdot v_i = 0$, one has:

$$\sum_{i=1}^m \|v_i\|_X^2 \leq O\left(\left(\frac{1+\log \mathsf{d}}{\lambda_2(\mathcal{L})}\right)^2\right) \cdot \sum_{i=1}^m \left\| v_i - \sum_{j=1}^m \frac{g_{ij} v_j}{\sqrt{\rho_G(i)\rho_G(j)}} \right\|_X^2. \tag{8}$$

Our goal is to show that for every $u_1, u_2, \ldots, u_m \in \mathbb{R}^d$, one has:

$$\sum_{i,j=1}^m \rho_G(i)\rho_G(j) \cdot \|u_i - u_j\|_X^2 = O\left(\left(\frac{1+\log \mathsf{d}}{\lambda_2(\mathcal{L}_G)}\right)^2\right) \cdot \sum_{i,j=1}^m g_{ij} \cdot \|u_i - u_j\|_X^2. \tag{9}$$

Without loss of generality, we can assume that $\sum_{i=1}^m \rho_G(i) \cdot u_i = 0$. We set $v_i = \sqrt{\rho_G(i)} \cdot u_i$. Hence, $\sum_{i=1}^m \sqrt{\rho_G(i)} \cdot v_i = 0$ and (8) applies. On the one hand, one has:

$$\begin{aligned}
\sum_{i,j=1}^m \rho_G(i)\rho_G(j) \cdot \|u_i - u_j\|_X^2 &\leq \sum_{i,j=1}^m \rho_G(i)\rho_G(j) \cdot \left(\|u_i\|_X + \|u_j\|_X\right)^2 \\
&\leq 2 \sum_{i,j=1}^m \rho_G(i)\rho_G(j) \cdot \left(\|u_i\|_X^2 + \|u_j\|_X^2\right) \\
&= 4 \sum_{i=1}^m \rho_G(i) \cdot \|u_i\|_X^2 = 4 \sum_{i=1}^m \|v_i\|_X^2.
\end{aligned} \tag{10}$$

On the other hand, one has:

$$\begin{aligned}
\sum_{i=1}^m \left\| v_i - \sum_{j=1}^m \frac{g_{ij} v_j}{\sqrt{\rho_G(i)\rho_G(j)}} \right\|_X^2 &= \sum_{i=1}^m \left\| \sum_{j=1}^m \frac{g_{ij}}{\sqrt{\rho_G(i)}} \cdot \left(\frac{v_i}{\sqrt{\rho_G(i)}} - \frac{v_j}{\sqrt{\rho_G(j)}} \right) \right\|_X^2 \\
&= \sum_{i=1}^m \left\| \sum_{j=1}^m \frac{g_{ij}}{\sqrt{\rho_G(i)}} \cdot (u_i - u_j) \right\|_X^2 \leq \sum_{i=1}^m \left(\sum_{j=1}^m \frac{g_{ij}}{\sqrt{\rho_G(i)}} \cdot \|u_i - u_j\|_X \right)^2 \\
&\leq \sum_{i,j=1}^m g_{ij} \cdot \|u_i - u_j\|_X^2,
\end{aligned} \tag{11}$$

where the third step is due to the triangle inequality, and the fourth step is due to Jensen's inequality. Combining (8), (10) and (11), we obtain (9). □

Now let us show the proof of Lemma 6.2. For this we will need to relate the geometry of X

and the Euclidean geometry. Let $H = (\mathbb{R}^d, \|\cdot\|_H)$ be a Hilbert space such that for every $v \in \mathbb{R}^d$, one has $\|v\|_H \le \|v\|_X \le \mathsf{d} \cdot \|v\|_H$. The existence of such a space follows immediately from the definition of Banach-Mazur distance. In particular, if T is a linear map such that $\|T\|_{X \to \ell_2^d} \le 1$ and $\|T^{-1}\|_{\ell_2^d \to X} \le \mathsf{d}$, we can define H by $\|x\|_H = \|Tx\|_2$. We define the normed space $V_H = (V, \|\cdot\|_{V_H})$ analogously to V_X: the norm $\|\boldsymbol{v}\|_{V_H}$ for $\boldsymbol{v} = (v_1, v_2, \dots, v_m) \in V$ is defined by $\|\boldsymbol{v}\|_{V_H} = \sqrt{\sum_{i=1}^m \|v_i\|_H^2}$. Clearly, for every $\boldsymbol{v} \in V$, one has:

$$\|\boldsymbol{v}\|_{V_H} \le \|\boldsymbol{v}\|_{V_X} \le \mathsf{d} \cdot \|\boldsymbol{v}\|_{V_H}. \tag{12}$$

Finally, we define $\widetilde{A} = \frac{A+I}{2}$ and $\widetilde{\mathcal{A}} = \frac{\mathcal{A}+\mathcal{I}}{2}$. Let us observe that $\mathcal{I} - \widetilde{\mathcal{A}} = \frac{1}{2}(\mathcal{I} - \mathcal{A})$, so $\|(\mathcal{I} - \mathcal{A})^{-1}\|_{V_X \to V_X} \le \frac{1}{2} \cdot \|(\mathcal{I} - \widetilde{\mathcal{A}})^{-1}\|_{V_X \to V_X}$, thus it is enough to show that

$$\|(\mathcal{I} - \widetilde{\mathcal{A}})^{-1}\|_{V_X \to V_X} = O\left(\frac{1 + \log \mathsf{d}}{\lambda_2(\mathcal{L}_G)}\right). \tag{13}$$

One can see that (13) is an immediate corollary of the following three statements together with (12). Let us note that Lemma 6.5 is the place where the logarithmic dependence on d shows up.

Claim 6.3. *One has* $\|\widetilde{\mathcal{A}}\|_{V_X \to V_X} \le 1$.

Claim 6.4. *One has* $\|\widetilde{\mathcal{A}}\|_{V_H \to V_H} \le 1 - \frac{\lambda_2(\mathcal{L}_G)}{2}$.

Lemma 6.5. *Let* $\|\cdot\|_P$ *and* $\|\cdot\|_Q$ *be two norms on* $\mathbb{R}^{d'}$ *such that for some* $\Phi \ge 1$ *and for every* $u \in \mathbb{R}^{d'}$ *one has* $\|u\|_Q \le \|u\|_P \le \Phi \cdot \|u\|_Q$. *Suppose that* $T \colon \mathbb{R}^{d'} \to \mathbb{R}^{d'}$ *is a linear map such that* $\|T\|_{P \to P} \le 1$ *and* $\|T\|_{Q \to Q} \le 1 - \varepsilon$ *for some* $0 < \varepsilon < 1$. *Then,* $\|(I - T)^{-1}\|_{P \to P} = O\left(\frac{1 + \log \Phi}{\varepsilon}\right)$.

Proof of Claim 6.3. For every $\boldsymbol{v} = (v_1, v_2, \dots, v_m) \in V$ one has

$$\begin{aligned}
\|\mathcal{A}\boldsymbol{v}\|_{V_X}^2 &= \sum_{i=1}^m \|(\mathcal{A}\boldsymbol{v})_i\|_X^2 = \sum_{i=1}^m \left\| \sum_{j=1}^m \frac{g_{ij} v_j}{\sqrt{\rho_G(i)\rho_G(j)}} \right\|_X^2 \le \sum_{i=1}^m \left(\sum_{j=1}^m \frac{g_{ij}}{\rho_G(i)} \left\| \sqrt{\frac{\rho_G(i)}{\rho_G(j)}} \cdot v_j \right\|_X \right)^2 \\
&\le \sum_{i=1}^m \sum_{j=1}^m \frac{g_{ij}}{\rho_G(i)} \left\| \sqrt{\frac{\rho_G(i)}{\rho_G(j)}} \cdot v_j \right\|_X^2 = \sum_{i=1}^m \sum_{j=1}^m \frac{g_{ij}}{\rho_G(j)} \|v_j\|_X^2 = \sum_{j=1}^m \|v_j\|_X^2 = \|\boldsymbol{v}\|_{V_X}^2,
\end{aligned}$$

where the third step is by the triangle inequality, and the fourth step is by Jensen's inequality. Hence, $\|\mathcal{A}\|_{V_X \to V_X} \le 1$. But this implies that $\|\widetilde{\mathcal{A}}\|_{V_X \to V_X} \le 1$ as well. □

Proof of Claim 6.4. Let us first observe that for every $u \in \mathbb{R}^m$ such that $\sum_{i=1}^m \sqrt{\rho_G(i)} \cdot u_i = 0$, one has

$$\|\widetilde{A}u\|_2 \le \left(1 - \frac{\lambda_2(\mathcal{L}_G)}{2}\right) \cdot \|u\|_2, \tag{14}$$

since $\widetilde{A}$ is positive semidefinite, the largest eigenvalue is 1, the corresponding eigenvector is $(\sqrt{\rho_G(i)})_{i=1}^m$, and the second largest eigenvalue is $1 - \lambda_2(\mathcal{L}_G)/2$.

The desired inequality reduces to (14) as follows. Since H is a Hilbert space, there exists an orthogonal basis $e_1, e_2, \dots, e_d \in \mathbb{R}^d$ such that for every $u \in \mathbb{R}^m$, one has $\|u\|_H^2 = \sum_{i=1}^m \langle u, e_i \rangle^2$. For

$1 \le i \le d$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_m) \in V$, define $\pi_i(\boldsymbol{v}) = (\langle v_1, e_i\rangle, \langle v_2, e_i\rangle, \ldots, \langle v_m, e_i\rangle) \in \mathbb{R}^m$. Then, $\|\boldsymbol{v}\|_{V_H}^2 = \sum_{i=1}^d \|\pi_i(\boldsymbol{v})\|_2^2$. One has:

$$\begin{aligned}\|\widetilde{\mathcal{A}}\boldsymbol{v}\|_{V_H}^2 &= \sum_{i=1}^d \|\pi_i(\widetilde{\mathcal{A}}\boldsymbol{v})\|_2^2 = \sum_{i=1}^d \|\widetilde{A}\pi_i(\boldsymbol{v})\|_2^2 \le \left(1 - \frac{\lambda_2(\mathcal{L}_G)}{2}\right)^2 \sum_{i=1}^d \|\pi_i(\boldsymbol{v})\|_2^2 \\ &= \left(1 - \frac{\lambda_2(\mathcal{L}_G)}{2}\right)^2 \|\boldsymbol{v}\|_{V_H}^2.\end{aligned}$$

□

Proof of Lemma 6.5. For every $k \ge 1$, one has

$$\|T^k\|_{P\to P} \le \Phi \|T^k\|_{Q\to Q} \le \Phi \cdot (1-\varepsilon)^k.$$

Thus, we can choose $k^* = O((\log 2\Phi)/\varepsilon)$ such that $\|T^{k^*}\|_{P\to P} \le 1/2$. Finally, we have:

$$\|(I-T)^{-1}\|_{P\to P} \le \sum_{k=0}^{\infty} \|T^k\|_{P\to P} \le k^* \cdot \sum_{i=0}^{\infty} \|T^{ik^*}\|_{P\to P} \le k^* \cdot \sum_{i=0}^{\infty} (1/2)^i = 2k^* = O\left(\frac{1+\log\Phi}{\varepsilon}\right)$$

as desired. □

Theorem 6.6. *For every normed space $X = (\mathbb{R}^d, \|\cdot\|_X)$ with $d_{\mathrm{BM}}(X, \ell_2^d) = \mathsf{d} \le \sqrt{d}$, and every $0 < \varepsilon < 1/2$, one has: $\Xi(X,\varepsilon) = O\left(\frac{1+\log \mathsf{d}}{\varepsilon^2}\right)$. In particular, one always has: $\Xi(X,\varepsilon) = O\left(\frac{\log d}{\varepsilon^2}\right)$.*

Proof. Let $R > 0$ be a parameter to be fixed later. Let $G \in \Delta(m)$ and let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in X^m$ be such that $\|x_i - x_j\|_X \le 1$ if $g_{ij} > 0$. Suppose that $\boldsymbol{x}$ has no $1/2$-dense ball of radius R. Then,

$$\sum_{i,j=1}^m \rho_G(i)\rho_G(j) \cdot \|x_i - x_j\|_X^2 \ge \frac{R^2}{2}. \tag{15}$$

On the other hand, we have:

$$\sum_{i,j=1}^m g_{ij} \cdot \|x_i - x_j\|_X^2 \le 1, \tag{16}$$

since $\|x_i - x_j\|_X^2 \le 1$ whenever $g_{ij} > 0$. Thus, combining (15), (16) and Theorem 6.1, we get: $\lambda_2(\mathcal{L}_G) = O\left(\frac{1+\log \mathsf{d}}{R}\right)$. Thus, by setting R to a large enough multiple of $\frac{1+\log \mathsf{d}}{\varepsilon^2}$ and using Cheeger's inequality (Theorem 2.2), we conclude that G has a cut with conductance at most ε. □

Note that Theorems 3.6 and 6.6, together with a standard discretization argument imply Theorem 1.9. Indeed, given a norm $\|\cdot\|_X$ on $\mathbb{R}^d$, and a radius R so that $\log R$ is polynomial in d, we can greedily find N points $x_1, \ldots x_N$ so that the balls $B_X(x_1,\gamma), \ldots, B_X(x_N,\gamma)$ cover $B_X(0,R)$, and $\log N = O(d \log(R/\gamma))$. We can then use Theorem 3.6 with the metric space of size N induced on $\{x_1, \ldots, x_N\}$ and the cutting modulus bound given in Theorem 6.6. We identify any set S in the collection $\mathcal{C}$ guaranteed by Theorem 3.6 with the union of the balls $B_X(x_i,\gamma)$ that cover the elements of S. It is easy to verify that any two points $u, v \in B_X(0,R)$ that lie at a distance at most

$1 - 2\gamma$ apart are separated by a uniformly random set in the subcollection $\mathcal{S}$ with probability at most 50ε. The guarantee of Theorem 1.9 then follows by a simple rescaling.

