A Mathematical Introduction to Fast and Memory Efficient Algorithms for Big Data

by

Mark Iwen (Scribed by Cullen Haselby, Fall Semester 2020)

Michigan State University

This is Version 0 – January 2, 2021: THESE NOTES ARE VERY ROUGH! Many proofs are sketchy (pun intended), typos abound, referencing is incomplete, and some topics are not completely finished. Use with caution and email me at will with complaints, questions, and suggestions. All feedback welcome. Most Importantly, thanks to all my Michigan State students in MTH 994 and CMSE 890 this Fall 2020 for providing feedback on these notes. And, many thanks in particular to Craig Gross for transcribing much of Chapter VI, and of course to Cullen Haselby for scribing everything else!
– Mark Iwen

Table of Contents

List of Algorithms

List of Figures

List of Tables

List of Appendices

Chapter I

Some Motivating Problems Involving Big Data

1.1 Approximate Counting (CMSE 890 Lecture 1)

As way of introduction, in order provide some interesting and relevant examples that are illustrative of the larger course content, we consider three by now quotidian problems in big data settings.

Counting objects is a common challenge in settings involving very large data sets. Memory efficient methods are needed in order to make object counts feasible for routine use on these data. This type of problem and the ensuing discussion will also serve as an introduction to some key ideas for the course. In it we see a deterministic, simple sounding task (counting in the case) which under further study shows the need for fast and memory efficient algorithms that give good approximations to well constructed statistics questions.

A formal statement of the problem is as follows: Given a sequence $\{z_j\}_{j=1}^N$ where $\forall j$, $z_j \in U$ and some item $w \in U$ of interest, count the number of occurrences of term w in the sequence $\{z_j\}_{j=1}^N$.

Goal. *Estimate the count of w occurring in the sequence using $\lceil \log_2 \lceil \log_2 N \rceil \rceil$ bits of memory. We require our estimate of the count be larger than the actual count, but no more than twice the actual count.*

The source of the overestimate error on the count will be made clear shortly. Examples abound for data sets for which counts of this sort are useful

Example 1.1.1. U is the set of all possible phone numbers, and $\{z_j\}_{j=1}^N$ is a list of phone numbers which have communicated with a particular cellphone tower over some period of time. The term w is a phone number of interest, perhaps a known spammer.

Example 1.1.2. U is all possible pairs of words in the English language. So `hello world` or `thank you` are members of U. The sequence $\{z_j\}_{j=1}^N$ is a list of all pairs of words that appear in emails contained in some user's inbox. The term w then could be `buy pepsi` which is of interest to perhaps stock traders or advertisers.

Example 1.1.3. U is all possible IP addresses and $\{z_j\}_{j=1}^N$ is a list that contains the originating IP address for all packets received by a certain router. The term w is the IP address of a server used by a movie streaming service of interest to an internet service provider.

We may wish to consider counts of many different terms w for say all cell-phone towers in a particular country, or all users of some particular email service. Clearly, the size of such data sets means that counts can be potentially very large. Since N is an integer, a priori, we would need (maximally) $\lfloor \log_2 N \rfloor + 1$ bits to store a count of each w.

Note. *We can store N using $\lfloor \log_2 N \rfloor + 1$ bits. We have $\lfloor \log_2 N \rfloor = k$ only if $k \leq \log_2 N < k+1$ if and only if $2^k \leq N < 2^{k+1}$. That is, $2^k \leq N < 2^{k+1}$ is the range of integers which requires $k+1$ bits. So the integers requiring 4-bits for example are 8 through 15. Depending on implementation there are other bits required to say, store the sign of the integer. For simplicity we say that storing an integer of*

size N requires $\lceil \log_2 N \rceil$ bits, though this may be off by one, or some other constant, depending on implementation.

A first, naive approach is to increment a counter after one scan of the sequence, and then store the logarithm of that count.

Algorithm 1.1.1 Naive Counter

Input: $\{z_j\}_{j=1}^{N}, w$
Output: approximate count of w in $\{z_j\}_{j=1}^{N}$
for $j = 1$ to N **do**
  if $z_j = w$ **then**
    $\tilde{w} \leftarrow \tilde{w} + 1$
  end if
end for
$E \leftarrow \lceil \log_2 \tilde{w} \rceil$

Since E is of size at most $\lceil \log_2 N \rceil$ it takes at most $\lceil \log_2 \lceil \log_2 N \rceil \rceil$ bits to store. Due to the information lost by taking the ceiling, we also have that $\tilde{w} \leq 2^E \leq 2\tilde{w}$.

Question 1.1.4. *Does E and the algorithm 1.1.1 achieve our goal?*

No. While it is true that E occupies the right number of bits, the counter itself $\tilde{w}$ would need to occupy possibly $\lceil \log_2 N \rceil$ bits when running the algorithm.

Another problem which is similar to counting objects is the distinct elements problem. Here we concerned with determining the number of distinct elements which appear in a given sequence, as opposed to the frequency.

Formally, given a sequence $\{z_j\}_{j=1}^{N}$ where $\forall j$, $z_j \in U$, and $|U| = D$ we wish to compute the cardinality of $\{z_j\}_{j=1}^{N}$ as a set.

Goal. *Estimate the number of distinct elements in a sequence using a number of bits independent of both N and D.*

One can imagine many different settings where such a count of distinct elements would be useful.

Example 1.1.5. U is the set of all possible phone numbers, and $\{z_j\}_{j=1}^N$ is a list of phone numbers which have communicated with a particular cellphone tower over some period of time. The cardinality of the sequence as a set would be the number of unique cellphones that used the tower.

Example 1.1.6. U is all possible pairs of words in the English language. The sequence $\{z_j\}_{j=1}^N$ is a list of all pairs of words that appear in a user's current email outbox. The cardinality of the sequence as a set would be an indicator of the variation of a given user's word choice in writing emails.

We consider two naive solutions to this problem, and observe how they do not achieve the stated goal.

Algorithm 1.1.2 Naive Distinct Elements by D-array

Input: $\{z_j\}_{j=1}^N$
Output: number of distinct elements in $\{z_j\}_{j=1}^N$
Let $A :=$ array of zeros of size D
for $j = 1$ to N **do**
  if $A[z_j] = 0$ **then**
    $A[z_j] := 1$
  end if
end for
$\|A\|_0$

Algorithm 1.1.3 Naive Distinct Elements by Sorted List

Input: $\{z_j\}_{j=1}^N$
Output: number of distinct elements in $\{z_j\}_{j=1}^N$
Let $L[j] := z_j$
Sort L
for $j = 1$ to $N - 1$ **do**
  if $L[j] = L[j + 1]$ **then**
    flag $L[j + 1]$ for removal
  end if
end for
$|L|$

However, neither of these algorithms meet the requirements of the stated goal. In the case of algorithm 1.1.2 the array A clearly occupies D bits. In the case of 1.1.3,

the list L needs N entries, and so will occupy at least N bits of memory. In a future lecture, we will study the Flajolet-Martin Algorithm which does solve the distinct element problem with constant memory.

The third problem we consider is Nearest Neighbor in $\mathbb{R}$. Here we have a set of points, and are presented with a query point and wish to return the closet point in our set to the query point, reckoned by a norm of interest. Formally, we have $S \subset R^D$, and query $\mathbf{y} \in R^D$ and compute $\mathbf{y}_{NN} = \arg\min \|\mathbf{x} - \mathbf{y}\|$. The set S has cardinality N which can be very large. Naturally we can extend this to k-nearest neighbors by returning a list of the k closest points.

A simple linear scan then of the set is perhaps the most obvious solution to the problem

Algorithm 1.1.4 Naive Nearest Neighbors

```
Input: S, y, ‖·‖
Output: y_NN
d = ∞
for x in S do
  if ‖x − q‖ < d then
    y_NN ← x
  end if
end for
```

This problem is a fundamental building block type of problem in many algorithms and data science applications.

Example 1.1.7. S is the a database of gray-scale images. A query point q is a novel image, we return the image that is closest to using the ℓ_1 norm

Example 1.1.8. S is a database of names of people who bought departing tickets from a given airport. A query point q is a name of a passenger of interest, we return the name that is closest to it using the Hamming distance.

Example 1.1.9. S is a database of users of a dating website. Each user has a vector

of different features, which is computed from data collected about their interests, hobbies, preferences, etc. A query point q represents a particular user, and developers for the website have engineered a norm which represents similarity between users. The closest point in S is recommended as a potential partner.

Since each of the N points in S needs to be compared to the query point, and calculating the norm of the difference depends on the dimension D of the space, this scan has $\mathcal{O}(ND)$ complexity. We will later study how to improve on this using good approximations.

Definition 1.1.10. For $p \geq 1$, the ℓ_p-norm of a vector in $\mathbf{x} \in \mathbb{C}^D$ is a map $\|\cdot\|_p : \mathbb{C}^D \to [0, \infty)$ defined by

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^{D} |x_j|^p \right)^{1/p}$$

if $p = \infty$ then $\|\mathbf{x}\|_\infty = \max_j |x_j|$

Homework 1.1.1. Given norms $\|\cdot\|_\dagger$ and $\|\cdot\|_\star$ on $\mathbb{C}^D$ and $\alpha, \beta \in [0, \infty)$, prove that $\|\mathbf{x}\|_+ = \alpha\|\mathbf{x}\|_\dagger + \beta\|\mathbf{x}\|_\star$ is also a norm in $\mathbb{C}^D$

Other applications that will be relevant to our study in this course are

1. Fast Monte Carlo integration approximation
2. Fast approximate solutions to classic numerical linear algebra problems in the big data setting such as
   (a) least square regression
   (b) matrix-matrix multiplication
   (c) Principal Component Analysis
3. Compressive sensing

4. Heavy Hitter problems

Heavy Hitter problems refer to cases where we want to find those values which occur most frequently in a large (streaming) sequence of data. For example, a seller such as Walmart may be interested in the hundred most purchased items across many different stores on a minute by minute or second by second time-frame. Another subtype of Heavy Hitter problem appears in group testing. Here, many specimens are collected and tested together in batches. So for example, 20 patients may submit specimens that are combined into batches containing samples from 5 different specimens, and say samples from each specimen are included in 3 different batches which are then tested for the disease. If the prevalence of the disease is sufficiently small, batching schemes can be designed to economize testing but still ensure identifiability of patients who have the disease.

1.2 Fast Function Approximation via Compressive Sensing (MTH 994 Lecture 1)

The main problem that the course addresses is as follows

Design an algorithm, i.e. a computable function, $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ where $m \leq N$ and a set of linear measurements $\Phi \in \mathbb{C}^{m \times N}$ for a given subset $\mathcal{F}_{\mathbf{p}} \subset \mathbb{C}^N$ with parameters $\mathbf{p} \in \mathbb{C}^r$ such that for all $(\mathbf{n}, \mathbf{x}) \in \boldsymbol{Z} \times \mathcal{F}_{\mathbf{p}}$ the following holds

$$\|\Delta(\Phi\mathbf{x} + \mathbf{n}) - \mathbf{x}\|_X \leq C_{\mathbf{p},X,Y} \inf_{\mathbf{y} \in F_{\mathbf{p}}} \|\mathbf{x} - \mathbf{y}\|_Y + \tilde{C}_{\mathbf{p},X,Y,Z} \|\mathbf{n}\|_Z + \epsilon_{\mathbf{p},X,Y,Z} \tag{1.1}$$

In effect, what we seek is a reconstruction or invertibility property for our algorithm, namely, $\Delta(\Phi\mathbf{x}) = \mathbf{x}$. We know that this property cannot hold in the generic case where $\mathbf{x} \in \mathbb{C}^N$ since $m \leq N$ and thus the null space of Φ will be at least of dimension $N - m$. So, the condition that $\mathbf{x} \in \mathcal{F}_{\mathbf{p}}$ makes the desired property possible,

and the nature of $\mathcal{F}_{\mathbf{p}}$ crucial to our understanding and solution to the problem.

Some key remarks for equation 1.1:

1. The algorithm $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ should be implementable in a manner that is fast, memory efficient, and robust to noise.

2. The norms $\|\cdot\|_{X,Y,Z}$ will usually be ℓ_p-norms for $p = 1, 2, \ldots$

3. $\mathbf{n} \in \mathbb{C}^m$ is arbitrary noise on the input $\Phi\mathbf{x}$. Deterministic or probabilistic pertubations to the input are both possible and the more general the case we can accommodate in our algorithm the better.

4. $\epsilon_{\mathbf{p},X,Y,Z} \in \mathbb{R}^+$ is a small error. This can be round-off error, though often in the sequel we will take it to be zero.

5. Constants like $C_{\mathbf{p},X,Y}$ are absolute in the sense that they are independent of any particular $\mathbf{x}$ and noise $\mathbf{n}$.

6. Often, we consider the compressed measurement case, where for $\Phi \in \mathbb{C}^{m\times N}$ we have $m \ll N$.

7. $\mathcal{F}_{\mathbf{p}} \subset \mathbb{C}^N$ will be some geometrically simple set parameterized by $\mathbf{p}$, such as

   (a) (Manifold) $\mathcal{M}_{[d,\tau]}$, a d-dimensional sub-manifold of $\mathbb{R}^N$ whose reach is bounded by τ. There are other possible parameters, such as volume or diameter which could be used to describe the geometry of the manifold.

   (b) (Compressed sensing) $K_s \subset \mathbb{C}^N$, where K_s is the set of s-sparse vectors in $\mathbb{C}^N$, i.e $\left\{\mathbf{x} \in \mathbb{C}^N | \|\mathbf{x}\|_0 \leq s\right\}$ which is equivalently $\bigcup_{S\subset[N],|S|=s} \operatorname{span}\left\{\mathbf{e}_j\right\}_{j\in S}$ where $\mathbf{e}_j$ are the standard basis vectors. That is, the span of vectors with s non-zero entries.

Note that when $\epsilon_{\mathbf{p},X,Y,Z} = 0$ and in the absence of noise, $\mathbf{n} = \mathbf{0}$, equation 1.1 implies the invertibility property, $\Delta(\Phi\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in \mathcal{F}_{\mathbf{p}}$, which is equivalent to the following, by the linearity of Φ

$$\mathbf{x} \neq \mathbf{y} \iff \Phi(\mathbf{x} - \mathbf{y}) \neq 0 \, \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}_{\mathbf{p}} \tag{1.2}$$

However, numerically 1.2 is not a tenable, realistic property to design around. So we define a stronger property which will imply 1.2.

This is known as the Johnson-Lindenstrauss (JL) embedding property. We say that Φ has the JL-property when $\exists \epsilon \in (0, 1)$ such that

$$|\|\Phi(\mathbf{x} - \mathbf{y})\|_X^2 - \|\mathbf{x} - \mathbf{y}\|_Y^2| \leq \epsilon \|\mathbf{x} - \mathbf{y}\|_Y^2, \, \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}_{\mathbf{p}} \tag{1.3}$$

Analyzing this property in terms of different spaces and matrices will occupy much of our subsequent study. We now conclude the introduction lecture with a discussion of function approximation, and how it relates to the key property 1.3 larger goals of the course.

Running Example. *Suppose $\mathcal{D} = [0, 1]^D$, the D-dimensional cube, and $f \in \mathcal{H}$ for $\mathcal{H} = L_\mu^2(\mathcal{D}, \mathbb{C})$, the separable Hilbert space of square intergrable complex valued functions on domain $\mathcal{D}$*

1. Pick a countable orthornormal basis $\mathcal{B}$ of $\mathcal{H}$.

$$\mathcal{B} = \{b_j\}_{j \in \mathbb{Z}^D}$$

For example, $\mathcal{B}$ is the Fourier basis. Note that through a Gram-Schmidt process we are guaranteed the existence of a maximal orthonormal set in seperable Hilbert space.

2. *Pick a finite subset* $\mathcal{B}' \subset \mathcal{B}$ *with* $|\mathcal{B}'| = N$*. For example,* $\mathcal{B}'$ *corresponds to some frequencies such as those in a hyperbolic cross, or frequencies in* $(\mathbb{Z} \cap [-M, M])^D$ *for some* $M \in [0, \infty)$

3. *Approximate* f *by its projection* $P_{\mathcal{B}'} f = \sum_{j \in \mathcal{I}} b_j \langle b_j, f \rangle$ *where* $\mathcal{I}$ *is the index set corresponding to the finite basis,* $\mathcal{I} = \left\{ j \in \mathbb{Z}^D | b_j \in \mathcal{B}' \right\}$*, so* $|\mathcal{I}| = N = |\mathcal{B}'|$

Given m input measurements $\langle a_1, f \rangle, \ldots, \langle a_m, f \rangle$ we can restate the example in terms of 1.3 by setting $\mathbf{x} \in \mathbb{C}^N$ to $x_j = \langle b_j, f \rangle$ and $\Phi \in \mathbb{C}^{m \times N}$, $\Phi_{\ell,j} = \langle a_\ell, b_j \rangle$, $\forall j \in \mathcal{I}$ such that each input measurement satisfies

$$\begin{aligned}
\langle a_\ell, f \rangle &= \langle a_\ell, P_{\mathcal{B}'} f \rangle + \underbrace{\langle a_\ell, (I - P_{\mathcal{B}'}) f \rangle}_{n_\ell} \\
&= \langle a_\ell, \sum_{j \in \mathcal{I}} b_j \langle b_j, f \rangle \rangle + n_\ell \\
&= \sum_{j \in \mathcal{I}} \langle a_\ell, b_j \rangle x_j + n_\ell \\
&= \sum_{j \in \mathcal{I}} \Phi_{\ell,j} x_j + n_\ell
\end{aligned}$$

Note the noise vector $\mathbf{n}$ is due to the truncation error incurred by using only a finite number of basis elements in this function approximation setting. So, if we also have that $\mathbf{x}$ is in (or near) $\mathcal{F}_p \in \mathbb{C}^N$ then the conclusion of the running example is indeed statement of the result 1.3.

Note that a large number of basis elements may be required to reduce error of the approximation, especially for high dimensional input space D. This means that computationally, function approximation may be intractable unless we use the structure of $\mathcal{F}_{\mathbf{p}}$ to compress the computation of $\sum_{j \in \mathcal{I}} \Phi_{\ell,j} x_j$.

Example 1.2.1 (Function Approximation, 1-D sparse Fourier Transform)**.** Suppose $f : [0,1] \to \mathbb{C}$, $f \in L^2([0,1],\mathbb{C}) \cap C^1([0,1])$. Choose the orthonormal basis $\mathcal{B} = \left\{e^{2\pi kix}\right\}_{k\in\mathbb{Z}}$ and select the finite subset $\mathcal{B}' = \left\{e^{2\pi kix}\right\}_{k\in[-N,N]\cap\mathbb{Z}}$.

Let the input measurements be point samples inside the domain, i.e. $a_\ell = \delta_{x_\ell}$ where $\delta_{x_\ell} = \delta(x - x_\ell)$ for $x_\ell \in [0,1]$, $\forall \ell \in [m]$. Choose $\mathcal{F}_{\mathbf{p}}$ to be $\mathcal{K}_s$, the s-sparse vectors in the Fourier basis, $x = \hat{f}\big|_{k\in[-N,N]\cap\mathbb{Z}}$ has at most s non-zero entries (alternatively we may relax this and say that the bulk of the energy of x is in at most s-entries).

Note that Φ has several constraints - it's entries are now taken from point evaluations of different basis functions at the different sample points x_ℓ; and yet we still require that it preserves the norms of vectors in $\mathcal{F}_{\mathbf{p}}$ as stated in 1.3. Additionally, in order to achieve an improvement in the speed of our algorithm, we need to improve on the usual Fast Fourier Transform sampling complexity. Instead of the bound $m \leq N \log N$ we want $m \leq s \log^C N$ where C is a small positive, absolute constant. In this way we can benefit from the sparsity of $\mathcal{F}_{\mathbf{p}}$. Lastly, we want to be able to recover x, i.e. find $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ that is able to run in $\mathcal{O}(s \log^C N)$ (in contrast to $\mathcal{O}(N \log N)$ required for the standard $FFT^{-1}(FFT(x))$)

Homework 1.2.1. Prove that 1.3 implies 1.2

Homework 1.2.2. Prove that 1.3 implies

$$|\|\Phi(\mathbf{x}-\mathbf{y})\|_X - \|\mathbf{x}-\mathbf{y}\|_Y| \leq \epsilon \frac{\|\mathbf{x}-\mathbf{y}\|_Y}{1+\sqrt{1-\epsilon}} \leq \frac{\epsilon\|\mathbf{x}-\mathbf{y}\|_Y}{2-\epsilon}$$

1.3 Tensor Applications

Definition 1.3.1. An n-mode or order-n tensor (or n-th order) is an n dimensional array of complex values, written as

$$\mathcal{A} \in \mathbb{C}^{I_0 \times I_1 \times \cdots \times I_{n-1}}$$

where $I_j \in \mathbb{N}, j \in [n]$. An n-mode tensor's entries are indexed by a vector $\mathbf{i} \in [I_0] \times [I_1] \times \cdots \times [I_{n-1}]$ where $(\mathcal{A})_{\mathbf{i}} = a_{\mathbf{i}} = a_{i_0,\ldots,i_{n-1}} \in \mathbb{C}$

Example 1.3.2 (1 and 2 mode tensors)**.** 1. A 1-mode tensor, $\mathbf{a} \in \mathbb{C}^{I_0}$ is a vector with entries $a_j \in \mathbb{C}$, $j \in [I_0]$. We will denote 1-mode tensors, vectors, the usual way with bolded lowercase letters

2. A 2-mode tensor, $A \in \mathbb{C}^{I_0 \times I_1}$ is a matrix with entries $a_{i_0,i_1} \in C$ for $i_0 \in [I_0], i_1 \in [I_1]$. Equivalently $a_{\mathbf{k}}$, $\mathbf{k} \in [I_0] \times [I_1]$. We will denote 2-mode tensors, matrices, the usual way with capital un-bolded letters.

We introduce some terminology that will be useful when describing tensors

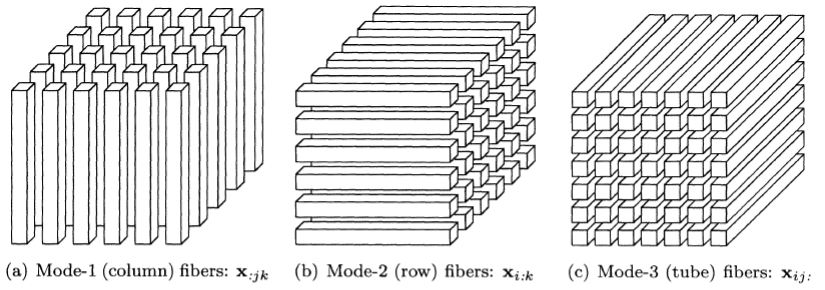
Definition 1.3.3 (Fiber)**.** Fibers are 1-dimensional subsets of an n-mode tensor. They are formed by fixing $n-1$ of the dimensions and then ranging over all indices in the remaining dimension. So for any $k \in [n]$, and $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{n-1}}$ then a k-mode fiber would be a vector $\mathbf{a} \in \mathbb{C}^{I_k}$ where indices $i_0, \ldots, i_{k-1}, i_{k+1}, \ldots, i_{n-1}$ are fixed, i.e. using `Matlab` notation

$$(\mathcal{A})_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{n-1}} = \mathbf{a}_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{n-1}}$$

So for example, given a matrix $A \in \mathbb{C}^{I_0 \times I_1}$ then $A_{i,:} = \mathbf{a}_{i,:} \in \mathbb{C}^{I_1}$ is a mode-2 fiber (i.e. row). A mode-1 fiber, $\mathbf{a}_{:,j}$ is a column of the matrix.

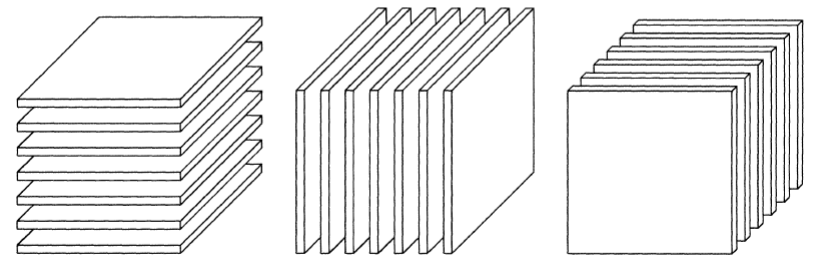
Definition 1.3.4 (Slice)**.** A matrix formed by varing 2 indices and fixing all other indices of a tensor. That is, suppose $j, k \in [n]$ where $j \neq k$ then

$$A = \mathcal{A}_{i_0,\ldots,i_{j-1},:,i_{j+1},\ldots,i_{k-1},:,i_{k+1},\ldots,i_{n-1}} \in \mathbb{C}^{I_j \times I_k}$$



(a) Mode-1 (column) fibers: $\mathbf{x}_{:jk}$ (b) Mode-2 (row) fibers: $\mathbf{x}_{i:k}$ (c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$

Fig. 2.1 *Fibers of a 3rd-order tensor.*



(a) Horizontal slices: $\mathbf{X}_{i::}$ (b) Lateral slices: $\mathbf{X}_{:j:}$ (c) Frontal slices: $\mathbf{X}_{::k}$ (or $\mathbf{X}_k$)

Fig. 2.2 *Slices of a 3rd-order tensor.*

Figure 1.1: Figure seen in [23]

Definition 1.3.5 (Sub-tensor)**.** A k-subtensor of an n-mode tensor ($k < n$) is denoted by a vector of length $n-k$ of indices and a set of k mode indices from the set $[n]$. That is given distinct $j_0, \ldots, j_{k-1} \in [n]$ and define vector $\mathbf{i} \in \bigotimes_{i \neq j_\ell}[I_i]$ of length $n-k$. Let

$$\mathcal{A}_{j_0,\ldots,j_{k-1},\mathbf{i}} \in \mathbb{C}^{I_{j_0} \times \cdots \times I_{j_{k-1}}}$$

Using this subtensor notation then a mode-k fiber is a subtensor $\mathcal{A}_{k,\mathbf{i}}$ where $k \in [n]$ and $\mathbf{i} \in [I_0] \times \ldots [I_{k-1}] \times [I_{k+1}] \times \cdots \times [I_{n-1}]$. There are $\prod_{\ell \neq k} I_\ell$ potentially different mode-k fibers, one for each possible $\mathbf{i}$.

A slice then is denoted $\mathcal{A}_{\ell,k,\mathbf{i}} \in \mathbb{C}^{I_\ell \times I_k}$. There are $\prod_{j\neq \ell,k} I_j$ slices of dimension $I_\ell \times I_k$.

Next we will discuss reshaping operators - this involves many different possible ways of changing the dimensions of tensors so that they have the same number of entries.

Definition 1.3.6 (Vectorization)**.** For $\mathcal{A} \in \mathbb{C}^{I_0\times\cdots\times I_{n-1}}$ $\text{vec}(\mathcal{A}) = \mathbf{a}$ where $\mathbf{a} \in \mathbb{C}^{\prod_{k=0}^{n-1} I_k}$

Definition 1.3.7 (Mode-k Flattening)**.** For $\mathcal{A} \in \mathbb{C}^{I_0\times\cdots\times I_{n-1}}$ the k-mode flattening is a matrix $A^{(k)} \in \mathbb{C}^{I_k \times \prod_{j\neq k} I_j}$. We have effectively made the k-th dimension into the rows of the matrix, and the columns are then the different mode-k fibers. In particular $\left(A^{(k)}\right)_{j,\ell} = \mathcal{A}_{\ell_1,\ldots,\ell_{k-1},j,\ell_{k+1},\ldots,\ell_{n-1}}$. The columns are the fibers $\mathcal{A}_{k,\mathbf{i}}$.

Definition 1.3.8 (Reshaping)**.** We can reshape an n-mode into any other m-mode tensor with a reshaping operation $R : \mathbb{C}^{I_0\times\cdots\times I_{n-1}} \to \mathbb{C}^{J_0\times\cdots\times J_{m-1}}$ provided

$$\prod_{j=0}^{n-1} I_j = \prod_{\ell=0}^{m-1} \tilde{J}_\ell$$

k-mode flattening and vectorization are two particular reshaping operations.

What is the underlying vector space we can use to study tensors? To answer that, we consider the following norm, inner-product, and operations on tensors:

Definition 1.3.9 (2-norm of a Tensor)**.** For $\mathcal{A} \in \mathbb{C}^{I_0\times\cdots\times I_{n-1}}$ then given any $k \in [n]$ we have

$$\|\mathcal{A}\|_2^2 = \|A^{(k)}\|_2^2 = \|\mathbf{a}\|_2^2 = \sum_{\mathbf{i}\in I} |a_{\mathbf{i}}|^2$$

where $I = [I_0] \times \cdots \times [I_{n-1}]$

Definition 1.3.10 (Inner-product)**.** For tensors $\mathcal{A}, \mathcal{B} \in \mathbb{C}^{I_0\times\cdots\times I_{n-1}}$ then

$$\langle \mathcal{A}, \mathcal{B}\rangle = \sum_{\mathbf{i}\in I} a_{\mathbf{i}}\bar{b}_{\mathbf{i}} = \langle \mathbf{a}, \mathbf{b}\rangle$$

that is, the inner-product of the vectorization of the tensors. Note that this is equivalent to $\langle A^{(k)}, B^{(k)} \rangle_{HS} = \text{Trace}(A^k (B^{(k)})^*)$, $\forall k \in [n]$, the Hilbert-Schmidt inner product for matrices

Addition and scalar multiplication work component-wise, i.e.

$$(\mathcal{A} + \mathcal{B})_{\mathbf{i}} = a_{\mathbf{i}} + b_{\mathbf{i}}$$

$$(\alpha \mathcal{A})_{\mathbf{i}} = \alpha a_{\mathbf{i}}, \, \forall \alpha \in \mathbb{C}$$

1.3.1 Restricted Inner and Matrix Products

Given long vectors, as may result from reshaping a tensors for example, we may perform inner products with smaller vectors by using well chosen samples or parts from the longer vectors.

Definition 1.3.11. k-mode product of $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ and $U \in \mathbb{C}^{J_k \times I_k}$ for $k \in [d]$ is a tensor in $\mathbb{C}^{I_0 \times \cdots \times I_{k-1} \times J_k \times I_{k+1} \times \cdots \times I_{d-1}}$ denoted by

$$(\mathcal{A} \times_k U)_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{d-1}} = U \mathcal{A}_{i_0,\ldots\ldots,i_{d-1}}$$

In other words, the k-mode product applies the matrix U to all the mode-k fibers of the tensor $\mathcal{A}$. For example, suppose $\mathcal{A} \in \mathbb{C}^{5 \times 3 \times 2}$ and $U \in \mathbb{C}^{4 \times 5}$. Then $\mathcal{A} \times_1 U \in \mathbb{C}^{4 \times 3 \times 2}$ where each of the mode-1 fibers is now the product of $U \mathcal{A}_{:,j,\ell}$, for some $j \in [3], \ell \in [2]$.

In the 2-mode tensor case, i.e., matrices, the usual matrix-matrix multiplication can be understood in terms of 1-mode tensor product.

$$\begin{aligned} AB &= A \left[\begin{array}{c|c|c|c} \mathbf{b}_1 & \mathbf{b}_2 & \ldots & \mathbf{b}_n \end{array} \right] \\ &= \left[\begin{array}{c|c|c|c} A\mathbf{b}_1 & A\mathbf{b}_2 & \ldots & A\mathbf{b}_n \end{array} \right] \\ &= B \times_1 A \end{aligned}$$

Lemma 1.3.12. $(\mathcal{A}+\mathcal{B})\times_k U = \mathcal{A}\times_k U + \mathcal{B}\times_k U$

Proof. For any $i_0 \in I_0, \ldots, i_{k-1} \in I_{k-1}, i_{k+1} \in I_{k+1}, \ldots, i_{d-1} \in I_{d-1}$ we have

$$\begin{aligned}[(\mathcal{A}+\mathcal{B})\times_k U]_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{d-1}} &= U(\mathcal{A}+\mathcal{B})_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{d-1}} \\ &= U\mathcal{A}_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{d-1}} + U\mathcal{B}_{i_0,\ldots,i_{k-1},:,i_{k+1},\ldots,i_{d-1}} \\ &= \mathcal{A}\times_k U + \mathcal{B}\times_k U\end{aligned}$$

□

Lemma 1.3.13 (Properties of k-mode products)**.** *Let $\mathcal{A},\mathcal{B} \in \mathbb{C}^{I_0,\ldots,I_{d-1}}$, $\alpha,\beta \in \mathbb{C}$, $U_\ell, V_\ell \in \mathbb{C}^{m_\ell \times I_\ell}$, $\forall \ell \in [d]$. Then*

1. $(\alpha\mathcal{A}+\beta\mathcal{B})\times_j U_j = \alpha(A\times_j U_j) + \beta(\mathcal{B}\times_j U_j)$
2. $\mathcal{A}\times_j (\alpha U_j + \beta V_j) = \alpha(\mathcal{A}\times_j U_j) + \beta(\mathcal{A}\times_j V_j)$ *that is, k-mode product is bilinear*
3. *If $j \neq \ell$ then*

$$(\mathcal{A}\times_j U_j)\times_\ell V_\ell = (\mathcal{A}\times_\ell V_\ell)\times_j U_j$$

   Note that the run-time complexity is the same regardless of the order one applies the k- or j-mode products

4. *If $W \in \mathbb{C}^{p\times m_j}$ then $(\mathcal{A}\times_j U_j)\times_j W = \mathcal{A}\times_j (WU_j) \in \mathbb{C}^{I_0\times I_{j-1}\times p\times I_{j+1}\times \ldots I_{d-1}}$*

Definition 1.3.14 (Kronecker Product)**.** The Kronecker product of two matrices $U \times \mathbb{C}^{m\times n}$ and $V \in \mathbb{C}^{p\times q}$ is a matrix

$$U\otimes V = \begin{pmatrix} u_{11}V & \ldots & u_{1n}V \\ \vdots & & \vdots \\ u_{m1}V & \ldots & u_{mn}V \end{pmatrix}$$

where $U\otimes V \in \mathbb{C}^{mp\times nq}$

Lemma 1.3.15. *Let $\mathcal{A} \in \mathbb{C}^{I_0,\ldots,I_{d-1}}$, $U_\ell \in \mathbb{C}^{m_\ell\times I_\ell}$ then*

1. $(\mathcal{A} \times_j U_j)^{(j)} = U_j \mathcal{A}^{(j)} \in \mathbb{C}^{m_j \times \prod_{\ell \neq j} I_\ell}$

2. $(\mathcal{A} \times_0 U_0 \times_1 U_1 \cdots \times_{d-1} U_{d-1})^{(j)} = U_j \mathcal{A}^{(j)} \left(U_{d-1} \otimes U_{d-2} \otimes \cdots \otimes U_{j+1} \otimes U_{j-1} \otimes \cdots \otimes U_0\right)^T$

we have assumed a column-major convention for matricization.