7 Algorithm for ℓ_p

In this section, we give an $O(p)$-ANN algorithm for ℓ_p norms when $2 < p \leq \infty$. We use the framework of Section 4 and a more refined bound on the cutting modulus of ℓ_p in order to achieve an improved approximation. The improved bound on the cutting modulus for ℓ_p norms will follow from a slight generalization of an argument due to Matoušek [Mat97]. We note that this improved bound on the cutting modulus can also be proved by generalizing the interpolation-based argument in [Nao14]. However, Matoušek's argument is more explicit, and allows us to relate Rayleigh quotients in the ℓ_p and ℓ_2 norms. Using this observation, we can show that in the case of ℓ_p norms, we can also bound the cutting modulus with respect to just balls and complements of boxes, rather than arbitrary finite sets of points. Because both are efficiently describable, we can derive an efficient data structure for $O(p)$-ANN over ℓ_p.

The algorithm presented achieves efficient query time and near-linear space. As in the case of the cell-probe algorithm of Section 4, the time for *preprocessing* is exponential in the dimension. During preprocessing, we consider a finite metric space of $\exp(O(d \log d))$ points discretizing the space, so executing the partitioning theorems takes $\exp(O(d \log d))$ time.

The goal of this section is to prove the following theorem.

Theorem 7.1. *For any $0 < \alpha < 1$, there exists a data structure solving c-ANN for ℓ_p with success probability $\frac{9}{10}$ with the following guarantees:*

- *the approximation is $c = O(p/\alpha)$,*
- *the query time of the data structure is $\mathrm{poly}(d) \cdot n^{\alpha}$, and*
- *the space of the data structure is $\mathrm{poly}(d) \cdot n^{1+\alpha}$.*

It suffices to give a data structure over some finite metric space of $N = \exp(O(d \log d))$ points, where pairwise distances are given by the ℓ_p norm. In particular, let X be a set of N points $x_1, \ldots, x_N \in \mathbb{R}^d$ formed by suitably discretizing the cube $[-s, s]^d$ where $s = O(d)$ (see Lemma 5.3); we consider the metric space $(X, \|\cdot\|_p)$. Theorem 7.1 follows from the following lemma.

Lemma 7.2. *For any $0 < \alpha < 1$, there exists a data structure solving c-ANN for $(X, \|\cdot\|_p)$ with success probability $\frac{9}{10}$, approximation $O(p/\alpha)$, query time $\mathrm{poly}(d) \cdot n^{\alpha}$ and space $\mathrm{poly}(d) \cdot n^{1+\alpha}$.*

In order to prove Lemma 7.2, we follow the framework from Section 4, which requires two tasks:

1. Show that the metric space $(X, \|\cdot\|_p)$ described above has $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(p/\varepsilon)$ for $\varepsilon = \Theta(\alpha)$, and an efficiently representable, or *succinct* (see the following definition) collection $\mathfrak{H}$ of subsets of X.

2. Give an analogous, efficient version of Theorem 3.6. In particular, the collection of subsets $S_1, \ldots, S_m \subset X$ in Theorem 3.6 will belong to a succinct collection $\mathfrak{S}$.

Definition 7.3 (Succinct Collections)**.** *We say that a collection $\mathfrak{S}$ of subsets of X is* b-succinct *if there exists a function $E\colon \mathfrak{S} \to \{0,1\}^b$, as well as an algorithm D running in time* $\mathrm{poly}(b)$ *taking inputs in $\{0,1\}^b \times X$ satisfying*

$$D(E(S), q) = \begin{cases} 1 & q \in S \\ 0 & q \notin S \end{cases} \qquad \forall S \in \mathfrak{S}.$$

We proceed to state the lemmas accomplishing steps 1 and 2 described above. Let $\mathfrak{H}$ be the collection of subsets of X induces by coordinate halfspaces. I.e. $\mathfrak{H}$ consists of sets of the type $\{x \in X : x_i \geq t\}$ or $\{x \in X : x_i \leq t\}$ for some $i \in [d]$ and $t \in \mathbb{R}$. We let $P \subset X$ be any set of n points in X. We aim to prove the following two lemmas.

Lemma 7.4 (Cutting Modulus for ℓ_p)**.** *We have $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(p/\varepsilon)$.*

By a quick inspection of the proof of Theorem 3.6, an efficient version of Theorem 3.6 would follow from the following efficient version of Lemma 3.7.

Lemma 7.5 (Efficient Lemma 3.7)**.** *There exists an a collection of subsets $\mathfrak{S}$ of X which is b-succinct for $b = \mathrm{poly}(d)$ such that for any matrix $G \in \Delta(N)$ where $g_{ij} > 0$ only if $\|x_i - x_j\|_p \leq 1$, either there exists a $\frac{1}{4}$-dense ball of radius $R = \Xi_{\mathfrak{H}}(X, \varepsilon)$, or there exists a subset $S \in \mathfrak{S}$ where:*

$$\frac{1}{3} \leq \rho_G(S) \leq \frac{3}{4} \qquad \text{and} \qquad \sum_{i \in S, j \notin S} g_{ij} \leq 2\varepsilon.$$

Lemma 7.5 implies Theorem 1.8 by an argument analogous to the proof of Theorem 3.6.

Proof of Theorem 7.1 given Lemmas 7.4 and 7.5. The data structure proceeds in a similar fashion to the cell-probe data structure in Section 4. The one modification is that the output $\mathcal{S}$ from sub-routine $\mathrm{MWU}(P)$ satisfies $\mathcal{S} \subset \mathfrak{S}$ for a b-succinct collection $\mathfrak{S}$ with $b = \mathrm{poly}(d)$. Therefore, $\textsc{ProcessCut}(P, S, \ell, v)$ stores the b bits, $E(S)$, in $v.S$. In $\textsc{QueryCut}(q, v)$, the algorithm executes $D(v.S, q)$ to evaluate the "if" statement in $\textsc{QueryCut}(q, v)$. □

Lemma 7.4 and Lemma 7.5 follow from a Rayleigh quotient inequality for ℓ_p norms, which we prove in the next subsection.

7.1 Rayleigh quotient inequality for ℓ_p spaces and proof of Lemma 7.4

Let $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^d)^m$ and $G \in \Delta(m)$. As in Section 3, let $\rho_G(i) = \sum_{j=1}^m g_{ij}$, and for simplicity in notation, we will write $\rho(i) = \rho_G(i)$ since the matrix G will be fixed.

For $k \in [d]$, consider d functions $F_k \colon \mathbb{R} \to \mathbb{R}$. We define a function $F \colon (\mathbb{R}^d)^m \to (\mathbb{R}^d)^m$ by applying to all m points the functions $F_1, \ldots, F_d$ coordinate-wise. In particular, for each $j = 1, \ldots, m$,

$$(F(\boldsymbol{x}))_j = (F_1(x_{j1}), F_2(x_{j2}), \ldots, F_d(x_{jd})) \in \mathbb{R}^d. \tag{17}$$

Additionally, for any $k \in [d]$, let $\pi_k \colon \mathbb{R}^d \to \mathbb{R}$ be the projection onto the kth coordinate. We also think of π_k as acting on a sequence of points by letting $\pi_k \colon (\mathbb{R}^d)^m \to \mathbb{R}^m$ be given by:

$$\pi_k(\boldsymbol{x}) = (x_{1k}, x_{2k}, \ldots, x_{mk}) \in \mathbb{R}^m. \tag{18}$$

Note that $\pi_k(F(\boldsymbol{x})) = F_k(\pi_k(\boldsymbol{x}))$.

We will use the Mazur map [Maz29] (see also [BL00]) between ℓ_p and ℓ_2 defined on $\mathbb{R}^d$ by $M_{p,2}(x)_i = \operatorname{sign}(x_i) \cdot |x_i|^{p/2}$ for every $i \in [d]$. Note that for any $x \in \mathbb{R}^d$, we have $\|M_{p,2}(x)\|_2^2 = \|x\|_p^p$. The following inequality gives a well-known estimate on the modulus of uniform continuity of the the Mazur map: for any $x, y \in \mathbb{R}^d$, we have

$$\|M_{p,2}(x) - M_{p,2}(y)\|_2^2 \leq \frac{p^2}{4} \|x - y\|_p^2 \cdot (\|x\|_p^p + \|y\|_p^p)^{1-\frac{2}{p}} \tag{19}$$

See Section 5.1 of [Nao14] for a proof of this inequality with the explicit constants above.

We state the following result of Matoušek [Mat97], generalized slightly to the case of non-negative symmetric matrices, and stated in terms of Rayleigh quotients.

Lemma 7.6. *For any $\boldsymbol{x} \in (\mathbb{R}^d)^m$ there exist d monotone functions $F_1, \ldots, F_d \colon \mathbb{R} \to \mathbb{R}$ such that:*

$$\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} \leq \mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_p^p) \cdot (\sqrt{2}p)^p.$$

Proof. For $k \in [d]$, we define the function $F_k(x) = \operatorname{sign}(x - z_k) \cdot |x - z_k|^{p/2}$ for some numbers $z_1, \ldots, z_d \in \mathbb{R}$ which we specify later. Note that each F_k is monotone. Let $\boldsymbol{y} = (y_1, \ldots, y_m) \in (\mathbb{R}^d)^m$ be given by $\boldsymbol{y} = F(\boldsymbol{x})$; we will show the bound relating $\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_p^p)$ and $\mathsf{R}(\boldsymbol{y}, G, \|\cdot\|_2^2)$.

For any coordinate $k \in [d]$, let $z_k \in \mathbb{R}$ be a real number such that

$$\sum_{i=1}^m \rho(i) y_{ik} = \sum_{i=1}^m \rho(i) F_k(x_{ik}) = 0.$$

The existence of such a number follows by the continuity of $\sum_{i=1}^m \rho(i) \operatorname{sign}(x_{ik} - z_k) \cdot |x_{ik} - z_k|^{p/2}$ in z_k, and the fact that this function goes to $+\infty$ as z_k tends to $-\infty$, and to $-\infty$ as z_k tends to $+\infty$. Let us denote then, for notational convenience, $\tilde{x}_i = x_i - z$, where z is the vector $(z_1, \ldots, z_d)$. We have $y_i = M_{p,2}(\tilde{x})$, so, $\|y_i\|_2^2 = \|\tilde{x}\|_p^p$ for all $i \in [m]$. It follows from the triangle inequality and

Cauchy-Schwarz that

$$\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|_p^p = \sum_{i=1}^{m}\rho(i)\left\|y_i - \sum_{j=1}^{m}\rho(j)y_j\right\|_2^2 \le \sum_{i=1}^{m}\sum_{j=1}^{m}\rho(i)\rho(j)\,\|y_i - y_j\|_2^2$$
$$= \frac{1}{\mathsf{R}(y, G, \|\cdot\|_2^2)}\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|y_i - y_j\|_2^2.$$

Applying inequality (19) to each term on the right hand side, we have

$$\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|_p^p \le \frac{1}{\mathsf{R}(y, G, \|\cdot\|_2^2)}\frac{p^2}{4}\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|\tilde{x}_i - \tilde{x}_j\|_p^2 \cdot (\|\tilde{x}_i\|_p^p + \|\tilde{x}_j\|_p^p)^{1-\frac{2}{p}}$$

By Hölder's inequality, we may write

$$\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|\tilde{x}_i - \tilde{x}_j\|_p^2 \cdot (\|\tilde{x}_i\|_p^p + \|\tilde{x}_j\|_p^p)^{1-\frac{2}{p}} \le \left(\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|\tilde{x}_i - \tilde{x}_j\|_p^p\right)^{2/p}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}(\|\tilde{x}_i\|_p^p + \|\tilde{x}_j\|_p^p)\right)^{1-\frac{2}{p}}$$
$$= \left(\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|\tilde{x}_i - \tilde{x}_j\|_p^p\right)^{2/p}\left(2\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|_p^p\right)^{1-\frac{2}{p}}$$

Combining the inequalities, and using $\tilde{x}_i - \tilde{x}_j = x_i - x_j$, we get

$$\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|^p \le \frac{1}{\mathsf{R}(y, G, \|\cdot\|_2^2)}\cdot\frac{p^2}{2^{(p+2)/p}}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|x_i - x_j\|^p\right)^{2/p}\left(\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|^p\right)^{1-\frac{2}{p}}.$$

Thus, we obtain:

$$\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|^p \le \frac{1}{\mathsf{R}(y, G, \|\cdot\|_2^2)^{p/2}}\cdot\frac{p^p}{2^{(p+2)/2}}\cdot\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|x_i - x_j\|^p. \tag{20}$$

Finally, we have that:

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\rho(i)\rho(j)\|x_i - x_j\|^p = \sum_{i=1}^{m}\sum_{j=1}^{m}\rho(i)\rho(j)\|\tilde{x}_i - \tilde{x}_j\|^p \le 2^{p+1}\sum_{i=1}^{m}\rho(i)\|\tilde{x}_i\|^p, \tag{21}$$

and combining (20) and (21), we obtain the desired result. □

With the Rayleigh quotient inequality from Lemma 7.6, we can easily prove Lemma 7.4.