In order for the identity seen in Lemma 1.3.15 to hold, we need specify our precise matricization convention.

In our convention, the entry at location $(i_0, \ldots, i_{d-1})$ in $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ is located in the matrix as entry (i_n, j) where

$$j = \sum_{\substack{k=0 \\ k \neq n}}^{d-1} i_k J_k$$

where

$$J_k = \prod_{\substack{m=0 \\ m \neq n}}^{k-1} I_m$$

set $J_k = 1$ if the index of the product above is empty.

Example 1.3.16 (3-mode)**.** The following example appears in [23]: Consider a tensor $\mathcal{A} \in \mathbb{R}^{3 \times 4 \times 2}$, the frontal slices are as follows

$$\mathcal{A}_{:,:,0} = \mathcal{A}_0 = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}, \mathcal{A}_{:,:,1} = \mathcal{A}_1 = \begin{pmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{pmatrix}$$

The three different unfoldings are then as follows. Consider $n = 0$,

$$\mathcal{A}^{(0)} = \begin{pmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{pmatrix}$$

Note

$$J_0 = \prod_{\substack{m=0 \\ m \neq 0}}^{0-1} I_m = 1, J_1 = \prod_{\substack{m=0 \\ m \neq 0}}^{1-1} I_m = 1, J_2 \prod_{\substack{m=0 \\ m \neq 0}}^{2-1} I_m = I_1 = 4$$

Now to see how to locate a particular entry, note that $\mathcal{A}_{1,2,1} = 20$, so in our unfolding $\mathcal{A}^{(k)}_{(i,j)}$ we can simply copy the index that corresponds to the n-th mode, i.e. the first here, $i = 1$. To find the column, compute $j = \sum_{\substack{k=0 \\ k\neq n}}^{d-1} i_k J_k = 2(1) + 1(4) = 6$. So our entry with value 20 is in location $(1, 6)$.

Consider $n = 1$,

$$\mathcal{A}^{(1)} = \begin{pmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{pmatrix}$$

Again, to locate our entry,$\mathcal{A}_{1,2,1} = 20$, we return the index on the n-th mode as our row, $i = 2$, and repeat the same calculation for j:

Note

$$J_0 = \prod_{\substack{m=0 \\ m\neq 1}}^{0-1} I_m = 1, J_1 = \prod_{\substack{m=0 \\ m\neq 1}}^{1-1} I_m = I_0 = 3, J_2 \prod_{\substack{m=0 \\ m\neq 1}}^{2-1} I_m = I_0 = 3$$

.

This time we leave out the J_1 factor: $j = \sum_{\substack{k=0 \\ k\neq n}}^{d-1} i_k J_k = 1(1) + 1(3) = 4$. So our entry with value 20 is located in $(2, 4)$.

Consider $n = 2$, $\mathcal{A}^{(2)}$

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \dots \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & \dots \end{pmatrix}$$

Again, to locate our entry,$\mathcal{A}_{1,2,1} = 20$, we return the index on the n-th mode as our row, $i = 1$, and repeat the same calculation for j:

Note

$$J_0 = \prod_{\substack{m=0 \\ m\neq 2}}^{0-1} I_m = 1, J_1 = \prod_{\substack{m=0 \\ m\neq 2}}^{1-1} I_m = I_0 = 3, J_2 \prod_{\substack{m=0 \\ m\neq 2}}^{2-1} I_m = I_0 = 3$$

.

This time we leave out the J_2 factor: $j = \sum_{\substack{k=0 \\ k\neq n}}^{d-1} i_k J_k = 1(1) + 2(3) = 7$. So our entry with value 20 is located in $(1, 7)$.

1.3.2 Low Rank Approximation

Our next topic concerns how different methods can be used to approximate tensors. Our first attempt will illustrate the need for better decompositions other than simply reshaping tensors into familiar objects.

Suppose we have q tensors of size $\mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$, $\mathcal{A}_1, \ldots, \mathcal{A}_q$. We will compress the tensors by performing PCA on the vectorized tensors. Our goal then in this case is to solve the minimization problem:

$$\sum_{j=1}^{m} \min_{S_j \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}} \|\mathcal{A}_j - \mathcal{S}_j\|_2^2$$

This is equivalent to

$$\sum_{j=1}^{m} \min_{\mathbf{s}_j \in \mathcal{S} \subseteq \mathbb{C}^{\prod_{m=0}^{d-1} I_m}} \|\mathbf{a}_j - \mathbf{s}_j\|_2$$

where $\text{vec}(\mathcal{A}_j) = \mathbf{a}_j \in \mathbb{C}^{\prod_{m=0}^{d-1} I_m}$. We can use the SVD of the following data matrix

$$\left[\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_q \end{array} \right] = U\Sigma V^* \in \mathbb{C}^{q \times \prod_{m=0}^{d-1} I_m}$$

Once we have the singular vectors, we can tensorize their outer product using the inverse of our vectorizing reshaping operation. That is

$$\mathcal{A}_j \approx \sum_{k=1}^{m} \sigma_{j,k} \mathcal{T}_k$$

where $\mathcal{T}_k$ are the tensors obtained from the principal directions (obtained from the singular vectors of SVD of the data matrix) and $\sigma_{j,k}$ are the appropriate principal

scores (again computed from the singular vectors and singular values of the data matrix).

What compression does this achieve? The space required for our original collection of tensors $\mathcal{A}_1, \ldots, \mathcal{A}_q \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ is $\mathcal{O}(q \prod_{m=0}^{d-1} I_m)$. After PCA, we need keep m basis tensors of the same dimension $\mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ and our coordinates or principal scores will also need to be stores and there are mq of these. So the space is $\mathcal{O}(m \prod_{m=0}^{d-1} + qm)$ which is unsatisfactory because the dependence on $\prod_{m=0}^{d-1}$ is unchanged. Additionally, there's no interpretable structure to the basis tensors $\mathcal{T}_k$. This motivates us then to look to another approach for decomposing (and therefore compressing) tensors.

Definition 1.3.17 (Rank one Tensor)**.** Given d-vectors $\mathbf{x}_j \in \mathbb{C}^{I_j}$ for $j \in [d]$, the outer-product

$$\mathcal{X} = \bigcirc_{j=0}^{d-1} \mathbf{x}_j \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$$

has entries given the product of corresponding entries of the vectors, i.e.

$$\mathcal{X}_{i_0,\ldots,i_{d-1}} = \left(\bigcirc_{j=0}^{d-1} \mathbf{x}_j \right)_{i_0,\ldots,i_{d-1}} = (\mathbf{x}_0)_{i_0} (\mathbf{x}_1)_{i_1} \ldots (\mathbf{x}_{d-1})_{i_{d-1}}$$

any d-mode tensor where it is possible to write it as such an outer product of d vectors is a rank one tensor.

Note that storing a rank one tensor means storing only the vector components, rather than all entries. This definition in the 2-mode case is the familiar rank one matrix case, for $\mathbf{u} \in \mathbb{C}^m, \mathbf{v} \in \mathbb{C}^N$

$$A = \mathbf{u} \circ \mathbf{v} = \mathbf{u} v^*$$

then matrix $A \in \mathbb{C}^{m \times N}$ is a rank one matrix.

Definition 1.3.18. $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ is a r tensor if it can be written as the sum of r rank one tensors, that is

$$\mathcal{A} = \sum_{\ell=0}^{r-1} \left(\bigcirc_{j=0}^{d-1} \mathbf{x}_j^{(\ell)} \right)$$

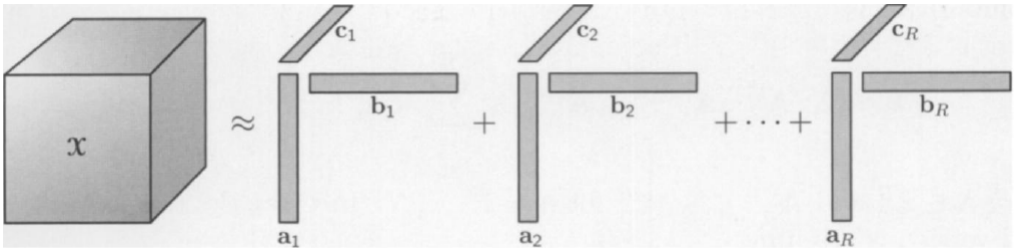


Figure 1.2: Schematically a rank R decomposition for a 3-mode tensor $\mathcal{X}$ as seen in [23]

Note that unlike the PCA example given above, this decomposition does not require us to consider a set of tensors; a single tensor will be decomposable in this fashion. Furthermore, each of the basis tensors in this case does have a simple structure - it can be stored as d vectors and so takes up $\mathcal{O}\left(r \sum_{j=0}^{d-1} I_j\right)$-space.

So with this definition, we ask then how, given a tensor $\mathcal{A}$ can we find its rank r decomposition. How to select or determine r is a question we will set aside for the time being.

Given $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$ we want to find a rank r approximation as

$$\arg \min_{\mathbf{x}_j^{(\ell)} \in \mathbb{C}^{I_j}, j \in [d], \ell \in [r]} \| \mathcal{A} - \sum_{\ell=1}^{r} \left(\bigcirc_{j=0}^{d-1} \mathbf{x}_j^{(\ell)} \right) \|$$

In the generic case, the above optimization problem is difficult. However, the base case of $d = 2$ leads to consider a method which in practice can yield good results, though its gaurauntees are not well understood.

In the event that $d = 2$ then the optimization problem is equivalent to

$$\min_{\alpha_\ell} \| A - \sum_{\ell=0}^{r-1} \alpha_l \mathbf{u}_\ell \mathbf{v}_\ell^* \|_F = \sqrt{\sum_{j=r}^{N-1} \sigma_j(A)}$$

where $\alpha_\ell = \sigma_\ell(A)$ and $\mathbf{u}_\ell, \mathbf{v}_\ell$ are the singular vectors of A. That is, the best rank r approximation to a matrix A is given by the leading r factors from the SVD.

So the idea then for our tensor decomposition is to reduce to the $d = 2$ case. Suppose for the time being that we have $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$, we know it is rank r and we know all but the first mode vectors in each of the r rank 1 factors. That is, we have

$$\bigotimes_{j>0}^{d-1} \mathbf{x}_j^{(\ell)}$$

for all $\ell \in [r]$. With this (mostly) complete factorization for $\mathcal{A}$, we can find the missing mode by solving a least squares problem.

So,

$$\mathcal{A} = \sum_{\ell=0}^{r-1} \left(\bigotimes_{j=0}^{d-1} \mathbf{x}_j^{(\ell)} \right)$$

Now consider the subtensor $\mathcal{A}_{([d]\setminus\{0\}, i_0)}$. This is the tensor found by fixing an index in the 0-th mode and varying all other indices. Naturally then there are I_0 such subtensors. For any $i_0 \in I_0$ the entries of the subtensor $\mathcal{A}_{([d]\setminus\{0\}, i_0)}$ are equal to

$$\sum_{\ell=0}^{r-1} \left(\mathbf{x}_0^{(\ell)} \right)_{i_0} \bigotimes_{j>0}^{d-1} \mathbf{x}_j^{(\ell)}$$

That is, we have a sum of products of scalar unknowns with $d-2$ outer product - and so after a careful rearrangement of elements, we will have a linear system:

$$\left[\begin{array}{c|c|c|c|c} & & & & \\ \operatorname{vec}\left(\bigcirc_{j=0}^{d-1}\mathbf{x}_j^{(0)}\right) & \operatorname{vec}\left(\bigcirc_{j=0}^{d-1}\mathbf{x}_j^{(1)}\right) & \dots & \operatorname{vec}\left(\bigcirc_{j=0}^{d-1}\mathbf{x}_j^{(r-1)}\right) & \end{array}\right]\left[\begin{array}{c} \left(\mathbf{x}_0^{(0)}\right)_{i_0} \\ \left(\mathbf{x}_0^{(1)}\right)_{i_0} \\ \vdots \\ \left(\mathbf{x}_0^{(0)}\right)_{i_0} \end{array}\right] = \dots$$

$$\left[\begin{array}{c} \\ \operatorname{vec}\left(\mathcal{A}_{([d]\setminus\{0\},i_0)}\right) \\ \\ \end{array}\right]$$

Let us denote B_0 as the $\prod_{j=1}^{d-1} I_j \times r$ matrix formed by using the vectorized $d-1$-mode outer products as columns. Note that $\mathcal{A}_{([d]\setminus\{0\},i_0)} = \mathcal{A}^{(k)}_{i_0,:}$, that is the vectorized subtensor is equal to the i_0-th row of the 0-mode unfolding of $\mathcal{A}$

So, in order to solve the missing unknown, we have I_0 overdetermined linear systems of the form $A^{(0)}_{:,i_0} = B_0\mathbf{x}$ to solve in order find all the unknowns. i.e. denoting $\mathbf{x}_\ell = \left(\mathbf{x}^{(\ell)_0}\right)_{i_0}$ for $\ell \in [r]$

$$\mathbf{x} = \arg\min_{\mathbf{y}\in\mathbb{C}^r} \|\mathbf{b} - B_0\mathbf{y}\|_2$$

Solving this for all $i_0 \in [I_0]$.

This can be formulated equivalently as follows

$$\left(\mathcal{A}^{(k)}\right)^T = B_k \left[\begin{array}{c|c|c|c} \mathbf{x}_k^{(0)} & \mathbf{x}_k^{(1)} & \dots & \mathbf{x}_k^{(r-1)} \end{array}\right]^T$$

where we have combined the I_k different vector least square fitting problems into one matrix least square fitting problem of the form

$$\arg\min_{X\in\mathbb{C}^{r\times I_k}} \left\| \left(\mathcal{A}^{(k)}\right)^T - B_k X \right\|_F^2$$

that is X will solve for all the k-mode missing factor vectors.

With this in hand, we are now ready to address the question of how to obtain the complete factorization of an arbitrary rank r tensor – our proceeding formulation only addressed how to find the k-mode missing factor vectors supposing all the other $r(d-1)$ factor vectors were known.

Algorithm 1.3.1 Alternating Least Squares Minimization

Input: $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}$
Output: $\left\{\mathbf{x}_j^{(\ell)}\right\}_{j\in[d],\ell\in[d]}$
Initialize $\left\{\mathbf{x}_j^{(\ell)}\right\}_{j\in[d],\ell\in[r]}$ randomly
for $i = 1$ to maximum iterations **do**
  for $k = 0$ to $d-1$ **do**
    $\left[\ \mathbf{x}_k^{(0)} \ \middle|\ \mathbf{x}_k^{(1)} \ \middle|\ \ldots \ \middle|\ \mathbf{x}_k^{(r-1)}\ \right] \leftarrow \arg\min_{X\in\mathbb{C}^{r\times I_k}} \left\|\left(\mathcal{A}^{(k)}\right)^T - B_k X\right\|_F^2$
  end for
end for
return $\left\{\mathbf{x}_j^{(\ell)}\right\}_{j\in[d],\ell\in[d]}$

Note that the above algorithm requires the solution of (d)`max_iterations` overdetermined least square problems - a potential bottleneck which can be mitigated by using fast approximate least square methods like the one described in Theorem 3.3.4.

Also note that the algorithm is a greedy algorithm an - its convergence properties are not well understood nor does it guarantee any type of global optimality.

Next we turn to another important Tensor decomposition method.

1.3.3 Tucker Rank and Decomposition

Definition 1.3.19. A tensor $\mathcal{A} \in \mathbb{C}^{I_0\times\ldots I_{d-1}}$ has $(r_1,\ldots,r_{d-1})$-Tucker rank if there exists a core tensor $\mathcal{C} \in \mathbb{C}^{r_0\times\ldots r_{d-1}}$ and matrices $U_j \in \mathbb{C}^{I_j\times r_j}$, $\forall j \in [d]$ such that

$$\mathcal{A} = \mathcal{C} \bigtimes_{j=0}^{d-1} U_j$$

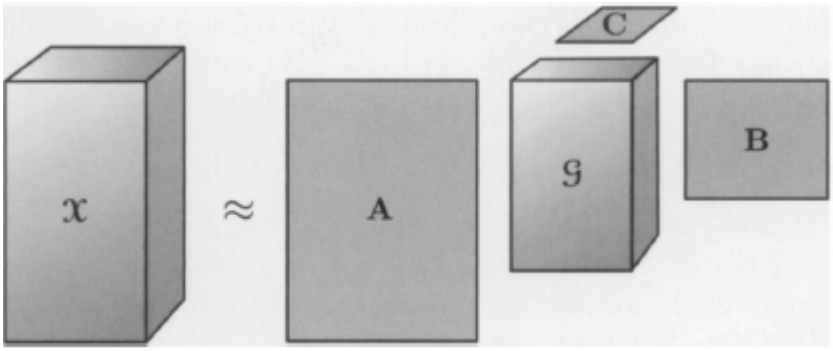


Figure 1.3: Schematically a Tucker decomposition for a 3-mode tensor $\mathcal{X}$ with core tensor $\mathcal{G}$ and factor matrices A, B, C as seen in [23]

Note the space requirement to store a Tucker decomposition of a tensor is $\mathcal{O}\left(\prod_{j=0}^{d-1} r_j + \sum_{j=0}^{d-1} I_j r_j\right)$ where the first term accounts for all the entries of the core tensor and the second term accounts for all entries of the factor matrices. Recall that $\mathcal{O}\left(\prod_{j=0}^{d-1} I_j\right)$ space is required to store the unfactored tensor, and so in the event that the Tucker rank is appreciably smaller than the original mode for at least some of the modes, the Tucker decomposition will occupy significantly less space.

Note that as a convention, U_j can be taken to have orthonormal columns - by orthonormalizing U_j, we can suitably alter $\mathcal{C}$.

To approximate a tensor with a low Tucker rank representation, we can solve the following optimiziation problem

$$\underset{\substack{\mathcal{C}\in\mathbb{C}^{r_0\times\cdots\times r_{d-1}} \\ \{U_j\}_{j\in[d]}, U_j^* U_j = I}}{\arg\inf} \|\mathcal{A} - \mathcal{C} \bigtimes_{j=0}^{d-1} U_j\|_2^2$$

when the number of modes is larger than 2, this optimization problem is difficult to solve. One approach is to use the SVD of each of the d different unfoldings of $\mathcal{A}$

Algorithm 1.3.2 Higher Order SVD

Input: $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}, (r_0, \ldots, r_{d-1})$
Output: $\mathcal{C}, \{U_j\}_{j \in [d]}$
Compute r_j-truncated SVD of mode-j unfolding of $\mathcal{A}$

$$A^{(j)} = U_j \Sigma_j V_j^*, \forall j \in [d]$$

$\mathcal{C} \leftarrow \mathcal{A} \bigtimes_{j=0}^{d-1} U_j^* \in \mathbb{C}^{r_0 \times \cdots \times r_{d-1}}$
return $\mathcal{C}, \{U_j\}_{j \in [d]}$

Note that for each unfolding, the full SVD has form

$$A^{(j)} = \underbrace{U}_{I_j \times I_j} \underbrace{\Sigma}_{I_j \times \prod_{j \neq k}} \underset{\prod_{j \neq k}}{I_k} \underbrace{V^*}_{I_k \times \prod_{j \neq k}} \underset{}{I_k}$$

the r_j-truncated SVD has form

$$A^{(j)} \approx \underbrace{U}_{I_j \times r_j} \underbrace{\Sigma}_{r_j \times r_j} I_k \underbrace{V^*}_{r_j \times \prod_{j \neq k}} I_k$$

This problem then is repeated for each of the d different modes. This is potentially a bottleneck computationally and so can likely benefit from fast approximations to the SVD as described in algorithm 3.3.1.

Now, to further improve the decomposition, we can take an approach like 1.3.1 and alternate, successively solving for U_j and iterate on this processes.

Algorithm 1.3.3 Higher Order Orthogonal Iteration

Input: $\mathcal{A} \in \mathbb{C}^{I_0 \times \cdots \times I_{d-1}}, (r_0, \ldots, r_{d-1})$
Output: $\mathcal{C}, \{U_j\}_{j \in [d]}$
Initalize $\left\{U_j^{(0)}\right\} \leftarrow HOSVD(\mathcal{A})$
for $i = 1$ to M **do**
  $\forall j$ update $U_j^{(i-1)}$ by computing

$$\left(A \bigtimes_{k \neq j} \left(U_k^{(i-1)}\right)^*\right)^{(j)} = U_j^{(i)} \Sigma_j^{(i)} \left(V^{(i)}\right)_j^*$$

end for
$\mathcal{C} \leftarrow A \bigtimes_{k \neq j} \left(U_k^{(M)}\right)^*$
return $\mathcal{C}, \left\{U_j^{(M)}\right\}_{j \in [d]}$

Next we will show one way in the Tucker and CP rank relate. First though we note a Lemma which shows that how the mode-k product of a CP rank one tensor with a matrix U can be expressed as another rank one tensor.

Lemma 1.3.20. *Let $\mathbf{x}_j \in \mathbb{C}^{I_j}$, $U_k \in \mathbb{C}^{m_j \times I_j}$ for all $j \in [d]$ then*

$$\left(\bigcirc_{j=0}^{d-1} \mathbf{x}_j\right) \times_k U_k = \left(\bigcirc_{j=0}^{k-1} \mathbf{x}_j\right) \bigcirc U_k \mathbf{x}_k \bigcirc \left(\bigcirc_{j=k+1}^{d-1} \mathbf{x}_j\right)$$

Proof. Note that the k-mode fibers the tensor $\left(\bigcirc_{j=0}^{d-1} \mathbf{x}_j\right)$ are scalar multiples of the same vector, $\mathbf{x}_k$

That is, the k-mode fiber indexed by $(\ell_0, \ldots, \ell_{k-1}, \ell k + 1, \ldots, \ell_{d-1})$ is

$$\left(\prod_{j \neq k}^{d-1} (\mathbf{x}_j)_{\ell_j}\right) \mathbf{x}_k$$

but $\left(\prod_{j \neq k}^{d-1} (\mathbf{x}_j)_{\ell_j}\right)$ is a scalar. So the identity follows now from noting the definition of the mode-k product. (Each column of the unfolding is a scalar multiple of the same vector, scalar commutes with matrix-vector multiplication) □

Theorem 1.3.21. *If $\mathcal{A} \in \mathbb{C}^{I_0 \times \ldots I_{d-1}}$ has Tucker rank $(r_0, \ldots, r_{d-1})$ then it has CP rank of at most $\prod_{j=0}^{d-1} r_j$*

Proof. The tensor has an exact Tucker decomposition, so

$$\mathcal{A} = \mathcal{C} \bigtimes_{j=0}^{d-1} U_j^*$$

Note that we can express any tensor in the standard basis; here the standard basis for tensors is a tensor with only one non-zero entry.

So for example for some $\ell \in [r_0] \times \cdots \times [r_{d-1}]$ the associated standard basis element is

$$\bigcirc_{j=0}^{d-1} \mathbf{e}_{\ell_j}$$

where $\mathbf{e}_{\ell_j}$ is the usual standard basis vector in $\mathbb{C}^{r_j}$. Denote $\mathcal{I} = [r_0] \times \cdots \times [r_{d-1}]$. Thus

$$\mathcal{C} = \sum_{\ell \in \mathcal{I}} \mathcal{C}_\ell \left(\bigcirc_{j=0}^{d-1} \mathbf{e}_{\ell_j} \right)$$

Now use this expression in the Tucker decomposition of $\mathcal{A}$

$$\begin{aligned}
\mathcal{A} &= \mathcal{C} \bigtimes_{j=0}^{d-1} U_j^* \\
&= \left[\sum_{\ell \in \mathcal{I}} \mathcal{C}_\ell \left(\bigcirc_{j=0}^{d-1} \mathbf{e}_{\ell_j} \right) \right] \bigtimes_{j=0}^{d-1} U_j^* \\
&= \sum_{\ell \in \mathcal{I}} \mathcal{C}_\ell \left(\bigcirc_{j=0}^{d-1} U_j^* \mathbf{e}_{\ell_j} \right) \\
&= \sum_{\ell \in \mathcal{I}} \mathcal{C}_\ell \bigcirc_{j=0}^{d-1} (U_j^*)_{\ell_j}
\end{aligned}$$

where we have used Lemma 1.3.20, and denoted the ℓ_j-th column of U_j as $(U_j)_{\ell_j}$. We therefore have the sum of rank one tensors. There are $\prod_{j=0}^{d-1} r_j$ possible values for ℓ and so we have a CP decomposition of that rank. This provides an upper bound on CP rank, since the decomposition may not be optimal. $\square$

Chapter II

A Review of Introductory Probability with Some Algorithmic Applications (CMSE 890 Lectures 2 – 4)

2.1 Probability Densities and Random Variables (CMSE 890 Lecture 2)

Definition 2.1.1. A non-negative function $p : \mathbb{R}^n \to [0, \infty)$ for which

$$\int_{\mathbb{R}^n} p(x)dx = 1$$

is a probability density function (pdf)

Definition 2.1.2 (Random Variable)**.** A random variable $X \in \mathbb{R}^n$ with probability density p represents a value in $\mathbb{R}^n$. It takes a particular value in set $S \subset \mathbb{R}^n$ with probability

$$\mathbb{P}\left[X \in S\right] = \mathbb{P}_X\left[S\right] = \int_S p(x)dx \in [0, 1]$$

Theorem 2.1.3 (Union Bound)**.** *Let $X \in \mathbb{R}^n$ be a random variable with probability density function p then*

$$\mathbb{P}_X\left[\bigcup_{j=1}^{k} S_j\right] \leq \sum_{j=1}^{k} \mathbb{P}_X(S_j),\ \forall k \geq 1,\ S_j \subset \mathbb{R}^n$$

equality holds when S_j are mutually disjoint

Proof. We proceed by induction. The base case is trivially true, $\mathbb{P}(S_1) \leq \mathbb{P}(S_1)$. The case for $k-1$ is thus

$$\mathbb{P}_X\left[\bigcup_{j=1}^{k-1} S_j\right] \leq \sum_{j=1}^{k-1} \mathbb{P}_X(S_j)$$

Note that for any two sets $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. So

$$\mathbb{P}_X\left[\bigcup_{j=1}^{k} S_j\right] = \mathbb{P}_X\left[\bigcup_{j=1}^{k-1} S_j\right] + \mathbb{P}_X\left[S_k\right] - \mathbb{P}_X\left[S_k \cap \left(\bigcup_{j=1}^{k-1} S_j\right)\right]$$

All probabilities are non-negative, so we have

$$\begin{aligned}\mathbb{P}_X\left[\bigcup_{j=1}^{k} S_j\right] &\leq \mathbb{P}_X\left[\bigcup_{j=1}^{k-1} S_j\right] + \mathbb{P}_X\left[S_k\right] \\ &\leq \sum_{j=1}^{k-1} \mathbb{P}_X(S_j) + \mathbb{P}_X(S_k) \\ &= \sum_{j=1}^{k} \mathbb{P}_X(S_j)\end{aligned}$$

□

Remark 2.1.4. We have suppressed details about the measurability of sets in theorem 2.1.3; which will in general not concern us, and we will assume that sets are measurable. Additionally, the result does hold for countably infinite sets, though our use for the theorem needs concern only finite sets.

Example 2.1.5 (Gaussian pdf)**.** The single valued gaussian or single valued normal pdf g is defined as

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

for mean μ and variance σ^2. When $\mu = 0$ and $\sigma^2 = 1$ we say g is the standard gaussian distribution.

Example 2.1.6 (Multivariate Gaussian)**.** The multivariate Gaussian or multivariate normal (MVN) is a widely used joint probability density function. The pdf for a MVN in D dimensions is

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

where $\mu = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^D$ is the mean vector and $\mathbf{\Sigma} = \text{Cov}[\mathbf{x}] \in \mathbb{R}^{D\times D}$ is the covariance matrix. The covariance matrix is a symmetric, positive definite matrix defined by

$$\text{Cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T\right]$$

when $\mu = \mathbf{0}$ and $\mathbf{\Sigma} = I$ we say the distribution is the standard multivariate Guassian.

Example 2.1.7 (Standard Folded Normal)**.** The standard folded normal distribution is denoted by $|X|$, where $X \sim \mathcal{N}(0, 1)$. The pdf is defined in the following way, for $[a, b] \in [0, \infty)$ we have

$$\mathbb{P}_{|X|}\left[[a, b]\right] = \mathbb{P}_X\left[[a, b] \cup [-b, -a]\right] = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx + \frac{1}{\sqrt{2\pi}}\int_{-b}^{-a} e^{-\frac{x^2}{2}}dx$$

and the pdf is zero for negative values.

Next we state the definition of a dirac delta function, which we will use to describe discrete probability distributions using the same framework as the continuous case.

Definition 2.1.8 (Dirac Delta)**.** A dirac delta is a generalized function denoted by $\delta(x)$. We take it as part of the definition that for any smooth function $f : \mathbb{R}^n \to \mathbb{R}$, (referred to as a test function in the theory of distributions) and any $c \in \mathbb{R}^n$ the following holds

$$\int_S f(x)\delta(x - c)dx = \begin{cases} f(c) & c \in S \\ 0 & c \notin S \end{cases}$$

Though we will often refer to δ as a function, it is in fact a generalized function and not a standard function. It is only well defined in terms of how it affects other functions when integrated against them.

So, to define a discrete probability distribution we start with a finite set of values $\{a_j\}_{j=1}^m$ where $a_j \in (0, 1]$ which satisfy

$$\sum_{j=1}^m a_j = 1$$

and then define the distribution p as

$$p(x) = \sum_{j=1}^{m} a_j \delta(x - c_j)$$

usually we understand a_j as $\mathbb{P}[X = c_j]$ for the associated discrete random variable X. Using this notation for example to find $\mathbb{P}_X[S]$ for some $S \subset \mathbb{C}^N$ we have

$$\begin{aligned}
\mathbb{P}_X[S] &= \int_S \sum_{j=1}^{m} \mathbb{P}[X = c_j]\, \delta(x - c_j) dx \\
&= \sum_{j=1}^{m} \int_S \mathbb{P}[X = c_j]\, \delta(x - c_j) dx \\
&= \sum_{j=1}^{m} \mathbb{P}[X = c_j]
\end{aligned}$$

using the definition 2.1.8 after interchanging summation and integration.

Example 2.1.9 (Fair Coin Discrete)**.** Let X be the value of a fair flipped coin where a heads represents a value of one and a tails represents zero. The probability density is given by

$$p(x) = \frac{1}{2}\delta(x - 0) + \frac{1}{2}\delta(x - 1) = \mathbb{P}[X = 0]\, \delta(x - 0) + \mathbb{P}[X = 1]\, \delta(x - 1)$$

Homework 2.1.1. Show that g from Example 2.1.5 is a probability density function.

Homework 2.1.2. Write the probability distribution for the result of a fair six-sided die roll.

2.2 Independence

Definition 2.2.1 (Independence)**.** Two random variables $X_1 \in \mathbb{R}^m$ and $X_2 \in \mathbb{R}^n$ are independent, denoted $X_1 \perp X_2$, if the joint distribution for all $\mathbf{x} \in X_1, \mathbf{y} \in X_2$ can be written as a product of marginals. That is, for $p : \mathbb{R}^m \times \mathbb{R}^n \to [0, \infty)$

$$p(\mathbf{x}, \mathbf{y}) = p(x_1, \ldots, x_m, y_1, \ldots, y_n) = p_1(x_1, \ldots, x_m) p_2(y_1, \ldots, y_n) = p(\mathbf{x}) p(\mathbf{y})$$

Example 2.2.2 (Independent Fair Coin Tosses)**.** Suppose $X_1 \in \{0,1\}$ and $X_2 \in \{0,1\}$ are the random variables corresponding to the values of two different coin tosses. If the coin tosses are independent, then

$$p(X_1 = 1, X_2 = 0) = p_1(X_1 = 1)p_2(X_2 = 0)$$

Definition 2.2.3. Two random variables $X_1, X_2 \in \mathbb{R}^n$ are identically distributed if their corresponding densities p_1, p_2 satisfy $p_1 = p_2$.

If two variables are independent and identically distributed we abbreviate as i.i.d.

Example 2.2.4 (Multivariate Gaussian as Vector of i.i.d Gaussian Random Variables)**.** Two entries $x_\ell, x_j \in \mathbb{R}$ of $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$ are independent whenever $\ell \neq j$ since

$$\begin{aligned} p(\mathbf{x}) &= p(x_1, \ldots, x_N) \\ &= \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2}\right) \\ &= \prod_{j=1}^{N} \left[\frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{x_j^2}{2}\right)\right] \\ &= \prod_{j=1}^{N} p_j(x_j) \end{aligned}$$

where p_j refers to the standard normal for a single value, x_j

Definition 2.2.5. the expectation of a random variable $X \in \mathbb{R}^n$ is defined as

$$\mathbb{E}[X] = \int_{\mathbb{R}^n} xp(x)dx$$

Theorem 2.2.6 (Linearity of Expectation)**.** *Let $\alpha, \beta \in \mathbb{R}$ and $X, Y \in \mathbb{R}^n$ be random variables. Then*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$$

Proof. Using definition 2.2.5 we have

$$\begin{aligned}
\mathbb{E}[\alpha X + \beta Y] &= \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} (\alpha x + \beta y)\, p(x,y)dxdy \\
&= \int_{\mathbb{R}^n} \alpha x \left(\int_{\mathbb{R}^n} p(x,y)dy\right) dx + \int_{\mathbb{R}^n} \beta y \left(\int_{\mathbb{R}^n} p(x,y)dx\right) dy \\
&= \alpha \int_{\mathbb{R}^n} x p_1(x)dx + \beta \int_{\mathbb{R}^n} y p_2(y)dy \\
&= \alpha \mathbb{E}\left[X\right] + \beta \mathbb{E}\left[Y\right]
\end{aligned}$$

Where we have used Fubini's theorem to write the iterated integral for $\mathbb{R}^n \times \mathbb{R}^n$ and to change the order of integration to recover the marginals. □

Note that the above theorem is stated with no consideration for the independence of X_1 and X_2.

Note that for $X_1 \in \mathbb{R}^m$ and $X_2 \in \mathbb{R}^n$ with densities p_1, p_2 then the joint density $p : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ has the property that

$$\begin{aligned}
p_1(x) &= \int_{\mathbb{R}^n} p(x,y)dy \\
p_2(x) &= \int_{\mathbb{R}^m} p(x,y)dx
\end{aligned}$$

As a matter of terminology, when we want to emphasize their relationship to the joint distribution, we call p_1 and p_2 the marginal distributions for p.

Functions of random variables are themselves random variables, whose probabilities and expectation may be computed by integrating over the corresponding pre-image. So if $X \in \mathbb{R}^n$ is a random variable with pdf p and $f : \mathbb{R}^n \to \mathbb{R}^m$, then $f(X)$ is a random variable where for $S \subset \mathbb{R}^m$

$$\mathbb{P}_{f(X)}[S] = \mathbb{P}_X[f^{-1}(S)]$$

here $f^{-1}(S) = \{x \in \mathbb{R}^n | f(x) \in S\}$. Note that f may or may not be invertible, the set we've given is well defined regardless. To compute expectation of a function of a random variable, we follow the definition and write

$$\mathbb{E}\left[f(X)\right] = \int_{R^n} f(x)p(x)dx$$

Theorem 2.2.7 (Markov's inequality)**.** *Let $f : \mathbb{R}^n \to [0, \infty)$, $a \geq 0$ and $X \in \mathbb{R}^n$ by a random vector for $n \geq 1$. Then*

$$P_{f(X)}\left[[a, \infty]\right] = P\left[f(X) \geq a\right] \leq \frac{\mathbb{E}\left[f(X)\right]}{a}$$

Proof. We restrict the domain of integration and use a lower-bound on the function to achieve the inequality

$$\begin{aligned}
\mathbb{E}\left[f(X)\right] &= \int_{\mathbb{R}^n} f(y)p(y)dy \\
&\geq \int_{\underbrace{\{y|f(y) \geq a\}}_{f^{-1}([a,\infty))}} f(y)p(y)dt \\
&\geq \int_{f^{-1}([a,\infty))} ap(y)dt \\
&= a\int_{f^{-1}([a,\infty))} p(y)dt \\
&= aP_{f(X)}\left[[a, \infty)\right]
\end{aligned}$$

a re-arrangement of terms then yields the desired result. □

Using the Markov inequality we can quantify the following claim: A positive random variable is unlikely to deviate too much from its expectation. To see why, consider $X \in \mathbb{R}^+$ and a function that does not change the value of X, $f(X) = |X|$ and then using the constant $a\mathbb{E}[X]$ in from the Markov Inequality, we see $\forall a \geq$

$$P\left[X \geq a\mathbb{E}[X]\right] \leq \frac{\cancel{\mathbb{E}[x]}}{\cancel{a\mathbb{E}[X]}}\frac{1}{a}$$

Now we turn to some useful bounds associated with the variance of random variables.

Definition 2.2.8. Let $X \in \mathbb{R}^n$. Then $\mathrm{Var}\left[X\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)\left(X - \mathbb{E}\left[X\right]\right)^T\right]$ is a matrix in general (known as the covariance matrix). When $n = 1$ this simplifies to $\mathrm{Var}\left[X\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]$

Theorem 2.2.9 (Chebyshev's Inequality)**.** *Let $X \in \mathbb{R}$ have $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var[X] > 0$. Then $\forall k \geq 0$*

$$P\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

Proof. Using the Markov inequality on the positive random variable $|X - \mu|^2$ for constant $a = k^2\sigma^2$, and the definition of variance we obtain

$$\begin{aligned} P\left[|X - \mu| \geq k\sigma\right] &= P\left[|X - \mu|^2 \geq k^2\sigma^2\right] \\ &\leq \frac{\mathbb{E}\left[|X - \mu|^2\right]}{k^2\sigma^2} \\ &= \frac{\mathbb{E}\left[(X - \mu)^2\right]}{k^2\sigma^2} \\ &= \frac{\mathrm{Var}\left[X\right]}{k^2\sigma^2} \\ &= \frac{1}{k^2} \end{aligned}$$

□

Now we turn to a key fact about random variables and controlling variance: averaging across several i.i.d. copies of a random variable decreases the variance. We make this precise in the following theorem

Theorem 2.2.10. *Let $X_1, \ldots, X_n \in \mathbb{R}$ be i.i.d. random variables with expectation μ and variance σ^2. Then*

$$Var\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{\sigma^2}{n}, \quad \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \mu$$

Proof. By the linearity of expectation, the second result is immediate:

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[X_j\right] = \frac{n\mu}{n} = \mu$$

For variance we use the equality $\mathrm{Var}\left[X\right] = \mathbb{E}[X^2] - \left(E[X]\right)^2$ on the sum of random variables:

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2\right] - \left(\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right]\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{j=1}^{n} X_j^2 + \sum_{j\neq k} X_j X_k\right] - \mu^2 \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left[X_j^2\right] + \frac{1}{n^2}\sum_{j\neq k}\mathbb{E}\left[X_j X_k\right] - \frac{1}{n}\mu^2 - \frac{(n-1)}{n}\mu^2 \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_j^2\right] - \mu^2\right) + \frac{1}{n^2}\sum_{j\neq k}\mathbb{E}\left[X_j\right]\mathbb{E}\left[X_k\right] - \frac{(n-1)}{n}\mu^2 \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\left(\sigma^2\right) + \frac{1}{n^2}\sum_{j\neq k}\mu^2 - \frac{(n-1)}{n}\mu^2 \\
&= \frac{\sigma^2}{n} + \frac{n^2-n}{n^2}\mu^2 - \frac{(n-1)}{n}\mu^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

where we have used linearity of expectation, and independence of the variables. □

Before we proceed to the next useful probability bound, we recall some inequalities from calculus which will feature in the sequel

Fact 2.2.11. *For all $x \in \mathbb{R}$,*

$$1 + x \leq e^x$$

Which we can in turn use to see that for $c \in \mathbb{R}$:

$$\left(1 + \frac{c}{x}\right) \leq e^{c/x} \implies \left(1 + \frac{c}{x}\right)^x \leq e^c$$

Fact 2.2.12. *For all $y \in (0, \infty)$,*

$$y \ln y \geq y - 1$$

Using this then, we have the following result: Let $x \geq a > 0$ and $c \in (-a, \infty)$. Then

$$\left(1 + \frac{c}{a}\right)^a \leq \left(1 + \frac{c}{x}\right)^x$$

With these in hand, we turn now yet another bound which relates how a sum of random variables deviates from its mean.

Theorem 2.2.13 (Chernoff Inequality)**.** *Let $I_1, \ldots, I_n \in \mathbb{R}$ by independent discrete random variables with probability densities $p_j(x) = \lambda_j \delta(x-1) + (1-\lambda_j)\delta(x)$ where $\lambda_j \in (0,1)$. Let $Y = \sum_{j=1}^n I_j$ with $\mu = \mathbb{E}[Y] = \sum_{j=1}^n \lambda_j$. Then, for $w \in (0,1)$ we have*

$$P\left[Y < (1-w)\mu\right] < \left[\frac{e^{-w}}{(1-w)^{(1-w)}}\right]^\mu$$

Proof. Suppose $t > 0$, then the following are equivalent

$$P\left[Y < (1-w)\mu\right] = P\left[-tY > -t(1-w)\mu\right] = P\left[e^{-tY} > e^{-t(1-w)\mu}\right]$$

We now apply the Markov inequality to the right hand side to obtain

$$P\left[Y < (1-w)\mu\right] \leq \frac{\mathbb{E}\left[\exp\left(-tY\right)\right]}{\exp\left(-t(1-w)\mu\right)}$$

Note that since the random variables are independent, we can write the sum in the exponent as a product:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-tY\right)\right] &= \mathbb{E}\left[\exp\left(-t\sum_{j=1}^{n} I_j\right)\right] = \\
&= \prod_{j=1}^{n} \mathbb{E}\left[\exp\left(-tI_j\right)\right] \\
&= \prod_{j=1}^{n} \int_{\mathbb{R}} \exp\left(-tI_j\right)\left[\lambda_j delta(x-1) + (1-\lambda_j)\delta(x)\right] dx \\
&= \prod_{j=1}^{n} \left[\lambda_j \exp\left(-t\right) + (1-\lambda_j)\exp\left(-t(0)\right)\right] \\
&= \prod_{j=1}^{n} \left[1 + \lambda_j\left(e^{-t}-1\right)\right] \\
&\leq \prod_{j=1}^{n} e^{-\lambda_j(e^{-t}-1)} \\
&= e^{-\sum_{j=1}^{n}\lambda_j(e^{-t}-1)} \\
&= e^{\mu(e^{-t}-1)}
\end{aligned}$$

where we have used 2.2.11 where $x \leftarrow \lambda_j(e^{-t}-1)$. Combining this with our earlier Markov inequality result, and substituting $t = -\ln(1-w)$ we have obtain

$$\begin{aligned}
P\left[Y < (1-w)\mu\right] &\leq \frac{e^{\mu(e^{-t}-1)}}{\exp\left(-t(1-w)\mu\right)} \\
&= \frac{e^{\mu(e^{\ln(1-w)}-1)}}{\exp\left(\ln(1-w)(1-w)\mu\right)} \\
&= \left(\frac{e^{-w}}{(1-w)^{(1-w)}}\right)^{\mu}
\end{aligned}$$

which is our desired result. □

Homework 2.2.1. Suppose p_1 and p_2 are pdfs of independent random variables on $\mathbb{R}^m$ and R^n. Prove that their product is always a pdf in $R^m \times R^n$.

Homework 2.2.2. Consider

1. Suppose that $X \sim \mathcal{N}(\mu, \sigma)$. Show $\mathbb{E}[X] = \mu$

2. Suppose that X is the random variable of Example 2.1.9. Show $\mathbb{E}[X] = 1/2$

Homework 2.2.3. Show that if $X, Y \in \mathbb{R}$ are independent random variables then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Homework 2.2.4. Show that $\mathrm{Var}[X] = \mathbb{E}[XX^T] - \mathbb{E}[X]\left(\mathbb{E}[X]\right)^T$ and that for $n = 1$, $\mathrm{Var}[X] = \mathbb{E}[X^2] - \left(E[X]\right)^2$

Homework 2.2.5. Show that if $X \in \mathbb{R}$ is a gaussian random variable then $\mathrm{Var}[X] = \sigma^2$. Show that if $X \in \{0, 1\}$, is uniform random variable (e.g. fair coin flip) then $\mathrm{Var}[X] = \frac{1}{4}$.

Homework 2.2.6. If $a \in \mathbb{R}$ show that $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$.

Homework 2.2.7 (Strong Law of Large Numbers)**.** Use the Chebyshev inequality to argue that for any set of i.i.d. random variables $\{X_j\}_{j=1}^n$ with finite mean μ and variance then

$$P\left[\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} X_j = \mu\right] = 1$$

2.3 Monte Carlo Integration and Median of Means

We now turn to an algorithmic application involving these probabilistic ideas and results: Monte Carlo integration. This is useful in cases where f has no closed form expression.

Let $f : [0,1]^N \to \mathbb{R}$. Choose $X_1, \ldots, X_m \in [0,1]^N$ are i.i.d uniform random variables. We seek to estimate the integral,

$$J = \int_{[0,1]^N} f(x)dx$$

To begin estimating this, we introduce the random variable Z, which is the sum of function evaluations at the i.i.d. uniform points X_j

$$Z = \frac{1}{m}\sum_{j=1}^{m} f(X_j)$$

By Theorem 2.2.10 the mean of the sum is unchanged from each of the addends

$$\mathbb{E}[Z] = \mathbb{E}[f(X_j)] = \int_{[0,1]^n} f(y)dy \tag{2.1}$$

That is, $J = \mathbb{E}[Z]$. Furthermore,

(2.2)

$$\mathrm{Var}[Z] = \frac{1}{m}\mathrm{Var}[f(X_j)] = \frac{1}{m}\mathbb{E}\left[f(X_j)^2 - J^2\right] \leq \frac{1}{m}\mathbb{E}\left[f(X_j)^2\right] = \frac{1}{m}\int_{[0,1]^n} |f(y)|^2 dy = \frac{1}{m}\|f\|_2^2$$

Lemma 2.3.1. *Choose $\epsilon > 0$. If $m \geq \frac{10}{\epsilon^2}$ then*

$$P[|z - J| < \epsilon\|f\|_2] \geq 0.9$$

where

$$J = \int_{[0,1]^N} f(x)dx$$

Proof. Note from complementary events we have

$$P[|z - J| < \epsilon\|f\|_2] = 1 - P[|z - J| \geq \epsilon\|f\|_2]$$

As we saw in the previous discussion, we have that $|z - J| = |z - \mathbb{E}[z]|$ and $\|f\|_2 \leq (m\mathrm{Var}[z])^{1/2}$. Thus after noting $\mathbb{R}\backslash[J-\epsilon\|f\|_2, J+\epsilon\|f\|_2] \subset \mathbb{R}\backslash[J-\epsilon\,(m\mathrm{Var}[z])^{1/2}, J+$

$(m\text{Var}[z])^{1/2}]$ apply Chebyshev's inequality and the hypothesis to obtain

$$P\left[|z-J| \geq \epsilon \|f\|_2\right] \leq P\left[|z-J| \geq \epsilon \left(m\text{Var}[z]\right)^{1/2}\right] \leq \frac{1}{m\epsilon^2} \leq \frac{1}{\left(\frac{10}{\epsilon^2}\right)\epsilon^2} = 0.1$$

So in turn using the complement of the event $|z-J| \geq \epsilon \|f\|_2$, we have the desired result. □

Using Lemma 2.3.1, we can justify the statement that error of Monte Carlo Integration decays $\mathcal{O}(\frac{1}{\sqrt{m}})$. That is, solving for ϵ in the hypothesis of Lemma 2.3.1 we have $\epsilon = \sqrt{\frac{10}{m}}$.

We now turn to a method by which we can increase the likelihood of a our estimate being within a desired error bound beyond the guarantee seen in Lemma 2.3.1. To that end we introduce some notation, building from what we saw in the proof of Lemma 2.3.1.

Let

$$Z_k = \frac{1}{m}\sum_{j=1}^{m} f(X_{k,j}) \tag{2.3}$$

where $X_{k,j} \in [0,1]^N, 1 \leq j \leq m, 1 \leq k \leq K$ are all i.i.d uniform random variables. We have described repeating the experiment K times and gathering these estimators in the independent random variables $\{Z_k\}_{k=1}^K$. We then use the median of these estimators:

Furthermore, let

$$I_k = \begin{cases} 1 & |Z_k - J| \leq \epsilon \|f\|_2 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

These are indicator functions for whether a given estimate Z_k is within the desired error bound ϵ to the target integral J. Note that since the random variables Z_k are independent, then so are the discrete random variables I_k.

How big then does K need to be to achieve some desired likelihood q of a sufficiently accurate estimator $\tilde{Z}$ (estimator is defined in equation 2.5)? The following Lemma provides an answer: when K is larger than $C \log \frac{1}{q}$, most of our estimates will be accurate within the error bound with probability at least $1 - q$.

Lemma 2.3.2. *$\exists C \in [0, \infty)$ so that when $K \geq C \log \frac{1}{q}$ then*

$$P\left[\sum_{k=1}^{K} I_k \leq K/2\right] \leq q$$

for any arbitrary $q \in (0, 1)$, provided $m \geq 10/\epsilon^2$

Proof. Lemma 2.3.1 implies that $\tilde{p} = P[I_k = 1] \geq 0.9$. Theorem 2.2.13 however provides us a means to bound the sum of indicator variables of this sort. Noting $K/2 = (1 - (1 - 1/2\tilde{p}))K\tilde{p}$ and $K\tilde{p} = 0.9K = \mathbb{E}\left[\sum_{k=1}^{K} I_k\right]$

$$\begin{aligned}
P\left[\sum_{k=1}^{K} I_k \leq K/2\right] &= P\left[\sum_{k=1}^{K} I_k \leq (1 - (1 - 1/2\tilde{p}))K\tilde{p}\right] \\
&< \left[\frac{\exp\left(-(1 - 2/\tilde{p})\right)}{(1/2\tilde{p})^{(1/2\tilde{p})}}\right]^{K\tilde{p}} \\
&= \left(\sqrt{2\tilde{p}} \exp\left(-(\tilde{p} - 1/2)\right)\right)^{K} \\
&\leq \left(\sqrt{2} e^{-0.4}\right)^{K} \\
&\leq 0.95^{K}
\end{aligned}$$

where we have used $w \leftarrow (1 - 1/2\tilde{p})$ for Theorem 2.2.13. So

$$0.95^{K} \leq q \implies K \geq -\log(0.95) \log(1/q)$$

and we have the desired result. □

Theorem 2.3.3 (Median of Means Estimation for Monte Carlo)**.** *Let $\epsilon, q \in (0, 1)$ and define Z_k as in 2.3 with $m \geq 10/\epsilon^2$ and K an odd integer such that $K \geq C \log(1/q)$*

where C is the constant from Lemma 2.3.2. Let $J = \int_{[0,1]^N} f(x)dx$ for $f : [0,1]^N \to \mathbb{R}$. Define $\tilde{Z}$:

$$\tilde{Z} = median\{Z_1, \ldots, Z_K\} \tag{2.5}$$

Then

$$\left|\tilde{Z} - J\right| \leq \epsilon \|f\|_2 \tag{2.6}$$

hold with probability at least $(1-q)$. The total number of function evaluations of f is $\mathcal{O}\left(\log(1/q)/\epsilon^2\right)$

Proof. The proof follows from Lemmas 2.3.1 and 2.3.2 along with Lemma 2.3.4 and is left as an exercise. □

Lemma 2.3.4. *Let K be odd. If $\sum_{k=1}^{K} I_k > \frac{K}{2}$ then $\tilde{Z}$ satisfies the inequality 2.6 with the desired probability.*

Proof. First, because K is odd, $\tilde{z} \in \{z_1, \ldots, z_2\}$. For eventual contradiction, suppose $\tilde{z} \leq \text{Int}(f) - \epsilon\|f\|_2$. This means at least $\frac{k-1}{2} + 1$ elements of the set $\{z_1, \ldots, z_2\}$ do not satisfy the inequality, and thus the corresponding indicator variables I_j are zero. However $\frac{k-1}{2} + 1 > \frac{k}{2}$ which contradicts $\sum_{k=1}^{K} I_k > \frac{K}{2}$ so the assumption that $\tilde{z} \leq \text{Int}(f) - \epsilon\|f\|_2$ cannot hold.

Similarly, we can arrive at another contradiction should we assume $\tilde{z} \geq \text{Int}(f) + \epsilon\|f\|_2$. □

Homework 2.3.1. Re-prove Lemma 2.3.1 for $f : [0,1]^N \to \mathbb{C}$. (Hint: $\mu = \mathbb{E}[a+ib] = \mathbb{E}[a] + i\mathbb{E}[b]$ and $\text{Var}[a+ib] = \mathbb{E}[(a+ib-\mu)\overline{(a+ib-\mu)}]$)

Homework 2.3.2. Complete a proof of Theorem 2.3.3

2.4 Conditioning

Lemma 2.4.1. *Let $X \in \mathbb{R}^N$ and $Y \in \mathbb{R}^M$ be random variables with joint density $p : \mathbb{R}^{N+M} \to \mathbb{R}^+$. Then if $f : \mathbb{R}^{M+N} \to \mathbb{R}^+$, $S \subseteq \mathbb{R}^D$, and $T \subseteq \mathbb{R}^M$, we have that*

$$P\left[f(X,Y) \in S \text{ and } Y \in T\right] = \int_T \int_{S_y} p(x,y)dxdy$$

where $S_y = \left\{x \in \mathbb{R}^N | f(x,y) \in S\right\} \subseteq \mathbb{R}^N$

Proof. □ ToDo

Note that the inner integral that appears above can be understood as a conditional probability

$$\int_{S_y} p(x,y)dx = P\left[Y = y | f(x,y) \in S\right] P\left[Y = y\right]$$

Lemma 2.4.2. *If $f(X,Y) \in S$ implies that $y \in T$ then*

$$P\left[f(X,Y) \in S\right] = P\left[f(X,Y) \in S,\, y \in T\right]$$

Proof. With the hypothesis, $f(X,Y) \in S$ implies that $y \in T$, we have that $f^{-1}(S) \subseteq \mathbb{R}^N \times T$. Equivalently $f^{-1}(S) \cap (\mathbb{R}^N \times T) = f^{-1}(S)$.

$$\begin{aligned} P\left[f(X,Y) \in S\right] &= P\left[(X,Y) \in f^{-1}(S)\right] \\ &= P\left[(X,Y) \in f^{-1}(S) \cap (\mathbb{R}^N \times T)\right] \\ &= P\left[(X,Y) \in f^{-1}(S),\, Y \in T\right] \qquad = P\left[f(X,Y) \in S,\, Y \in T\right] \end{aligned}$$

□

2.5 Closure of Gaussian Random Variables Under Linear Transforms

Gaussian random variables are closed under linear transformations. This is a special property that in general does not hold for random variables. Consider two

independent coin flips; the addition of these random variables is no longer a random variable of coin flips; we have outcomes which are not possible for a single coin flip, e.g. getting two heads or a sum of 2 is not possible with a single coin flip.

Gaussian random variables on the other hand do yield new Guassian random variables under these operations.

Lemma 2.5.1. *Let $X \sim \mathcal{N}(0, \sigma_x^2)$ and $Y \sim \mathcal{N}(0, \sigma_y^2)$ be two independent mean 0 real Gaussian random variables with variance σ_x^2 and σ_y^2 respectively. Then*

$$X + Y \sim \mathcal{N}(0, \sigma_x^2 + \sigma_y^2)$$

Proof. Suffice to show for an arbitrary interval $S = [a, b]$ that

$$P[X + Y \in S] = \frac{1}{\sqrt{2\pi(\sigma_x^2 + \sigma_y^2)}} \int_a^b \exp\left(-\frac{x^2}{2(\sigma_x^2 + \sigma_y^2)}\right) dx$$

If we consider $T = \mathbb{R}$ and apply Lemma 2.4.2 we obtain

$$P[X + Y \in S] = P[X + Y \in S \cap Y \in \mathbb{R}]$$

Now we apply Lemma 2.4.1, where f is the function defined as $f(x, y) = x + y$ and so $S_y = \{x \in \mathbb{R} | f(x, y) \in [a, b]\} = [a - y, b - y]$:

$$\begin{aligned}
P[X + Y \in S \cap Y \in \mathbb{R}] &= \int_T \int_{S_y} p(x, y) dx dy \\
&= \int_T \int_{S_y} p(x) p(y) dx dy \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \int_{\mathbb{R}} \int_{a-y}^{b-y} \exp\left(-\frac{x^2}{2\sigma_x^2}\right) \exp\left(-\frac{y^2}{2\sigma_y^2}\right) dx dy \\
&= \frac{1}{2\pi\sigma_x\sigma_y} \int_{\mathbb{R}} \int_a^b \exp\left(-\frac{(z-y)^2}{2\sigma_x^2}\right) \exp\left(-\frac{y^2}{2\sigma_y^2}\right) dz dy
\end{aligned}$$

Where we use a change of variable $z = x + y$, thus the bounds $x = a - y$ and $x = b - y$ become $z = a$ and $z = b$ respectively in the last line

Note that,

$$
\begin{aligned}
-\frac{(z-y)^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2} &= -\frac{\sigma_y^2}{2\sigma_x^2\sigma_y^2}z^2+\frac{2\sigma_y^2}{2\sigma_x^2\sigma_y^2}zy-\frac{\sigma_x^2+\sigma_y^2}{2\sigma_x^2\sigma_y^2}y \\
&= -\left(y\sqrt{\frac{\sigma_x^2+\sigma_y^2}{2\sigma_x^2\sigma_y^2}}-\frac{z}{2\sigma_x^2}\sqrt{\frac{2\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2}}\right)^2+\left(\left(\frac{1}{2\sigma_x^2}\sqrt{\frac{2\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2}}\right)^2-\frac{\sigma_y^2}{2\sigma_x^2\sigma_y^2}\right)z^2 \\
&= -\frac{1}{2}\left(y\sqrt{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}}-\frac{z}{\sigma_x^2}\sqrt{\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2}}\right)^2+\left(-\frac{1}{2(\sigma_x^2+\sigma_y^2)}\right)z^2
\end{aligned}
$$

Now let $u = y\sqrt{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}}-\frac{z}{\sigma_x^2}\sqrt{\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2}}$ and note $\frac{du}{dy}=\sqrt{\frac{\sigma_x^2+\sigma_y^2}{\sigma_x^2\sigma_y^2}}$ and switch the bounds of integration, rewriting the integral we have

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}}\int_a^b \exp\left(-\frac{z^2}{2(\sigma_x^2+\sigma_y^2)}\right)\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp\left(-\frac{u^2}{2}\right)du\right)dz \\
&= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}}\int_a^b \exp\left(-\frac{z^2}{2(\sigma_x^2+\sigma_y^2)}\right)dz
\end{aligned}
$$

Where we have noted that the inner integral must equal one. This now matches the pdf of $\mathcal{N}(0,\sigma_x^2+\sigma_y^2)$ as desired.

□

We can use Lemma 2.5.1 to understand the probability density of random Gaussian vectors. That is for $\mathbf{g}\sim\mathcal{N}(\mathbf{0},I)\in\mathbb{R}^N$ we can use the result with induction to conclude

$$\langle \mathbf{g},\mathbf{x}\rangle = \sum_{j=1}^{N} g_j x_j \sim \mathcal{N}(\mathbf{0},\|x\|_2^2)$$

Homework 2.5.1. Let $X\sim\mathcal{N}(0,\sigma^2)\in\mathbb{R}$. Show that $aX\sim\mathcal{N}(0,a^2\sigma^2)$, $\forall a\in[0,\infty)$

2.6 Locality Sensitive Hashing and (c, r)-Nearest Neighbor Problem

Recall the nearest-neighbor problem: Given $S = \{\mathbf{x}_0, \ldots, \mathbf{x}_{p-1}\} \subseteq \mathbb{R}^D$ find $f_{NN} : [p] \to [p]$ such that $\|\mathbf{x}_j - \mathbf{x}_{f_{NN}(j)}\|_2 = \min_{\mathbf{y} \in S} \|\mathbf{x}_j - \mathbf{y}\|_2$, $\forall j \in [p]$

Recall from chapter 1 that the complexity is $\mathcal{O}(p^2 D)$ for the linear scan solution, shown in 1.1.4. In this (exact) solution method, we compute all pairwise distances $\|\mathbf{x}_i - \mathbf{x}_j\|_2$, $\forall i, j \in [p]$.

However, in many applications either p or D can be large, which means the exact computation using this straightforward method becomes computationally intractable. We will sacrifice accuracy in order to achieve faster results. Now consider the following variant of the nearest neighbor problem.

Definition 2.6.1 ((c, r)-Nearest Neighbor Problem)**.** (c, r)-NN problem: Given $S = \{\mathbf{x}_0, \ldots, \mathbf{x}_{p-1}\} \subseteq \mathbb{R}^D$ find $f : [p] \to [p] \cup \{-1\}$ so that both:

1. $d(\mathbf{x}_j, \mathbf{x}_{f(j)}) \leq cr$, $\forall j \in [p]$ such that $\exists i \in [p]$ with $d(\mathbf{x}_j, \mathbf{x}_i) \leq r$.

   The idea is that if there is a point that is r-close to $\mathbf{x}_j$, then the assignment function will return a point that is almost as close. The possible error in the nearest neighbor to $\mathbf{x}_j$ is quantified by c.

2. $f(j) = -1$ if $\nexists i \in [p]$ with $d(\mathbf{x}_j, \mathbf{x}_i) \leq cr$.

   In other words, if there is no point that is cr-close to the query point $\mathbf{x}_j$, then the assignment function will indicate this fact by returning -1.

The diagram shows schematically how this new problem simplifies nearest neighbor. The query point in the diagram is $\mathbf{x}_j$. If $\mathbf{x}_k$ is within a distance r of the query point, then our assignment function can return either $\mathbf{x}_k$ or $\mathbf{x}_i$. If there are no points within cr distance to the query (i.e. remove points $\mathbf{x}_j$ and $\mathbf{x}_k$) then the function returns -1. If there is a point within cr but no point within r (i.e. remove only

$\mathbf{x}_k$) then there is no requirement that the function assign any particular value to the nearest neighbor (e.g. returning i or -1 are possible)

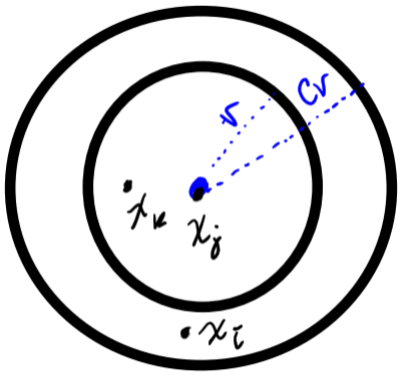


Figure 2.1: (c, r)-NN with query point $\mathbf{x}_j$

Note that for this, and most other examples, we will concern ourselves with the Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$.

Goal. *As a floor to solution run-time, we should at least read in all the data, which takes $\Omega(pD)$-time.*

To understand this lower bound on run-time, consider what might happen by withholding points from consideration in a solution. There is no way then to guarantee that the withheld points are not nearest neighbors to a given query point in this scenario. A pause to recall some

Definition 2.6.2 ($\mathcal{O}, \Omega, \Theta$ complexity)**.** 1. Let $g : \mathbb{R} \to \mathbb{R}^+$. We say g is $\mathcal{O}(h)$, $h : \mathbb{R} \to \mathbb{R}^+$, if $\exists C, x_0 \in \mathbb{R}$ such for all $y > x_0$ $g(y) \leq Ch(y)$.

2. Let $g : \mathbb{R} \to \mathbb{R}^+$. We say g is $\Omega(\tilde{h})$, $\tilde{h} : \mathbb{R} \to \mathbb{R}^+$, if $\exists C, x_0 \in \mathbb{R}$ such for all $y > x_0$ $g(y) \geq C\tilde{h}(y)$.

3. If g is $\mathcal{O}(h)$ and $\Omega(h)$ then g is $\Theta(h)$.

In order to achieve our goal of improving on nearest neighbor beyond $\mathcal{O}(p^2 D)$, we

will take our set $S \subset \mathbb{R}^D$ and project each of the points onto a random vector and find nearest neighbors of the projections which are lower dimensional. Schematically,

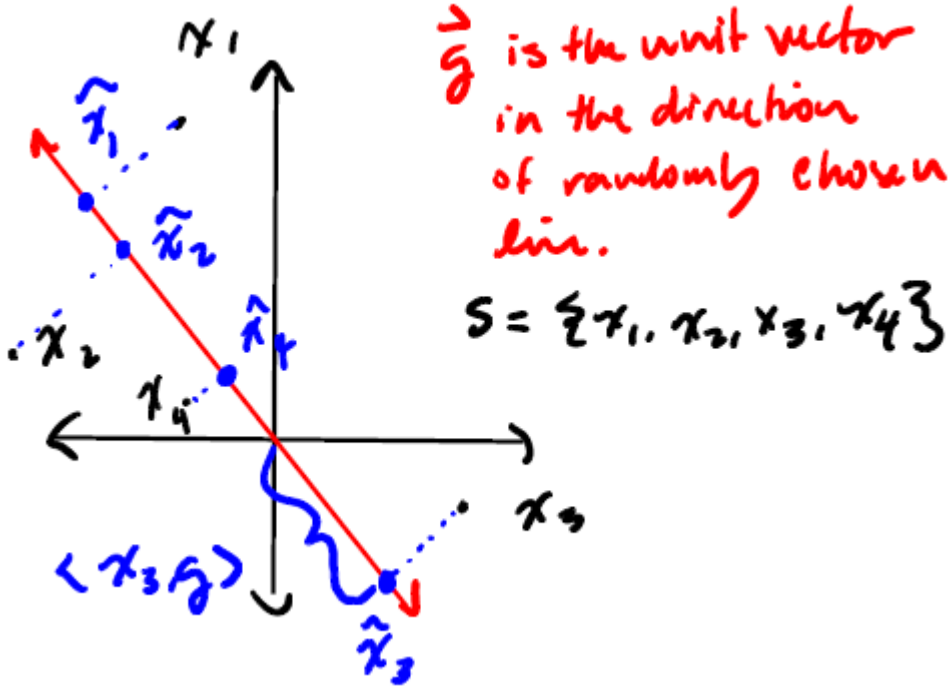


Figure 2.2: (c, r)-NN using 2D data set

Where the procedure is to generate a random line in the direction of $\mathbf{g}$, and then

1. Project all points of S onto $\mathbf{g}$, i.e. calculate $\langle \mathbf{g}, \mathbf{x}_1 \rangle, \langle \mathbf{g}, \mathbf{x}_2 \rangle, \ldots$
2. Sort the distances $\{\langle \mathbf{g}, \mathbf{x}_j \rangle\}_{j=1}^4$
3. Read off nearest neighbors from the sorted list

So, using the schematic our assignment function could be

$$f(1) = 2, f(2) = 1, f(3) = 4, f(4) = 2$$

Now let's consider complexity. The inner product of a point with a random vector takes on order D operations and must be performed for all points, so step one takes on order PD-time. Sorting lists is a well understood problem in computer science

and can be accomplished on order $P \log D$ time. Finally, scanning the list for a nearest neighbor takes P time, so overall our complexity is $\mathcal{O}(PD + P \log D + p)$.

Definition 2.6.3 (Locality Sensitive Hash Function)**.** We call a random function $h : \mathbb{R}^D \to \mathbb{Z}$ a Locality Sensitive Hash function if $\exists p_1, p_2 \in (0,1), p_1 > p_2$, so that the following properties hold for any two fixed points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$

1. If $\|\mathbf{x} - \mathbf{y}\| < r$ then $h(\mathbf{x}) = h(\mathbf{y})$ with at least probability p_1

2. If $\|\mathbf{x} - \mathbf{y}\| > cr$ then $h(\mathbf{x}) = h(\mathbf{y})$ with probability at most p_2

So a LSH function will hash similar points to the same integer and points which are dissimilar to different integers. We now consider a particular example of such a function

Example 2.6.4. Fix two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$. We define a hash function $h : \mathbb{R}^D \to \mathbb{Z}$ as follows

$$h(\mathbf{x}) = \left\lfloor \frac{\langle \mathbf{g}, \mathbf{x} \rangle + u}{w} \right\rfloor$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I)$ and $u \sim \text{Uniform}(0, w)$ and w is fixed positive number.