Corollary 7.7. *Let $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^d)^m$ be any set of m points, where:*

- *$g_{ij} > 0$ only if $\|x_i - x_j\|_p \le 1$, and*
- *$\sum_{i=1}^{m}\sum_{j=1}^{m}\rho(i)\rho(j)\|x_i - x_j\|_p^p \ge \frac{(\sqrt{2}p)^p}{\varepsilon^{p/2}}$,*

then $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \le \varepsilon$.

Proof. We simply note that $\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_p^p) \le \dfrac{\varepsilon^{p/2}}{(\sqrt{2}p)^p}$, so $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \le \varepsilon$ by Lemma 7.6. □

Lemma 7.8. *Let* $\boldsymbol{x} = (x_1, \dots, x_m) \in (\mathbb{R}^d)^m$ *be any sequence of* m *points where* $g_{ij} > 0$ *only if* $\|x_i - x_j\|_p \le 1$. *Then,*

- *either* $\boldsymbol{x}$ *has a* $\frac{1}{2}$*-dense* ℓ_p*-ball of radius* $R = O(p/\sqrt{\varepsilon})$, *or*
- $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \le \varepsilon$.

Proof. If $\sum_{i=1}^m \sum_{j=1}^m \rho(i)\rho(j)\|x_i - x_j\|_p^p \ge \dfrac{(\sqrt{2}p)^p}{\varepsilon^{p/2}}$, then we have $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \le \varepsilon$, so assume otherwise. Consider the distribution $\mathcal{D}$ supported on $[m]$ given by sampling i with probability $\rho(i)$. Then,

$$\mathop{\mathbb{E}}_{i,j\sim\mathcal{D}}[\|x_i - x_j\|_p^p] = \sum_{i=1}^m \sum_{j=1}^m \rho(i)\rho(j)\|x_i - x_j\|_p^p \le \frac{(\sqrt{2}p)^p}{\varepsilon^{p/2}}.$$

Then, there exists a fixed index $i \in [m]$ where

$$\mathop{\mathbb{E}}_{j\sim\mathcal{D}}[\|x_i - x_j\|_p^p] \le \frac{(\sqrt{2}p)^p}{\varepsilon^{p/2}},$$

and by Markov's inequality,

$$\Pr_{j\sim\mathcal{D}}\left[\|x_i - x_j\|_p^p \ge 2 \cdot \frac{(\sqrt{2}p)^p}{\varepsilon^{p/2}}\right] \le \frac{1}{2}.$$

Thus, we conclude that there exists an ℓ_p-ball of radius $R = \dfrac{2^{1/2+1/p}p}{\sqrt{\varepsilon}} = O(p/\sqrt{\varepsilon})$ around x_i, $B_p(x_i, R)$, such that $\sum_{x_j \in B_p(x_i,R)\cap P} \rho(j) \ge \frac{1}{2}$. □

Proof of Lemma 7.4. By Lemma 7.8 applied with $2\varepsilon^2$, if $\boldsymbol{x}$ does not have a $\frac{1}{2}$-dense ball of radius $R = O(p/\varepsilon)$, then $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \le 2\varepsilon^2$. Then there must exist some coordinate $k \in [d]$ for which

$$\mathsf{R}(F_k(\pi_k(\boldsymbol{x})), G, |\cdot|^2) = \mathsf{R}(\pi_k(F(\boldsymbol{x})), G, |\cdot|^2) \le 2\varepsilon^2.$$

We can apply Cheeger's inequality (Theorem 2.2) to $F_k(\pi_k(\boldsymbol{x}))$ and we get that there exists a real number t such that the set $S = \{j \in [m] : F_k(x_{jk}) \le t\}$ satisfies $\Phi_G(S) \le \varepsilon$. Observe that, since F_k is a monotone function, we can equivalently write $S = \{j \in [m] : x_{jk} \le F_k^{-1}(t)\}$, which is the set induced on $\boldsymbol{x}$ by $H = \{x \in X : x_k \le F^{-1}(t)\} \in \mathfrak{H}$. Thus, we have $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(p/\varepsilon)$. □

7.2 Proof of Lemma 7.5

For the metric space $(X, \|\cdot\|_p)$, we let $\mathcal{V} \subset \mathbb{R}$ be the set of values the coordinates of X take. In particular, $\mathcal{V}$ contains all real numbers $[-O(d), O(d)]$ with $\mathrm{poly}(d)$ bits of precision. Note that $|\mathcal{V}| \le \exp(O(d \log d))$.

Definition 7.9. *Consider d tuples of values in $\mathcal{V}$, $V = \{(v_{k,0}, v_{k,1})\}_{k=1}^{d} \in (\mathcal{V} \times \mathcal{V})^d$ where $v_{k,0}, v_{k,1} \in \mathcal{V}$. Then we say* $\mathrm{Box}(V) \subset X$ *is the set given by:*

$$\mathrm{Box}(V) = \{y \in X : \forall k \in [d], v_{k,0} < y_k < v_{k,1}\}.$$

We make the following observation.

Fact 7.10. *For any $V_1 \in (\mathcal{V} \times \mathcal{V})^d$ and $H \in \mathfrak{H}$, there exists $V_3 \in (\mathcal{V} \times \mathcal{V})^d$ such that:*

$$\mathrm{Box}(V_1) \cap H = \mathrm{Box}(V_3).$$

Consider the collection $\mathfrak{S}$ composed of all complements of boxes in X, i.e.,

$$\mathfrak{S} = \left\{\overline{\mathrm{Box}(V)} \subset X : V \in (\mathcal{V} \times \mathcal{V})^d\right\}.$$

Lemma 7.11. *The collection $\mathfrak{S}$ is b-succinct for $b \leq O(d^2 \log d)$.*

Proof. For each set $S \in \mathfrak{S}$, we can write the $2d$ values of $\mathcal{V}$ which form the values of $v_{k,0}$ and $v_{k,1}$ for each $k \in [d]$ so $S = \overline{\mathrm{Box}(V)}$. In order to determine whether a point $q \in S$, we simply check all d coordinates and compare them to $v_{k,0}$ and $v_{k,1}$ to determine if $q \notin \mathrm{Box}(V)$. □

Proof of Lemma 7.5. We closely follow the proof of Lemma 3.7. The proof gives an iterative procedure which maintains a set S which is initially empty. At each iteration, points are added to the set S. In any particular iteration, we consider a matrix $\widetilde{G}$, which is the submatrix of G given by the rows and columns of points in $\overline{S}$, and rescaled so that $\widetilde{G} \in \Delta(|\overline{S}|)$. Consider one iteration of the procedure of Lemma 3.7, and let $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^d)^m$ be the points remaining in $\overline{S}$. Suppose there exists some $V \in (\mathcal{V} \times \mathcal{V})^d$ with $\overline{S} = \mathrm{Box}(V)$; initially, this is true, and we will maintain this invariant throughout the procedure.

By Lemma 7.4, either there exists an ℓ_p ball of radius $O(p/\varepsilon)$ which is $\frac{1}{2}$-dense with respect to $\widetilde{G}$, or there exists a set $H \in \mathfrak{H}$ such that $H_{\boldsymbol{x}} = \{j \in [m] : x_j \in H\}$ satisfies $\Phi_{\widetilde{G}}(H_{\boldsymbol{x}}) \leq \varepsilon$. Similarly to Lemma 7.8, if there exists a $\frac{1}{2}$-dense ball with respect to $\widetilde{G}$, then that ball is $\frac{1}{4}$-dense with respect to G, and we are done. Assume then that this is not the case. Without loss of generality, $\rho_{\widetilde{G}}(H_{\boldsymbol{x}}) \leq \frac{1}{2}$, or, otherwise, we can replace H with its complement, which is also an element of $\mathfrak{H}$. The set $\overline{S} \setminus H_{\boldsymbol{x}}$ is the intersection of $\mathrm{Box}(V)$ and the complement $\overline{H}$ of H, which is also induced by a box. The proof of Lemma 3.7 then considers letting $S \leftarrow S \cup H_{\boldsymbol{x}}$, so by the above observation, $\overline{S}$ is still represented as a box. □

8 Algorithm for Schatten-p

We will first consider the case of $p > 2$, since we obtain slightly better dependence on α. We then discuss the case $p \in [1, 2]$. Let $(\mathbb{R}^{d\times d}, \|\cdot\|_{S_p})$ be the normed space over matrices $x \in \mathbb{R}^{d\times d}$ with norm $\|x\|_{S_p} = \left(\sum_{k=1}^{d} |\lambda_k(x)|^p\right)^{1/p}$, where $|\lambda_1(x)| \geq |\lambda_2|(x) \geq \cdots \geq |\lambda_d|$ are the d eigenvalues of x.

Note that $\|x\|_{S_2} = \|x\|_2$, where we consider x as a vector in $\mathbb{R}^{d^2}$. For this reason, in this section we will identify S_2 (as a norm on $\mathbb{R}^{d\times d}$) and ℓ_2 (as a norm on $\mathbb{R}^{d^2}$).

Theorem 8.1. *Fix some $0 < \alpha < 1$ and $p > 2$. There exists a data structure solving c-ANN for S_p with success probability $\frac{9}{10}$ with the following guarantees:*

- *the approximation is $c = O(p/\alpha)$.*
- *the query time of the data structure is $\mathrm{poly}(d^p) \cdot n^\alpha$, and*
- *the space of the data structure is $\mathrm{poly}(d^p) \cdot n^{1+\alpha}$.*

Theorem 8.1 for the case of $p = O(1)$ actually implies Theorem 7.1, with weaker space and query time bounds, by embedding points $x \in \ell_p$ into diagonal matrices in S_p. However, we divide the presentation since Theorem 8.1 presents its unique set of obstacles to overcome. At a high level, we follow a similar structure to Section 7. We consider the metric space (X, d_{S_p}) given by $N = \exp(O(d \log d))$ points (matrices) taken by rounding all entries of matrices with $\|x\|_{S_p} \le O(d)$ to $\mathrm{poly}(d)$ many bits (see Lemma 5.3). In particular, Theorem 8.1 follows from the following lemma.

Lemma 8.2. *Fix any $0 < \alpha < 1$, there exists a data structure solving c-ANN for (X, d_{S_p}) with success probability $\frac{9}{10}$, approximation $O(p/\alpha)$, query time $\mathrm{poly}(d^p) \cdot n^\alpha$ and space $\mathrm{poly}(d^p) \cdot n^{1+\alpha}$.*

Similarly to Section 7, we show $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(p/\varepsilon)$ for a succinct collection $\mathfrak{H}$, and give an efficient version of Lemma 3.7. We set $R = \Xi(X, \varepsilon)$. We state the efficient version of Lemma 3.7, whose statement is similar to the statement of Lemma 7.5; however, we will follow with a brief discussion of the difficulties encountered in S_p.