In the following then, we regard h to be a function of the two random variables of u and $\mathbf{g}$ where the points themselves $\mathbf{x}, \mathbf{y}$ are fixed. The key questions then to consider are what are p_1 and p_2 in 2.6.4?

By Lemma 2.4.2, since the event $h(\mathbf{x}) = h(\mathbf{y})$ implies $|\langle \mathbf{g}, \mathbf{x} \rangle - \langle \mathbf{g}, \mathbf{y} \rangle| < w$, we have

$$P_{u,\mathbf{g}}\left[h(\mathbf{x}) = h(\mathbf{y})\right] = P\left[h(\mathbf{x}) = h(\mathbf{y}) \cap |\langle \mathbf{g}, \mathbf{x} \rangle - \langle \mathbf{g}, \mathbf{y} \rangle| < w\right]$$

and so applying Lemma 2.4.1 we have the following equivalent expression as an

integral

$$P\left[h(\mathbf{x}) = h(\mathbf{y}) \cap \lfloor \langle \mathbf{g}, \mathbf{x} \rangle - \mathbf{g}, \mathbf{y} \rfloor < w\right] =$$
$$\int_0^w P\left[h(\mathbf{x}) = h(\mathbf{y}) || \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle| = z\right] P\left[|\langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle| = z\right] dz$$

Let's consider each of the probabilities that appear in the integrand,

- The probability below is a probability of only the random variable u since all other quantities are fixed.

  $$P\left[h(\mathbf{x}) = h(\mathbf{y}) || \langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle| = z\right]$$

  So, given a particular $\mathbf{x}, \mathbf{y}$ and $z \in [0, w]$ such that $|\langle \mathbf{g}, \mathbf{x} - \mathbf{y} \rangle| = z$ we need to determine the probability that an offset u results in $\langle \mathbf{g}, \mathbf{x} \rangle$ and $\langle \mathbf{g}, \mathbf{y} \rangle$ falling into the same "bin" of width w, i.e. they are hashed to same integer $h(\mathbf{x})$. Without loss of generality, suppose $a = \langle \mathbf{g}, \mathbf{x} \rangle - wh(\mathbf{x}) < \langle \mathbf{g}, \mathbf{y} \rangle - wh(\mathbf{x}) = b$. So, what then is the probability that $h(\mathbf{x}) = h(\mathbf{y})$ as a function of u, given z? This reduces to considering which offsets results in a segment of length z being contained entirely in a segment of length w - and due to the periodic nature of moving bin boundaries, it suffices to consider the case when $a = 0$. To see why this is, the diagram shows three possible scenarios for an offset u.

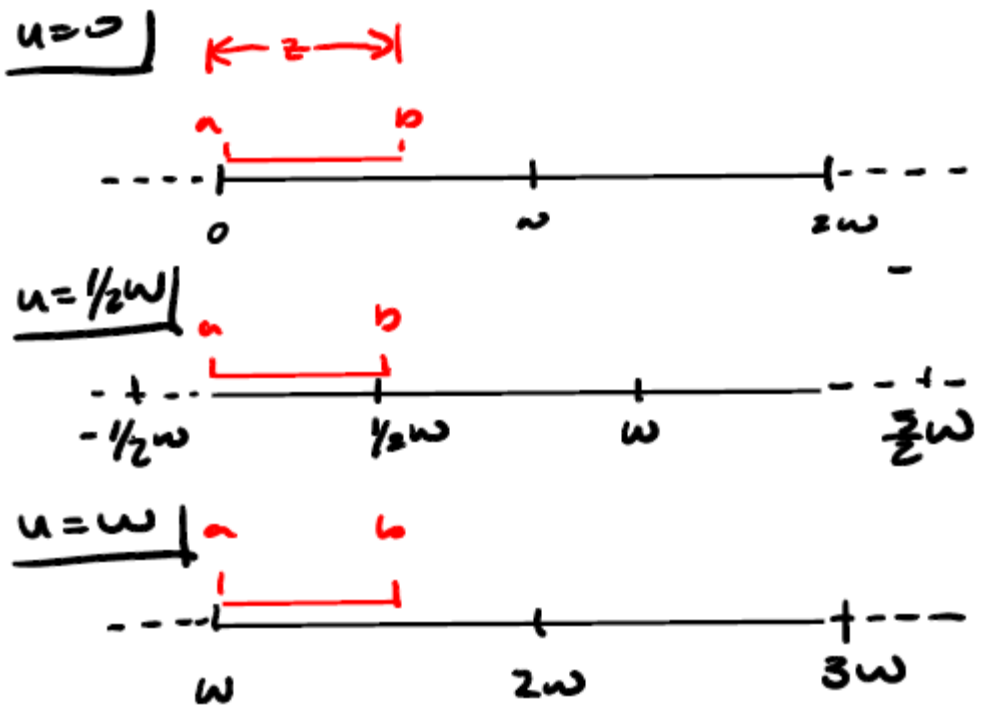


Figure 2.3: $P\left[h(\mathbf{x}) = h(\mathbf{y})||\langle \mathbf{g}, \mathbf{x} - \mathbf{y}\rangle| = z\right]$ as relationship between fixed segment of length z and bin boundaries controlled by u

When $u = 0$, the bin boundary of bin 0 is aligned with a. As we imagine u taking values from 0 to w we see the bin boundaries take all possible locations before ending back in a position where a is again exactly aligned with a bin boundary (now the bin corresponding to 1). So, for a non-zero a the starting condition is different, but the overall "movie" is the same. We need then consider for what proportion of "frames" for this movie the segment of length z is entirely contained in a single bin. If the movie is of length w then for the first z frames, the segment is split between two bins, i.e. with probability $\frac{z}{w}$ the segment is split between bins and $h(\mathbf{x}) \neq h(\mathbf{y})$. The complementary event then $1 - \frac{z}{w} = \frac{w-z}{w}$ is the probability that the segment is contained in a single bin, and thus $h(\mathbf{x}) = h(\mathbf{y})$

- We now consider

$$P\left[|\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle| = z\right]$$

We know from Lemma 2.5.1 and subsequent discussion that $\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle \sim \mathcal{N}(\mathbf{0}, \|\mathbf{x}-\mathbf{y}\|_2^2)$. Accounting for absolute values then, the probability is given by

$$\frac{\sqrt{2}}{\|\mathbf{x}-\mathbf{y}\|_2\sqrt{\pi}} \exp\left(-\frac{z^2}{2\|\mathbf{x}-\mathbf{y}\|_2^2}\right)$$

Combining the results then above, conducting a change of variables and an integration by parts, we have for $n = \|\mathbf{x}-\mathbf{y}\|_2$ then an expression which computes the probability that two points hash to the same integer as a function of the distance between the points:

$$\begin{aligned} p_w(n) &= \int_0^w P\left[h(\mathbf{x}) = h(\mathbf{y}) | |\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle| = z\right] P\left[|\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle| = z\right] dz \\ &= \frac{2}{\sqrt{\pi}} \int_0^{\frac{w}{n\sqrt{2}}} e^{-z^2} dz + \sqrt{\frac{2}{\pi}} \frac{n}{w} \left[e^{-\left(\frac{w}{n\sqrt{2}}\right)^2} - 1\right] \end{aligned}$$

Taking the derivative with respect to n we have

$$\begin{aligned} \frac{d}{dn} p_w(n) &= \frac{d}{dn}\left[\frac{2}{\sqrt{\pi}} \int_0^{\frac{w}{n\sqrt{2}}} e^{-z^2} dz + \sqrt{\frac{2}{\pi}} \frac{n}{w} \left[e^{-\left(\frac{w}{n\sqrt{2}}\right)^2} - 1\right]\right] \\ &= -\frac{2}{\sqrt{\pi}} \frac{w}{\sqrt{2}n^2} e^{\left(\frac{w}{\sqrt{2}n}\right)^2} + \sqrt{\frac{2}{\pi}} \frac{1}{w} \left[e^{-\left(\frac{w}{n\sqrt{2}}\right)^2} - 1\right] + \sqrt{\frac{2}{\pi}} \frac{n}{2} \frac{2w^2}{2n^3} \left[e^{-\left(\frac{w}{n\sqrt{2}}\right)^2}\right] \\ &= \sqrt{\frac{2}{\pi}} \frac{1}{w} \left[e^{-\left(\frac{w}{n\sqrt{2}}\right)^2} - 1\right] \end{aligned}$$

We can see that $\frac{d}{dn} p_w(n) < 0$ which is to say that the function is monotonically decreasing, thus when $n = \|\mathbf{x}-\mathbf{y}\|_2 < r$ we have that $p_w(n) < p_w(r)$. That is $p_w(r) = p_1$ from Definition 2.6.3. Furthermore, when for $c > 1$ we have $\|\mathbf{x}-\mathbf{y}\|_2 > cr$ we know that $p_w(cr) < p_w(r)$. That is $p_w(cr) = p_2$ from Definition 2.6.3.

We have shown quantitatively that the probability of hashing to the same integer is greater if the points are close and smaller if they are farther away. The following Lemma summarizes then what we have demonstrated.

Lemma 2.6.5. *Let* $\mathbf{g} \sim \mathcal{N}(0, I), u \sim ([0, w])$ *and* $w \in \mathbb{R}^+$. *Then* $h(\mathbf{x}) = \left\lfloor \frac{\langle \mathbf{g}, \mathbf{x} \rangle + u}{w} \right\rfloor$ *is a LSH function* $\forall r \in \mathbb{R}^+$ *and* $c \in (0, 1)$ *with respect to Euclidean distance. It has* $p_1 = p_w(r) > p_2(cr) = p_2$ *where*

$$p_w(n) = erf\left(\frac{w}{\sqrt{2}n}\right) + \sqrt{\frac{2}{\pi}}\frac{n}{w}\left[e^{-\left(\frac{w}{\sqrt{2}n}\right)^2} - 1\right]$$

At present, our hashing function only addresses what happens for two fixed vectors $\mathbf{x}$ and $\mathbf{y}$. How can we use this function to build a hashing scheme which works for a large data set with possibly billions of data points? We will accomplish this by repeating the hash using independent draws of the random variables and using the collection of these to hash the entire set.

Definition 2.6.6. Let $g_k : S \to \mathbb{Z}^k$ be a new LSH function created by using k i.i.d LSH functions of the type in Example 2.6.4, i.e. $h_1, \ldots, h_k$, $g_k(\mathbf{x}) = (h_1(\mathbf{x}, \ldots, h_k(\mathbf{x})))$

Definition 2.6.7. When $g_k : S \to \mathbb{Z}^k$ has the properties, for some fixed $\mathbf{x} \in S$

1. If for $\mathbf{y} \in S$ where $d(\mathbf{x}, \mathbf{y}) \geq rc$ then $g_k(\mathbf{x}) \neq g_k(\mathbf{y})$.

2. If for at least one $\mathbf{y} \in S$ it happens that $d(\mathbf{x}, \mathbf{y}) \leq r$ then $g_k(\mathbf{x}) = g_k(\mathbf{y})$.

then it is a satisfactory LSH function for $\mathbf{x} \in S$.

In other words, g_k does not hash a dissimilar points to the same thing as the query and also if there is some point close to our query, then g_k should hash a close point and query to the same thing. Note that we now have a more expansive criteria for our hash than in Example 2.6.4; it requires something hold true for all other points in the set, not simply a pair.

Definition 2.6.8. For $\mathbf{x} \in S$ we denote the nearest neighbor of $\mathbf{x}$ as

$$\mathbf{x}^* = \arg\min_{\substack{\mathbf{y}\in S \\ \mathbf{x}\neq\mathbf{y}}} d(\mathbf{x}, \mathbf{y})$$

We wish to know bounds on the probability that g_k fails to meet each of the properties in Definition 2.6.7. Consider the first property, it fails when there exists a point which is far away that nevertheless hashes to the same vector as $\mathbf{x}$. In order to hash to the same vector in $\mathbb{Z}^k$ all k component hashes need to incorrectly hash "far" points to the same integer. For each component that happens with probability p_2. Thus:

$$\begin{aligned} P\left[g_k \text{ fails 1 }\right] &= P\left[\exists \mathbf{y} \in S \text{ s.t. } g_k(\mathbf{x}) = g_k(\mathbf{y}),\, d(\mathbf{x}, \mathbf{y}) \geq rc\right] \\ &\leq |S| P\left[g_k(\mathbf{x}) = g_k(\mathbf{y}),\, d(\mathbf{x}, \mathbf{y}) \geq rc\right] \\ &\leq |S| p_2^k \end{aligned}$$

where we have used a union bound.

How can we bound the probability that the second property in Definition 2.6.6 fails? The complementary event is that all points which are close hash to the same vector in $\mathbb{Z}^k$. This means that component-wise the k hashes all need to hash $\mathbf{x}$ and $\mathbf{y}$ to the same integer, which happens with probability p_1 for each component. Thus:

$$\begin{aligned} P\left[g_k \text{fails 2}\right] &= P\left[\exists \mathbf{y} \in S g_k \text{ s.t. } (\mathbf{x}) \neq g_k(\mathbf{y}),\, d(\mathbf{x}, \mathbf{y}) < r\right] \\ &\leq 1 - |S| P\left[g_k(\mathbf{x}) = g_k(\mathbf{y}),\, d(\mathbf{x}, \mathbf{y}) \geq rc\right] \\ &\leq 1 - p_1^k \end{aligned}$$

The probability that g_k is a satisfactory LSH then can be bounded from below by

$$P\left[g_k\text{satisfies 1 and 2}\right] = 1 - P\left[g_k\text{fails 1 or 2}\right]$$

$$\geq 1 - P\left[g_k\text{fails 1}\right] - P\left[g_k\text{fails 2}\right] \quad \geq 1 - |S|p_2^k - (1 - p_1^k)$$

$$= p_1^k \left(1 - |S| \left(\frac{p_2}{p_1}\right)^k\right)$$

Let $k = \log_{p_1/p_2}(2|S|)$ and $\rho = \frac{\log p_1}{\log p_2}$, then this simplifies as follows:

$$\begin{aligned}
p_1^k \left(1 - |S| \left(\frac{p_2}{p_1}\right)^k\right) &= p_1^{\left(\log_{p_1/p_2}(2|S|)\right)} \left(1 - |S| \left(\frac{p_2}{p_1}\right)^{\left(\log_{p_1/p_2}(2|S|)\right)}\right) \\
&= p_1^{\left(\log_{p_1/p_2}(2|S|)\right)} \left(1 - |S| \left(\frac{p_1}{p_2}\right)^{\left(\log_{p_1/p_2}\left(\frac{1}{2|S|}\right)\right)}\right) \\
&= p_1^{\left(\log_{p_1/p_2}(2|S|)\right)} \left(1 - |S| \left(\frac{1}{2|S|}\right)\right) \\
&= \frac{1}{2} p_1^{\left(\log_{p_1/p_2} 2|S|\right)} \\
&= \frac{1}{2} \left[p_1^{\left(\frac{\log_{p_1}(2|S|)}{\log_{p_1}\left(\frac{p_1}{p_2}\right)}\right)}\right] \\
&= \frac{1}{2} \left[p_1^{\left(\log_{p_1}(2|S|)\right)}\right]^{\frac{1}{\frac{\log\left(\frac{p_1}{p_2}\right)}{\log p_1}}} \\
&= \frac{1}{2} \left[2|S|\right]^{\frac{\frac{\log p_1}{\log p_2}}{\frac{\log p_1}{\log p_2} - 1}} \\
&= \frac{1}{2} \left[2|S|\right]^{\frac{\rho}{1-\rho}}
\end{aligned}$$

We gather then this in the following lemma

Lemma 2.6.9. *If $k = \log_{p_1/p_2}(2|S|)$ and $\rho = \frac{\log p_1}{\log p_2}$ then g_k as in Definition 2.6.6 will be a satisfactory LSH as described in Definition 2.6.7 for any given $\mathbf{x} \in S$ with probability at least*

$$\frac{1}{2} \left[2|S|\right]^{\frac{\rho}{1-\rho}}$$

Lemma 2.6.10. *If we generate* $L \geq 2(2|S|)^{\frac{\rho}{1-\rho}} \log\left(\frac{|S|}{1-\sigma}\right)$ *i.i.d. hash functions* $g_k^i : S \to \mathbb{Z}^k, j = 1, \ldots, L$ *with* $k = \log_{p_1/p_2}(2|S|)$, $\rho = \frac{\log p_1}{\log p_2}$ *then the following will hold with probability at least* σ*:*

$\forall x \in S$, $\exists \ell \in [L]$ *s.t.* g_k^ℓ *is a satisfactory LSH in the sense of Definition 2.6.7 for* x

Proof. Let $\delta = \frac{1}{2}\left(\frac{1}{2|S|}\right)^{\frac{\rho}{1-\rho}}$ and fix $x \in S$. The probability that a g_k^i will be unsatisfactory for x is at most $(1-\delta)$ by Lemma 2.6.9. Thus the probability that all g_k^i will fail to be satisfactory is at most $(1-\delta)^L$. We can use Fact 2.2.11 to conclude that

$$(1-\delta)^L \leq e^{-\delta L} \leq e^{\delta 2(2|S|)^{\frac{\rho}{1-\rho}} \log\left(\frac{|S|}{1-\sigma}\right)} = e^{\log\left(\frac{1-\sigma}{|S|}\right)} = \frac{1-\sigma}{|S|}$$

So the probability that for every x all hash functions fail to be satisfactory is bounded by the union of $|S|$ such probabilities seen above, i.e. $1-\sigma$. The complementary event, that for every x at least one hash doesn't fail, is then at least σ. □

We now have the results needed to construct a randomized algorithm that solves (c, r)-NN problem

Algorithm 2.6.1 LSH for (c, r)-NN

Input: $S \subseteq \mathbb{R}^D$, $d(x, y) = \|\mathbf{x} - \mathbf{y}\|_2$
Output: $f : S \to S \cup \{\infty\}$ a (c, r)-NN map
for $\mathbf{x}$ in S **do**
  $\forall \ell \in L$ compute $g_k^\ell(\mathbf{x})$
end for
$\forall \mathbf{x} \in S, f(\mathbf{x}) = (\infty, \ldots, \infty)$
for each g_k^ℓ, $\ell \in [L]$ **do**
  for each n in $g_k^\ell(S)$ where $\left|(g_k^\ell)^{-1}(n)\right| \geq 2$ **do**
    for each $\mathbf{x}$ in $\left(g_k^\ell\right)^{-1}(n)$ **do**
      choose any $\mathbf{y}$ in $\left(g_k^\ell\right)^{-1}(n) \setminus \{x\}$
      if $\|\mathbf{x} - \mathbf{y}\|_2 < \min\{\|\mathbf{x} - f(\mathbf{x})\|_2, cr\}$ **then**
        $f(\mathbf{x}) \leftarrow \mathbf{y}$
      end if
    end for
  end for
end for

The first for loop has $\mathcal{O}(DKL|S|)$ run-time, since each element of S must be projected with an inner-product of length D, K-times for each of the L hash functions.

Next we move onto the bottom block, after the first end for. We analyze the run-time from the inside our: the inner if statement requires comparison of norms and so will require $\mathcal{O}(D)$ comparisons. The two most inner for loops could in the worst case scenario iterate over $|S|$ vectors, each of which needs to compared on K entries (in the case that all vectors hash to the same vector $\mathbf{n}$) and therefore overall the inner loop has worst case run-time complexity $\mathcal{O}(|S|DK)$. The outer-loop iterates L times, and so overall the entire loop has at most $\mathcal{O}(DKL|S|)$

Recall that algorithm 1.1.4 finds exact answers in $\mathcal{O}(D|S|^2)$ (which would naturally solve the (c, r)-NN problem as well). So for what types of problems does this represent a cost-savings? From our Lemma 2.6.10 and analysis we have that if

$$\begin{aligned} K &\geq \log_{p_1/p_2}(2|S|) \\ L &\geq 2\,(2|S|)^{\frac{\rho}{1-\rho}} \log\left(\frac{|S|}{1-\sigma}\right) \end{aligned}$$

then with probability at least σ our algorithm should produce a satisfactory solution to the (c, r)-nearest neighbor problem. That is for complexity

$$\mathcal{O}\left(D|S|^{1+\frac{\rho}{1-\rho}} \log\left(\frac{|S|}{1-\sigma}\right) \log_{p_1/p_2}(2|S|)\right)$$

and if we fix $w = 3r$ and $c = 3$ in the definition of the component hash functions h then $\frac{\rho}{1-\rho} \approx 0.449$ and $p1/p2 \geq 1.99$ so in this scenario we have saved something on the order of $|S|^{1/2}$ from the naive solution, which represents a significant savings for large sets.

Theorem 2.6.11. *Choose $\sigma \in (0,1)$, $S \subset \mathbb{R}^D$. Then $\forall r > 0$, $(3, r)$-NN problem can be solved for S with respect to Euclidean distance with probability σ in time*

$$\mathcal{O}\left(D|S|^{1.5} \log\left(\frac{|S|}{1-\sigma}\right) \log_{1.99}(2|S|)\right)$$

Note. *1. Having $\rho = \frac{\log p_1}{\log p_2}$ small is crucial. The following result is described in [9]: $\forall q \in (0,2]$ and $r \in \mathbb{R}^+, \delta.c \in (1,\infty) \exists$ an LSH function $h : \mathbb{R}^D \to \mathbb{Z}$ with respect to $\|\cdot\|_q$ having $\rho \leq \delta \max(c^{-q}, c^{-1})$*

2. In [1] we have the following result which shows a near optimal result for Euclidean distance: $\exists$ a LSH with respect to $\|\cdot\|_2$, $\forall r \in \mathbb{R}^+$,$c \in (1,\infty)$ that has

$$\rho = \frac{1}{c^2} + \mathcal{O}\left(\frac{\log\log|S|}{\log^{1/3}|S|}\right)$$

for any given $S \subset \mathbb{R}^D$

3. In [26] we have a lower-bound on ρ. It states that for large D there exists an r and $p_2 \geq 2^{-\mathcal{O}(D)}$ for which $\rho \geq \frac{0.462}{c^q}$ for any LSH with respect to $\|\cdot\|_q$, $\forall c, q \geq 1$.

Homework 2.6.1. Use the definition of Cauchy random variables and discussion below to prove that $h : \mathbb{R}^D \to \mathbb{Z}$ in 2.6.4 is still a Locality Hashing Function $\forall w, r \in \mathbb{R}^+$ and $c \in (1,\infty)$ with respect to $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1$ when each entry of $\mathbf{g}$ is an i.i.d Cauchy random variable with density $f_{0,1}(x)$ from 2.6.12.

Definition 2.6.12 (Cauchy Random Variable)**.** A Cauchy random variable is real value $X \sim \text{Cauchy}(x_0, \gamma_0)$ that has the probability density function

$$f_{x_0,\gamma}(x) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)} = \frac{1}{\pi}\left(\frac{\gamma}{(x-x_0)^2+\gamma^2}\right)$$

So for example when $x_0 = 0, \gamma = 1$, $f_{0,1}(x) = \frac{1}{\pi(1+x^2)}$

Interestingly, the mean and variance of Cauchy random variable are undefined, as can be seen by writing the associated integrals.

Consider two independent Cauchy random variables $X \sim \text{Cauchy}(x_0, \gamma_0)$ and $Y \sim \text{Cauchy}(y_0, \delta_0)$. Then

1. $kX + L \sim \text{Cauchy}(kx_0 + L, |k|\gamma_0)$
2. $X + Y \sim \text{Cauchy}(x_0 + y_0, \gamma_0 + \delta_0)$

These two properties constitute the stable distribution property.

Homework 2.6.2. For $c = 3$ and $w = 3r$ verify that $\frac{\rho}{1-\rho} \approx 0.449$ and $\frac{p_1}{p_2} \geq 1.99$. Can you improve for any arbitrary r?

Homework 2.6.3. Let

$$\frac{\min_{s \in S} \|\mathbf{x}^* - \mathbf{x}\|_2}{\max_{s \in S} \|\mathbf{x}\|_2} < 1$$

choose $\sigma \in (0, 1)$ such that , $\frac{\sigma}{\log_{4/3}(R)} < 1$. Prove that we can solve a sequence of $(3, r)$-NN problems to get $f_{NN}^A : S \to S$ satisfying

$$\|f_{NN}^A(\mathbf{x}) - \mathbf{x}\|_2 \leq 4\|\mathbf{x} - \mathbf{x}^*\|_2$$

for all $\mathbf{x}^* \in S$ with probability at least σ in

$$\mathcal{O}\left(D|S|^{1.5} \log\left(\frac{|S| \log_{4/3}\left(\frac{1}{R}\right)}{1-\sigma}\right) \log_{1.99}(|S|) \log_{1.99}\left(\frac{1}{R}\right)\right)$$

2.7 Approximate Counting with Few Bits

Given a sequence of $z_0, \ldots, z_{n-1} \in \{0, 1\}$ count the number of times that a 1 appears in the sequence. We would like to use only $c_1 \lceil \log \lceil \log n \rceil \rceil$ number of bits to store our estimate of the counting problem, where c_1 is an absolute constant (independent of n).

We introduce some notation:

- Let $T_j = \sum_{\ell=1}^{j} z_\ell$. This is the true count of the one's in the given sequence.
- $\forall j \in [n]$ output X_j such that $|X_j - T_j| \leq c_2 T_j$ where c_2 should be a (small) absolute constant independent of j.

Morris Algorithm

Algorithm 2.7.1 Morris' Algorithm

Input: $z_0, z_1, \ldots, z_{n-1} \in \{0, 1\}$
Output: $X_j \approx T_j = \sum_{\ell=0}^{j} z_j$, $\forall j \in [n]$
$Y_{-1} = 0$
for $j = 1, \ldots, n-1$ **do**
  if $z_j = 0$ **then**
    $Y_j \leftarrow Y_{j-1}$
  else
    $B \sim \begin{cases} 1 & \text{with probability } 2^{-Y_{j-1}} \\ 0 & \text{with probability } 1 - 2^{-Y_{j-1}} \end{cases}$
    $Y_j \leftarrow Y_{j-1} + B$
  end if
  $X_j \leftarrow 2^{Y_j} - 1$
end for

To see why the memory usage fits our stated goal, consider $2^{Y_{n-1}} - 1 \leq c_2 T_{n-1}$ which implies $Y_{n-1} \leq \log_2 (c_2 T_{n-1} + 1)$ which takes $\lceil \log \lceil \log (c_2 T_{n-1} + 1) \rceil \rceil$.

Note that $Y_0, \ldots, Y_{n-1}$ is an example of Markov chain because the process is "memoryless" - the next state only depends on the current state.

In order to simplify analysis, we will consider the subsequences that correspond to the actual events of interest: Let $z_{i_1}, \ldots, x_{i_{\tilde{n}}}$ be the members of the sequence where $z_j = 1$. We denote then $\tilde{Y}_k = Y_{j_k}$ for $k = 1, \ldots, \tilde{n} - 1$. This corresponds then to our estimates at the different points in the stream where the event of interest has occured.

In practice generating the random variable B with accuracy that accounts for potentially very small values $2^{-Y_{j-1}}$ itself could torpedo the project of making a low bit counter - however the efficient generation of random numbers with high accuracy is a involved topic outside our scope. In this course we'll take it for granted that it

can be accomplished.

For the following lemmas we take the random variables to be defined as described in Algorithm 2.7.1

Lemma 2.7.1. *Let $m, j \in \mathbb{N}$, such that $m \geq 1$, $j \geq 0$. Then*

$$\mathbb{E}\left[2^{m\tilde{Y}_j}\right] = (2^m - 1)\,\mathbb{E}\left[2^{(m-1)\tilde{Y}_{j-1}}\right] + \mathbb{E}\left[2^{m\tilde{Y}_{j-1}}\right]$$

Proof. We use the definition of expectation and then rearrange terms to get our desired result:

$$\begin{aligned}
\mathbb{E}\left[2^{m\tilde{Y}_j}\right] &= \sum_{i\in\mathbb{Z}} 2^{mi} P\left[\tilde{Y}_j = i\right] \\
&= \sum_{i\in\mathbb{Z}} 2^{mi}\left(P\left[B_j = 1\right] P\left[\tilde{Y}_{j-1} = i-1\right] + P\left[B_j = 0\right] P\left[\tilde{Y}_{j-1} = i\right]\right) \\
&= \sum_{i\in\mathbb{Z}} 2^{mi}\left(\frac{1}{2^{i-1}} P\left[\tilde{Y}_{j-1} = i-1\right] + \left(1 - \frac{1}{2^i}\right) P\left[\tilde{Y}_{j-1} = i\right]\right) \\
&= \sum_{i\in\mathbb{Z}} 2^m 2^{(m-1)(i-1)} \frac{1}{2^{i-1}} P\left[\tilde{Y}_{j-1} = i-1\right] + \sum_{i\in\mathbb{Z}} \left(2^{mi} - 2^{(m-1)i}\right) P\left[\tilde{Y}_{j-1} = i\right] \\
&= 2^m \sum_{i\in\mathbb{Z}} 2^{(m-1)\tilde{Y}_{j-1}} \frac{1}{2^{i-1}} P\left[\tilde{Y}_{j-1} = i-1\right] + \sum_{i\in\mathbb{Z}} \left(2^{m\tilde{Y}_{j-1}} - 2^{(m-1)\tilde{Y}_{j-1}}\right) P\left[\tilde{Y}_{j-1} = i\right] \\
&= 2^m \mathbb{E}\left[2^{(m-1)\tilde{Y}_{j-1}}\right] + \mathbb{E}\left[2^{m\tilde{Y}_{j-1}} - 2^{(m-1)\tilde{Y}_{j-1}}\right] \\
&= (2^m - 1)\,\mathbb{E}\left[2^{(m-1)\tilde{Y}_{j-1}}\right] + \mathbb{E}\left[2^{m\tilde{Y}_{j-1}}\right]
\end{aligned}$$

□

Lemma 2.7.2.

$$E\left[\tilde{X}_j\right] = j,\ \forall j = 0, \ldots, \tilde{n} \leq n$$

Proof. The proof is left as an exercise □

Lemma 2.7.3.

$$Var\left[\tilde{X}_j\right] = \frac{1}{2}\left(j^2 - j\right),\ \forall j = 0, \ldots, \tilde{n} \leq n$$

Proof. The proof is left as an exercise □

We now will describe how to use a median of means technique to produce an estimate of T_j that has good accuracy with high probability.

By applying Chebyshev's Inequality along with the results of Lemma 2.7.2 and 2.7.3,

$$\begin{aligned} P\left[\left|\tilde{X}_j - \mathbb{E}\left[\tilde{X}_j\right]\right| \geq kj\right] &= P\left[\left|\tilde{X}_j - j\right| \geq kj\frac{\text{Var}\left[\tilde{X}_j\right]}{\text{Var}\left[\tilde{X}_j\right]}\right] \\ &\leq \frac{1}{k^2}\frac{\frac{1}{2}\left(j^2 - j\right)}{j^2} \\ &\leq \frac{1}{2k^2} \end{aligned}$$

To decrease the variance of our estimator we will average L i.i.d draws from Algorithm 2.7.1 labeled $\tilde{X}_j^\ell$ and define the mean of these counts as

$$\overline{X}_j = \frac{1}{L}\sum_{\ell=1}^{L} \tilde{X}_j^\ell$$

By Theorem 2.2.10 we can calculate in the same manner above that

$$P\left[\left|\overline{X}_j - j\right| \geq kj\right] \leq \frac{1}{2k^2 L}$$

Now in a manner similar to Lemma 2.3.1, by setting $L \geq 5/\epsilon^2$ (relabeling k as ϵ) and noting complementary events we obtain

$$P\left[\left|\overline{X}_j - j\right| \leq \epsilon j\right] > 0.9$$

As was done previously during the discussion of Monte Carlo integration, we will consider repeating the experiment of finding means and use this collection's median to estimate the desired quantity.

Let

$$I_i = \begin{cases} 1 & \text{if } \left|\overline{X}_j^i\right| < \epsilon j \\ 0 & \text{otherwise} \end{cases}$$

where $\overline{X}_j^i$ are i.i.d copies of $\overline{X}_j$. Set $i = 1, \ldots, c\log(1/q)$ where $c \approx -\log(0.95)$ and $q \in (0,1)$. Then $\left|\text{median}\left\{\overline{X}_j^1, \ldots, \overline{X}_j^{c\log(1/q)}\right\} - j\right| < \epsilon j$ with probability at least $1-q$.

Theorem 2.7.4 (Morris Algorithm)**.** *Let $L = 5/\epsilon^2$ and $I = \tilde{c}\log\left(\frac{m}{q}\right)$ where $q, \epsilon \in (0,1)$, $m \in \mathbb{N}$ and $\tilde{c} \in \mathbb{R}^+$ is an absolute constant. Then there exists an approximate counting estimator X_j such that*

$$T_j \leq X_j \leq \frac{1+\epsilon}{1-\epsilon} T_j$$

holds for any m values of j with probability at least $1-q$. Furthermore the estimator will use at most

$$\mathcal{O}\left(LI \log\log n\right)$$

bit with probability at least $1 - \frac{LI}{n^2}$ for $c \in \mathbb{R}^+$ for $c \in \mathbb{R}^+$

Homework 2.7.1. Use induction and Lemma 2.7.1 to prove Lemma 2.7.2

Homework 2.7.2. Use induction and Lemmas 2.7.1 and 2.7.2 to prove Lemma 2.7.3

Homework 2.7.3. Use the discussion in this section to write a formal proof for Theorem 2.7.4.

2.8 Distinct Elements

Given a sequence $z_1, \ldots, z_N \in U$ where $|U| = M$, we want to know how many unique elements are in the sequence, i.e. the cardinality of the sequence as a set. We want to accomplish this using $\emptyset\left(\min\{N, M\}\right)$ bits of memory. We will show that, under idealized conditions this can be accomplished using $\mathcal{O}(\log M)$ memory.