Lemma 8.3. *There exists a collection of subsets $\mathfrak{S}$ of X which is b-succinct for $b = \mathrm{poly}(d^p)$ such that for any matrix $G \in \Delta(N)$ where $g_{ij} > 0$ only if $\|x_i - x_j\|_{S_p} \le 1$, either there exist a $\frac{1}{4}$-dense ball of radius R with respect to G, or there exists a subset $S \in \mathfrak{S}$ where:*

$$\frac{1}{3} \le \rho_G(S) \le \frac{3}{4} \qquad \text{and} \qquad \sum_{i\in S, j\notin S} g_{ij} \le 2\varepsilon.$$

At the heart of the algorithm for ℓ_p lies a Rayleigh quotient inequality for ℓ_p spaces (Lemma 7.6) showing that for any $\boldsymbol{x} \in (\mathbb{R}^d)^m$, there exist d *monotone* functions $F_1, \dots, F_d \colon \mathbb{R} \to \mathbb{R}$ which are applied *coordinate-wise* to every point so:

$$\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_p^p) \le \mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2) \cdot (\sqrt{2}p)^p.$$

Since the functions $F_1, \dots, F_d$ acted on the points coordinate-wise and were monotone, the proof of Lemma 7.5 claimed that coordinate-cuts of $F(\boldsymbol{x})$ corresponded to coordinate cuts of the original points $\boldsymbol{x}$. This has two advantages: 1) the querying algorithm requires no knowledge of the map F used, and 2) the possibly unbounded unions of coordinate cuts form the complement of boxes, which are efficiently described.

For the case of S_p, we give a similar inequality to Lemma 7.6. At a high level, we say that for any $\boldsymbol{x} \in (\mathbb{R}^{d\times d})^m$ and any matrix $G \in \Delta(m)$, there exists a map $F\colon \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ (which depends on $\boldsymbol{x}$, is not applied coordinate-wise, and is not monotone) such that:

$$\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p) \leq (O(p))^p \, \mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} + \eta,$$

where η is a small error term which depends on $\boldsymbol{x}$, G, and F. Similarly to the case of ℓ_p, the data structure divides the points in $\boldsymbol{x}$ according to a coordinate cut S *after applying the map* F. This means that the succinct collection $\mathfrak{S}$ must encode F in the description of the cut. In particular, the algorithm $D(E(S), q)$ will first evaluate $F(q)$ (which it decodes from $E(S)$), and then checks some coordinate of $F(q)$ to determine if $q \in S$.

One issue that arises is that Lemma 3.7 may produce sets S which are given by an unbounded union of $\widetilde{S}$ (one for each iteration in the proof of Lemma 3.7). In the case of Lemma 7.5, these unions formed complements of boxes, so we simply stored the box; however, storing the description of each map F in each iteration becomes too expensive.

Therefore, the data structure must balance the number of functions F to store with the error term η. More specifically, the data structure begins the partitioning procedure by storing the first map F found, and continues using the same map F until the error term η becomes too large. When the error term becomes too large, we re-compute and store a new map F. With this procedure, we show that it suffices to store $(O(d))^p$ many maps for each set, which gives us the succinct collection.

8.1 Rayleigh quotient inequality for S_p, $p > 2$

Let $x \in \mathbb{R}^{d\times d}$ be a matrix. We write $|x| = (x^T x)^{1/2} \in \mathbb{R}^{d\times d}$. The map used is the natural generalization of F from Lemma 7.6 to the setting of matrices. For $z \in \mathbb{R}^{d\times d}$, the map F_z applies the (non-commutative) Mazur map from S_p into S_2, or, equivalently, ℓ_2, after re-centering by the matrix z, i.e.,

$$F_z(x) = M_{p,2}(x - z),$$

where $M_{p,q}(x) = x|x|^{p/q-1}$. Here we are overloading the notation for the Mazur map from the previous section. Note that applying the (non-commutative map) just defined to a diagonal matrix is equivalent to applying the (commutative) map from the previous section to the vector of diagonal entries.

We will use the following lemma of Ricard [Ric15], generalizing (19), for the specific case of mapping S_p into ℓ_2.

Lemma 8.4 (Lemma 2.6 in [Ric15])**.** *If $1 \leq p \leq 2$ and $x, y \in \mathbb{R}^{d\times d}$ are matrices, then:*

$$\|M_{p,2}(x) - M_{p,2}(y)\|_2 \leq O(1) \cdot \|x - y\|_{S_p}^{p/2},$$

and

$$\|M_{2,p}(x) - M_{2,p}(y)\|_{S_p} \leq O(1) \cdot \|x - y\|_2 \cdot \left(\|x\|_2^{2/p-1} + \|y\|_2^{2/p-1}\right),$$

and if $2 < p$ *and* $x, y \in \mathbb{R}^{d\times d}$ *are matrices, then:*

$$\|M_{p,2}(x) - M_{p,2}(y)\|_2 \leq O(p) \cdot \|x-y\|_{S_p} \left(\|x\|_{S_p}^{p/2-1} + \|y\|_{S_p}^{p/2-1} \right),$$

and

$$\|M_{2,p}(x) - M_{2,p}(y)\|_{S_p} \leq O(1) \cdot \|x-y\|_2^{2/p}.$$

For the remainder of the subsection, we consider any sequence of matrices $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^{d\times d})^m$, as well as any matrix $G \in \Delta(m)$. For simplicity, we again write $\rho(i) = \rho_G(i)$. We first prove a general statement on how the values of $\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p)$ and $\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)$ relate in terms of the value of z. The proof follows by adapting the argument of Matoušek [Mat97] given in Lemma 7.6 with the estimates from Ricard [Ric15] in Lemma 8.4. We state a somewhat more general version which we need in our algorithm.

Lemma 8.5. *For any* $z \in \mathbb{R}^{d\times d}$*, let* $\delta \in \mathbb{R}^{d\times d}$ *be the matrix given by* $\delta = \sum_{i=1}^m \rho(i) F_z(x_i)$*. Then, we have:*

$$\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} \left(1 - \frac{2^{p-1}\|\delta\|_2^p}{\left(\sum_{i=1}^m \rho(i)\|x_i - z\|_{S_p}^p\right)^{p/2}} \right) \leq (O(p))^p \cdot \mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p).$$

Proof. For simplicity in the notation, we let $\tilde{x} = x - z$. So $\delta \in \mathbb{R}^{d\times d}$ is the matrix given by

$$\delta = \sum_{j=1}^m \rho(j) M_{p,2}(\tilde{x_j}).$$

Note that $\|\tilde{x}\|_{S_p}^p = \|M_{p,2}(\tilde{x})\|_2^2$, so we write:

$$\begin{aligned}
\sum_{i=1}^m \rho(i)\|\tilde{x_i}\|_{S_p}^p &= \sum_{i=1}^m \rho(i) \left\|M_{p,2}(\tilde{x_i})\right\|_2^2 \\
&= \sum_{i=1}^m \rho(i) \left\|M_{p,2}(\tilde{x_i}) - \delta + \delta\right\|_2^2 \\
&\leq 2\sum_{i=1}^m \rho(i) \left\| M_{p,2}(\tilde{x_i}) - \sum_{j=1}^m \rho(j) M_{p,2}(\tilde{x_j}) \right\|_2^2 + 2\left\|\delta\right\|_2^2. \qquad (22)
\end{aligned}$$

We now focus on the first term of the right-hand side, where by the triangle inequality and

Cauchy-Schwarz, we may write:

$$\begin{aligned}\left\| M_{p,2}(\tilde{x_i}) - \sum_{j=1}^{m} \rho(j) M_{p,2}(\tilde{x_j}) \right\|_2^2 &= \left\| \sum_{j=1}^{m} \rho(j) \left(M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j}) \right) \right\|_2^2 \\ &\leq \left(\sum_{j=1}^{m} \rho(j) \left\| M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j}) \right\|_2 \right)^2 \\ &\leq \sum_{j=1}^{m} \rho(j) \left\| M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j}) \right\|_2^2 . \end{aligned} \tag{23}$$

Combining the right-hand side of (23) and (22), with the definition of $\mathsf{R}(F(\boldsymbol{x}), G, \|\cdot\|_2^2)$, we obtain:

$$\sum_{i=1}^{m} \rho(i) \|\tilde{x_i}\|_{S_p}^p \leq 2 \cdot \frac{1}{\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)} \sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \left\| M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j}) \right\|_2^2 + 2\|\delta\|_2^2. \tag{24}$$

Applying Lemma 8.4 for the case $p > 2$ to the first term on the right-hand side, as well as Hölder's inequality,

$$\begin{aligned}\sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \|M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j})\|_2^2 &\leq O(p^2) \sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \|\tilde{x_i} - \tilde{x_j}\|_{S_p}^2 \left(\|\tilde{x_i}\|_{S_p}^{p/2-1} + \|\tilde{x_j}\|_{S_p}^{p/2-1} \right)^2 \\ &\leq O(p^2) \left(\sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \|\tilde{x_i} - \tilde{x_j}\|_{S_p}^p \right)^{\frac{2}{p}} \\ &\quad \times \left(\sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \left(\|\tilde{x_i}\|_{S_p}^{p/2-1} + \|\tilde{x_j}\|_{S_p}^{p/2-1} \right)^{\frac{2p}{p-2}} \right)^{\frac{p-2}{p}} \\ &= O(p^2) \left(\sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \|x_i - x_j\|_{S_p}^p \right)^{\frac{2}{p}} \left(\sum_{i=1}^{m} \rho(i) \|\tilde{x_i}\|_{S_p}^p \right)^{\frac{p-2}{p}} . \end{aligned} \tag{25}$$

By combining (24) and (25), and dividing by $\left(\sum_{i=1}^{m} \rho(i) \|\tilde{x_i}\|_{S_p}^p \right)^{\frac{p-2}{p}}$, we have:

$$\left(\sum_{i=1}^{m} \rho(i) \|\tilde{x_i}\|_{S_p}^p \right)^{\frac{2}{p}} \leq \frac{O(p^2)}{\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)} \left(\sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij} \|x_i - x_j\|_{S_p}^p \right)^{\frac{2}{p}} + \frac{2\|\delta\|_2^2}{\left(\sum_{i=1}^{m} \rho(i) \|\tilde{x_i}\|_{S_p}^p \right)^{\frac{p-2}{p}}},$$

and, therefore, we have:

$$\sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p \leq \left(\frac{O(p^2)}{\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)} \left(\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|x_i - x_j\|_{S_p}^p \right)^{\frac{2}{p}} + \frac{2\|\delta\|_2^2}{\left(\sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p\right)^{\frac{p-2}{p}}} \right)^{\frac{p}{2}}$$
$$\leq \frac{(O(p))^p}{\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2}} \sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|x_i - x_j\|_{S_p}^p + \frac{2^{p-1}\cdot\|\delta\|_2^p}{\left(\sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p\right)^{p/2-1}}.$$

Rearranging the terms, we get

$$\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} \left(1 - \frac{2^{p-1}\cdot\|\delta\|_2^p}{\left(\sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p\right)^{p/2}} \right) \leq (O(p))^p \cdot \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\|x_i - x_j\|_{S_p}^p}{\sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p}.$$

Additionally, we have:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \rho(i)\rho(j)\|x_i - x_j\|_{S_p}^p = \sum_{i=1}^{m}\sum_{j=1}^{m} \rho(i)\rho(j)\|\tilde{x_i} - \tilde{x_j}\|_{S_p}^p \leq \sum_{i=1}^{m}\sum_{j=1}^{m} \rho(i)\rho(j)\left(\|\tilde{x_i}\|_{S_p} + \|\tilde{x_j}\|_{S_p}\right)^p$$
$$\leq 2^p \sum_{i=1}^{m} \rho(i)\|\tilde{x_i}\|_{S_p}^p. \tag{26}$$

Combining the inequalities and recalling the definition of $\mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p)$ gives the desired result. □

Given Lemma 8.5, we prove the following lemma which allows us to pick a particular matrix $z \in \mathbb{R}^{d\times d}$ whose "error term" is small. In the case of ℓ_p we could show this by an elementary application of the intermediate value theorem. However, in the case of S_p, our map is no longer applied coordinatewise, and the existence of of a good z is non-trivial.