Definition 2.8.1 (Perfect Hash)**.** A function $h : U \to [0, 1]$ with the properties that $\forall a \in U$

1. $h(a)$ is a uniform random variable in $[0, 1]$

2. $h(a)$ is independent of $h(b)$ for $a \neq b$

There are practical limitations to achieving a perfect hash efficiently, though for applications "near" perfect may be sufficient. If the dictionary was indeed small in comparison to the size of the sequence, $M \ll N$, and/or we were to repeat the procedure many times, then we could create and store a relatively small number t of random arrays of length M. This would require $\mathcal{O}(tM \log M)$ memory. This solution is not entirely satisfactory, though for now we will take for granted that such a hash can be effected. In the following algorithm we will compute L collections of K estimators averaged and use the median of these means to produce an accurate estimate with high probability.

Algorithm 2.8.1 Flajolet-Martin Algorithm

Input: $z_0, z_1, \ldots, z_N \in [M]$, KL i.i.d. perfect hash functions $h_{(k,\ell)} : [M] \to [0, 1]$
Output: $\tilde{E}$ estimate of $|\{z_1, \ldots, z_N\}| = \tilde{n}$

$E_{(k,\ell)} \leftarrow 1$, $\forall k \in [K], \ell \in [L]$
for $j = 1, \ldots, N$ **do**
  for $\ell = 1, \ldots, L$ **do**
    for $k = 1, \ldots, K$ **do**
      $E_{(k,\ell)} \leftarrow \min\left(E_{(k,\ell)}, h_{(k,\ell)}(z_j)\right)$
    end for
    $E_\ell \leftarrow \frac{1}{K} \sum_{k=1}^{K} E_{(k,\ell)}$
  end for
end for
$E \leftarrow \text{median}\,(E_1, \ldots, E_L)$
$\tilde{E} \leftarrow \frac{1}{E} - 1$

We then proceed to analyze the estimators' expectation and variance (in addition to the by now routine median of means method) to see why the algorithm gives good estimates with high probability for the number of distinct elements.

$\forall k, \ell$ we have that $E_{(k,\ell)} = \min\left\{h_{(k,\ell)}(z_j)\right\}_{j\in[N]}$, which is the minimum value of $\tilde{n}$ i.i.d. uniform random variables $u_i, \ldots, u_{\tilde{n}} \in [0,1]$

Lemma 2.8.2. *The probability density of* $E_{(k,\ell)} = \min\{u_1, \ldots, u_{\tilde{n}}\}$ *where* $u_1, \ldots, u_{\tilde{n}}$ *are i.i.d uniform in interval* $[0,1]$ *is* $p(x) = \tilde{n}(1-x)^{\tilde{n}-1}$

Proof. Using the cumulative density function definition for the random variable $E_{(k,\ell)}$, complementary events and independence we obtain

$$\begin{aligned} F(x) = \int_0^x p(y)dy &= P\left[\min\{u_1, \ldots, u_{\tilde{n}}\} \in [0,x]\right] \\ &= 1 - P\left[\min\{u_1, \ldots, u_{\tilde{n}}\} \in (x,1]\right] \\ &= 1 - P\left[u_1 \in (x,1], \ldots, u_{\tilde{n}} \in (x,1]\right] \\ &= 1 - \prod_{\ell=1}^{\tilde{n}} P\left[u_j \in (x,1]\right] \\ &= 1 - (1-x)^{\tilde{n}} \end{aligned}$$

Note that by the fundamental theorem of calculus $\frac{d}{dx}F(x) = p(x)$ we then obtain our desired result. □

The above lemma is an example of a simple order statistic that uses uniform random variables.

Lemma 2.8.3. *If* $E_{(k,\ell)} = \min\{u_1, \ldots, u_{\tilde{n}}\}$ *as in Lemma 2.8.2 then* $\forall k, \ell$

$$\mathbb{E}\left[E_{(k,\ell)}\right] = \frac{1}{\tilde{n}+1},\ Var\left[E_{(k,\ell)}\right] = \frac{\tilde{n}}{(\tilde{n}+1)^2(\tilde{n}+2)} < \frac{1}{(\tilde{n}+1)^2}$$

Proof. Proof is left as an exercise to the reader □

Once we have the expectation and variance of a single estimator $E_{(k,\ell)}$ we can use reasoning similar to that seen in Monte Carlo integration and Morris' Algorithm

to combine estimates in a median of means scheme to achieve the following overall estimate with

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} E_{(k,\ell)}\right] = \frac{1}{\tilde{n}+1},\ \text{Var}\left[\frac{1}{K}\sum_{k=1}^{K} E_{(k,\ell)}\right] \leq \frac{1}{K(\tilde{n}+1)^2}$$

By choosing $K = \frac{10}{\epsilon^2}$ and using Chebyshev inequality then we have that

$$P\left[\left|\frac{1}{K}\sum_{k=1}^{K} E_{(k,\ell)} - \frac{1}{\tilde{n}+1}\right| < \epsilon\left(\frac{1}{\tilde{n}+1}\right)\right] > 0.9$$

Now, choosing $L \geq c\log\left(\frac{1}{q}\right)$ and using the sum of indicator variables and Chernoff inequality, we can argue that the majority of the estimators will be within our chosen error bound with high probability

$$P\left[\left|E - \frac{1}{\tilde{n}+1}\right| < \epsilon\left(\frac{1}{\tilde{n}+1}\right)\right] \geq 1-q$$

for $q \in (0,1)$

Theorem 2.8.4 (Flajolet-Martin Algorithm)**.** *Choose $\epsilon, q \in (0,1)$. Then Algorithm 2.8.1 will output an estimate E satisfying*

$$\frac{\tilde{n}}{1+\epsilon} - \frac{\epsilon}{1+\epsilon} < \frac{1}{E} - 1 < \frac{\tilde{n}}{1-\epsilon} + \frac{\epsilon}{1-\epsilon}$$

with probability at least $1-q$.

Proof. The formal proof is left as an exercise to the reader. □

Homework 2.8.1. Prove Lemma 2.8.3

Homework 2.8.2. Using the discussion in this section, formalize a proof of Theorem 2.8.4

2.8.1 Practical, Better Hashing

Storing arrays of M uniformly generated random numbers to use as hashes is a problem if $M \gg 1$. We may also object to this approach on the grounds that reusing

or having foreknowledge about the array would torpedo the randomness aspect that we are relying on to estimate our distinct element count in Algorithm 2.8.1. An adversary given access to the hash could effect any count or miscount he wanted by choosing the sequence of inputs (or simply by re-ordering them).

Instead let us consider the following procedure to generate a usable and tractable hash

1. Select a prime number p
2. Choose two random, independent uniformly distributed integer values $a, b \in \{0, \ldots, p-1\}$
3. Compute for $x \in [p]$ hash to the value $h_a(x)$ in the following way

$$h_a(x) = \frac{(ax+b) \mod p}{p}$$

Note that $h_a(x) \in [0, 1)$

Lemma 2.8.5. *$h_a(x)$ is uniformly distributed in $\left\{0, \frac{1}{p}, \ldots, \frac{p-1}{p}\right\} \subset [0, 1]$*

Proof. Fix $j \in [p]$, we wish to show that $P[h_a(x) = j/p] = 1/p$.

First assume $x = 0$. It follows that $P[h_a(x) = j/p] \iff P[b = j]$. We have by the hypothesis that b is drawn uniformly from $[p]$ and thus the probability is equal to $1/p$.

Next, assume $x \neq 0$. Note that since p is prime, $\mathbb{Z}_p$, the ring of integers modulo p, is a field. In particular it has no zero-divisors. Therefore x^{-1} exists and so $P[h_a(x) = j/p] = P[a \cong x^{-1}(j-b) \mod p]$

Note that since $\mathbb{Z}_p$ is a field and $x \neq 0$, for any choice of $u \in [p]$ there exists a

unique element $v \in [p]$ such that $u \cong x^{-1}(j-v)$. So

$$\begin{aligned} P\left[a \cong x^{-1}(j-b) \mod p\right] &= \sum_{u=0}^{p-1} P[a=u]\, P[b=v] \\ &= \sum_{u=0}^{p-1} \left(\frac{1}{p}\right)\left(\frac{1}{p}\right) \\ &= p\left(\frac{1}{p^2}\right) \\ &= \frac{1}{p} \end{aligned}$$

which is to say $h_a(x)$ is uniformly distributed in $\left\{0, \frac{1}{p}, \ldots, \frac{p-1}{p}\right\}$ □

Lemma 2.8.6. *Let $x, y \in [p]$ such that $x \neq y$. Then*

$$P\left[h_a(x) = j/p, h_a(y) = \frac{\ell}{p}\right] = \frac{1}{p^2}, \forall j, \ell \in [p]$$

thus $h_a(x)$ and $h_b(y)$ are pairwise independent.

Proof. Assume without loss of generality $x \neq 0$, thus $\exists x^{-1} \in \mathbb{Z}_p$

$$\begin{aligned} P\left[h_a(x) = j/p, h_a(y) = \frac{\ell}{p}\right] &= P\left[a = x^{-1}(j-b) \mod p, b = \ell - ay \mod p,\right] \\ &= P\left[a = x^{-1}(j-\ell+ay) \mod p, b = \ell - x^{-1}(j-b)y \mod p,\right] \\ &= P\left[a = x^{-1}(j-\ell)(1-x^{-1}y)^{-1} \mod p, b = (\ell - x^{-1}jy)(1-x^{-1})^{-1}\right. \\ &= \frac{1}{p^2} \end{aligned}$$

Thus $h_a(x)$ and $h_a(y)$ are pairwise independent □

Independence of a finite set of random variables means that any subset of the random variables should have the product property. That is given random variables $X_1, \ldots, X_N$ and some sub-sequence of length k, $j_1, \ldots, j_k \in [N]$ then

$$P[X_{j_1} = a_1, \ldots, X_{j_k} = a_k] = \prod_{\ell=1}^{k} P[X_{j_\ell} = a_\ell]$$

Observe that the hash function described in this section, and the resulting random variables of the type $h_a(x_1), \ldots, h_a(x_n)$ have the product property only for pairs. As we will see, it does not hold for other subsets, in particular and set that has three or more distinct members. To see why the hash is not three-wise independent, consider $P\left[h_a(x) = j/p, h_a(y) = \ell/p, h_a(z) = f/p\right]$. We can show that once two values are known, because h_a is a line, all subsequent values will be deterministic. That is, from the proof of Lemma 2.8.1 we have that

$$a = x^{-1}(j - \ell)(1 - x^{-1}y)^{-1} \mod p$$
$$b = (\ell - x^{-1}yj)(1 - x^{-1}y)^{-1} \mod p$$

So then

$$h_a(z) = \frac{az + b \mod p}{p} = \frac{(1 - x^{-1}y)^{-1}\left[(j - \ell)z + (\ell - x^{-1}yj)\right]}{p} = f_{x,y,z,\ell,j}$$

I.e. once we've selected values for j and ℓ there is only one possible value that $h_a(z)$ can hash to - it is completely determined by those values. So

$$P\left[h_a(x) = j/p, h_a(y) = \ell/p, h_a(z) = f/p\right] = \begin{cases} \frac{1}{p^2} & \text{if } f = (1 - x^{-1}y)^{-1}\left[(j - \ell)z + (\ell - x^{-1}yj)\right] \\ 0 & \text{otherwise} \end{cases}$$

If the values were in fact independent, we would need that the probability for any particular three values as $1/p^3$. Hash functions that have k-wise independence exist, however are beyond the scope of this course. We conclude our discussion of this hash function with a consideration for how large a prime p we should choose. In the example of a perfect hash h we have that $P\left[h(x) = h(y)\right] = 0$ when $x \neq y$. Using

our imperfect hash h_a, we see that for $x \neq y \in [M]$

$$\begin{aligned} P\left[h_a(x) = h_a(y)\right] &= \sum_{j=0}^{p-1} P\left[h_a(x) = \frac{j}{p}, h_a(y) = \frac{j}{p}\right] \\ &= \sum_{j=0}^{p-1} \frac{1}{p^2} \\ &= \frac{1}{p} \end{aligned}$$

We can use this result for a fixed pair of values x, y to union bound the probability that at least two distinct elements in our dictionary of possible inputs hash to the same value

$$\begin{aligned} P\left[\exists x, y \in [M] h_a(x) = h_a(y)\right] &\leq \binom{M}{2} \frac{1}{p} \\ &= \frac{M(M-1)}{2p} \end{aligned}$$

So if $p \geq \frac{M(M-1)}{2q}$ for some $q \in (0,1)$ then

$$\begin{aligned} P\left[\nexists x, y \in [M] h_a(x) = h_a(y)\right] &= 1 - P\left[\exists x, y \in [M] h_a(x) = h_a(y)\right] \\ &\geq 1 - q \end{aligned}$$

In light of this, a common heuristic for choosing how large to make the prime is $p \geq M^3$. Note that the probability calculation shown above is the same type of calculation involved in the birthday problem.

Chapter III

A Break from Probability: Linear Johnson-Lindenstrauss (LJL) Emeddings as Deterministic Objects with Applications in Numerical Linear Algebra (MTH 994 Lectures 2 – 4 & 6) & (CMSE 890 Lecture 5)

3.1 Johnson-Lindenstrauss Maps (MTH 994 Lecture 2)

Definition 3.1.1 (ϵ-JL map)**.** A matrix $\Phi \in \mathbb{C}^{m\times N}$ is a linear ϵ-JL map of a set $S \subset \mathbb{C}^N$ into $\mathbb{C}^m$ if

$$\|\Phi\mathbf{x}\|_2^2 = (1+\epsilon_x)\|\mathbf{x}\|_2^2$$

holds for some $\epsilon_x \in (-\epsilon, \epsilon)$ for all $\mathbf{x} \in S$

Definition 3.1.2 (Set difference)**.** Let $\tilde{S} \subset \mathbb{C}^N$, then the set difference of $\tilde{S}$ denoted $\tilde{S} - \tilde{S}$ is $\left\{\mathbf{x}-\mathbf{y} | \mathbf{x}, \mathbf{y} \in \tilde{S}\right\} \in \mathbb{C}^N$

Note. *When $\epsilon \in (0,1)$, and Φ a ϵ-JL map, then Φ satisfies 1.3 for $x \in \mathcal{F}_{\mathbf{p}} - \mathcal{F}_{\mathbf{p}}$, $X = \ell^2(\mathbb{C}^m)$ and $Y = \ell^2(\mathbb{C}^N)$.*

Perhaps surprisingly, there are simple ways to construct ϵ-JL maps which, other than an upperbound on the cardinality, do not depend on any particular property of S. We will see that by drawing Φ as a random matrix, in a variety of ways, independent of S will result in Φ being a ϵ-JL map for S so long as $m \geq C\log(|S|)$, where C is a (mild) absolute constant.

Theorem 3.1.3. *Let $S \subset \mathbb{C}^N$ be finite. Then there exists a linear ϵ-JL map $\Phi \in \mathbb{C}^{m\times N}$ of the set S into $\mathbb{C}^m$ where $m \leq \frac{C}{\epsilon^2}\log|S|$, and $C \in (0,\infty)$ is an absolute constant independent of both (S and ϵ). Furthermore, Φ may be generated with high probability for any $S \subset \mathbb{C}^N$ given only knowledge of $|S|$, the cardinality of the set.*

We will delay the proof of this theorem until lecture 5, and instead take it as given and work to understand its meaning and consequences.

In order to explain the meaning of the theorem, consider the following two-player game between Alice and Bob. This game will proceed in two phases. In the first phase, Alice selects a finite subset S of the space $\mathbb{C}^N$. The content of S is known only to Alice, however the dimension N of the space and the cardinality $|S|$ is information available to Bob. In the second phase, Alice provides an error bound $\epsilon \in (0,1)$ and Bob must generate an ϵ-JL map Φ for Alice's set S which has at most m rows where $m \leq \frac{C}{\epsilon^2}\log|S|$ rows.

Bob wins the game if Φ is an ϵ-JL map of S into $\mathbb{C}^m$, otherwise Alice wins.

At first gloss, we may suppose that Alice has the advantage in this game. After-all, she determines in whatever way she may wish a set S and keeps most of that information secret. Bob has the seemingly more difficult task of mapping a set he knows little about to a lower dimensional space with a distortion error specified by his opponent. To abuse a metaphor, Bob has to draw a faithful, recognizable picture of an object that Alice dreams while he is blindfolded. Shouldn't Alice be able to come up with a set and error for which this is difficult to achieve? Surprisingly, Theorem 3.1.3 states that with high probability Bob will win the game simply by generating a random matrix $\Phi \in \mathbb{C}^{m\times N}$, no matter the set S Alice produces.

Also, note that should $m \geq N$ then the existence of an ϵ-JL map Φ is of little practical use and we could in that case construct isometric embeddings trivially. We

are interested in ranges of values ϵ and $|S|$ which lead to compression, i.e. $m \leq N$.

Example 3.1.4 (Generating ϵ-JL maps)**.** The following random processes generate Φ as in Theorem 3.1.3 with high probability. Note that these are data oblivious maps - they do not depend on any property of the set S other than the cardinality.

Each entry $\Phi_{(i,j)}$ is a i.i.d where

$$\Phi_{(i,j)} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1) = \mathcal{N}(0, \frac{1}{m})$$

Each entry $\Phi_{(i,j)}$ is a i.i.d

$$\Phi_{(i,j)} \sim \begin{cases} \frac{1}{\sqrt{m}} & \text{with probability } 1/2 \\ -\frac{1}{\sqrt{m}} & \text{with probability } 1/2 \end{cases}$$

The matrix Φ may be a sparse JL matrix containing at most $\mathcal{O}(\epsilon^{-1} \log |S|)$ nonzero entries in each column, and thus there are proportionally $\mathcal{O}(\epsilon)$ nonzero entries of Φ

The above examples have a drawback in that they require $\mathcal{O}(nM)$ memory to store and do not permit a fast matrix-vector multiply - afterall a matrix with independent random values does not have a structure we can exploit to save on storing or in applying to vectors. Our next example will show how we may achieve maps which do have better performance. First however we will need a theorem.

Theorem 3.1.5. *Let $U \in \mathbb{C}^{N\times N}$ be a unitary matrix with entries bounded by $\max_{n,k\in[N]} |U_{k,n}| \leq \frac{K}{\sqrt{N}}$. Let $R \in \mathbb{C}^{m\times N}$ be a matrix obtained by selecting m-rows from the $N \times N$ identity matrix i.i.d. uniformly at random and let $D \in \mathbb{R}^{N\times N}$ be a diagonal matrix with i.i.d ± 1 Radamacher random values on its diagonal. Then*

$$\sqrt{\frac{N}{m}}RUD$$

will be an ϵ-JL map of any given $S \subset \mathbb{R}^N$ into $\mathbb{C}^m$ with probability at least $1 - p - N^{-\ln^3 N}$ provided that

$$m \geq c\frac{K^2}{\epsilon^2} \log\left(\frac{4|S|}{p}\right) \log^4 N$$

where $c \in \mathbb{R}^+$ is an absolute constant.

Note that D can be applied in $\mathcal{O}(N)$ time to a vector in $\mathbf{x} \in \mathbb{R}^N$, since it involves scanning the length of the vector and changing signs of some entries. The matrix R can be applied to a vector of length N, $UD\mathbf{x}$ in $\mathcal{O}(N)$-time since it involves scanning the length of the vector and discarding values. Applying U then, since generically it should require at least reading in the inputs should be at least $\mathcal{O}(N)$. Therefore the overall complexity is governed principally by U. If the unitary matrix U admits a fast matrix-vector, like for example using the Fast Fourier transforms to effect a Discrete Fourier transform, then U can be applied to $D\mathbf{x}$ in $\mathcal{O}(N \log N)$-time.

In practice, the $\log^4 N$ factor in the bound of m is often ignored with no change in performance.

The failure probability bound $p + N^{-\ln^3 N}$ is a result of union bounding over two events – $\frac{1}{\sqrt{m}}RF$ failing to have the yet undefined RIP property with probability at most $N^{-\ln^3 N}$; the details of that can be found in Theorem 4.3.6, and the probability of at most p that $\sqrt{\frac{N}{m}}RFD$ fails to be an $\frac{\epsilon}{2}$-JL map; details of which can be found in Theorem 4.4.4. Here, since we generally concern ourselves with $N \gg 1$, the failure probability can be made suitably small.

Example 3.1.6. Let F be a unitary discrete Fourier transform matrix

$$F_{n,k} = \frac{1}{\sqrt{N}} e^{\frac{2\pi i n k}{N}}$$

If we take $U = F$ and R and D be the matrices described in Theorem 3.1.5 then

$\sqrt{\frac{N}{m}}RFD$ is a JL map where

$$\max_{n,k} |F_{n,k}| = \frac{1}{\sqrt{N}}$$

so $K = 1$. The Fast Fourier Transform (FFT) can be used to apply F to any vector $\mathbf{x} \in \mathbb{R}^N$ in $\mathcal{O}(N \log N)$. This is the proto-typical "Fast JL Matrix."

We now consider a direct application of JL-maps to the Nearest Neighbor problem introduced in 1.1.4. Recall in the Nearest Neighbor problem we want to find for each element in a set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$ another element which is closest with respect to $\|\cdot\|_2$. Here we consider the case where $D \gg \log N$. The naive solution requires $\mathcal{O}(DN^2)$-time. How can we improve on this using a JL map? Note that $S - S$ as in Definition 3.1.2 has the following bound on its cardinality

$$|S - S| \leq N^2 - N + 1$$

Let $\Phi \in \mathbb{C}^{m \times D}$ is an ϵ-JL map of $S - S$ with $m \geq \mathcal{O}\left(\frac{\log |S|}{\epsilon^2}\right)$. Suppose then we apply the map to all elements of S, so $S' = \{\Phi \mathbf{x}_j | \mathbf{x}_j \in S\}$ and then perform naive nearest neighbors on the set S'.

Finding set S' takes $\mathcal{O}\left(\frac{ND}{\epsilon^2} \log N\right)$-time and naive nearest neighbors on S' takes $\mathcal{O}\left(\frac{N^2}{\epsilon^2} \log N\right)$-time. When $D \gg \log N$, note

$$\frac{N}{\epsilon^2} \log N(D + N) < N^2 D$$

Note that it possible to combine the JL compression and LSH solution approach to (c, r)-NN for faster speed ups.

Note. *A linear ϵ-JL map Φ of $S \subset \mathbb{C}^N$ with $\epsilon \in (0, 1)$ must have $S \cap Ker\Phi = \emptyset$. The requirement that the null space of the map is disjoint from the set S can be stated more precisely as uniformly bounding the operator norm on S.*

Definition 3.1.7. Let $\Phi \in \mathbb{C}^{m\times N}$ and $S \setminus \{0\} \subset \mathbb{C}^N$ be nonempty. The operator norm of Φ on S denoted $\|\Phi\|_{S,2\to 2}$ is defined as

$$\|\Phi\|_{S,2\to 2} = \sup_{\mathbf{x}\in S\setminus\{0\}} \frac{\|\Phi\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Lemma 3.1.8. *$\|\Phi\|_{S,2\to 2}$ is a semi-norm if $S \setminus \{0\} \neq \emptyset$ and it is an norm if S contains N linearly independent vectors.*

Lemma 3.1.9. *$\|\Phi\|_{S,2\to 2} = \|\Phi\|_{2\to 2}$ if $\left\{\frac{\mathbf{y}}{\|\mathbf{y}\|_2}|\mathbf{y} \in S \setminus \{0\}\right\} = \overline{B(0,1)} \setminus B(0,1) = \delta B(0,1)$*

Lemma 3.1.10. *If Φ is an ϵ-JL map of S then the following inequalities hold*

$$\|\Phi\|_{S,2\to 2} \leq \sqrt{1+\epsilon}, \inf_{\mathbf{y}\in S\setminus\{0\}} \frac{\|\Phi\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \geq \sqrt{1-\epsilon}, \sup_{\mathbf{y}\in S\setminus\{0\}} \frac{|\langle(\Phi^*\Phi - I)\mathbf{y},\mathbf{y}\rangle|}{\|\mathbf{y}\|_2^2} \leq \epsilon$$

Norm preserving maps of certain sets which are geometrically related to S can also preserve the geometry of S itself.

Lemma 3.1.11. *Let $S \subset \mathbb{C}^N$ and $\epsilon \in (0,1)$. If $\Phi \in \mathbb{C}^{m\times N}$ is an ϵ-JL map of S' into $\mathbb{C}^m$, where*

$$S' = \left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + i\frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - i\frac{\mathbf{y}}{\|\mathbf{y}\|_2}|\mathbf{x},\mathbf{y} \in S\right\}$$

will satisfy $\forall \mathbf{x},\mathbf{y} \in S$

$$|\langle\Phi\mathbf{x},\Phi\mathbf{y}\rangle - \langle\mathbf{x},\mathbf{y}\rangle| \leq 4\epsilon\|\mathbf{x}\|_2\|\mathbf{y}\|_2 \tag{3.1}$$

Proof. Consider the case where $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$ then the inequality holds because $0 \leq 0$.

So next we suppose $\mathbf{x},\mathbf{y} \neq 0$. Consider the normalizations $\mathbf{u} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \mathbf{v} = \frac{\mathbf{y}}{\|\mathbf{y}\|_2}$. The polarization identity relates inner products with norms. Observe,

$$
\begin{aligned}
|\langle \Phi\mathbf{u}, \Phi\mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle| &= \left| \frac{1}{4} \sum_{\ell=0}^{3} i^{\ell} \left(\|\Phi\mathbf{u} + i^{\ell}\Phi\mathbf{v}\|_2^2 - \|\mathbf{u} + i^{\ell}\mathbf{v}\|_2^2 \right) \right| \\
&= \left| \frac{1}{4} \sum_{\ell=0}^{3} i^{\ell} \epsilon_{\mathbf{u},\mathbf{v},\ell} \|\mathbf{u} + i^{\ell}\mathbf{v}\|_2^2 \right| \\
&\leq \frac{1}{4} \sum_{\ell=0}^{3} \epsilon \left(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2 \right)^2 \\
&= 4\epsilon
\end{aligned}
$$

□

Note that constructing Φ still only depends on the original information about the cardinality of S since $|S'| \leq 4|S|^2$, and so we can apply Theorem 3.1.3 whether we have $|S|$ or $|S'|$.

Lemma 3.1.11 and Theorem 3.1.3 imply that $\exists \Phi \in \mathbb{C}^{m \times N}$ a ϵ-JL map where the inequality 3.1 holds for any choice of S provided that

$$m = \frac{c}{\epsilon^2} \log\left(4|S|^2\right)$$

Should we wish to construct a ϵ-JL map with fast matrix-vector multiply, we can use an RFD matrix like that in Example 3.1.6, and our sketching dimension m is

$$m = \frac{c}{\epsilon^2} \log\left(32|S|^2\right) \log^4 N$$

We can use 3.1.11 to implement a fast approximate matrix multiplication algorithm. We will show how to use any type of ϵ-JL map to achieve the speed-up.

Lemma 3.1.12. *Let $V \in \mathbb{C}^{N \times p}$ and $U \in \mathbb{C}^{N \times q}$ have unit ℓ^2-normalized columns. Suppose that $\Phi \in \mathbb{C}^{m \times N}$ satisfies Equation 3.1 from Lemma 3.1.11 where $S =$*

$\{\mathbf{u}_j | \mathbf{u}_j = U[:, j]\} \cup \{\mathbf{v}_j | \mathbf{v}_j = V[:, j]\}$. *Then*

$$\left|(V^*\Phi^*\Phi U - V^*U)_{k,j}\right| \leq 4\epsilon, \, \forall k \in [p], \forall j \in [q]$$

Proof. Note that $|S| = p + q$ thus $S' = 4(p+q)^2$ from Lemma 3.1.11. Furthermore note that

$$\Phi V = \begin{pmatrix} | & | & \dots & | \\ \Phi\mathbf{v}_1 & \Phi\mathbf{v}_2 & \dots & \Phi\mathbf{v}_p \\ | & | & \dots & | \end{pmatrix}, \, \Phi U = \begin{pmatrix} | & | & \dots & | \\ \Phi\mathbf{u}_1 & \Phi\mathbf{u}_2 & \dots & \Phi\mathbf{u}_q \\ | & | & \dots & | \end{pmatrix}$$

So note then that $\left((\Phi V)^* \Phi U\right)_{k,j} = \langle \Phi\mathbf{v}_k, \Phi\mathbf{u}_j \rangle$. Therefore for all choices of k, j and given Lemma 3.1.11 we have

$$\begin{aligned} \left|(V^*\Phi^*\Phi U - V^*U)_{k,j}\right| &= |\langle \Phi\mathbf{v}_k, \Phi\mathbf{u}_j \rangle - \langle \mathbf{v}_k, \mathbf{u}_j \rangle| \\ &\leq 4\epsilon \|\mathbf{v}_k\|_2 \|\mathbf{u}_j\|_2 \\ &= 4\epsilon \end{aligned}$$

□

Note that we know there exists $\Phi \in \mathbb{C}^{m \times N}$ that satisfies the needed inequality from Lemma 3.1.11 such that

$$m = \mathcal{O}\left(\epsilon^{-2} \log\left(\max(p, q)^2\right)\right)$$

Theorem 3.1.13 (Fast Matrix-Matrix Multiply)**.** *Let $A \in \mathbb{C}^{p \times N}$ and $B \in \mathbb{C}^{N \times q}$ have SVDs given by $A = U_1\Sigma_1 V^*$ and $B = U\Sigma_2 V_2^*$ and suppose that $\Phi \in \mathbb{C}^{m \times N}$ satisfies the conditions of Lemma 3.1.12 for U and V. Then*

$$\|A\Phi^*\Phi B - AB\|_F \leq 4\epsilon \|A\|_F \|B\|_F$$

Proof. We will expand the quantity of interest according the SVD of the factors A and B

$$\begin{aligned}
\|A\Phi^*\Phi B - AB\|_F &= \|U_1\Sigma_1 V^*\Phi^*\Phi U\Sigma_2 V_2^* - U_1\Sigma_1 V^* U\Sigma_2 V_2^*\|_F \\
&= \|U_1\Sigma_1\left(V^*\Phi^*\Phi U - V^*U\right)\Sigma_2 V_2^*\|_F \\
&= \|\Sigma_1\left(V^*\Phi^*\Phi U - V^*U\right)\Sigma_2\|_F \\
&= \sqrt{\sum_{k=1}^{p}\sum_{j=1}^{q}(\Sigma_1)_{k,k}^2\,|V^*\Phi^*\Phi U - V^*U|_{k,j}^2\,(\Sigma_2)_{j,j}^2} \\
&\leq \sqrt{\sum_{k=1}^{p}\sum_{j=1}^{q}\sigma_k(A)^2(4\epsilon)^2\sigma_j(B)^2} \\
&= 4\epsilon\sqrt{\sum_{k=1}^{p}\sigma_k(A)^2}\sqrt{\sum_{j=1}^{q}\sigma_j(B)^2} \\
&= 4\epsilon\|A\|_F\|B\|_F
\end{aligned}$$

□

What are the savings in runtime then if we wish to approximate matrix-matrix multiplication in this way? To simplify the comparison suppose $p, q = N$ (or at comparable at any rate). Usual matrix multiplication then consists of computing N^2 entries, each consisting of the inner product of two N dimensional vectors, i.e. $\mathcal{O}(N^3)$. If we use Theorem 3.1.13 then there are three major operations to consider

1. Compute the product ΦB which takes $\mathcal{O}(mN^2)$
2. Compute the product $\Phi A*$ which takes $\mathcal{O}(mN^2)$. Conjugate transposition takes at most $\mathcal{O}(N^2)$ operations
3. Compute $\left(A\Phi^*\right)\left(\Phi B\right)$ which takes $\mathcal{O}(nN^2)$

We have seen from Lemma 3.1.11 and Theorem 3.1.3 that $m = m = \mathcal{O}\left(\epsilon^{-2}\log N\right)$. So the total runtime for the approximate matrix-matrix multiplication is $\mathcal{O}(\frac{N^2}{\epsilon^2}\log N)$.

Lemma 3.1.14. *Let $S \subset \mathbb{R}^N$ and $\epsilon \in (0,1)$, then an ϵ-JL map $\Phi \in \mathbb{C}^{m\times N}$ of the set*

$$S' = \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} + \frac{\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2} | \mathbf{x}, \mathbf{y} \in S \right\}$$

will satisfy $\forall \mathbf{x}, \mathbf{y} \in S$

$$|\Re\left(\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle\right) - \langle \mathbf{x}, \mathbf{y}\rangle| \leq 2\epsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

Up to this point, Theorem 3.1.3 and the subsequent discussion has dealt with finite sets of points S. We now turn to the question of whether these results can be applied to infinite sets. We will begin by building new from old; infinite sets which are constructed from finite ones.

Definition 3.1.15 (Cones)**.** The conical region generated by $S \subset \mathbb{C}^N$ is

$$\text{cone}(S) = \{\alpha x | x \in S, \alpha \in \mathbb{C}\}$$

As an immediate consequence of the linearity, an ϵ-JL map Φ of set S, will be an ϵ-JL map of cone(S) and vice-versa.

Another infinite set of importance is the convex hull of a set of points.

Definition 3.1.16 (Convex Hulls)**.** The convex hull of a $S \subset \mathbb{C}^N$ is

$$\text{conv}(S) = \bigcup_{j=1}^{\infty} \left\{ \sum_{\ell=1}^{j} \alpha_\ell \mathbf{x}_\ell | x_1, \ldots, x_j \in S,\ \alpha_1, \ldots, \alpha_j \in [0,1] \text{s.t.} \sum_{\ell=0}^{N} \alpha_\ell = 1 \right\}$$

We have in this next theorem that the infinitude of points in the convex hull can always be reduced to a finite number of points from the original set. That is that each point in conv(S) where $S \subset \mathbb{R}^N$ can be expressed as a convex combination of at most $N+1$ point from S.

Theorem 3.1.17 (Caratheodory)**.** *Given $S \in R^N$, $\forall \mathbf{x} \in conv(S)$, $\exists \mathbf{y}_1, \ldots, \mathbf{y}_{\tilde{N}}$, $\tilde{N} = \min(|S|, N+1)$, such that $\mathbf{x} = \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{y}_\ell$ for some $\alpha_1, \ldots, \alpha_{\tilde{N}} \in [0,1]$, $\sum_{\ell=1}^{\tilde{N}} \alpha_\ell = 1$.*

Theorem 3.1.18. *Suppose $S \subset \overline{B_{\ell^2}(\mathbf{0}, \gamma)} \subset \mathbb{R}^N$ and $\epsilon \in (0,1)$. Let $\Phi \in \mathbb{C}^{m \times N}$ be an $\left(\frac{\epsilon}{4\gamma^2}\right)$-JL map of the set S' defined as in Lemma 3.1.14, then*

$$|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| \leq \epsilon \tag{3.2}$$

$\forall \mathbf{x}, \mathbf{y} \in conv(S)$

Proof. Let $\mathbf{x}, \mathbf{y} \in \text{conv}(S)$. By Theorem 3.1.17, $\exists \{\mathbf{y}_i\}_{i=1}^{\tilde{N}}$, $\{\mathbf{x}_i\}_{i=1}^{\tilde{N}} \subset S$, $\{\alpha_\ell\}_{\ell=1}^{\tilde{N}}, \{\beta_\ell\}_{\ell=1}^{\tilde{N}} \subset [0,1]$ such that

$$\mathbf{x} = \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{x}_\ell, \ \mathbf{y} = \sum_{\ell=1}^{\tilde{N}} \beta_\ell \mathbf{y}_\ell$$

So,

$$\begin{aligned}
|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| &= \left| \sum_{\ell=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \alpha_\ell \beta_j \langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle \right| \\
&\leq 4 \sum_{\ell=1}^{\tilde{N}} \sum_{j=1}^{\tilde{N}} \alpha_\ell \beta_j \left(\frac{\epsilon}{4\gamma^2} \right) \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \\
&\leq \epsilon \left(\sum_{\ell=1}^{\tilde{N}} \alpha_\ell \right) \left(\sum_{j=1}^{\tilde{N}} \beta_j \right) \\
&= \epsilon
\end{aligned}$$

where we have used the embedding error $\left(\frac{\epsilon}{4\gamma^2}\right)$ and the fact that all norms of vectors in this case will be less than γ □

Corollary 3.1.19. *If $\inf_{x \in conv(S)} \|x\|_2 \geq 1$ then Φ as in Theorem 3.1.18 will also be an ϵ-JL map of $conv(S)$ into $\mathbb{C}^m$*

Proof. Consider 3.2 with $\mathbf{x} = \mathbf{y}$ to obtain

$$\left| \|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \epsilon \leq \epsilon \|\mathbf{x}\|_2^2$$

which is the ϵ-JP property. □

For points near zero, for example any points in the intersection $S \cap B_{\ell^2}(\mathbf{0}, \epsilon)$, Theorem 3.1.18 does not provide useful relative errors bounds. This occurs because the left hand side of 3.2 can be very small for these vectors near zero relative to the fixed ϵ.

Recall, S is finite and (usually) $\text{conv}(S)$ is infinite.

However, with the Corollary 3.1.19, when $\mathbf{x} = \mathbf{y}$ and $x \in \text{conv}(S) \setminus B_{\ell^2}(\mathbf{0}, \alpha)$ then we can achieve an ϵ-JL embedding of the convex hull less the ball about zero, $\text{conv}(S) \setminus B_{\ell^2}(\mathbf{0}, \alpha)$, by applying a $\left(\frac{\epsilon}{\alpha^2}\right)$-JL to S.

Homework 3.1.1. If B is the closed unit ball for any norm on $\mathbb{C}^N$, show that $B - B$ is the ball of radius 2.

Homework 3.1.2. Prove lemma 3.1.8

Homework 3.1.3. Prove lemma 3.1.9

Homework 3.1.4. Prove lemma 3.1.10

Homework 3.1.5. Prove lemma 3.1.14

Homework 3.1.6. Show that Theorem 3.1.18 still holds if $S \in \overline{B_{\ell^2}(\mathbf{0}, \gamma)} \subset \mathbb{C}^N$.

Homework 3.1.7. Let $A \in \mathbb{C}^{m \times N/2}$ be an ϵ-JL map of $T \cup S \subset \mathbb{C}^{N/2}$. Then for $\mathbf{x}_1, \mathbf{x}_2 \in S \cup T$, $g : \mathbb{C}^N \to \mathbb{C}^m$, defined by $g(\mathbf{x}_1, \mathbf{x_2}) = (A\mathbf{x_1}, A\mathbf{x_2})$ is an ϵ-JL embedding of $(S \times T) \cup (T \times S) \cup (S \times S) \cup (T \times T)$

Homework 3.1.8. Fix $\epsilon \in (0, 1)$ and let $A \in \mathbb{C}^{\tilde{m} \times N}$ be a ϵ-JL map of $(S - S) \cup S$ and $G \in \mathbb{C}^{m \times \tilde{m}}$ be an ϵ-JL embedding of $A(S) \subset \mathbb{C}^{\tilde{m}}$, $S \subset \mathbb{C}^N$ then

1. $|A(S)| = |S|$

2. GA is a 3ϵ-JL embedding of S into $\mathbb{C}^m$

3.2 Covering Numbers of Balls (MTH 994 Lecture 3)

Next we turn to covering numbers, which will enable us to apply ϵ-JL maps to more general infinite sets. We begin then with these definitions.

Definition 3.2.1 (δ-cover)**.** Let $T \subseteq \mathbb{C}^N$. A δ-cover of T with respect to norm $\|\cdot\|_X$ is a subset of $S \subseteq T$ such that $\forall \mathbf{x} \in T$, $\exists \mathbf{y} \in S$ with $\|\mathbf{x} - \mathbf{y}\|_X < \delta$ where

$$T \subseteq \bigcup_{\mathbf{y} \in S} B_X(\mathbf{y}, \delta)$$

Note that $B_X(\mathbf{y}, \delta)$ is the open ball with center $\mathbf{y} \in \mathbb{C}^N$ and radius δ with respect to the norm $\|\cdot\|_X$. Usually it will be clear from context the space and norm, and so we'll simplify notation and write instead $B(\mathbf{y}, \delta)$

Definition 3.2.2 (δ-covering Number)**.** The δ-covering number of $T \subseteq \mathbb{C}^N$, denoted $C_\delta^X(T)$ with respect to $\|\cdot\|_X$ is the smallest integer such that there exists a δ-cover $S \subseteq T$ where $|S| = C_\delta^X(T)$. If no such integer exists we say that $C_\delta^X(T) = \infty$

Definition 3.2.3 (δ-packing)**.** Let $T \subseteq \mathbb{C}^N$. A δ-packing of T with respect to norm $\|\cdot\|_X$ is a subset of $S \subseteq T$ such that $\forall \mathbf{x}, \mathbf{y} \in S$, $\mathbf{x} \neq \mathbf{y}$ with $\|\mathbf{x} - \mathbf{y}\|_X \geq \delta$ then

$$B_X(\mathbf{x}, \delta/2) \cap B_X(\mathbf{y}, \delta/2) = \emptyset$$

Definition 3.2.4 (δ-packing Number)**.** The δ-packing number of $T \subseteq \mathbb{C}^N$, denoted $P_\delta^X(T)$ with respect to $\|\cdot\|_X$ is the largest integer such that there exists a δ-packing $S \subseteq T$ where $|S| = P_\delta^X(T)$. If no such integer exists we say that $P_\delta^X(T) = \infty$

Lemma 3.2.5. *Let $T \subseteq \mathbb{C}^N$ and $\delta \in (0, \infty)$. Then*

$$P_{2\delta}^X(T) \leq C_\delta^X(T) \leq P_\delta^X(T)$$

Proof. Let $P_{2\delta} \subset T$ be a maximal 2δ-packing of T and $C_\delta \subseteq T$ be a minimal δ-cover of T. Each point $\mathbf{x} \in P_{2\delta}$ is closest to a different point $\mathbf{y} \in C_\delta$. To see this, suppose to the contrary that for $\mathbf{x}_1, \mathbf{x}_2, \in P_{2\delta}, \mathbf{x}_1 \neq \mathbf{x}_2$ there was some point $\mathbf{y} \in C_\delta$ such that $\mathbf{x}_1, \mathbf{x}_2 \in B(\mathbf{y}, \delta)$ this implies that $\mathbf{y} \in B_X(\mathbf{x}_1, \delta) \cap B_X(\mathbf{x}_2, \delta)$ which is a contradiction.

So, since each point in P_δ can be identified with at least one point in C_δ. We thus define an injection $f : P_{2\delta} \to C_\delta$ where $f(\mathbf{x}) = \mathbf{y}$, $x \in B(\mathbf{y}, \delta)$. Since f is an injection, we have that the cardinality of C_δ must be equal to or larger than $P_{2\delta}$, which is equivalent to the left hand side of the desired inequality.

Next, suppose P_δ is a maximal δ-packing of T. Now suppose for eventual contradiction that there exists a point $\mathbf{y} \in T, \mathbf{y} \notin P_\delta$ such that $\|\mathbf{x} - \mathbf{y}\| \geq \delta$, $\forall x \in P_\delta$. This implies that $B_X(\mathbf{x}, \delta/2) \cap B_X(\mathbf{y}, \delta/2) =$. Thus $P_\delta \cup \{\mathbf{y}\}$ is a δ-packing of T. This contradicts that P_δ is maximal. So, for all points $\mathbf{y} \in T$, there is $\mathbf{x} \in P_\delta$ such that $\|\mathbf{x} - \mathbf{y}\| \leq \delta$, which is to say P_δ is a δ-covering of T, and therefore the cardinality of P_δ is equal to or larger than the δ-covering number for T. This is the right hand side of the desired inequality. □

Lemma 3.2.6. *Let $T \subseteq \mathbb{R}^N$ and $\delta \in (0, \infty)$. Furthermore let B denote the unit ball $B_X(0, 1)$ in $\mathbb{R}^N$ with respect to some norm $\|\cdot\|_X$. Then*

$$\left(\frac{1}{\delta}\right)^N \frac{Vol(T)}{Vol(B)} \leq C_\delta^X(T) \leq P_\delta^X(T) \leq \left(\frac{2}{\delta}\right)^N \frac{Vol\left(T + \left(\frac{\delta}{2}\right) B\right)}{Vol(B)}$$

holds, where $Vol(T) = \int_T 1 dV$, the Lebesgue measure of T in $\mathbb{R}^N$.

Note that addition of sets is syntactical sugar for set difference of certain sets, i.e. $T + S = T - (-S) = \{t + s | \forall t \in T, s \in S\}$

Proof. Suppose C_δ is a minimal δ cover of T. By definition then of δ-cover

$$T \subseteq \bigcup_{\mathbf{y} \in C_\delta} B(\mathbf{y}, \delta)$$

So, using sub-additivity of measurable sets, translation invariance, and scaling we have

$\text{Vol}(T) \leq \text{Vol}\left(\bigcup_{\mathbf{y} \in C_\delta} B(\mathbf{y}, \delta)\right) \leq C_\delta^X \text{Vol}\left(B(\mathbf{y}, \delta)\right) = C_\delta^X \delta^N \text{Vol}\left(B(0,1)\right)$

Rearranging terms, we obtain the left hand side of the desired inequality

$$\left(\frac{1}{\delta}\right)^N \frac{\text{Vol}\left(T\right)}{\text{Vol}\left(B\right)} \leq C_\delta^X(T)$$

Now suppose P_δ is a maximal δ-packing of T. It follows that

$$\bigcup_{\mathbf{y} \in P_\delta} B(\mathbf{y}, \delta/2) \subset T + B(0, \delta/2)$$

Since the balls that make up the δ-packing of T are disjoint, we have that their measure is additive. Again, using translation invariance and scaling, this implies

$$\text{Vol}\left(\bigcup_{\mathbf{y} \in P_\delta} B(\mathbf{y}, \delta/2)\right) = P_\delta^X(T) \left(\frac{\delta}{2}\right)^N \text{Vol}\left(B(\mathbf{0}, 1)\right) \leq \text{Vol}(T + B(0, \delta/2))$$

Which after rearranging terms matches the right hand side of the desired inequality □

Corollary 3.2.7. $\left(\frac{1}{\delta}\right)^N \leq C_\delta^X\left(B\right) \leq \left(1 + \frac{2}{\delta}\right)^N$ *for all norms* $\|\cdot\|_X$ *on* $\mathbb{R}^N$

Proof. We can apply lemma 3.2.6 where $T = B(0,1)$. So,

$$\left(\frac{1}{\delta}\right)^N \frac{\cancel{\text{Vol}\left(B\right)}}{\cancel{\text{Vol}\left(B\right)}} \leq C_\delta^X(T)$$

yields the first half of the corollary. Next, note that using $T = B(0,1)$ we see that

$$B(0,1) + B\left(0, \frac{\delta}{2}\right) \subseteq B\left(0, \left(1 + \frac{\delta}{2}\right)\right)$$

Note that by scaling, we have the following for the volume calculation

$$\text{Vol}\left(B\left(0,\left(1+\frac{\delta}{2}\right)\right)\right) = \left(1+\frac{\delta}{2}\right)^N \text{Vol}(B(0,1))$$

Putting this into the previous lemma, we see

$$C_\delta^X(T) \leq \left(\frac{2}{\delta}\right)^N \left(1+\frac{\delta}{2}\right)^N \frac{\cancel{\text{Vol}(B(0,1))}}{\cancel{\text{Vol}(B)}} = \left(1+\frac{2}{\delta}\right)^N$$

Which completes the second inequality □

Note that if we add the assumption in the corollary that $\delta \in (0,1)$ then we can bound $\left(1+\frac{2}{\delta}\right) \leq \left(\frac{1}{\delta}+\frac{2}{\delta}\right) = \frac{3}{\delta}$, for a more concise, though less tight bound.

Corollary 3.2.8. *If $S \subseteq \overline{B_X(0,1)} \subset \mathbb{R}^N$ then*

$$C_\delta^X(T) \leq \left(1+\frac{2}{\delta}\right)^N$$

Proof. By observing

$$\text{Vol}\left(S + B\left(\mathbf{0},\frac{\delta}{2}\right)\right) \leq \text{Vol}\left(\overline{B_X(0,1)} + \overline{B\left(\mathbf{0},\frac{\delta}{2}\right)}\right)$$

and applying the same reasoning as corollary 3.2.7, we achieve the result. □

Homework 3.2.1. Consider the identification of $\mathbb{C}^N$ with $\mathbb{R}^{2N}$ given by the map

$$f: \begin{pmatrix} x_1 \\ \vdots \\ x_{2N} \end{pmatrix} \to \begin{pmatrix} x_1 + ix_2 \\ \vdots \\ x_{2N-1} + ix_{2N} \end{pmatrix}$$

1. Verify that $f: \mathbb{R}^{2N} \to \mathbb{C}^N$ is a bijection with

$$f^{-1}(c\mathbf{z}) = \begin{pmatrix} \Re(c)\Re(z_1) - \Im(c)\Im(z_1) \\ \Re(c)\Im(z_1) + \Im(c)\Re(z_1) \\ \vdots \end{pmatrix}, \forall c \in \mathbb{C}, \mathbf{z} \in \mathbb{C}^N$$

2. Show that $f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ and $f(c\mathbf{x}) = cf(\mathbf{x})$ both hold $\forall \mathbf{x}, \mathbf{y}\mathbb{R}^{2N}$ and $c \in \mathbb{R}$

3. Verify that $\|f(\mathbf{x})\|_2 = \|\mathbf{x}\|_2$, $\forall \mathbf{x} \in \mathbb{R}^{2N}$, i.e. f is an isometry

4. Let $\|\cdot\|_X$ be a norm on $\mathbb{C}^N$ over $\mathbb{C}$. Prove that $\|\cdot\|_{X'} : \mathbb{R}^{2N} \to [0,\infty)$ defined by $\|\cdot\|_{X'} = \|f(\mathbf{x})\|_X$ is a norm on $\mathbb{R}^{2N}$ over R.

5. Prove that $f\left(B_{X'}(\mathbf{y}, r)\right) = B_X\left(f(\mathbf{y}), r\right)$ for all $\mathbf{y} \in \mathbb{R}^{2N}$ and $r \in [0,\infty)$

Homework 3.2.2. For $T \subset \mathbb{C}^N$ we have $\mathrm{Vol}(T) = \mathrm{Vol}(f^{-1}(T))$. Modify the proofs of Lemma 3.2.6 and Corollary 3.2.7 to prove that

$$\left(\frac{1}{\delta}\right)^{2N} \leq C_\delta^X\left(B_X\right) \leq \left(\frac{3}{\delta}\right)^{2N}$$

for all norms $\|\cdot\|_X$ on $\mathbb{C}^N$ and $\delta \in (0,1)$

3.3 JL Subspace Embeddings and the Restricted Isometry Property (MTH 994 Lecture 4)

In this section we describe how we can embed a cover of the unit ball in a subspace and in turn prove that this map will in fact embed the entire subspace. To this end, we first fix some notation

Consider the ambient space $\mathbb{C}^N$. We denote an r-dimensional linear subspace of some orthonormal basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$ as

$$\mathcal{L}_{\mathcal{B}}^r = \{a\mathbf{x} | \alpha \in \mathbb{C}, \mathbf{x} \in S_{\mathcal{B}}^r\} \subset \mathbb{C}^N$$

Where $S_{\mathcal{B}}^r$ denotes the r-dimensional unit sphere with respect to the basis $\mathcal{B}$

$$S_{\mathcal{B}}^r = \left\{ \mathbf{x} \in \mathbb{C}^N | \mathbf{x} = \sum_{j=1}^{r} c_j \mathbf{b}_j, \mathbf{c} \in \mathbb{C}^r \text{ s.t. } \|\mathbf{c}\|_2 = 1 \right\}$$

Note that it is possible to represent any r-dimensional subspace as some $\mathcal{L}_{\mathcal{B}}^r$. Our strategy for proving the main result of this section will be then to embed a sufficiently dense cover of $S_{\mathcal{B}}^r \subset \mathbb{C}^N$

Theorem 3.3.1 (Subspace Embeddings)**.** *Fix $\epsilon \in (0,1)$. Let $\mathcal{L}_{\mathcal{B}}^r \subset \mathbb{C}^N$ be a r-dimensional subspace of $\mathbb{C}^N$ with respect to some orthonormal basis $\mathcal{B}$ and furthermore let $C \subset S_{\mathcal{B}}^r$ be a minimal $\left(\frac{\epsilon}{16}\right)$-cover of $S_{\mathcal{B}}^r \subset \mathcal{L}_{\mathcal{B}}^r$. Then if $\Phi \in \mathbb{C}^{m\times N}$ is an $\left(\frac{\epsilon}{2}\right)$-JL map of C into $\mathbb{C}^m$ it will also satisfy*

$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2, \forall \mathbf{x} \in \mathcal{L}_{\mathcal{B}}^r \tag{3.3}$$

Proof. Note that $S_{\mathcal{B}}^r$ is a compact set, and therefore will contain its maximal element $\mathbf{x}$. That is, $\exists \mathbf{x} \in S_{\mathcal{B}}^r$ such that

$$\|\Phi\mathbf{x}\|_2 = \|\Phi\|_{S_{\mathcal{B}}^r, 2\to 2} = \|\Phi\|_{L_{\mathcal{B}}^r} = \gamma$$

Let $\mathbf{y} \in C$ be such that $\|\mathbf{x}-\mathbf{y}\|_2 < \epsilon/16$. Then, after noting that $\mathbf{x}$ and $\mathbf{y}$ both are of unit norm, we have by the triangle inequality and $\left(\frac{\epsilon}{2}\right)$-JL property of Φ:

$$\begin{aligned}
\gamma - 1 &= \|\Phi\mathbf{x}\|_2 - \|\mathbf{x}\|_2 \\
&\leq \|\Phi\mathbf{y}\|_2 + \|\Phi(\mathbf{x}-\mathbf{y})\|_2 - \|\mathbf{x}\|_2 \\
&\leq \left(1+\frac{\epsilon}{2}\right)^{1/2}\|\mathbf{y}\|_2 + \frac{\epsilon}{2}\gamma - 1
\end{aligned}$$

After rearranging terms and using the inequality $(a+b)^2 \geq a^2+b^2$ and the fact that $\epsilon^2 < \epsilon$, we have

$$\gamma \leq \frac{1+\epsilon/4}{1-\epsilon/16} = 1+\epsilon/3$$

That is, we have shown that $\|\Phi\mathbf{x}\|_2^2 < 1+\epsilon$, $\forall x \in S_{\mathcal{B}}^r$. This establishes the required upper bound for our desired result.

For the lower bound, let $\beta = \inf_{\mathbf{z}\in S_{\mathcal{B}}^r \setminus \{0\}} \|\Phi\mathbf{z}\|_2$. The infimum is included in the set since the set is compact and the function continuous. Thus there exists $\mathbf{x} \in S_{\mathcal{B}}^r$,

with $\|\Phi\mathbf{x}\|_2 = \beta$. Now we take a point in the cover, $\mathbf{y} \in C$. So $\|\mathbf{x} - \mathbf{y}\|_2 < \epsilon/16$. Now we use the reverse triangle inequality to observe:

$$\begin{aligned}
\beta - 1 = \|\Phi\mathbf{x}\|_2 - 1 &\geq \|\Phi\mathbf{y}\|_2 - \|\Phi(\mathbf{x}-\mathbf{y})\|_2 - 1 \\
&\geq \left(1 - \frac{\epsilon}{2}\right)^{1/2} \|\mathbf{y}\|_2 - \gamma\left(\frac{\epsilon}{16}\right) - 1 \\
&\geq \left(1 - \frac{\epsilon}{2}\right) - \left(1 + \frac{\epsilon}{3}\right)\left(\frac{\epsilon}{16}\right) - 1 \\
&\geq \left(1 - \frac{\epsilon}{3}\right) - \left(1 + \frac{\epsilon}{3}\right)\left(\frac{\epsilon}{16}\right) - 1 \\
&= 1 - \frac{\epsilon}{3} - \frac{\epsilon}{16} - \frac{\epsilon^2}{48} - 1 \\
&\geq 1 - \frac{\epsilon}{3} - \frac{\epsilon}{16} - \frac{\epsilon}{48} - 1 \\
&\geq 1 - \frac{5\epsilon}{12} - 1
\end{aligned}$$

So $\beta \geq 1 - \frac{5\epsilon}{12} > 1 - \epsilon$ which is the desired lower bound. Having shown the inequality 3.3 holds for $\mathbf{x} \in S^r_{\mathcal{B}}$, we have that the inequality holds for $\mathcal{L}^r_{\mathcal{B}}$ by re-scaling and reducing to the previous case. □

1. Since $\mathcal{L}^r_{\mathcal{B}} - \mathcal{L}^r_{\mathcal{B}} \subseteq \mathcal{L}^r_{\mathcal{B}}$, so by merit of Φ being a JL map of $\mathcal{L}^r_{\mathcal{B}}$ it is also a JL map of the set difference of $\mathcal{L}^r_{\mathcal{B}}$ with itself, and so Lemma 3.1.11 applies, and we have that Φ approximately preserves inner products of vectors in the subspace. So both norms and angles are preserved, and so in some sense the geometry of the subspace is preserved by the map.

2. Use covering number bound in Lemma 3.2.7, we can conclude that $|C| \leq \left(\frac{48}{\epsilon}\right)^{2r}$, where C is a minimal cover of the r dimensional unit sphere. Thus we can construct Φ where

$$m = \frac{c}{\epsilon^2} \log |C| \leq \frac{\tilde{c}r}{\epsilon^2} \log \frac{48}{\epsilon}$$

Note we have optimal dependence on r. Note that the construction of Φ is oblivious - we do not need to know anything special about $\mathcal{L}_{\mathcal{B}}^r$; an upper on r will suffice in order to construct Φ

3. Fast $\sqrt{\frac{N}{m}}RFD$ ϵ-JL matrices have the row requirement

$$m = \frac{cr}{\epsilon^2}\log\left(\frac{48}{\epsilon}\right)\log^4 N$$

Corollary 3.3.2. *Let $\epsilon \in (0,1)$. There exists an ϵ-JL map $\Phi \in \mathbb{C}^{m\times N}$ of any given r-dimensional subspace $\mathcal{L}_{\mathcal{B}} \subset \mathbb{C}^N$ with $m \leq \frac{Cr}{\epsilon^2}\log\left(1+\frac{32}{\epsilon}\right)$ where $C \in \mathbb{R}^+$ is an absolute constant (independent of all $r, \epsilon, m, N, \mathcal{L}_{\mathcal{B}}^r$)*

The proof using Corollary 3.2.8, Theorems 3.1.3 and 3.3.1 is left as an exercise.

Recall the following property of orthonormal matrices:

Suppose $B \in \mathbb{C}^{N\times r}$ is the matrix formed by writing the orthonormal basis elements of $\mathcal{L}_{\mathcal{B}}^r$ as columns:

$$B = \begin{pmatrix} | & | & & | \\ b_1 & b_2 & \dots & b_r \\ | & | & & | \end{pmatrix}$$

Then $\|B^*\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$, $\forall \mathbf{x} \in \mathcal{L}_{\mathcal{B}}^r$.

In terms of the subject matter of this course, we can say that B is a 0-JL map. So why then are ϵ-JL maps of interest, if there is a common, well understood way to find lossless embeddings? In many settings however, we do not have detailed information about the subspace - for example we cannot easily find r linearly independent points, or in general sampling the space is costly.

Corollary 3.3.2 is an oblivious embedding; this means that we do not need to know what $\mathcal{L}_{\mathcal{B}}^r$ is in order to embed it into $\mathbb{C}^r$ accurately. One useful application that

fits this setting is finding the (approximate) principle Eigenspace for huge matrices. Another application is a fast, approximate solution to least squares problems.

Application 3.3.3 (Overdetermined Least Squares)**.** In the overdetermined least squares problem, we are tasked with finding a ℓ_2 minimizer $\mathbf{y}_{\min}$ to the matrix equation $A\mathbf{x} = b$ where $A \in \mathbb{C}^{N\times n}$, $N > n$, $\mathbf{b} \in \mathbb{C}^N$.

$$\mathbf{y}_{\min} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2$$

When $\mathbf{b}$ is in the range of A, then we will be able to find a solution which solves the equation exactly. However, if $\mathbf{b}$ does not lie in the range of A then $\mathbf{y}_{\min}$ will be the closest vector to $\mathbf{b}$ in the range of A.

A standard, classic solution approach to this uses QR decomposition and takes $\mathcal{O}(Nn^2)$ time. See Lecture 11 in [31]. We seek then then a solution approach which improves this runtime for $N \gg n$.

Theorem 3.3.4. *There exists universal constants $\bar{c}, c'$ such that a fast JL embedding matrix, $\sqrt{\frac{N}{m}}RFD \in \mathbb{C}^{m\times N}$ with*

$$m = \bar{c}(n+1)\ln\left(\frac{c'}{\epsilon}\right)\ln^4 N$$

will satisfy

$$(1-\epsilon)\|A\mathbf{y} - \mathbf{b}\|_2 \leq \sqrt{\frac{N}{m}}\|RFDA\mathbf{y} - RFD\mathbf{b}\|_2 \leq (1+\epsilon)\|A\mathbf{y} - \mathbf{b}\|_2$$

$\forall \mathbf{y} \in \mathbb{R}^n$ *with probability at least* $1 - p - N^{-\ln^3 N}$

Proof. Let $\mathcal{B} = \{\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{b}\}$ be the $n+1$ orthonormalized columns of A as well as $\mathbf{b}$. As before, let $\mathcal{L}_{\mathcal{B}}^{n+1}$ be the linear subspace spanned by the basis, and $S_{\mathcal{B}}^{n+1}$ the unit ball in the subspace. Let $C \subset S_{\mathcal{B}}^{n+1}$ be a minimal $\left(\frac{\epsilon}{16}\right)$ cover as in Theorem 3.3.1, and so $|C| \leq \left(\frac{48}{\epsilon}\right)^{n+1}$

Theorem 3.3.1 then implies that so long as Φ is a $\frac{\epsilon}{2}$-JL map of C then for each $\mathbf{y} \in \mathbb{C}^m$ we have $A\mathbf{y} - \mathbf{b} \in \text{span}(\mathcal{B})$ or equivalently $A\mathbf{y} - \mathbf{b} \in \mathcal{L}_{\mathcal{B}}^{n+1}$. So then for any such $\mathbf{y}$

$$(1-\epsilon)\|A\mathbf{y} - \mathbf{b}\|_2^2 \leq \sqrt{\frac{N}{m}}\|RFDA\mathbf{y} - RFD\mathbf{b}\|_2^2 \leq (1+\epsilon)\|A\mathbf{y} - \mathbf{b}\|_2^2$$

with probability at least $1 - p - N^{-\ln^3 N}$ □

Denote $\tilde{F} = \sqrt{\frac{N}{m}}RFD$, the fast JL matrix from Theorem 3.3.4. How good is the approximation to the original least squares problem? Observe that $\tilde{F}A \in \mathbb{C}^{m\times n}$ and $\tilde{F}b \in \mathbb{C}^m$ we have then a compressed minimization problem,

$$\mathbf{y}'_{\min} = \arg\min_{\mathbf{z}\in\mathbb{C}^n} \|\tilde{F}A\mathbf{z} - \tilde{F}\mathbf{b}\|_2$$

By Theorem 3.3.4 we have

$$\begin{aligned}(1-\epsilon)\|A\mathbf{y}'_{\min} - \mathbf{b}\|_2^2 &\leq \|\tilde{F}A\mathbf{y}'_{\min} - \tilde{F}\mathbf{b}\|_2^2 \\ &\leq \|\tilde{F}A\mathbf{y}_{\min} - \tilde{F}\mathbf{b}\|_2^2 \\ &\leq (1+\epsilon)\|A\mathbf{y}_{\min} - \mathbf{b}\|_2^2\end{aligned}$$

Therefore,

$$\|A\mathbf{y}'_{\min} - \mathbf{b}\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\|A\mathbf{y}_{\min} - \mathbf{b}\|_2$$

Similarly we can bound from below,

$$\begin{aligned}(1+\epsilon)\|A\mathbf{y}'_{\min} - \mathbf{b}\|_2^2 &\geq \|\tilde{F}A\mathbf{y}'_{\min} - \tilde{F}\mathbf{b}\|_2^2 \\ &\geq (1-\epsilon)\|A\mathbf{y}'_{\min} - \mathbf{b}\|_2^2 \\ &\geq (1-\epsilon)\|A\mathbf{y}_{\min} - \mathbf{b}\|_2^2\end{aligned}$$

Thus the solution to the approximate answer has the following error bounds

$$\sqrt{\frac{1-\epsilon}{1+\epsilon}}\|A\mathbf{y}_{\min}-\mathbf{b}\|_2 \leq \|A\mathbf{y}'_{\min}-\mathbf{b}\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\|A\mathbf{y}_{\min}-\mathbf{b}\|_2$$

The runtime of the approximate solution depends on multiplying A and b by $\tilde{F}$ and then also solving the compressed minimization problem. That is,

1. Computing $\tilde{F}A \in \mathbb{C}^{m\times n}$ can be computed in $\mathcal{O}\left(nN\log N\right)$ time
2. Computing $\tilde{F}\mathbf{b}$ can be computed in $\mathcal{O}\left(N\log N\right)$ time
3. Solving least squares problem $\arg\min_{\mathbf{z}\in\mathbb{C}^n}\|\tilde{F}A\mathbf{z}-\tilde{F}\mathbf{b}\|_2$ can be solved in $\mathcal{O}(mn^2)$ using QR factorization for example. Substituting in $m=\bar{c}(n+1)\ln\left(\frac{c'}{\epsilon}\right)\ln^4 N$ we have $\mathcal{O}(n^3\log^4 N)$ (constants depending on ϵ and p are collapsed)

So the total runtime to find $\mathbf{y}'_{\min}$ is $\mathcal{O}\left(nN\log N+n^3\log^4 N\right)$. Recall the classic solution requires $\mathcal{O}(n^2N)$. Therefore when $\log N \lesssim n \lesssim N\log^{-4}N$ we achieve a speedup by using the approximate solution approach. For example, should $n=\sqrt{N}$ then we have a runtime of $\mathcal{O}\left(N^{1.5}\log^4 N\right)$ for the approximate solution and $\mathcal{O}(N^2)$ for the classical solution approach.

Consider the SVD of the $A\in\mathbb{C}^{N\times n}$ where $N\gg n$ and A has full rank

$$A=U\Sigma V^*=\begin{pmatrix} | & | & & | \\ | & | & & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_N \\ | & | & & | \\ | & | & & | \end{pmatrix}\begin{pmatrix} \sigma_1(A) & & \\ & \ddots & \\ & & \sigma_n(A) \\ \vdots & & \vdots \\ 0 & & 0 \end{pmatrix}\begin{pmatrix} — & \mathbf{v}_1^* & — \\ & \vdots & \\ — & \mathbf{v}_n^* & — \end{pmatrix}$$

Definition 3.3.5 (Condition Number)**.** If $A\in\mathbb{C}^{N\times n}$ then the condition number $\kappa(A)$ is defined as

$$\kappa(A)=\frac{\sigma_1(A)}{\sigma_r(A)}$$

where r is the rank of matrix A. If A is full rank and $N\geq n$ as above, then $r=n$

It is the case that the stability of least square solution approaches depend on the condition number of the matrix, e.g. conjugate gradient descent. If we wish to use iterative methods to solve the compressed least square problem then we naturally need to know how the JL map effects the condition number of A. That is, how does $\kappa(\tilde{F}A)$ compare to the original matrix $\kappa(A)$? By Theorem 3.3.4

$$\begin{aligned}\sigma_n(\tilde{F}A) &= \arg \min_{\substack{\mathbf{y}\in\mathbb{C}^n \\ \|\mathbf{y}\|_2=1}} \|\tilde{F}A\|_2 \\ &\geq \arg \min_{\substack{\mathbf{y}\in\mathbb{C}^n \\ \left(\sqrt{1-\epsilon}\right)\|\mathbf{y}\|_2=1}} \|A\|_2 \\ &= \left(\sqrt{1-\epsilon}\right)\sigma_n(A)\end{aligned}$$

Similarly,

$$\begin{aligned}\sigma_1(\tilde{F}A) &= \arg \max_{\substack{\mathbf{y}\in\mathbb{C}^n \\ \|\mathbf{y}\|_2=1}} \|\tilde{F}A\|_2 \\ &\leq \arg \min_{\substack{\mathbf{y}\in\mathbb{C}^n \\ \|\mathbf{y}\|_2=1}} \left(\sqrt{1+\epsilon}\right)\|A\|_2 \\ &= \left(\sqrt{1+\epsilon}\right)\sigma_1(A)\end{aligned}$$

Thus

$$\kappa(\tilde{F}A) \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\kappa(A)$$

So we see that for many reasonable choices of ϵ that the condition number of the sketched matrix will be close to the condition number of the original matrix. So there are faster options for solving the least squares problem using the compressed system of equations. To summarize a theorem found in [30]:

It is possible to compute a minimizer $\mathbf{y}''_{\min}$ such that

$$\|A\mathbf{y}''_{\min} - \mathbf{b}\|_2 \leq (1+\epsilon)\|A\mathbf{y}_{\min} - \mathbf{b}\|_2$$

using a preconditioned gradient method in runtime $\mathcal{O}\left(\left(\log m + \log\frac{1}{\epsilon}\right)Nn + n^2m\right)$.