Lemma 8.6. *There exists a matrix $z_0 \in \mathbb{R}^{d\times d}$ with $\|z_0\|_{S_p} \leq 2^{O(p)} \cdot \max_{i\in[m]} \|x_i\|_{S_p}$ such that*

$$\delta = \sum_{i=1}^{m} \rho(i) F_{z_0}(x_i) = 0 \in \mathbb{R}^{d\times d}.$$

Proof. Let $R = \max_{i\in[m]} \|x_i\|_{S_p}$. Define the following map $f\colon \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$:

$$f(z) = M_{2,p}\left(\sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) \right).$$

Claim 8.7. *For every z, one has:*

$$\|f(z) - z\|_{S_p} \leq O(1)\cdot R^{2/p} \cdot (\|z\|_{S_p} + R)^{1-2/p}.$$

Proof. We simply follow the computation, applying Lemma 8.4 and the triangle inequality as

necessary.

$$\begin{aligned}
\|f(z)-z\|_{S_p}^{p/2} &= \left\| M_{2,p}\left(\sum_{i=1}^{m}\rho(i)M_{p,2}(z-x_i)\right) - M_{2,p}(M_{p,2}(z))\right\|_{S_p}^{p/2} \\
&\le 2^{O(p)}\cdot\left\|\sum_{i=1}^{m}\rho(i)M_{p,2}(z-x_i) - M_{p,2}(z)\right\|_2 \\
&= 2^{O(p)}\cdot\left\|\sum_{i=1}^{m}\rho(i)(M_{p,2}(z-x_i) - M_{p,2}(z))\right\|_2 \\
&\le 2^{O(p)}\cdot\sum_{i=1}^{m}\rho(i)\cdot\|M_{p,2}(z-x_i) - M_{p,2}(z)\|_2 \\
&\le 2^{O(p)}\cdot\sum_{i=1}^{m}\rho(i)\cdot\|x_i\|_{S_p}\cdot\left(\|z-x_i\|_{S_p}^{p/2-1} + \|z\|_{S_p}^{p/2-1}\right) \\
&\le 2^{O(p)}\cdot\sum_{i=1}^{m}\rho(i)\cdot\|x_i\|_{S_p}\cdot\left(\|z\|_{S_p} + \|x_i\|_{S_p}\right)^{p/2-1} \\
&\le 2^{O(p)}\cdot R\cdot\left(\|z\|_{S_p} + R\right)^{p/2-1},
\end{aligned}$$

where the second and the fifth step are due to Lemma 8.4. Thus,

$$\|f(z)-z\|_{S_p} \le O(1)\cdot R^{2/p}\cdot\left(\|z\|_{S_p} + R\right)^{1-2/p}.$$

□

Claim 8.8. *Let r be a positive integer and $h\colon \mathbb{R}^r \to \mathbb{R}^r$ be a continuous map such that for some norm $\|\cdot\|$ on $\mathbb{R}^r$ one has:*

$$\|h(w) - w\| = o(\|w\|)$$

as $w \to \infty$. Then h is surjective.

Proof. Let $x_0 \in S^r$ be an arbitrary point. Let $\tau\colon \mathbb{R}^r \to S^r \setminus \{x_0\}$ be a homeomorphism. Consider $\widetilde{h} : S^r \to S^r$ defined as follows: $\widetilde{h}(x_0) = x_0$, and $\widetilde{h} = \tau \circ h \circ \tau^{-1}$ on $S^r \setminus \{x_0\}$. Since $h(w) \to \infty$ as $w \to \infty$, the function $\widetilde{h}$ is continuous. Let us show that $\widetilde{h}$ is homotopic to the identity via:

$$\widetilde{F}(t,x) = tx + (1-t)\widetilde{h}(x).$$

It is enough to check that $\widetilde{F}$ is continuous in (t, x_0) for every $0 \le t \le 1$. For this it is sufficient to check that

$$F(t,w) = tw + (1-t)h(w)$$

converges to ∞ as $x \to \infty$ uniformly in $0 \le t \le 1$. We have:

$$\|F(t,w)\| = \|w + (1-t)(h(w)-w))\| \ge \|w\| - \|h(w)-w\| \ge (1-o(1))\|w\|.$$

Thus, $\widetilde{h}$ is homotopic to the identity. This implies that $\widetilde{h}$ is surjective, since any continuous map $S^r \to S^r$ homotopic to the identity is surjective (see, e.g., Section 2.2 of [Hat02] for the proof). Since $\widetilde{h}(x_0) = x_0$, h is surjective as well. □

Combining Claims 8.7 and 8.8, we conclude that f is surjective. Since $M_{2,p}(t) = 0$ iff $t = 0$, we get the existence of z_0 such that

$$\sum_{i=1}^{m} \rho(i) M_{p,2}(x_i - z_0) = 0.$$

From Claim 8.7 it follows that $\|z_0\|_{S_p} \leq 2^{O(p)} \cdot R$, as required. □

We may now state a corollary which will be used in the algorithm, which simply follows from Lemma 8.5 and Lemma 8.6.

Corollary 8.9. *There exists a matrix $z_0 \in \mathbb{R}^{d\times d}$ with $\|z_0\|_{S_p} \leq 2^{O(p)} \max_{i\in[m]} \|x_i\|_{S_p}$ such that:*

$$\mathsf{R}(F_{z_0}(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} \leq (O(p))^p \cdot \mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p).$$

Let $\mathfrak{H}$ be the collection of sets of the type $z + M_{2,p}(\{x \in \mathbb{R}^{d\times d} : x_{ij} \geq t\})$ or $z + M_{2,p}(\{x \in \mathbb{R}^{d\times d} : x_{ij} \leq t\})$ for $z \in \mathbb{R}^{d\times d}$, some $i, j \in [d]$, and $t \in \mathbb{R}$. The bound $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(p/\varepsilon)$ now follows from Corollary 8.9 in the same fashion as in the proof of Lemma 7.4.

Finally, we show that Lemma 8.6 is somewhat robust, and allows for small deviations from z_0 when there are no dense balls in $\boldsymbol{x}$.

Lemma 8.10. *Suppose $\boldsymbol{x}$ does not contain a $\frac{1}{2}$-dense ball of radius $10p$ with respect to G, and all points x_i have $\|x_i\|_{S_p} \leq O(d)$. Let $z_0 \in \mathbb{R}^{d\times d}$ with $\|z_0\|_{S_p} \leq 2^{O(p)} \cdot d$ be any matrix with*

$$\delta_0 = \sum_{i=1}^{m} \rho(i) F_{z_0}(x_i) \qquad \textit{satisfying} \qquad \|\delta_0\|_2 \leq 1,$$

and $z \in \mathbb{R}^{d\times d}$ be any matrix with $\|z - z_0\|_2 \leq \frac{1}{d^{O(p)}}$. Then we have:

$$\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)^{p/2} \leq (O(p))^p \cdot \mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p).$$

Proof. If there is no $\frac{1}{2}$-dense ball of radius $10p$ in $\boldsymbol{x}$ with respect to G, then

$$\sum_{i=1}^{m} \rho(i) \|x_i - z\|_{S_p}^p \geq \frac{(10p)^p}{2}. \tag{27}$$

Let $\delta = \sum_{i=1}^m \rho(i) F_z(x_i)$, then we have:

$$\begin{aligned}
\|\delta - \delta_0\|_2 &\leq \sum_{i=1}^m \rho(i) \left\| M_{p,2}(x_i - z) - M_{p,2}(x_i - z_0) \right\|_2 \\
&= \sum_{i=1}^m \rho(i) \left\| M_{p,2}(x_i - z_0 + (z_0 - z)) - M_{p,2}(x_i - z_0) \right\|_2 \\
&\leq O(p) \|z_0 - z\|_{S_p} \left(\|x_i - z\|_{S_p}^{p/2-1} + \|x_i - z_0\|_{S_p}^{p/2-1} \right) \\
&\leq O(p) \cdot \|z_0 - z\|_{S_p} \left(O(d)\right)^p \leq 1,
\end{aligned}$$

and therefore, $\|\delta\|_2 \leq 2$. Combining this fact, along with (27) and Lemma 8.5 gives the desired inequality. □

The following lemma follows in a similar fashion to Lemma 7.8 using Lemma 8.10.

Lemma 8.11. *Let $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^{d\times d})^m$ be any set of m points where $g_{ij} > 0$ only if $\|x_i - x_j\|_{S_p} \leq 1$, and the conditions of $z_0 \in \mathbb{R}^{d\times d}$, $\delta_0 \in \mathbb{R}^{d\times d}$ and $z \in \mathbb{R}^{d\times d}$ in Lemma 8.10 are satisfied. Then,*

- *either there exists a S_p-ball B of radius $R = O(p/\varepsilon)$ such that $\sum_{x_j \in B_p \cap \boldsymbol{x}} \rho(j) \geq \frac{1}{2}$, or*
- $\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2) \leq \varepsilon^2$.

8.2 Proof of Lemma 8.3

Let $\mathcal{Z}$ be the set of matrices $z \in \mathbb{R}^{d\times d}$ with $\|z\|_{S_p} \leq 2^{O(p)} \cdot d$ and each entry of z rounded to precision $\frac{1}{d^{O(p)}}$. In particular, we have that for every matrix $z_0 \in \mathbb{R}^{d\times d}$ with $\|z_0\|_{S_p} \leq 2^{O(p)} \cdot d$ there exists some matrix $z \in \mathcal{Z}$ with $\|z - z_0\|_{S_p} \leq \frac{1}{d^{O(p)}}$. Note that $|\mathcal{Z}| \leq \exp(\text{poly}(d^p))$. Similarly to Subsection 7.2, we let $\mathcal{V} \subset \mathbb{R}$ be the set of values the entries of matrices $F_z(x)$ take when $z \in \mathcal{Z}$ and $x \in X$, and note $|\mathcal{V}| \leq \exp(\text{poly}(d^p))$.

Definition 8.12. *Let $\tau = \text{poly}(d^p)$. We consider the collection $\mathfrak{S}$ of subsets of X given by all sets*

$$S = \bigcup_{t=1}^{\tau} S^{(t)},$$

where for each $t \in [\tau]$, there is some $V \in (\mathcal{V} \times \mathcal{V})^{d\times d}$ such that

$$S^{(t)} = \{x \in X : F_z(x) \notin \text{Box}(V), z \in \mathcal{Z}\}.$$

Lemma 8.13. *The collection $\mathfrak{S}$ is b-succinct for $b \leq \text{poly}(d^p)$.*

Proof. The encoding $E(S)$ is given by encoding the τ values of $z^{(t)} \in \mathcal{Z}$ and $V^{(t)} \in (\mathcal{V} \times \mathcal{V})^{d\times d}$ defining the sets $S^{(t)}$. Then, the decoding algorithm $D(E(S), q)$ goes through each $t = 1, \ldots, \tau$, and checks if $q \in S^{(t)}$ by checking if $F_{z^{(t)}}(q) \in \text{Box}(V^{(t)})$ in time $\text{poly}(d^p)$. □

We have the following analogue of Lemma 7.5.

Lemma 8.14. *For the collection $\mathfrak{S}$ of subsets of X described above, we have that for any matrix $G \in \Delta(N)$ where $g_{ij} > 0$ only if $\|x_i - x_j\|_{S_p} \le 1$, either there exists a $\frac{1}{4}$-dense ball of radius $R = O(p/eps)$, or there exists a subset $S \in \mathfrak{S}$ where:*

$$\frac{1}{3} \le \rho_G(S) \le \frac{3}{4} \qquad \text{and} \qquad \sum_{i \in S, j \notin S} g_{ij} \le 2\varepsilon.$$

Proof. We will prove the lemma via an iterative procedure similar to the one in the proof of Lemma 7.5. We start with $\boldsymbol{x}$ containing all points in X, and use Lemma 8.6 in order to find some $z_0 \in \mathbb{R}^{d \times d}$ with $\|z_0\|_{S_p} \le 2^{O(p)} \cdot d$ with $\delta_0 = 0 \in \mathbb{R}^{d \times d}$ (where δ_0 is defined as in Lemma 8.10).

- If there exists a $\frac{1}{2}$-dense ball in $\boldsymbol{x}$ of radius $O(p/\varepsilon)$ with respect to G, we return this dense ball, and we are done.
- Otherwise, we let $z^{(1)} \in \mathcal{Z}$ be the matrix z_0 fixed to bounded precision, and by Lemma 8.11, $\mathsf{R}(F_{z^{(1)}}(\boldsymbol{x}), G, \|\cdot\|_2^2) \le \varepsilon^2$. Thus, by Cheeger's inequality, we may find a set $\widetilde{S} = \{x \in \overline{S} : F_{z^{(1)}}(x)_{ij} \le t\}$ for some $i, j \in [d]$ and $t \in \mathbb{R}$, so that $\Phi_G(\widetilde{S}) \le \sqrt{2}\varepsilon$. If $\rho_G(\widetilde{S}) > \frac{1}{2}$, we replace $\widetilde{S}$ with its complement. Then we let $S \leftarrow S \cup \widetilde{S}$. Note that by a similar argument to Lemma 7.5, as long as the map $F_{z^{(1)}}$ is fixed, the union of sets $\widetilde{S}$ is the complement of a box after transforming points by $F_{z^{(1)}}$.