Application 3.3.6 (Fast Principle Component Analysis)**.** In this section we will describe how to use JL maps to compress and solve an approximation to PCA - we are interested in finding good low rank approximation to an arbitrary matrix A. First however we will need some lemmas dealing with how singular values are affected by JL maps.

Lemma 3.3.7. *If $\Phi \in \mathbb{C}^{m\times N}$ satisfies*

$$\sqrt{1-\epsilon}\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq \sqrt{1+\epsilon}\|\mathbf{x}\|_2$$

when $\epsilon \in (0,1)$, $\forall x \in S \subset \mathbb{C}^N$, then

$$(1-\epsilon)\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq (1+\epsilon)\|\mathbf{x}\|_2$$

equivalently, if $\|x\|_2 = 1$, then

$$|\|\Phi\mathbf{x}\|_2 - 1| \leq \epsilon$$

Proof. Note that since $1-\epsilon < 1$, we have that $1-\epsilon \leq \sqrt{1-\epsilon}$. Since $1+\epsilon > 1$ we have that $\sqrt{1+\epsilon} \leq 1+\epsilon$. That is,

$$(1-\epsilon)\|\mathbf{x}\|_2 \leq \sqrt{1-\epsilon}\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq \sqrt{1+\epsilon}\|\mathbf{x}\|_2 \leq (1+\epsilon)\|\mathbf{x}\|_2$$

if $\|\mathbf{x}\|_2 = 1$ then

$$(1+\epsilon) \leq \|\Phi\mathbf{x}\|_2 \leq (1-\epsilon) \iff -\epsilon \leq \|\Phi\mathbf{x}\|_2 - 1 \leq \epsilon \iff |\|\Phi\mathbf{x}\|_2 - 1| \leq \epsilon$$

□

Lemma 3.3.8. *Let $V \in \mathbb{C}^{N\times r}$ be a matrix with orthonormal columns, and suppose that $\Phi \in \mathbb{C}^{m\times N}$ is an ϵ-JL map of the column space of V into $\mathbb{C}^M$ (i.e. $\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$, Φ is an ϵ-JL map of $\mathcal{L}^r_{\mathcal{B}}$ as in Theorem 3.3.1). Then ΦV is full rank with*

$$(1-\epsilon) \leq \sigma_j(\Phi V) = \sigma_j(V^*\Phi^*) \leq (1+\epsilon)$$

$\forall j \in [r]$.

Proof. Consider $\sigma_1(\Phi V)$: Choose $\mathbf{y} \in \mathbb{C}^r$, $\|\mathbf{y}\|_2 = 1$ to be the right singular vector of ΦV that corresponds to $\sigma_1(\Phi V)$. i.e. if $\Phi V = U\Sigma Y^*$ is the SVD of ΦV then $\Phi V \mathbf{y}_1 = \sigma_1 \mathbf{u}_1$ where $\mathbf{y}_1, \mathbf{u}_1$ are the right and left singular vectors; the first column of Y, U respectively. Note that $\|\Phi V \mathbf{y}_1\|_2 = \|\sigma_1 \mathbf{u}_1\|_2 = |\sigma_1| \|\mathbf{u}_1\|_2 = \sigma_1$

Naturally $V\mathbf{y}$ is a vector in the column space of V, i.e. $V\mathbf{y} \in \mathcal{L}^r_{\mathcal{B}}$ and since V has orthonormal columns, $\|V\mathbf{y}\|_2 = \|\mathbf{y}\|_2 = 1$. So on one hand, by definition of $\mathbf{y}$ we have $\sigma_1(\Phi V) = \|\Phi V\|_{2\to 2} = \|\Phi V \mathbf{y}\|_2$. On the other, since Φ is an ϵ-JL map of the column space of V, and by Lemma 3.3.7 we have $\|\Phi V \mathbf{y}\|_2 \leq 1 + \epsilon$. Thus $\sigma_1(\Phi V) \leq 1 + \epsilon$.

Now consider the smallest singular value $\sigma_r(\Phi V)$. Let $\mathbf{y} \in \mathbb{C}^r$ be the right singular vector that corresponds to $\sigma_r(\Phi V)$. Similarly as before, $\|\Phi V \mathbf{y}\|_2 = \sigma_r(\Phi V)$, since $V\mathbf{y} \in \mathcal{L}_{\mathcal{B}^r}$ then by Lemma 3.3.7 $(1 - \epsilon) \leq \|\Phi V \mathbf{y}\|_2 = \sigma_r(\Phi V)$.

Since the singular values are ordered then, we have

$$(1 - \epsilon) \leq \sigma_j(\Phi V) \leq (1 + \epsilon)$$

$\forall j \in [r]$. Observe that $(\Phi V)^* = V^* \Phi^* = (U\Sigma Y^*)^* = Y\Sigma^* U^*$ and so we have that $Y\Sigma^* U^*$ is the unique SVD of $V^*\Phi^*$, but $\Sigma^*_{jj} = \Sigma_{jj}$ and so we see that the singular values are unchanged by conjugate transpose. Finally, observe that $\sigma_r \neq 0$ and thus ΦV must have full rank. □

Since the operator norm is equal to the largest singular value, we can see from Lemma 3.3.8 that $\|\Phi V\|_{2\to 2} \approx 1$, i.e. ΦV is close to a matrix which has its columns from a unitary matrix.

Lemma 3.3.9. *Suppose $\Phi \in \mathbb{C}^{m\times N}$ is an ϵ-JL map for some set $S \subset \mathbb{C}^N$. The operator norm $\|\Phi\|_{2\to 2}$*

1. *If* Φ *is i.i.d.*

$$(\Phi)_{ij} = \begin{cases} \frac{1}{\sqrt{m}} & \textit{probability } 1/2 \\ -\frac{1}{\sqrt{m}} & \textit{probability } 1/2 \end{cases}$$

then

$$\|\Phi\|_{2\to 2} \leq \|\Phi\|_F = \sqrt{\frac{mN}{m}} = \sqrt{N}$$

2. *If* $\Phi = \sqrt{\frac{N}{m}} RUD$ *as in Example 3.1.6, then*

$$\|\Phi\|_{2\to 2} \leq \sqrt{\frac{N}{m}} \|R\|_{2\to 2} \|U\|_{2\to 2} \|D\|_{2\to 2} = \sqrt{\frac{N}{m}}$$

3. *If* Φ *has i.i.d.* $\mathcal{N}(0,1)$ *entries then*

$$\|\Phi\|_{2\to 2} \leq \sqrt{\frac{mN}{m}} \sup_{i,j} |\Phi_{ij}| \leq 2\sqrt{N}\sqrt{2\ln(mN)}$$

with probability at least $1/2$

4. *If* Φ *has the RIP of* (s,ϵ) *then* $\|\Phi\|_{2\to 2} \leq \sqrt{1+\epsilon}\left(\sqrt{\frac{N}{s}}+1\right)$

Proof. 1. Note $\|\Phi\|_{2\to 2} = \sigma_1(\Phi) \leq \sqrt{\sum_{j=1}^{r} \sigma_j(\Phi)^2} = \|\Phi\|_F = \sqrt{\sum_{i,j} \Phi_{ij}}^2 = \sqrt{\sum_{i,j} \frac{1}{m}}\sqrt{N}$.

2. Note that $\|AB\|_{2\to 2} \leq \|A\|_{2\to 2}\|B\|_{2to2}$ for any compatible matrices A, B. Here R, U, D all have operator norm 1.

3. Use Markov's inequality and Theorem 5.1.15

4. see Lemma 4.4.3

□

Definition 3.3.10 (Pseudo-inverse of Matrix)**.** Suppose that $A \in \mathbb{C}^{p\times q}$. Then the pseudo-inverse of this matrix is denoted $A^\dagger$ and can be written in terms of the SVD

of $A = U\Sigma V^*$ as follows:

$$A^\dagger = V\Sigma^{-1}U^* = \begin{pmatrix} | & | & & | \\ | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_q \\ | & | & & | \\ | & | & & | \end{pmatrix} \begin{pmatrix} 1/\sigma_1(A) & & \\ & \ddots & \\ & & 1/\sigma_r(A) \end{pmatrix} \begin{pmatrix} — & \mathbf{u}_1^* & — \\ & \vdots & \\ — & \mathbf{u}_p^* & — \end{pmatrix}$$

Note that in a full SVD, matrix Σ will have the same dimensions as A - in particular it will generically be rectangular. This also means that Σ^{-1} is an abuse of notation, since the matrix is not guaranteed to be square. We denote Σ^{-1} here as the operation of placing the multiplicative inverse, when it exists, of the entries along the diagonal and the transposition of rows or columns of zeros when $p \neq q$. In this way $\Sigma\Sigma^{-1}$ when restricted to the columns and rows indexed by less than or equal to the rank will indeed yield the identity.

Additionally, when A is square then the pseudo-inverse and matrix inverse coincide $A^\dagger = A^{-1}$.

Lemma 3.3.11. *Under the conditions of Lemma 3.3.8*

$$\| (V^*\Phi^*)^\dagger \|_{2\to 2} = \| (\Phi V)^\dagger \|_{2\to 2} \leq \frac{1}{1-\epsilon}$$

Proof. Let $\Phi V = \underbrace{U}_{m\times r}\underbrace{\Sigma}_{r\times r}\underbrace{\tilde{V}^*}_{r\times r}$ be the (truncated) SVD of the matrix ΦV. Note

$$(\Phi V)^\dagger = \tilde{V}\Sigma^{-1}U$$

Where $\max_{k,\ell} (\Sigma^{-1})_{k\ell} = \sigma_r(\Phi V)^{-1} \leq (1-\epsilon)^{-1}$ by Lemma 3.3.8 □

We now have the necessary machinery to describe and analyze compressed PCA algorithm. First however though we establish some notation

Suppose

$$A = U'\Sigma'V' = \left(\begin{array}{c|c} U_1' & U_2' \end{array} \right) \left(\begin{array}{c|c} \Sigma_1' & 0 \\ \hline 0 & \Sigma_2' \end{array} \right) \left(\begin{array}{c|c} V_1' & V_2' \end{array} \right)^*$$

where $U_1' \in \mathbb{C}^{q \times r}$ has orthonormal columns, $\Sigma_1' \in \mathbb{C}^{r \times r}$ is a diagonal matrix with the first r singular values and $V_1' \in \mathbb{C}^{\times r}$ has orthonormal columns.

Note also that the optimal error achievable for a r rank approximation is equal to the $r+1$ singular value. That is.

$$\inf_{\text{Rank}(A_r)=r} \|A - A_r\|_{2\to 2} = \sigma_{r+1}(A)$$

and this approximation is possible by using the leading r singular vectors of the SVD.

Algorithm 3.3.1 Compressed PCA

Input: $A \in \mathbb{C}^{q \times p}, k \in \mathbb{N}, \Phi \in \mathbb{C}^{m \times p}$ an ϵ-JL of the column span of $V_1 \in \mathbb{C}^{p \times r}$. V_1 is a matrix of the r leading right singular vectors taken from the SVD of $Z = (AA^*)^k A$

Output: U, Σ, V^* as approximations to U_1', Σ_1', V_1'

Form $Y = (AA^*)^k A\Phi^* \in \mathbb{C}^{q \times m}$ by alternating multiplications of A and A^* against Φ^*

By stable QR decomposition, construct an orthonormal basis $\{\mathbf{b_1}, \ldots, \mathbf{b_m}\} \in \mathbb{C}^q$ for the range of Y. Let

$$P = [\mathbf{b}_1 \ldots \mathbf{b}_m] \in \mathbb{C}^{q \times m}$$

Let $B = P^*A \in \mathbb{C}^{m \times p}$

Compute the SVD of $B = \underbrace{\hat{U}}_{m\times m} \underbrace{\Sigma}_{m\times m} \underbrace{V^*}_{m\times p}$

Form $\tilde{U} = P\hat{U}$

return $\tilde{U}, \Sigma, V^*$

Note that $(AA^*)^k = U' (\Sigma')^k (U')^*$; so higher powers k will have the effect of increasing the ratio of the largest smallest singular value to the smallest which is computationally useful. We now turn to a runtime analysis of the algorithm: Forming $A\phi^*$ requires $\mathcal{O}(pq \log p)$ flops if we assume fast matrix-vector multiplication. Applying $(AA^*)^k$ requires $\mathcal{O}(km\|A\|_0)$ flops. QR decomposition to find P can be accomplished with $\mathcal{O}(m^2 q)$. Forming B requires $\mathcal{O}(m\|A\|_0))$ flops. The SVD of $B \in \mathbb{C}^{m \times p}$ can be found $\mathcal{O}(pm^2)$ and the matrix multiplication to find U requires

$\mathcal{O}(qm^2)$. Summing then we find a total runtime of

$$\mathcal{O}\left((p+q)m^2 + qp\log p + (k+1)m\|A\|_0\right)$$

Here we can see that the term $(p+q)m^2$ obtained from forming B and finding its SVD dominates the runtime of the algorithm if in fact A is sparse (i.e. $\|A\|_0 \ll pq$) and k is small. Compare this to a method which finds the SVD of the uncompressed matrix A at a cost of $\mathcal{O}(\max(p,q)^2\min(q)) = \mathcal{O}(p^2q)$ if we assume $p \geq q$. This shows us that the compressed version could obtain a speed up when $m < q$.

Theorem 3.3.12. *The rank m approximation, $\tilde{A} = U\Sigma V^*$ output of 3.3.1 will satisfy*

$$\|A - \tilde{A}\|_{2\to 2} \leq \sigma_{r+1}(A)\left[1 + \frac{\|\Phi\|_{2\to 2}^2}{(1-\epsilon)^2}\right]^{\frac{1}{4k+2}}$$

Proof. We are going to express the SVD of $Z = (AA^*)^k A$ in block form.

$$Z = (AA^*)^k A = U\Sigma V = \left(\begin{array}{c|c} U_1 & U_2 \end{array}\right)\begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\left(\begin{array}{c} V_1^* \\ \hline V_2 \end{array}\right)$$

where $V_1 \in \mathbb{C}^{p\times r}$, $V_2 \in \mathbb{C}^{p\times(q-r)}$, $\Sigma_1 \in \mathbb{C}^{r\times r}$, $\Sigma_2\mathbb{C}^{\min(p,q)-r\times\min(p,q)-r}$.

Recall from our algorithm description 3.3.1 that $Y = Z\Phi^*$. Let $\Omega_1 = V_1^*\Phi^*$ and $\Omega_2 = V_2^*\Phi$. Therefore, in block form we have

$$Y = Z\Phi^* = U\begin{pmatrix} \Sigma_1\Omega_1 \\ \Sigma_2\Omega_2 \end{pmatrix}$$

Proposition 3.3.13. *Let $A \in \mathbb{C}^{q\times p}$ with singular values $\sigma_j(A)$ and fix $k, r \geq 0$, $r \leq \min(p,q)$. Let $\Phi \in \mathbb{C}^{m\times p}$ be an ϵ-JL map of $V_1 \in \mathbb{C}^{p\times r}$, the r leading right singular vectors of $Z = (AA^*)^kA$. As above, let $\Omega_1 = V_1^*\Sigma_1^*$, $\Omega_2 = V_2^*\Sigma_2^*$ (where Ω_1 is full rank). If $P = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{C}^{q\times m}$ is an orthonormal matrix such that $P^*PY = Y$ then*

$$\|\left(I - PP^*\right)A\|_{2\to 2} \leq \|\left(I - PP^*\right)Z\|_{2\to 2}^{\frac{1}{2k+1}} \leq \left(\|\Sigma_2\|_{2\to 2}^2 + \|\Sigma_2\Omega_2\Omega_1^\dagger\|_{2\to 2}^2\right)^{\frac{1}{4k+2}}$$

Proof. See Theorem 9.2 and Theorem 9.1 from [15]. □

Note that $B = P^*A = \hat{U}\Sigma V^*$ and $\tilde{A} = \tilde{U}\Sigma V^* = P\hat{U}\Sigma V^* = PB = PP^*A$.

So

$$\|A - \tilde{A}\|_{2\to 2} = \|A - PP^*A\|_{2\to 2} = \|\left(I - PP^*\right)A\|_{2\to 2}$$

Now using Proposition 3.3.13, we look at the summands of interest. First, since Σ_2 is diagonal, we can read off the operator norm by taking the largest entry

$$\|\Sigma_2\|_{2\to 2}^2 = \sigma_{r+1}(Z)^2 = \sigma_{r+1}^{4k+2}$$

Now consider

$$\begin{aligned}
\|\Sigma_2\Omega_2\Omega_1^\dagger\|_{2\to 2} &\leq \|\Sigma_2\|_{2\to 2}\|\Omega_2\Omega_1^\dagger\|_{2\to 2} \\
&\leq \sigma_{r+1}(A)^{2k+1}\|\Omega_2\|_{2\to 2}\|\Omega_1^\dagger\|_{2\to 2} \\
&= \sigma_{r+1}(A)^{2k+1}\|V_2^*\Phi^*\|_{2\to 2}\|(V_1^*\Phi^*)^\dagger\|_{2\to 2} \\
&\leq \sigma_{r+1}(A)^{2k+1}\frac{\|V_2^*\|_{2\to 2}\|\Phi^*\|_{2\to 2}}{1-\epsilon} \\
&= \sigma_{r+1}(A)^{2k+1}\frac{\|\Phi^*\|_{2\to 2}}{1-\epsilon}
\end{aligned}$$

So we have then that

$$\|A - \tilde{A}\|_{2\to 2} \leq \left(\|\Sigma_2\|_{2\to 2}^2 + \|\Sigma_2\Omega_2\Omega_1^\dagger\|_{2\to 2}^2\right)^{\frac{1}{4k+2}} = \sigma_{r+1}(A)\left[1 + \frac{\|\Phi\|_{2\to 2}^2}{(1-\epsilon)^2}\right]^{\frac{1}{4k+2}}$$

□

Corollary 3.3.14. *Choose $\epsilon \in (0,1)$, $k \in \mathbb{N}$ then*

1. $\exists \Phi \in \mathbb{C}^{m\times p}$ with $\pm\frac{1}{\sqrt{m}}$ entries such that 3.3.1 yields

$$\|A - \tilde{A}\|_{2\to 2} \leq \sigma_{r+1}(A)\left[1 + \frac{p}{(1-\epsilon)^2}\right]^{\frac{1}{4k+2}}$$

2. $\exists \Phi \in \mathbb{C}^{m\times p}$ *with fast FFT-matrix-vector multiply such that*

$$\|A - \tilde{A}\|_{2\to 2} \leq \sigma_{r+1}(A)\left[1 + \frac{p}{m(1-\epsilon)^2}\right]^{\frac{1}{4k+2}}$$

where $m = \frac{Cr}{\epsilon^2}\log\left(\frac{48}{\epsilon}\right)\log^4 p$

Proof. Apply Lemma 3.5.6 □

Homework 3.3.1. Prove Corollary 3.3.2.

3.4 Best Achievable JL-maps by Orthogonal Projections

Consider as before the linear r-dimensional subspace $\mathcal{L}^r_{\mathcal{B}}$ with a given orthonormal basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$ and the matrix B formed by taking the basis elements and arranging them as columns in a matrix. We define the distance from a set $T \subset \mathbb{C}^N$ to $\mathcal{L}^r_{\mathcal{B}}$ as the maximum Euclidean distance between an element of T and its projection into $\mathcal{L}^r_{\mathcal{B}}$. Namely

$$d^\infty(T, \mathcal{L}^r_{\mathcal{B}}) = \sup_{\mathbf{x}\in T} \|\mathbf{x} - \Pi_{\mathcal{L}^r_{\mathcal{B}}}\mathbf{x}\|_2$$

where $\Pi_{\mathcal{L}^r_{\mathcal{B}}} = BB^*$ is the projection matrix from $\mathbb{C}^N$ into $\mathcal{L}^r_{\mathcal{B}}$. We will use this to establish a lower bound on the accuracy of JL embeddings of a given set.

Definition 3.4.1 (Kolmogorov Width)**.** The Kolmogorov width of $T \subset \mathbb{C}^N$ is

$$\begin{aligned} d^\infty_r(T) &= \inf_{\mathcal{L}^r_{\mathcal{B}}\in\Gamma_r} d^\infty(T, \mathcal{L}^r_{\mathcal{B}}) \\ &= \inf_{\mathcal{L}^r_{\mathcal{B}}\in\Gamma_r} \sup_{\mathbf{x}\in T} \|\left(I - BB^*\right)\mathbf{x}\|_2 \end{aligned}$$

Lemma 3.4.2. *If* $B^* \in \mathbb{C}^{m\times N}$ *has orthonormal rows and is an* ϵ*-JL map of* $T \subset \mathbb{C}^N$ *into* $\mathbb{C}^m$ *and we define* T' *as:*

$$T' = \left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2} | \mathbf{x} \in T\,\{\mathbf{0}\}\right\} \subset \mathbb{C}^N$$

then $\epsilon \geq \left(d^\infty_r(T')\right)^2$.

Proof. If B^* is an ϵ-JL map of T then

$$\left|\|B^*\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \epsilon\|\mathbf{x}\|_2^2, \forall \mathbf{x} \in T$$

it follows that

$$\left|\|B^*\mathbf{x}\|_2^2 - 1\right| \leq \epsilon, \forall \mathbf{x} \in T'$$

this holds for all $\mathbf{x} \in T'$ and so holds for the supremum

$$\sup_{vbx \in T'} \left|\|B^*\mathbf{x}\|_2^2 - 1\right| \leq \epsilon$$

and so by Homework 3.4.1 we have

$$\epsilon \geq \left(d_r^\infty(T')\right)^2$$

□

So, if we are able to estimate the Kolmogorov width of T then we can evaluate the nearness to optimal of a particular JL map. Algorithms do exist to approximate Kolmogorov width, which rely on sampling T'.

Homework 3.4.1. Let

$$T' = \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} | \mathbf{x} \in T \setminus \{\mathbf{0}\} \right\} \subset \mathbb{C}^N$$

and $B^* \in \mathbb{C}^{r \times N}$ has orthonormal rows. Prove that

$$\inf_{B \in \mathbb{C}^{r \times N}} \sup_{\mathbf{x} \in T'} \left|1 - \|B^*\mathbf{x}\|_2^2\right| = \left(d_r^\infty(T')\right)^2$$

3.5 Restricted Isometry Property (RIP)

The Restricted Isometry Property (RIP) is no more and no less than the norm preserving Johnson-Lindenstrauss property applied to the set of sparse vectors. Recall the definition for s-sparse vectors

$$K_s = \left\{\mathbf{x} \in \mathbb{C}^N | \|\mathbf{x}\|_0 \leq s\right\} = \bigcup_{S\subseteq[N]|S|\leq s} \text{span}\left\{\mathbf{e}_j\right\}_{j\in S}$$

Definition 3.5.1 ((s,ϵ)-RIP)**.** A matrix $\Phi \in \mathbb{C}^{m\times N}$ has the RIP of order (s,ϵ) if its an ϵ-JL map of K_s into $\mathbb{C}^m$. Equivalently

$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2, \forall \mathbf{x} \in K_s$$

Lemma 3.5.2. *The following hold for the set of s and t sparse vectors K_s and K_t*

1. $K_t \subseteq K_s, \forall t \leq s$
2. $K_t \pm K_s \subset K_{t+s}$
3. *Given a full rank diagonal matrix $D \in \mathbb{C}^{N\times N}$, we have $DK_s = K_s, \forall s$*

We now introduce some new notation so that we can state a result dealing with RIP of submatrices.

Definition 3.5.3. Given $S \subset [N]$ and $\Phi \in \mathbb{C}^{m\times N}$, we let $\Phi_S \in \mathbb{C}^{m\times|S|}$ be the submatrix Φ composed of the columns of Φ enumerated by S

Example 3.5.4. If

$$\Phi = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 1 & 3 & 4 & 5 \end{pmatrix}$$

then

$$\Phi_{\{1,3\}} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

or

$$\Phi_{\{1,3\}} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \end{pmatrix}$$

The differing definitions of the submatrix as either zero-filled is overloading the notation, however context should make it clear which version is being used. We sacrifice some consistency for convenience and brevity.

Definition 3.5.5 (Support Function)**.** $\text{supp} : \mathbb{C}^N \to [N]$ is defined by $\text{supp}(\mathbf{x}) = \{j \in [N] | x_j \neq 0\}$

Lemma 3.5.6. *If $\Phi \in \mathbb{C}^{m \times N}$ has the RIP of order (s, ϵ) then*

$$\|\Phi_S^* \Phi_S - I\|_{2 \to 2} \leq \epsilon$$

holds $\forall S \subset [N]$ with $|S| \leq s$

Proof. Φ has (s, ϵ)-RIP thus

$$\max_{\substack{S \subset [N] \\ |S| \leq s}} \sup_{\substack{\mathbf{y} \in \mathbb{C}^S \subset \mathbb{C}^N \\ \mathbf{y} \neq 0}} \frac{|\|\Phi_S \mathbf{y}\|_2^2 - \|\mathbf{y}\|_2^2|}{\|\mathbf{y}\|_2^2} \leq \epsilon$$

Equivalently, using properties of inner-products we have

$$\begin{aligned} \max_{\substack{S \subset [N] \\ |S| \leq s}} \sup_{\substack{\mathbf{y} \in \mathbb{C}^S \subset \mathbb{C}^N \\ \mathbf{y} \neq 0}} \frac{|\langle \Phi_S \mathbf{y}, \Phi_S \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle|}{\|\mathbf{y}\|_2^2} &= \max_{\substack{S \subset [N] \\ |S| \leq s}} \sup_{\substack{\mathbf{y} \in \mathbb{C}^S \subset \mathbb{C}^N \\ \mathbf{y} \neq 0}} \frac{|\langle (\Phi_S^* \Phi_S - I)\, \mathbf{y}, \mathbf{y} \rangle|}{\|\mathbf{y}\|_2^2} \\ &= \max_{\substack{S \subset [N] \\ |S| \leq s}} \|\Phi_S^* \Phi_S - I\|_{2 \to 2} \end{aligned}$$

Note that $\Phi_S^* \Phi_S - I$ is a Hermitian matrix, and so the maximum norm corresponds to the top eigenvector and eigenvalue of $\Phi_S^* \Phi_S - I$. This is the operator norm. $\square$

Restate in terms of singular values	ToDo

Theorem 3.5.7. *Suppose that* $\mathbf{x} \in K_s$ *and* $\mathbf{y} \in K_t$*, and that* Φ *has the RIP of order* $(s+t, \epsilon)$*. Then*

1. *If* $s = t$ *then*

$$|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| \leq 4\epsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

2. *If* $supp(\mathbf{x}) \cap supp(\mathbf{y}) = \emptyset$ *then*

$$|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle| \leq \epsilon \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

3. *If* $s = t$ *then*

$$|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle| \leq 4\epsilon$$

for all $\mathbf{x}, \mathbf{y} \in Conv(K_s \cap \overline{B(0,1)})$

Proof. 1. This follows from Lemma 3.1.11. Note that $K'_s \subset K_{2s} = K_{s+t}$, where K'_s is defined as seen in the lemma.

2. Let $S = \text{supp}(\mathbf{x}) \cup \text{supp}(\mathbf{y})$. Observe,

$$\begin{aligned}
|\langle \Phi\mathbf{x}, \Phi\mathbf{y} \rangle| &= |\langle \Phi_S \mathbf{x}_S, \Phi_S \mathbf{y}_S \rangle - \langle \mathbf{x}_S, \mathbf{y}_S \rangle| \\
&= |\langle (\Phi_S^* \Phi - I)\, \mathbf{x}_S, \mathbf{y}_S \rangle| \\
&\leq \| (\Phi_S^* \Phi - I)\, \mathbf{x}_S \|_2 \|\mathbf{y}_S\|_2 \\
&\leq \| (\Phi_S^* \Phi - I)\, \|_{2\to 2} \|\mathbf{x}_S\|_2 \|\mathbf{y}_S\|_2 \\
&\leq \epsilon \|\mathbf{x}_S\|_2 \|\mathbf{y}_S\|_2
\end{aligned}$$

where we have used Lemma 3.5.6 and Cauchy-Schwarz inequality.

3. Follows from 3.1.18, where in this case $\gamma = 1$

□

Homework 3.5.1. Prove Lemma 3.5.2.

Homework 3.5.2. Prove the following: If Φ has $(|S| + |T|, \epsilon)$-RIP and $S \cap T = \emptyset$ then $\|\Phi_T^* \Phi_S\|_{2\to 2} \leq \epsilon$

3.6 Towards Invertability of $\Delta(\Phi \mathbf{x}) = x$

In this section we week to answer the questions: how many rows are needed in Φ to Allow $\Delta(\Phi \mathbf{x}) = x$ for $\mathbf{x} \in K_s$? What algorithms Δ exist that let us get close to these lower bounds on the number of rows? Can they be computed efficiently?

Lemma 3.6.1. *If* $\mathbf{u} \in \mathbb{C}^k$ *and* $\mathbf{v} \in \mathbb{C}^k$ *satisfy*

$$\max_{i \in [k]} |u_i| \leq \min_{j \in [k]} |v_j|$$

then $\sqrt{k}\|\mathbf{u}\|_2 \leq \|\mathbf{v}\|_1$

Proof. Proof is left as an exercise to reader. □

Definition 3.6.2. If $\mathbf{v} \in \mathbb{C}^N$ and $S \subset [N]$ then the vector $\mathbf{v}_S$ will be of length N and contain zeros in the entries which are not in the set S, i.e.

$$(\mathbf{v}_S)_i = \begin{cases} v_i & i \in S \\ 0 & i \notin S \end{cases}$$

The following theorem states that RIP matrices have the null space property; where null space property can be thought of as a property that no sparse vectors are in the kernel of the given matrix. This is naturally a desirable property if we hope to be able to recover (invert) vectors which are sparse.

Theorem 3.6.3. *Suppose that $\Phi \in \mathbb{C}^{m\times N}$ has the $(2s,\epsilon)$-RIP property where $\epsilon \in (0,1/2)$. Then*

$$\|\mathbf{v}_S\|_1 \leq \frac{\epsilon}{1-2\epsilon}\|\mathbf{v}_{\overline{S}}\|_1 \tag{3.4}$$

for all vbv $\in Ker(\Phi)$ and for all $S \subseteq [N]$ with $|S| \leq s$

Proof. Without loss of generality we will consider the entries of an arbitrary member of the kernel $\mathbf{v}$ to be ordered in the following way

$$|v_1| \geq |v_2| \geq \cdots \geq |v_N|$$

We do not lose generality since at all points in the sequel we will be considering subsets of indices, which can be permuted anyway. We label sets of size s as follows: $S_0 = \{1,\ldots,s\}, S_1 = \{s+1,\ldots,2s\}$ and so on. That is, S_0 contains the top s-largest entries by absolute value. If 3.4 holds for $S = S_0$ then it must hold for all sets of size s since $\|\mathbf{v}_S\|_1$ is maximized when $S = S_0$ and also $\|\mathbf{v}_{\overline{S}}\|_1$ is minimized for $S = S_0$. That is, if we prove the property for the worst case, that will be sufficient.

Since Φ has the $(2s,\epsilon)$-RIP property, and $\mathbf{v}_{S_0} \in K_{2s}$, then

$$(1-\epsilon)\|\mathbf{v}_{S_0}\|_2^2 \leq \|\Phi\mathbf{v}_{S_0}\|_2^2 \implies \|\mathbf{v}_{S_0}\|_2^2 \leq \frac{1}{(1-\epsilon)}\langle\Phi\mathbf{v}_{S_0}, \Phi\mathbf{v}_{S_0}\rangle$$

Notice that $\Phi\mathbf{v}_{S_0} = -\Phi\mathbf{v}_{\overline{S_0}}$ since $\mathbf{0} = \Phi\mathbf{v} = \Phi\left(\mathbf{v}_{S_0} + \mathbf{v}_{\overline{S_0}}\right)$. Furthermore $\overline{S_0} = \bigcup_{j\geq 1} S_j$ and $S_j \cap S_i = \emptyset$ whenever $i \neq j$.