Claim 8.15. *Suppose $\rho_G(S) \le \frac{1}{d^{O(p)}}$, let $\widetilde{G}$ be the normalized matrix restricted on rows and columns in $\overline{S}$, and $\delta_0 = \sum_{i \in \overline{S}} \rho_{\widetilde{G}}(i) F_{z_0}(x_i)$. Then, $\|\delta_0\|_2 \le 1$.*

Proof. We simply note that for all $i \in \overline{S}$, $\rho_{\widetilde{G}}(i) \le \frac{\rho_G(i)}{1 - 2\rho_G(S)}$, since $\tilde{G}$ is given by removing from G the rows and columns in S. Thus, we have:

$$\begin{aligned}
\|\delta_0\|_2 &= \left\| \sum_{i=1}^m \rho_G(i) F_{z_0}(x_i) - \sum_{i \in S} \rho_G(i) F_{z_0}(x_i) + \sum_{i \in \overline{S}} (\rho_{\widetilde{G}}(i) - \rho_G(i)) F_{z_0}(x_i) \right\|_2 \\
&\le \left\| \sum_{i=1}^m \rho_G(i) F_{z_0}(x_i) \right\|_2 + \sum_{i \in S} \rho_G(i) \|F_{z_0}(x_i)\|_2 + \sum_{i \in \overline{S}} (\rho_{\widetilde{G}}(i) - \rho_G(i)) \|F_{z_0}(x_i)\|_2 \\
&\le \rho_G(S) \cdot d^{O(p)} + 2\rho_G(S) \rho_{\widetilde{G}}(\overline{S}) \cdot d^{O(p)} \le 1,
\end{aligned}$$

since $\sum_{i=1}^m \rho_G(i) F_{z_0}(x_i) = 0$ by Lemma 8.6, $\rho_G(S) \le \frac{1}{d^{O(p)}}$, and $\|F_{z_0}(x_i)\|_2 \le d^{O(p)}$ from Lemma 8.4, as well as the fact that $\|x_i\|_{S_p} \le O(d)$ and $\|z_0\|_{S_p} \le 2^{O(p)} \cdot d$. □

Thus, whenever $\rho_G(S) \ge \frac{1}{d^{O(p)}}$ for the first time, we let $S^{(1)} \leftarrow S$, and we recompute $z_0 \in \mathbb{R}^{d \times d}$ from Lemma 8.6 with $\boldsymbol{x}$ as the points in $X \setminus S^{(1)}$. We repeat this procedure for $S^{(2)}, \dots, S^{(\tau)}$, where each implication of Cheeger's inequality corresponds to $\boldsymbol{x} \in X^{|\overline{S}|}$ containing the points remaining in $\overline{S}$ with the Rayleigh quotient $\mathsf{R}(F_{z^{(t)}}(\boldsymbol{x}), \widetilde{G}, \|\cdot\|_2^2)$; thus, the sets in $S^{(t)}$ are the complements

of boxes after applying the map $F_{z^{(t)}}$ to all points. Since each $S^{(t)}$ has $\rho_G(S^{(t)}) \geq \frac{1}{d^{O(p)}}$ and once $\rho_G(S) > \frac{1}{3}$ we stop, $\tau \leq d^{O(p)}$. □

Lemma 7.5 implies the following space partitioning result by an argument analogous to the proof of Theorem 3.6.

Theorem 8.16. *Let $0 < \varepsilon < 1$, $2 < p < \infty$ and $R > 0$. Consider any dataset $P \subset \mathbb{R}^d$ of n $d \times d$ matrices lying in $B_{S_p}(0, R) = \{x \in \mathbb{R}^{d\times d} \mid \|x\|_{S_p} \leq R\}$. Either there is an S_p-ball of radius $O(p/\varepsilon)$ containing $\Omega(n)$ points from P, or there exists a distribution $\mathcal{D}$ over sets $S \subseteq \mathbb{R}^{d\times d}$ such that:*

1. *For every $u, v \in B_{S_p}(0, R)$ with $\|u - v\|_{S_p} \leq 1$, a random set $S \sim \mathcal{D}$ separates u and v with probability at most ε.*
2. *For every set S from the support of $\mathcal{D}$, the number of points in P lying in S is between $\Omega(n)$ and $(1 - \Omega(1)) \cdot n$.*
3. *Every set S in the support of $\mathcal{D}$ is the union of $\mathrm{poly}(d^p)$ sets of the type $\{x \in \mathbb{R}^{d\times d} : F_z(x) \notin B\}$, where $\|z\|_{S_p} = 2^{O(p)}d$ and B is a box in $\mathbb{R}^{d^2}$.*

8.3 The case of $1 \leq p \leq 2$

Lemma 8.17. *Let $1 \leq p \leq 2$. For any $z \in \mathbb{R}^{d\times d}$, let $\delta \in \mathbb{R}^{d\times d}$ be the matrix given by $\delta = \sum_{i=1}^m \rho(i) F_z(x_i)$. Then, we have:*

$$\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)\left(1 - \frac{2\|\delta\|_2^2}{\sum_{i=1}^m \rho_G \|x_i - z\|_p^p}\right) \leq O(1) \cdot \mathsf{R}(\boldsymbol{x}, G, \|\cdot\|_{S_p}^p).$$

Proof. This proof is very similar to the proof of Lemma 8.5. We simply note that we may use Lemma 8.4 for the case $1 \leq p \leq 2$. In particular, this means that up to (24), both proofs follow the same inequalities. Using Lemma 8.4, we conclude:

$$\begin{aligned}\sum_{i=1}^m \rho(i)\|\tilde{x_i}\|_{S_p}^p &\leq \frac{2}{\cdot \mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)} \sum_{i=1}^m \sum_{j=1}^m g_{ij} \left\|M_{p,2}(\tilde{x_i}) - M_{p,2}(\tilde{x_j})\right\|_2^2 + 2\|\delta\|_2^2. \\ &\leq \frac{O(1)}{\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2)} \sum_{i=1}^m \sum_{j=1}^m g_{ij} \left\|x_i - x_j\right\|_{S_p}^p + 2\|\delta\|_2^2,\end{aligned}$$

which by (26), gives the desired inequality after rearranging the terms. □

Lemma 8.18. *There exists a matrix $z_0 \in \mathbb{R}^{d\times d}$ with $\|z_0\|_{S_p} \leq O(1) \cdot \max_{i\in[m]} \|x_i\|_{S_p}$, such that*

$$\delta = \sum_{i=1}^m \rho(i) F_{z_0}(x_i) = 0 \in \mathbb{R}^{d\times d}.$$

Proof. Similarly to the proof of Lemma 8.6, let $f\colon \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ be the map given by:

$$f(z) = M_{2,p}\left(\sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i)\right).$$

We will similarly prove that this map is surjective to conclude that there exists some z_0 with $\sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) = 0$. We have:

$$\begin{aligned}
\|f(z) - z\|_{S_p} &= \left\| M_{2,p}\left(\sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i)\right) - M_{2,p}(M_{p,2}(z)) \right\|_{S_p} \\
&\le O(1) \left\| \sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) - M_{p,2}(z) \right\|_2 && (28) \\
&\quad \times \left(\left\| \sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) \right\|_2^{2/p-1} + \|M_{p,2}(z)\|_2^{2/p-1} \right), && (29)
\end{aligned}$$

where we used Lemma 8.4. We first bound the term in (28).

$$\begin{aligned}
\left\| \sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) - M_{p,2}(z) \right\|_2 &\le \sum_{i=1}^{m} \rho(i) \left\| M_{p,2}(z - x_i) - M_{p,2}(z) \right\|_2 && (30) \\
&\le O(1) \sum_{i=1}^{m} \rho(i) \|x_i\|_{S_p}^{p/2}, && (31)
\end{aligned}$$

where we used the triangle inequality in (30) and Lemma 8.4 in (31). We now bound both summands in (29). The first term of (29) has:

$$\begin{aligned}
\left\| \sum_{i=1}^{m} \rho(i) M_{p,2}(z - x_i) \right\|_2^{2/p-1} &\le \left(\sum_{i=1}^{m} \rho(i) \|M_{p,2}(z - x_i)\|_2 \right)^{2/p-1} && (32) \\
&\le \left(O(1) \sum_{i=1}^{m} \rho(i) \|z - x_i\|_{S_p}^{p/2} \right)^{2/p-1} && (33) \\
&\le O(1) \left(\left(\sum_{i=1}^{m} \rho(i) \|z - x_i\|_{S_p} \right)^{p/2} \right)^{2/p-1} && (34) \\
&\le O(1) \left(\|z\|_{S_p}^{1-p/2} + R^{1-p/2} \right). && (35)
\end{aligned}$$

In (32), we used the triangle inequality, followed by Lemma 8.4 in (33). Then we used the fact that $\frac{1}{2} \le \frac{p}{2} \le 1$ and concavity of $t^{p/2}$ in (34). Finally, (35) follows from the triangle inequality. The second term of (29) has:

$$\begin{aligned}
\|M_{p,2}(z)\|_2^{2/p-1} &\le \left(O(1) \cdot \|z\|_{S_p}^{p/2} \right)^{2/p-1} \\
&\le O(1) \cdot \|z\|_{S_p}^{1-2/p}, && (36)
\end{aligned}$$

by Lemma 8.4. Putting (28) and (29) together with (31), (35), and (36), we obtain:

$$\|f(z) - z\|_{S_p} \leq O(1) \cdot R + O(1) \cdot R^{p/2} \cdot \|z\|_{S_p}^{1-p/2}.$$

Thus, we may similarly apply Claim 8.8 to conclude the lemma. □

Note that there is a slight difference in the dependence on the powers of the Rayleigh quotients. Thus, we derive the following lemma, which shows that $\Xi_{\mathfrak{H}}(X, \varepsilon) = O(1/\varepsilon^{2/p})$.

Lemma 8.19. *Suppose $\boldsymbol{x} = (x_1, \ldots, x_m) \in (\mathbb{R}^d)^m$ be any set of m points where $g_{ij} > 0$ only if $\|x_i - x_j\|_p \leq 1$, and the conditions of $z_0 \in \mathbb{R}^{d\times d}$, $\delta_0 \in \mathbb{R}^{d\times d}$ and $z \in \mathbb{R}^{d\times d}$ in Lemma 8.10 are satisfied. Then,*

- *either there exists a S_p-ball B of radius $R = O(1/\varepsilon^{1/p})$ such that $\sum_{x_j \in B_p \cap \boldsymbol{x}} \rho(j) \geq \frac{1}{2}$, or*
- *$\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2) \leq \varepsilon$.*

Proof. Suppose $\sum_{i=1}^m \sum_{j=1}^m \rho(i)\rho(j)\|x_i - x_j\|_p^p \geq C/\varepsilon$ for a high enough constant, then $\mathsf{R}(F_z(\boldsymbol{x}), G, \|\cdot\|_2^2) \leq \varepsilon$; so assume otherwise. In the same way as in the proof of Lemma 7.8 and Lemma 8.10, we may conclude there exists a $\frac{1}{2}$-dense ball with respect to G of radius $O(1/\varepsilon^{1/p})$. □

Thus, using the same partitioning procedure as in Lemma 8.3, we obtain the following theorem.

Theorem 8.20. *Fix some $0 < \alpha < 1$ and $p \in [1, 2]$. There exists a data structure solving c-ANN for S_p with success probability $\frac{9}{10}$ with the following guarantees:*

- *the approximation $c = O(1/\alpha^{2/p})$.*
- *the query time of the data structure is $\mathrm{poly}(d) \cdot n^{\alpha}$, and*
- *the space of the data structure is $\mathrm{poly}(d) \cdot n^{1+\alpha}$.*

9 Lower bounds

9.1 General norms do not admit succinct collections

The goal of this section is to rule out algorithms for general norms which proceed by a generalization of the simple algorithm from Section 7. In particular, statements of the form of Lemma 7.5 cannot be true for general norms unless, the collection $\mathfrak{S}$ has high description complexity, or the collection $\mathfrak{S}$ depends on the norm.

Fix a dimension d to at least a large enough constant. We let $X \subset S^{d-1}$ be a set of points with the following properties:

- The set $|X| = N = 2^{d^{0.1}}$, and
- For every $x_1, x_2 \in X$, we have $|\langle x_1, x_2\rangle| \leq \frac{1}{d^{1/4}}$.

The set exists by Lemma 6.2 of [ANN+17]. Our lower bound is captured by the following theorem.

Theorem 9.1 (No succinct collections exist)**.** *Let $b = 2^{d^{o(1)}}$, $R = \frac{d^{1/4}}{10}$, and some $\gamma \geq 2^{-d^{0.01}}$. For any collection $\mathfrak{S}$ of subsets of $X = \{x_1, \ldots, x_N\}$ of size $|\mathfrak{S}| \leq 2^b$, there exists a normed space $(\mathbb{R}^d, \|\cdot\|)$ and a matrix $G \in \Delta(N)$ where $g_{ij} > 0$ only if $\|x_i - x_j\| \leq 1$, satisfying the following:*

- *there is no γ-dense ball of radius R with respect to G;*
- *for any $S \in \mathfrak{S}$, if $\delta = \min\{\rho_G(S), \rho_G([N] \setminus S)\} \geq 2^{-d^{0.01}}$, we have that $\sum_{i \in S, j \notin S} g_{ij} = \Omega(\delta\gamma)$.*

Proof. We choose the norm $(\mathbb{R}^d, \|\cdot\|)$ randomly and prove that any fixed set $S \in \mathfrak{S}$ works only with extremely low probability. By a union bound, this implies there exists a norm satisfying the conditions of the theorem.

Consider constructing the norm $(\mathbb{R}^d, \|\cdot\|)$ as follows: we pick a random subset $C \subset [N]$ of size $\gamma N - 1$. For any $y \in \mathbb{R}^d$, we let:

$$\|y\| = \frac{d^{1/4}}{2} \cdot \max_{i \notin C} |\langle x_i, y \rangle|.$$

For any $i, j \in C$, $\|x_i - x_j\| \leq \|x_i\| + \|x_j\| \leq 1$. If $i \notin C$ and $j \in X$ with $i \neq j$, $\|x_i - x_j\| \geq \frac{d^{1/4}}{2}(1 - \frac{1}{d^{1/4}})$. We let $G \in \Delta(N)$ be the matrix given by:

$$g_{ij} = \begin{cases} \frac{1}{N} \cdot \frac{1}{|C|} & i, j \in C \\ \frac{1}{N} & i = j \notin C \\ 0 & \text{o.w} \end{cases}.$$

We note that there are no γ-dense balls of radius R with respect to G. This is because a ball of radius R with more than 1 point must contain no points from $[N] \setminus C$, and $\rho_G(C) \leq \frac{1}{N}(\gamma N - 1) < \gamma$. Now, consider any $S \in \mathfrak{S}$, and let $\delta = \rho_G(S) \leq \frac{1}{2}$ (otherwise, we consider $X \setminus S$). By Chernoff bound, we have $\frac{\gamma\delta}{2} \cdot N \leq |S \cap C| \leq \frac{3\gamma\delta}{2} \cdot N$ with probability at least $1 - e^{-\Omega(\gamma\delta N)}$. In that case, whenever $i \in S \cap C$ and $j \in C \setminus S$, the edge (i, j) is cut with $g_{ij} = \frac{1}{N} \cdot \frac{1}{|C|}$. Thus,

$$\sum_{i \in S, j \notin S} g_{ij} \geq |S \cap C| \cdot (|C| - |S \cap C|) \cdot \frac{1}{N} \cdot \frac{1}{|C|} \geq \frac{\gamma\delta}{2} \cdot N \left(\gamma N - 1 - \frac{3\delta\gamma}{2} \cdot N\right) \frac{1}{\gamma N^2} \geq \Omega(\delta\gamma).$$

We can union bound over all $S \in \mathfrak{S}$ to conclude that for all $S \in \mathfrak{S}$, $\sum_{i \in S, j \notin S} g_{ij} = \Omega(\gamma\delta)$ since $b \ll \gamma\delta N$ when $\gamma, \delta \geq 2^{-d^{0.01}}$. □

In order to interpret Theorem 9.1, we compare it to Lemma 7.5. Lemma 7.5 claims that for any set of points X and $G \in \Delta(|X|)$ (with $g_{ij} > 0 \Rightarrow \|x_i - x_j\|_p \leq 1$), there is a $\Omega(1)$-dense ball of radius $O(\frac{p}{\varepsilon})$, or there exists a balanced set from a collection of $2^{O(d \log d)}$ sets which cuts an ε-fraction of edges with respect to G. In the notation of Theorem 9.1, this corresponds to setting γ and δ to constants; in this case, if there are no $\Omega(1)$-dense balls of radius R, $\Omega(1)$-fraction of edges are cut with respect to G when $R = \Omega(d^{1/4})$. The notable fact is that we cannot tradeoff R and the fraction

of edges cut, as we do in Lemma 7.5. At a high level, this rules out $\log^{O(1)} d$-ANN for any norm by balanced sets or $\Omega(1)$-dense balls from a fixed succinct collections. These include hyperplane cuts, coordinate cuts, and box cuts.

Theorem 9.1 does not apply to partitions which depend on the norm. For example, Theorem 9.1 does not rule out the collection of balls in the norm, i.e., $\mathfrak{S} = \{B_X(x, r) : x \in X, r \in [\text{poly}(d)]\}$. It also does not rule out succinct cuts after applying a norm-dependent transformation, which occurs in Lemma 8.3.

9.2 Lower bound for random partitions

Let $G = (V, E)$ be a degree-m spectral expander with $|V| = N$ vertices, and let (V, d_G) be the metric space where $i, j \in V$, $d_G(i, j)$ is the length of the shortest path between i and j. Consider the distributions $\mathcal{D}$ supported on datasets, and for a dataset P consider the distribution $\mathcal{Q}(P)$ supported on queries, where:

- $P \sim \mathcal{D}$ where P is an n-point dataset, and $q \sim \mathcal{Q}(P)$ is a query.
- $P = \{p_1, \ldots, p_n\} \sim \mathcal{D}$ has $p_i \sim V$ uniformly for each $i \in [n]$,
- $q \sim \mathcal{Q}(P)$ is sampled by first picking $i \sim [n]$ and choosing q to be a neighbor of p_i uniformly at random. Let $p^* = p_i \in P$ denote the near-neighbor of q.

Definition 9.2. *We say the dataset $P \subset V$ is c-separated in (V, d_G) if*

$$\min_{\substack{p_1, p_2 \in P \\ p_1 \neq p_2}} d_G(p_1, p_2) \geq c.$$

Lemma 9.3. *Consider a small constant $0 < \gamma_1 \leq \frac{1}{2 \log m}$, and let $N \geq n^5$ and $c \leq \gamma_1 \log N$, then when $P \sim \mathcal{D}$, P is c-separated with probability at least 0.99.*

Proof. For each point $v \in V$, the set $B_G(v, c) = \{x \in V : d_G(v, x) \leq c\}$ contains $|B_G(v, c)| \leq m^c \leq N^{\gamma_1 \log m}$ points in V. Since $n^2 \ll N^{1 - \gamma_1 \log m}$, with high probability over $(P, q) \sim \mathcal{D}$, P is c-separated. □

Lemma 9.4. *Let $\delta \geq n^{-o(1)}$ and $\gamma_2 > 0$ be any fixed constant. Let $\mathcal{S}$ be a collection of $2^{n^{1-\gamma_2}}$ partitions of V. With probability 0.99 over $P \sim \mathcal{D}$, every $S \in \mathcal{S}$ satisfies the following:*

- *Setting $p_2(S) = \Pr_{u,v \sim V}[S(u) = S(v)]$, we have:*

$$\Pr_{i,j \sim [n]}[S(p_i) = S(p_j)] \geq p_2(S) - \delta + \frac{1}{n},$$

- *and $\Pr_{q \sim \mathcal{Q}(P)}[S(p^*) = S(q)] \leq 1 - \lambda_2(\mathcal{L}_G)\,(1 - p_2(S)) + \delta$,*

Proof. Consider a fixed $S \in \mathcal{S}$ which partitions V into $t \geq 2$ parts, $S_1, \ldots, S_t \subset V$, where $\alpha_k = \frac{|S_k|}{N}$ for $k \in [t]$. Note that

$$p_2(S) = \mathop{\mathbb{E}}_{P}\left[\Pr_{\substack{i,j\sim[n]\\ i\neq j}}[S(p_i) = S(p_j)]\right] = \sum_{k=1}^{t} \alpha_k^2.$$

For any fixed dataset P, changing the value of one point p_j changes $\Pr_{\substack{i,j\sim[n]\\ i\neq j}}[S(p_i) = S(p_j)]$ by at most $\frac{2}{n}$. Therefore, letting

$$p_2(S, P) = \Pr_{\substack{i,j\sim[n]\\ i\neq j}}[S(p_i) = S(p_j)],$$

we obtain by McDiarmid's inequality that:

$$\Pr_{P_i}\left[|p_2(S, P) - p_2(S)| \geq \delta\right] \leq 2e^{-\frac{\delta^2 n}{2}}.$$

Note that $\Pr_{i,j\sim[n]}[S(p_i) = S(p_j)] = p_2(S, P) + \frac{1}{n}$, since with probability $1/n$, $i = j$. Additionally, since the sparsity of a cut is lowerbounded by $\lambda_2(\mathcal{L}_G)$,

$$\begin{aligned}
\Pr_{(u,v)\sim E}[S(u) \neq S(v)] = \frac{2|\partial S|}{Nm} &= \frac{\sum_{k=1}^{t} |\partial S_k|}{Nm} \\
&\geq \frac{1}{Nm}\lambda_2(\mathcal{L}_G) \sum_{k=1}^{t} \left(\frac{(\alpha_i Nm)(Nm - \alpha_i Nm)}{Nm}\right) \\
&\geq \lambda_2(\mathcal{L}_G)\left(1 - p_2(S)\right).
\end{aligned}$$

For $i \in [n]$ and $v \in V$, let $X_{i,v}$ be the indicator random variable that $p_i = v$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{P\sim\mathcal{Q}}\left[\Pr_{q\sim\mathcal{Q}(P)}[S(p^*) \neq S(q)]\right] &= \frac{1}{n \cdot m} \sum_{(u,v)\in\partial S} \sum_{i=1}^{n} \left(\Pr[X_{i,u} = 1] + \Pr[X_{i,v} = 1]\right) \\
&\geq \lambda_2(\mathcal{L}_G)\left(1 - p_2(S)\right).
\end{aligned}$$

Finally, letting

$$p^*(S, P) = \Pr_{q\sim\mathcal{Q}(P)}[S(p^*) \neq S(q)],$$

since changing a dataset point changes $p^*(S, P)$ by at most $\frac{1}{n}$, we apply McDiarmid's inequality again to obtain:

$$\Pr_{P\sim\mathcal{D}}\left[|p^*(S, P) - \lambda_2(\mathcal{L}_G)\left(1 - p_2(S)\right)| \geq \delta\right] \leq 2e^{-2\delta^2 n}.$$

By the setting of δ and γ_2, we may union bound over all $S \in \mathcal{S}$. □

Theorem 9.5. *There exists a positive $\varepsilon > 0$ such that the following holds. For every positive integers d, k there exists a distribution $\mathcal{D}$ over n-point datasets in ℓ_∞^d, where $n = d^k$, such that the following holds for $c = \Omega(\log d)$.*

- *With probability at least 0.99, a dataset sampled according to $\mathcal{D}$ is pairwise $(c+1)$-separated;*
- *Let $\mathcal{R}$ to be a collection of partitions of $\mathbb{R}^d$ of size $|\mathcal{R}| \le 2^{n^{1-\Omega(1)}}$. Then, with probability at least 0.99 over the dataset P sampled according to $\mathcal{D}$, there is no random partition $\mathcal{R}_P$ supported on the partitions from $\mathcal{R}$ that has the following two properties:*
  - *For every $q \in \mathbb{R}^d$, for which there exists $p \in P$ with $\|p - q\|_\infty \le 1$, one has*

$$\Pr[\textit{a partition sampled from } \mathcal{R}_P \textit{ separates } p \textit{ and } q] \le \varepsilon;$$

  - *For every $q \in \mathbb{R}^d$, one has:*

$$\Pr[\textit{a partition sampled from } \mathcal{R}_P \textit{ separates } p \textit{ and } q, \textit{ where } p \in P \textit{ uniformly}] \ge 1/10.$$

We take a 3-regular expander G with $N = n^{100} = d^{100k}$ vertices and embed it into ℓ_∞^d with distortion $O(k)$ using the result from [Mat97]. We assume that all the edges have length at most 1, and the distances are contracted by a factor at most $O(k)$. Then a dataset is obtained by sampling n independent images of the vertices of $G = (V, E)$.

The diameter of G is $\Theta(\log N) = \Theta(k \log d)$. If we sample n vertices, then with high probability they will be $\Omega(k \log d)$-separated. After the embedding, the corresponding points are $(c+1)$-separated, since the distortion is $O(k)$. Consider the following distribution of the queries: we choose uniformly one of the n datapoints, and choose a random adjacent vertex of G. Clearly, the marginal distribution of the queries is uniform.

Consider a fixed partition $\mathcal{P} \in \mathcal{R}$ of $\mathbb{R}^d$, which induces a partition of the image of G under the embedding. Define $p_2^*(\mathcal{P}) = \Pr_{x,y\sim V}[\mathcal{P}(x) = \mathcal{P}(y)]$, and $p_1^*(\mathcal{P}) = \Pr_{x\,\sim\, V,\ y\,\sim\, \text{neighbor of } x}[\mathcal{P}(x) = \mathcal{P}(y)]$. Then, using that G is an expander, we can show that $p_1 \le 1 - \Omega(1 - \sqrt{p_2})$.

Now consider a subsampled dataset $P \sim \mathcal{D}$. Define $p_1(\mathcal{P}, P)$ and $p_2(\mathcal{P}, P)$ similar to $p_1^*(\mathcal{P})$ and $p_2^*(\mathcal{P})$, but we sample vertices in P, not in the whole V.

By the McDiarmid inequality, $p_2(\mathcal{P}, P)$ is within additive 0.001 from $p_2^*(\mathcal{P})$ with probability at least $1 - 2^{-\Omega(n)}$. At the same time, $1 - p_1(\mathcal{P}, P)$ is within a factor of three from $1 - p_1^*(\mathcal{P})$ with probability at least $1 - 2^{-\Omega(n)}$. Combining with the above estimate for p_1^* and p_2^*, we have that with probability at least $1 - 2^{-\Omega(n)}$, we cannot have both $p_2(\mathcal{P}, P) \le 9/10$ and $p_1(\mathcal{P}, P) \ge 1 - \varepsilon$ for a sufficiently small $\varepsilon > 0$.

Taking union bound over the whole collection $\mathcal{R}$, we get that with high probability for all the partitions we have the above statement. Thus, by Yao's min-max principle, we get the required statement.

To get a lower bound $\Omega(p)$ for the ℓ_p space, we do exactly the same as above, but we embed an expander with distortion $O\left(\frac{\log N}{p}\right)$, and thus the diameter we are getting is $\Omega(p)$.

10 Acknowledgments

We thank Richard Peng and Tselil Schramm for useful discussions.

The first-named author was supported in part by the Simons Foundation (#491119), NSF grants CCF-1617955, CCF-1740833, and Google Research Award. The second-named author was supported in part by the NSF grant CCF-1412958, the Packard Foundation and the Simons Foundation. The third-named author was supported in part by the NSF grant CCF-1740425 and NSERC Discovery Grant. The fourth-named author was supported in part by the Simons Junior Fellowship. The fifth-named author was supported in part by the NSF grants CCF-1563155, CCF-1420349, CCF-1149257, CCF-1423100), and the NSF Graduate Research Fellowship (DGE-16-44869).

The work was done in part while the third-named author was visiting Simons Institute for the Theory of Computing and while the fourth-named author was a graduate student at MIT and a postdoc at Columbia University. This work was was carried out under the auspices of the Simons Algorithms and Geometry (A&G) Think Tank.

References

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing*, 8(1):121–164, 2012.

[AI06] Alexandr Andoni and Piotr Indyk. Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS '2006)*, pages 459–468, 2006.

[AI17] Alexandr Andoni and Piotr Indyk. Nearest Neighbors in High-Dimensional Spaces. In Jacob E. Goodman, Joseph O'Rourke, and Csaba D. Tóth, editors, *Handbook of Discrete and Computational Geometry*, pages 1133–1153. CRC Press LLC, 2017.

[AIK09] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Overcoming the ℓ_1 Non-Embeddability Barrier: Algorithms for Product Metrics. In *Proceedings of the 20th ACM-SIAM Symposium on Discrete Algorithms (SODA '2009)*, pages 865–874, 2009.

[AINR14] Alexandr Andoni, Piotr Indyk, Huy L. Nguyen, and Ilya Razenshteyn. Beyond Locality-Sensitive Hashing. In *Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms (SODA '2014)*, pages 1018–1028, 2014. Available as arXiv:1306.1547.

[AIR18] Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. Approximate nearest neighbor search in high dimensions. In *Proceedings of ICM 2018 (to appear)*, 2018.

[AKR15] Alexandr Andoni, Robert Krauthgamer, and Ilya Razenshteyn. Sketching and Embedding are Equivalent for Norms. In *Proceedings of the 47th ACM Symposium on the Theory of Computing (STOC '2015)*, pages 479–488, 2015. Available as arXiv:1411.2577.

[ALRW17] Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten. Optimal Hashing-based Time–Space Trade-offs for Approximate Near Neighbors. In *Proceedings of the 28th ACM-SIAM Symposium on Discrete Algorithms (SODA '2017)*, pages 47–66, 2017. Available as arXiv:1608.03580.

[And09] Alexandr Andoni. *Nearest Neighbor Search: the Old, the New, and the Impossible*. PhD thesis, MIT, 2009.

[And10] Alexandr Andoni. Nearest neighbor search in high-dimensional spaces. Invited talk at the Workshop on Barriers in Computational Complexity II, http://www.mit.edu/~andoni/nns-barriers.pdf, 2010.

[ANN+17] Alexandr Andoni, Huy L. Nguyen, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten. Approximate Near Neighbors for General Symmetric Norms. In *Proceedings of the 49th ACM Symposium on the Theory of Computing (STOC '2017)*, pages 902–913, 2017. Available as arXiv:1611.06222.

[AR15] Alexandr Andoni and Ilya Razenshteyn. Optimal Data-Dependent Hashing for Approximate Near Neighbors. In *Proceedings of the 47th ACM Symposium on the Theory of Computing (STOC '2015)*, pages 793–801, 2015. Available as arXiv:1501.01062.

[AR16] Alexandr Andoni and Ilya Razenshteyn. Tight Lower Bounds for Data-Dependent Locality-Sensitive Hashing. In *Proceedings of the 32nd International Symposium on Computational Geometry (SoCG '2016)*, pages 9:1–9:11, 2016. Available as arXiv:1507.04299.

[Bal97] Keith Ball. *An Elementary Introduction to Modern Convex Geometry*, volume 31 of *MSRI Publications*. Cambridge University Press, 1997.

[BBC+17] Jaroslaw Blasiok, Vladimir Braverman, Stephen R. Chestnut, Robert Krauthgamer, and Lin F. Yang. Streaming Symmetric Norms via Measure Concentration. In *Proceedings of the 49th ACM Symposium on the Theory of Computing (STOC '2017)*, 2017. Available as arXiv:1511.01111.

[BG15] Yair Bartal and Lee-Ad Gottlieb. Approximate Nearest Neighbor Search for ℓ_p-Spaces ($2 < p < \infty$) via Embeddings. Available as arXiv:1512.01775, 2015.

[BKL06] Alina Beygelzimer, Sham Kakade, and John Langford. Cover Trees for Nearest Neighbor. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '2006)*, pages 97–104, 2006.

[BL00] Yoav Benyamini and Joram Lindenstrauss. *Geometric Nonlinear Functional Analysis. Vol. 1*, volume 48 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2000.

[Cha02] Moses Charikar. Similarity Estimation Techniques from Rounding Algorithms. In *Proceedings of the 34th ACM Symposium on the Theory of Computing (STOC '2002)*, pages 380–388, 2002.

[Che69] Jeff Cheeger. A Lower Bound for the Smallest Eigenvalue of the Laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pages 195–199, 1969.

[Chu96] F. R. K. Chung. Laplacians of graphs and Cheeger's inequalities. In *Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993)*, volume 2 of *Bolyai Soc. Math. Stud.*, pages 157–172. János Bolyai Math. Soc., Budapest, 1996.

[Cla99] Kenneth L. Clarkson. Nearest neighbor queries in metric spaces. *Discrete and Computational Geometry*, 22(1):63–93, 1999.

[Glu81] Efim D. Gluskin. Diameter of the Minkowski Compactum is Approximately Equal to n. *Functional Analysis and Its Applications*, 15(1):57–58, 1981.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[HR10] Moritz Hardt and Guy N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 61–70. IEEE Computer Society, 2010.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the 30th ACM Symposium on the Theory of Computing (STOC '1998)*, pages 604–613, 1998.

[Ind01] Piotr Indyk. On Approximate Nearest Neighbors under ℓ_∞ Norm. *Journal of Computer and System Sciences*, 63(4):627–638, 2001.

[Ind02] Piotr Indyk. Approximate Nearest Neighbor Algorithms for Fréchet Distance via Product Metrics. In *Proceedings of the 18th ACM Symposium on Computational Geometry (SoCG '2002)*, pages 102–106, 2002.

[Ind04] Piotr Indyk. Approximate Nearest Neighbor under Edit Distance via Product Metrics. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA '2004)*, pages 646–650, 2004.

[IT03] Piotr Indyk and Nitin Thaper. Fast Color Image Retrieval via Embeddings. Workshop on Statistical and Computational Theories of Vision (at ICCV), 2003.

[Joh48] Fritz John. Extremum Problems with Inequalities as Subsidiary Conditions. In *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*, pages 187–204. Interscience Publishers, Inc., New York, N. Y., 1948.

[KL04] Robert Krauthgamer and James R. Lee. Navigating Nets: Simple Algorithms for Proximity Search. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA '2004)*, pages 798–807, 2004.

[KR02] David R. Karger and Matthias Ruhl. Finding Nearest Neighbors in Growth-Restricted Metrics. In *Proceedings of the 34th ACM Symposium on the Theory of Computing (STOC '2002)*, pages 741–750, 2002.

[LNW14] Yi Li, Huy L. Nguyên, and David P. Woodruff. On Sketching Matrix Norms and the Top Singular Vector. In *Proceedings of the 25th ACM-SIAM Symposium on Discrete Algorithms (SODA '2014)*, pages 1562–1581, 2014.

[LW16a] Yi Li and David P. Woodruff. On Approximating Functions of the Singular Values in a Stream. In *Proceedings of the 48th ACM Symposium on the Theory of Computing (STOC '2016)*, pages 726–739, 2016.

[LW16b] Yi Li and David P. Woodruff. Tight Bounds for Sketching the Operator Norm, Schatten Norms, and Subspace Embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 19th International Workshop, APPROX '2016, and 20th International Workshop, RANDOM '2016*, pages 39:1–39:11, 2016.

[LW17] Yi Li and David P. Woodruff. Embeddings of Schatten Norms with Applications to Data Streams. In *Proceedings of the 44th International Colloquium on Automata, Languages and Programming (ICALP '2017)*, pages 60:1–60:14, 2017.

[Mat97] Jiří Matoušek. On Embedding Expanders into ℓ_p Spaces. *Israel Journal of Mathematics*, 102:189–197, 1997.

[Maz29] S. Mazur. Une remarque sur l'homÃľomorphie des champs fonctionnels. *Studia Mathematica*, 1(1):83–85, 1929.

[Mil99] Peter Bro Miltersen. Cell Probe Complexity – a Survey. In *Advances in Data Structures*, 1999.

[MN14] Manor Mendel and Assaf Naor. Nonlinear Spectral Calculus and Super-Expanders. *Publications Mathématiques de l'IHÉS*, 119(1):1–95, 2014.

[MN15] Manor Mendel and Assaf Naor. Expanders with Respect to Hadamard Spaces and Random Graphs. *Duke Mathematical Journal*, 164(8):1471–1548, 2015.

[Nao14] Assaf Naor. Comparison of Metric Spectral Gaps. *Analysis and Geometry in Metric Spaces*, 2:1–52, 2014.

[Nao17] Assaf Naor. A Spectral Gap Precludes Low-Dimensional Embeddings. In *Proceedings of the 33rd International Symposium on Computational Geometry (SoCG '2017)*, pages 50:1–50:16, 2017.

[Nao18] Assaf Naor. Metric dimension reduction: a snapshot of the Ribe program. In *Proceedings of ICM 2018 (to appear)*, 2018.

[Ngu14] Huy L. Nguyên. *Algorithms for High Dimensional Data.* PhD thesis, Princeton University, 2014. Available as http://arks.princeton.edu/ark:/88435/dsp01b8515q61f.

[NPS18] Assaf Naor, Gilles Pisier, and Gideon Schechtman. Impossibility of Dimension Reduction in the Nuclear Norm. In *Proceedings of the 29th ACM-SIAM Symposium on Discrete Algorithms (SODA '2018)*, 2018.

[NR06] Assaf Naor and Yuval Rabani. On Approximate Nearest Neighbor Search in ℓ_p, $p > 2$. Manuscript, available on request, 2006.

[NS07] Assaf Naor and Gideon Schechtman. Planar Earthmover is not in L_1. *SIAM Journal on Computing*, 37(3):804–826, 2007.

[OR07] Rafail Ostrovsky and Yuval Rabani. Low Distortion Embedding for Edit Distance. *Journal of the ACM*, 54(5):23:1–23:16, 2007.

[Raz17] Ilya Razenshteyn. *High-Dimensional Similarity Search and Sketching: Algorithms and Hardness.* PhD thesis, Massachusetts Institute of Technology, 2017.

[Ric15] Éric Ricard. Hölder Estimates for the Noncommutative Mazur Map. *Archiv der Mathematik*, 104(1):37–45, 2015.

[Spi15] Daniel A. Spielman. Conductance, the Normalized Laplacian, and Cheeger's Inequality. Lecture Notes, 2015.

[Yao81] Andrew Chi-Chih Yao. Should tables be sorted? *Journal of the ACM*, 28(3):615–628, 1981.