So we have then that

$$\begin{aligned}
\|\mathbf{v}_{S_0}\|_2^2 &\leq \frac{1}{(1-\epsilon)}\langle \Phi\mathbf{v}_{S_0}, \Phi\mathbf{v}_{S_0}\rangle \\
&= \frac{1}{(1-\epsilon)}\langle \Phi\mathbf{v}_{S_0}, -\Phi\mathbf{v}_{\overline{S_0}}\rangle \\
&= \frac{1}{(1-\epsilon)}\langle \Phi\mathbf{v}_{S_0}, -\Phi\left(\sum_{j\geq 1}\mathbf{v}_{S_j}\right)\rangle \\
&= \frac{1}{(1-\epsilon)}\sum_{j\geq 1}\langle \Phi\mathbf{v}_{S_0}, -\Phi\mathbf{v}_{S_j}\rangle \\
&\leq \frac{\epsilon}{(1-\epsilon)}\sum_{j\geq 1}\|\mathbf{v}_{S_0}\|_2\|\mathbf{v}_{S_j}\|_2
\end{aligned}$$

where we have used Theorem 3.5.7, since the support of the vectors $\mathbf{v}_{S_i}$ and $\mathbf{v}_{S_j}$ are disjoint for $i \neq j$ and they both s-sparse. Dividing by $\|\mathbf{v}_{S_0}\|_2$, noting that Lemma 3.6.1 implies that $\|\mathbf{v}_{S_j}\|_2 \leq \frac{1}{\sqrt{s}}\|\mathbf{v}_{S_{j-1}}\|$ and that the ℓ^1-norm for vectors with disjoint supports is additive, we obtain

$$\begin{aligned}
\|\mathbf{v}_{S_0}\|_2 &\leq \frac{\epsilon}{(1-\epsilon)}\sum_{j\geq 1}\|\mathbf{v}_{S_j}\|_2 \\
&\leq \frac{\epsilon}{\sqrt{s}\,(1-\epsilon)}\sum_{j\geq 1}\|\mathbf{v}_{S_j}\|_1 \\
&\leq \frac{\epsilon}{\sqrt{s}\,(1-\epsilon)}\left(\|\mathbf{v}_{S_0}\|_1 + \|\mathbf{v}_{\overline{S_0}}\|_1\right)
\end{aligned}$$

Note that by Holder's inequality $\|\mathbf{v}_{S_0}\|_1 = \|\mathbf{v}_{S_0}\mathbf{1}_{S_0}\|_1 \leq \|\mathbf{1}_{S_0}\|_2\|\mathbf{v}_{S_0}\|_2 = \sqrt{s}\|\mathbf{v}_{S_0}\|_2$

So we have the inequality

$$\|\mathbf{v}_{S_0}\| \leq \frac{\epsilon}{1-\epsilon}\left(\|\mathbf{v}_{S_0}\|_1 + \|\mathbf{v}_{\overline{S_0}}\|_1\right)$$

which is equivalent to 3.4 after a rearrangement of terms. □

Lemma 3.6.4. *Given $S \subseteq [N]$ and $\mathbf{x}, \mathbf{z} \in \mathbb{C}^N$ then*

$$\| (\mathbf{x} - \mathbf{z})_{\overline{S}} \|_1 \leq \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \| (\mathbf{x} - \mathbf{z})_S \|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1$$

Theorem 3.6.5. *Suppose that $\exists \rho \in (0,1)$ such that $\Phi \in \mathbb{C}^{m \times N}$ satisfies*

$$\|\mathbf{v}_S\|_1 \leq \rho \|\mathbf{v}_{\overline{S}}\|_1, \, \forall \mathbf{v} \in ker(\Phi), \, \forall S \subseteq [N], \, |S| \leq s \tag{3.5}$$

Then any $\mathbf{z}^{\#} \in \mathbb{C}^N$ satisfying

$$\|\mathbf{z}^{\#}\|_1 \text{ is minimal over all } \mathbf{z} \in \mathbb{C}^N \text{ where } \Phi\mathbf{z} = \Phi\mathbf{x} \tag{3.6}$$

will approximate $\mathbf{x} \in \mathbb{C}^N$ near optimally in the sense that

$$\|\mathbf{x} - \mathbf{z}^{\#}\|_1 \leq \frac{2(1+\rho)}{1-\rho} \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_1$$

Proof. Let $S \subseteq [N]$, $|S| = s$ such that

$$\|\mathbf{x}_{\overline{S}}\|_1 = \inf_{\|\mathbf{z}\|_0 \| \leq s} \|\mathbf{x} - \mathbf{z}\|_1$$

That is, we can minimize the error by matching the s-largest entries of $\mathbf{x}$. Now consider $\mathbf{z}^{\#}$ from equation 3.6 in the theorem statement. We have by hypothesis $\Phi\mathbf{z}^{\#} = \Phi\mathbf{x}$ and so $\mathbf{x} - \mathbf{z}^{\#} \in \text{Ker}(\Phi)$. Note that $\|\mathbf{z}^{\#}\|_1 \leq \|\mathbf{x}\|_1$ so using Lemma 3.6.4 and noting the null-space property we obtain

$$\begin{aligned} \| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_{\overline{S}} \|_1 &\leq \|\mathbf{z}\|_1 - \|\mathbf{x}\|_1 + \| (\mathbf{x} - \mathbf{z})_S \|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 \\ &\leq \| (\mathbf{x} - \mathbf{z})_S \|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 \\ &\leq \rho \| (\mathbf{x} - \mathbf{z})_{\overline{S}} \|_1 + 2\|\mathbf{x}_{\overline{S}}\|_1 \end{aligned}$$

After a rearrangement of terms then we have

$$\| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_{\overline{S}} \|_1 \leq \frac{2}{1-\rho} \|\mathbf{x}_{\overline{S}}\|_1$$

Which, after using that vectors with non-intersecting support have additive ℓ^1-norms and another application of the null space property and bound above we obtain

$$\begin{aligned}
\|\mathbf{x} - \mathbf{z}^{\#}\|_1 &= \| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_S \|_1 + \| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_{\overline{S}} \|_1 \\
&\leq \rho \| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_{\overline{S}} \|_1 + \| \left(\mathbf{x} - \mathbf{z}^{\#}\right)_{\overline{S}} \|_1 \\
&\leq \rho \frac{2}{1-\rho} \|\mathbf{x}_{\overline{S}}\|_1 + \frac{2}{1-\rho} \|\mathbf{x}_{\overline{S}}\|_1 \\
&= \frac{2(1+\rho)}{1-\rho} \|\mathbf{x}_{\overline{S}}\|_1 \\
&= \frac{2(1+\rho)}{1-\rho} \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_1
\end{aligned}$$

□

Note. *1. If $\mathbf{x} \in K_s$ then $\mathbf{z}^{\#}$ will reconstruct $\mathbf{x}$ exactly. I.e. we have a Δ, Φ such that $\Delta(\Phi\mathbf{x}) = \mathbf{x}$ for $x \in K_s$.*

2. Matrices $\Phi \in \mathbb{C}^{m \times N}$ with random sub-gaussian entries are ϵ-JL maps for K_s and thus have the (s, ϵ)-RIP property with high probability so long as the number of rows m satisfy

$$m \geq c \frac{s}{\epsilon^2} \ln\left(\frac{N}{s}\right)$$

Matrices with the RIP property have the null space property.

3. Δ as the basis pursuit solution to 3.6 is a computationally efficient way to find the projection of any $\mathbf{x}$ into K_s

Theorem 3.6.6. *Suppose that $\Phi \in \mathbb{C}^{m \times N}$ has the $(2s, \epsilon)$-RIP for $\epsilon < 4/\sqrt{41} \approx 0.6246$. Then, for any $\mathbf{x} \in \mathbb{C}^N$ and $\mathbf{y} \in \mathbb{C}^m$ with $\|\Phi\mathbf{x} = \mathbf{y}\|_2 \leq \eta$, a solution $\mathbf{x}^{\#}$ of*

$$\min_{\mathbf{z} \in \mathbb{C}^N} \|\mathbf{z}\|_1$$

such that $\|\Phi\mathbf{z}-\mathbf{y}\|_2 \leq \eta$ *approximates* $\mathbf{x}$ *with errors*

$$\|\mathbf{x}-\mathbf{x}^{\#}\|_1 \leq c\sigma_s(\mathbf{x})_1 + D\sqrt{s}\eta$$
$$\|\mathbf{x}-\mathbf{x}^{\#}\|_2 \leq \frac{c}{\sqrt{s}}\sigma_s(\mathbf{x})_1 + D\eta$$

where $\sigma_s = \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x}-\mathbf{z}\|_1$, *and* c, d *are constants that only depend on* ϵ

Proof. The proof can be found in [12] as proof for Theorem 6.12. □

We have then using results shown 4.2.8 we claim, $\exists \Phi \in \mathbb{R}^{m\times N}$ with $m \leq cs\log\left(\frac{N}{s}\right)/\epsilon^2$ rows and $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ such that $\Delta(\Phi\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in K_s$

Homework 3.6.1. Prove Lemma 3.6.1.

Homework 3.6.2. Prove Lemma 3.6.4.

3.7 Lower Bounds on Sketching Dimension that Satisfy Recoverability

Can we produce bounds for the optimal value of m? That is what is the smallest number of rows, m such that $\exists \Phi \in \mathbb{C}^{m\times N}$ and $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ which satisfies $\Delta(\Phi\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in K_s$?

The following lemma provides a (unsatisfactory) lower bound.

Lemma 3.7.1. *If* $\Delta(\Phi\mathbf{x}) = \mathbf{x}$, $\forall \mathbf{x} \in K_s$ *then* $\Phi \in \mathbb{C}^{m\times N}$ *must have* $m \geq 2s$.

Proof. Proof left as an exercise to reader. □

Can we achieve a lower bound of the form $m \geq cs\log\left(\frac{N}{s}\right)$?

Lemma 3.7.2. *Given* $s, N \in \mathbb{N}$, $s < N$, $\exists n \geq \left(\frac{N}{4s}\right)^{\frac{s}{2}}$ *subsets* $s_1, \ldots, s_n \subset [N]$ *such that* $\forall \ell, j \in [n], \ell \neq j$

1. $|s_j| = s$

2. $|s_j \cap s_\ell| < \frac{s}{2}$

Proof. Let $s \leq N/4$ and let $\mathcal{C} = \{S \subseteq [N] || S| = s\}$. Choose $S_1 \in \mathcal{C}$. Let $\mathcal{C}_1 \subset \mathcal{C}$ be such that $S \in \mathcal{C}_1 \iff |S \cap S_1| \geq s/2$. That is I have fixed a set of indices and defined a collection which contains only elements which are sufficiently dissimilar to that fixed set. We will bound the cardinality of this set.

$$\begin{aligned} |\mathcal{C}_1| &= \sum_{k=\lceil s/2 \rceil} \binom{s}{k}\binom{N-s}{s-k} \\ &\leq 2^s \max \lceil s/2 \rceil \leq k \leq s \binom{N-s}{s-k} \\ &\leq 2^s \binom{N-s}{\lfloor s/2 \rfloor} \end{aligned}$$

The binomial coefficient is maximized when $k = \lceil s/2 \rceil$. We will continue to construct sets in this way:

1. Choose $s1 \in \mathcal{C}, \mathcal{C}_1$ as described above
2. While $|\mathcal{C} \setminus \bigcup_{\ell=1}^{n} C_\ell| > 0$
3. Choose any $s_{n+1} \in \mathcal{C} \setminus \bigcup_{\ell=1}^{n} C_\ell$
4. Let $\mathcal{C}_{n+1}$ be such that $S \in \mathcal{C}_{n+1} \iff |S \cap S_{n+1} \geq s/2|$
5. $n \leftarrow n+1$

We now have a way to estimate how many such sets can be obtained

$$n \geq \frac{|\mathcal{C}|}{\max_{1 \leq i \leq n} |\mathcal{C}_i|} \geq \frac{\binom{N}{s}}{2^s \binom{N-s}{\lfloor s/2 \rfloor}} \geq \left(\frac{N}{4s}\right)^{\frac{s}{2}}$$

□

Definition 3.7.3. Given $p, q \geq 1$ a matrix $\Phi \in \mathbb{C}^{m \times N}$ and function $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ we say the pair (Φ, Δ) is (ℓ^q, ℓ^p) instance optimal of order s with constant $c > 0$ if

$$\|\mathbf{x} - \Delta(\Phi\mathbf{x})\|_q \leq \frac{c}{s^{1/p-1/q}} \left(\inf_{\mathbf{z} \in K_s} \|\mathbf{x} - \mathbf{z}\|_p \right)$$

$\forall \mathbf{x} \in \mathbb{C}^N$.

If $p = q$ then we say that (Δ, Φ) is ℓ^p instance optimal. We note that this definition matches the equation 1.1 where $\mathcal{F}_{\mathbf{p}} = K_s, C_{\mathbf{p},X,Y} = \frac{c}{s^{1/p-1/q}}$ and where the error term $\epsilon_{\mathbf{p},X,Y,Z} = 0$.

We have from Theorem 3.6.6 that ℓ^1 minimization and matrices generated with random sub-gaussian entries satisfy definition 3.7.3 for $q = 1, 2$ and $p = 1$ where $m \leq cs \log \frac{N}{s}$

Theorem 3.7.4. *If (Φ, Δ) is ℓ^1-instance optimal of order s with constant c then $m \geq \tilde{c}s \ln\left(\frac{eN}{s}\right)$ where $\tilde{c} \in \mathbb{R}^+$ only depends on c.*

Proof. By Lemma 3.7.2, we have that there are n subsets $S_1, \ldots, S_n$ where $|S_j| = s, |S_i \cap S_j| < s/2$ if $i \neq j$ and $n \geq \left(\frac{N}{4s}\right) \frac{s}{2}$.

Define vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{C}^N$ by $(\mathbf{x}_j)_k = \not\Vdash_{S_j}(k)$. Note that $\|\mathbf{x}_j\|_1 = 1$ and since no two vectors have more than $s/2$ entries in their support that overlap $\|\mathbf{x}_j - \mathbf{x}_i\|_1 > 1$ whenever $i \neq j$.

We denote B_1^N as the open unit ℓ^1-ball in $\mathbb{C}^N$. By $\mathbf{x}_j + \rho B_1^N$ we mean the translated values of ρB_1^N - all points in an open ball of radius ρ translated by $\mathbf{x}_j$.

Claim: $\Phi\left(\mathbf{x}_j + \rho B_N^1\right) \cap \Phi\left(\mathbf{x}_i + \rho B_N^1\right) = \emptyset$ if $= \rho = (2(c+1))^{-1}$

Naturally, $\Phi\left(\mathbf{x}_i + \rho B_N^1\right) \subset \Phi\left(\mathbb{C}^N\right)$ and so if $d = \dim \Phi\left(\mathbf{x}_i + \rho B_N^1\right)$ then $d \leq m$. We prove the claim by contradiction. Suppose for eventual contradiction that

$\exists \mathbf{z}_i, \mathbf{z}_j \in \rho B_N^1$ such that $\Phi(\mathbf{x}_i + \mathbf{z}_i) = \Phi(\mathbf{x}_j + \mathbf{z}_j)$ where $i \neq j$. Note however that

$$\begin{aligned}
\|\mathbf{x}_i - \mathbf{x}_j\|_1 &= \|\mathbf{x}_i + \mathbf{z}_i - \Delta(\Phi(\mathbf{x}_i + \mathbf{z}_i)) - (\mathbf{x}_j + \mathbf{z}_j - \Delta(\Phi(\mathbf{x}_j + \mathbf{z}_j))) - \mathbf{z}_j + \mathbf{z}_i\|_1 \\
&\leq \|\mathbf{x}_i + \mathbf{z}_i - \Delta(\Phi(\mathbf{x}_i + \mathbf{z}_i))\|_1 + \|(\mathbf{x}_j + \mathbf{z}_j - \Delta(\Phi(\mathbf{x}_j + \mathbf{z}_j)))\|_1 + \|\mathbf{z}_j\|_1 + \|\mathbf{z}_i\|_1 \\
&< c\|\mathbf{z}_i\|_1 + c\|\mathbf{z}_j\|_1 + 2\rho \\
&< 2(c+1)\rho \\
&= 1
\end{aligned}$$

which is a contradiction, and thus the intersection of the images of the sets must be empty. That is $\left\{\Phi\left(\mathbf{x}_j + \rho B_1^N\right)\right\}_{j=1}^n$ is a collection of disjoint sets, and so we will be able to use their volumes additively

$$\bigcup_{j=1}^{n} \Phi\left(\mathbf{x}_j + \rho B_1^N\right) \subseteq \Phi\left((1+\rho) B_1^N\right)$$

Using linearity of Φ and the additivity of dis-joint sets, we have

$$\sum_{j=1}^{n} \mathrm{Vol}\left(\Phi\left(\mathbf{x}_j + \rho B_1^N\right)\right) \leq \mathrm{Vol}\left((1+\rho)\,\Phi\left(1 + \rho B_1^N\right)\right)$$

Noting that $\mathrm{Vol}(\rho T) = T^{2d}$ for $T \subset \mathbb{C}^d$, $\dim T = d$ and applying translational invariance of measure we have

$$n\rho^{2d}\mathrm{Vol}\left(\Phi\left(B_1^N\right)\right) \leq (1+\rho)^{2d}\mathrm{Vol}\left(\Phi\left(B_1^N\right)\right)$$

Combining then our hypothesis and a rearrangement of the above inequality we have

$$\left(\frac{N}{4s}\right)\frac{s}{2} \leq n \leq (1 + \frac{1}{\rho})^{2d} = (2c+3)^{2d} \leq (2c+3)^{2m}$$

We observe that according to Lemma 3.7.1 we know $2\frac{m}{s} \geq 4$

$$\begin{aligned}\ln\left(\frac{eN}{s}\right) &\leq \ln\left(\frac{e^3}{4}\frac{eN}{s}\right)\\ &= \ln\left(\frac{N}{4s}\right) + \ln\left(e^4\right)\\ &= 4\frac{m}{s}\ln(2c+3) + 4\\ &\leq \frac{m}{s}\left(4\ln(2c+3)+2\right)\end{aligned}$$

So after a rearrangement of terms and labeling $\tilde{c} = \left(4\ln(2c+3)+2\right)^{-1}$ we obtain the desired bound

$$m \geq \tilde{c}s\ln\left(\frac{eN}{s}\right)$$

□

The upshot of Theorem 3.7.4 is that the ℓ^1-minimization and random matrices of the type we've described are near optimal, up to a scaling in the constant term.

Our aim now is to generalize Theorem 3.7.4 for any choice of norm $p, q \geq 1$. First though we show that the null space property is essential to (ℓ^p, ℓ^q) instance optimalality.

Lemma 3.7.5. *Let $p \geq q \geq 1$ and $\Phi \in \mathbb{C}^{m\times N}$ be given. If there exists a map $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ such that (Φ, Δ) is (ℓ^p, ℓ^q)-optimal of order s with constant c then*

$$\|\mathbf{v}\|_q \leq \frac{c}{s^{1/p-1/q}}\sigma_{2s}(\mathbf{v})_p,\ \forall \mathbf{v} \in Ker(\Phi) \tag{3.7}$$

where $\sigma_{2s} = \inf_{\|\mathbf{z}\|_0\leq 2s}\|\mathbf{v}-\mathbf{z}\|_p$.

Conversely, if inequality 3.7 holds for Φ then there exists a map $\Delta' : \mathbb{R}^m \to \mathbb{R}^N$ such that (Φ, Δ) is (ℓ^p, ℓ^q)-instance optimal or order s with constant $2c$.

Proof. Suppose (Φ, Δ) is (ℓ^p, ℓ^q)-optimal of order s with constant c. Given $\mathbf{v} \in \text{Ker}(\Phi)$, let S be the index of set of the s largest entries of $\mathbf{v}$. Note $-\mathbf{v}_S \in K_s$. Instance optimality implies that

$$-\mathbf{v}_S = \Delta\left(\Phi\left(-\mathbf{v}_S\right)\right)$$

Also, $\mathbf{v} \in \text{Ker}(\Phi)$ implies $\Phi(-\mathbf{v}_S) = \Phi(\mathbf{v}_{\overline{S}})$. So $-\mathbf{v}_S = \Delta\left(\Phi\left(\mathbf{v}_{\overline{S}}\right)\right)$ and therefore, using the instance optimality of (Φ, Δ)

$$\begin{aligned}
\|\mathbf{v}\|_q &= \|\mathbf{v}_{\overline{S}} + \mathbf{v}_S\|_q \\
&= \|\mathbf{v}_{\overline{S}} - \Delta\left(\Phi\left(\mathbf{v}_{\overline{S}}\right)\right)\|_q \\
&\leq \frac{c}{s^{1/p-1/q}} \sigma_s(\mathbf{v}_{\overline{S}})_p \qquad\qquad =\leq \frac{c}{s^{1/p-1/q}} \sigma_{2s}(\mathbf{v})_p
\end{aligned}$$

Which matches the desired result for this direction of the implication.

Now suppose the inequality 3.7 holds for matrix Φ. Define

$$\Delta'(\mathbf{y}) = \text{argmin}_{\mathbf{z} \in \mathbb{C}^N} \sigma_s(\mathbf{z})_p \text{ subject to } \Phi\mathbf{z} = \mathbf{y}$$

Note that $\Phi\left(\Delta'\left(\Phi\mathbf{x}\right)\right) = \Phi\mathbf{x}$ since the constraints in the optimization problem are that for $\Delta'(\mathbf{y})$, we are constrained to $\mathbf{z} \in \mathbb{C}^N$ where $\Phi\mathbf{z} = \mathbf{y}$. So by choosing $\mathbf{y}$ equal to something in the range of Φ, $\mathbf{y} = \Phi\mathbf{x}$, we must have that $\Phi\left(\Delta'\left(\Phi\mathbf{x}\right)\right) = \Phi\mathbf{z} = \Phi\mathbf{x}$.

So, the vector we define $\mathbf{v} = \mathbf{x} - \Delta'(\Phi\mathbf{x})$ must then be in the kernel, $\mathbf{v} \in \mathrm{Ker}(\Phi)$.

$$\begin{aligned}
\|\mathbf{v}\|_q &= \|\mathbf{x} - \Delta'(\Phi\mathbf{x})\|_q \\
&\leq \frac{c}{s^{1/p-1/q}}\sigma_{2s}(\mathbf{x} - \Delta'(\Phi\mathbf{x}))_p \\
&= \frac{c}{s^{1/p-1/q}} \inf_{\mathbf{z}\in K_{2s}} \|\mathbf{x} - \Delta'(\Phi\mathbf{x}) - \mathbf{z}\|_p \\
&= \frac{c}{s^{1/p-1/q}} \inf_{\mathbf{z}_1,\mathbf{z}_2\leq s} \|\mathbf{x} - \Delta'(\Phi\mathbf{x}) - \mathbf{z}_1 - \mathbf{z}_2\|_p \\
&\leq \frac{c}{s^{1/p-1/q}} \inf_{\mathbf{z}_1\leq s} \|\mathbf{x} - \mathbf{z}_1\|_p + \inf_{\mathbf{z}_2\leq s} \|\Delta'(\Phi\mathbf{x}) - \mathbf{z}_2\|_p \\
&\leq \frac{c}{s^{1/p-1/q}} \inf_{\mathbf{z}_1\leq s} \|\mathbf{x} - \mathbf{z}_1\|_p + \inf_{\mathbf{z}_2\leq s} \|\mathbf{x} - \mathbf{z}_2\|_p \\
&= \frac{c}{s^{1/p-1/q}} (\sigma_s(\mathbf{x})_p + \sigma_s(\mathbf{x})_p) \\
&= \frac{2c}{s^{1/p-1/q}}\sigma_s(\mathbf{x})_p
\end{aligned}$$

Where we have used in the last inequality above that Δ' returns the vector $\mathbf{v}$ with minimal error residue after removing s entries such that $\Phi\mathbf{v} = \Phi\mathbf{x}$ so certainly we only do worse in terms of error by replacing $\Delta'(\Phi\mathbf{x})$ with $\mathbf{x}$ i.e. $\mathrm{argmin}_{v\in\mathbb{C}^N}\sigma_s(\mathbf{z}) \leq \sigma_s(\mathbf{x})$. □

Because of its utility and importance, inequality 3.7 is referred to as the mixed null space property for a matrix Φ.

Theorem 3.7.6. *Given $q \geq 1$, if a pair of measurement matrix and recovery function (Φ, Δ) is (ℓ^q, ℓ^1)-instance optimal of order s with constant c then m the number of rows of the matrix Φ must satisfy*

$$m \geq \tilde{c}s \ln\left(\frac{eN}{s}\right)$$

for some $\tilde{c}$ which depends only on the constant c

Proof. Consider the following: If there is an (ℓ^q, ℓ^1)-instance optimal pair (Φ, Δ) then there exists a function Δ' such that (Φ, Δ') is ℓ^1-instance optimal and thus, by Theorem 3.7.4 the number of rows m must satisfy

$$m \geq \tilde{c} s \ln\left(\frac{eN}{s}\right)$$

So we are able to reduce our current theorem to the previous ℓ^1 result if we are able to show the existence of Δ'. We will do this by making use of the mixed null space property equivalence detailed in Lemma 3.7.5.

Consider $\mathbf{v} \in \text{Ker}(\Phi)$. Since (Φ, Δ) is instance optimal with constant c, by the Lemma

$$\|\mathbf{v}\|_q \leq \frac{c}{s^{1-1/q}} \sigma_{2s}(\mathbf{v})_1 \tag{3.8}$$

Let $S \subset [N]$ be such that V_S contains the $2s$ largest magnitude entries of $\mathbf{v}$. Using Holder's inequality then we have

$$\begin{aligned} \|\mathbf{v}_S\|_1 &= \langle \text{sgn}(\mathbf{v}) \mathbb{1}_S, \mathbf{v} \rangle \\ &\leq \|\mathbb{1}_S\|_{\frac{q}{q-1}} \|\mathbf{v}\|_q \\ &= (2s)^{1-\frac{1}{q}} \frac{c}{s^{1-1/q}} \sigma_{2s}(\mathbf{v})_1 \qquad \leq 2c\sigma_{2s}(\mathbf{v})_1 \end{aligned}$$

On the other hand $\|\mathbf{v}_{\overline{S}}\|_1 = \sigma_{2s}$ since we choose S to include the largest $2s$ entries. Thus

$$\|\mathbf{v}\|_1 = \|\mathbf{v}_S\|_1 + \|\mathbf{v}_{\overline{S}}\|_1 \leq (2c+1)\sigma_{2s}(\mathbf{v})_1$$

And so using the null space property equivalence from Lemma **??** we have the desired bound on m □

We conclude this section with some general remarks about how to interpret and understand the main theorems.

- Matrices with the RIP along with basis pursuit are a way to achieve instance optimalality.
- To ensure that a matrix has the RIP, we need to have the number of rows on the order of $cs \log(N/s)$.
- A matrix with RIP is just an instance of an ϵ-JL map for the set $K_s \subset \mathbb{C}^N$.
- We have a particular example of a finite set of $\mathbb{C}^N$ with the lower bound $cs \log(N/s)$. This then suggests that the lower bound for m is needed to ensure an embedding of an arbitrary subset $S \subset \mathbb{C}^N$. In [25] the authors prove the result: $\exists S \subset \mathbb{C}^N$ such that an ϵ-JL map of S into $\mathbb{C}^m$ requires $m \geq c\epsilon^{-2} \log(|S|)$.
- The result in [25] does not restrict to any particular type of JL map - for example nonlinear maps. The surprising conclusion is that matrices do not incur a penalty beyond scaling in the constant in terms of size of the sketching dimension.

Homework 3.7.1. Prove Lemma 3.7.1. **Hint**: consider a sub-matrix of Φ which interacts with the support of $\mathbf{x} - \mathbf{y}$.

CMSE 890 Lecture 5 and the MTH 994 Lectures should be merged – there is a lot of overlap. Ask me if you need to discuss the best organization.

Chapter IV

Probability Strikes Back: Randomized Constructions of Oblivious LJL Embeddings and More (MTH 994 Lectures 5 & 7) & (CMSE Lecture 6)

4.1 Useful General Purpose Probability Inequalities (MTH 994 Lecture 5)

Theorem 4.1.1 (Cramer's Theorem)**.** *Let $X_1, \ldots, X_m$ be a sequence of independent real-valued random variables with cumulant generating functions*

$$C_{X_\ell}(\theta) = \ln\left(\mathbb{E}\left[e^{\theta X_\ell}\right]\right)$$

where $\ell \in [m]$. Then, $\forall t > 0$

$$P\left[\sum_{\ell=1}^{m} X_\ell \geq t\right] \leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{\ell=1}^{m} C_{X_\ell}(\theta)\right\}\right)$$

Proof. By Markov's inequality, for any $\theta > 0$, and independence of the random

variables we have:

$$\begin{aligned} P\left[\sum_{\ell=1}^{m} X_\ell \geq t\right] &= P\left[\exp\left(\theta \sum_{\ell=1}^{m} X_\ell\right) \geq \exp\left(\theta t\right)\right] \\ &\leq e^{-\theta t}\mathbb{E}\left[\exp\left(\theta \sum_{\ell=1}^{m} X_\ell\right)\right] \\ &= e^{-\theta t}\prod_{\ell=1}^{m}\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right] \\ &= \exp\left(\ln\left(e^{-\theta t}\prod_{\ell=1}^{m}\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right]\right)\right) \\ &= \exp\left(-\theta t + \sum_{\ell=1}^{m} C_{X_\ell}(\theta)\right) \end{aligned}$$

Since the above holds for all $\theta > 0$, it will hold for the infimum, which matches our desired outcome □

Theorem 4.1.2 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_m$ be a sequence of independent random variables such that $\mathbb{E}[X_\ell] = 0$ and $|X_\ell| \leq B_\ell$, $\forall \ell \in [m]$. Then, $\forall t > 0$*

$$P\left[\sum_{\ell=1}^{m} X_\ell \geq t\right] \leq \exp\left(\frac{-t^2}{2\sum_{\ell=1}^{m} B_\ell^2}\right)$$

and so

$$P\left[\left|\sum_{\ell=1}^{m} X_\ell\right| \geq t\right] \leq 2\exp\left(\frac{-t^2}{2\sum_{\ell=1}^{m} B_\ell^2}\right)$$

Proof. We first estimate the moment generating function $\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right]$ and then apply Cramer's Theorem.

For some $\tilde{t}_\ell > 0$, we can write each of the random variables as some combination of its bounds:

$$X_\ell = \tilde{t}(-B_\ell) + (1 - \tilde{t})B_\ell$$

Solving for $\tilde{t}_\ell$ we have

$$\tilde{t}_\ell = \frac{B_\ell - X_\ell}{2B_\ell} \in [0,1]$$

Since $\exp(\theta x)$, $\theta > 0$ is a convex function and so we have the bound

$$\begin{aligned}\exp(\theta X_\ell) &\leq \tilde{t}\exp(-\theta B_\ell) + (1-\tilde{t})\exp(\theta B_\ell) \\ &= \left(\frac{B_\ell - X_\ell}{2B_\ell}\right)\exp(-\theta B_\ell) + \left(\frac{B_\ell + X_\ell}{2B_\ell}\right)\exp(\theta B_\ell)\end{aligned}$$

and so taking the expectation and recalling $\mathbb{E}[X_\ell] = 0$, we obtain the moment generating function and the following bound

$$\begin{aligned}\mathbb{E}\left[\exp(\theta X_\ell)\right] &\leq \frac{1}{2}\left[\exp(-\theta B_\ell) + \exp(\theta B_\ell)\right] \\ &= \sum_{k=0}^{\infty}\frac{(\theta B_\ell)^{2k}}{(2k)!} \\ &\leq \sum_{k=0}^{\infty}\frac{(\theta B_\ell)^{2k}}{2^k k!} \\ &= \exp\left(\frac{\theta^2 B_\ell^2}{2}\right)\end{aligned}$$

Apply Cramer's Theorem 4.1.1 with $\theta = \frac{t}{\sum_{\ell=1}^m B_\ell^2}$ and bound C_{X_ℓ} by $\frac{\theta^2 B_\ell^2}{2}$ to obtain

$$P\left[\sum_{\ell=1}^m X_\ell \geq t\right] \leq \exp\left(\frac{-t^2}{2\sum_{\ell=1}^m B_\ell^2}\right)$$

□

Definition 4.1.3 (Radamacher Random Variable)**.** A random variable X such that

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

Note that the expectation of such a variable is 0.

Corollary 4.1.4. *Let $a \in \mathbb{R}^m$ and $\mathbf{X} = (X_1, \ldots, X_m)$ be a random vector with i.i.d. Radamacher entries. Then, $\forall u > 0$, we have*

$$P\left[\left|\sum_{\ell=1}^m a_\ell X_\ell\right| \geq u\|\mathbf{a}\|_2\right] \leq 2\exp\left(-\frac{u^2}{2}\right)$$

The corollary follows from an application of Hoeffding's inequality.

Theorem 4.1.5 (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_m$ be a sequence of independent random variables such that $\mathbb{E}[X_\ell] = 0$ such that $\forall n \geq 2$*

$$\mathbb{E}\left[|X_\ell|^n\right] \leq n! R^{n-2} \sigma_\ell^2 / 2, \, \forall \ell \in [m]$$

for some constants $R > 0$ and $\sigma_\ell > 0, \ell \in [m]$. Then $\forall t > 0$

$$P\left[\sum_{\ell=1}^{m} X_\ell \geq t\right] \leq 2 \exp\left(\frac{-t^2/2}{\sigma^2 + Rt}\right)$$

where $\sigma^2 = \sum_{\ell=1}^{m} \sigma_\ell^2$

Proof. First we bound the moment generating function $\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right]$ by expanding the exponential function into a series and applying the linearity of expectation after exchanging integration and summation (Fubini and Dominated Convergence):

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right] &= 1 + \theta \mathbb{E}\left[X_\ell\right] + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}\left[X_\ell^n\right]}{n!} \\
&= 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}\left[X_\ell^n\right]}{n!} \\
&\leq 1 + \frac{\sigma_\ell^2 \theta^2}{2} \sum_{n=0}^{\infty} (\theta R)^n \\
&\leq 1 + \frac{\sigma_\ell^2 \theta^2}{2} \left(1 - R\theta\right)^{-1} \\
&\leq \exp\left(\frac{\sigma_\ell^2 \theta^2}{2(1 - R\theta)}\right)
\end{aligned}$$

where we require that $0 < R\theta < 1$ to get convergence of the geometric series. Apply Cramer's theorem then using this bound on the cumulant density functions to obtain

$$\begin{aligned}
P\left[\left|\sum_{\ell=1}^{m} X_\ell\right| \geq t\right] &\leq 2 \inf_{\theta \in (0, R^{-1})} \exp\left(-\theta t + \sum_{\ell=1}^{m} \frac{\sigma_\ell^2 \theta^2}{2(1 - R\theta)}\right) \\
&= 2 \inf_{\theta \in (0, R^{-1})} \exp\left(-\theta t + \frac{\sigma^2 \theta^2}{2(1 - R\theta)}\right)
\end{aligned}$$

Choosing $\theta = \frac{t}{\sigma^2 + Rt} < \frac{1}{R}$ then leads to our desired final bound. □

Lemma 4.1.6. $\mathbb{E}[|X|^n] \leq n \int_0^\infty P[|X| \geq t] t^{n-1} dt$, $\forall n > 0$

Proof.

$$\begin{aligned}
\int_\Omega |X|^n p(x) dx &= \int_\Omega \left(\int_0^\infty \mathbb{1}_{\{0 \leq y \leq |X|^n\}} dy \right) p(x) dx \\
&= \int_0^\infty \int_\Omega P\left[|X|^n \geq y\right] dy && \text{Fubini's Theorem} \\
&= n \int_0^\infty \int_\Omega P\left[|X|^n \geq t^n\right] t^{n-1} dt && \text{change of variable } y = t^n \\
&= n \int_0^\infty \int_\Omega P\left[|X| \geq t\right] t^{n-1} dt
\end{aligned}$$

□

We now introduce some definitions about a large class of random variables which will help us prove in a general way many important and useful results for JL maps.

Definition 4.1.7 (Sub-exponential Random Variable)**.** We say that $X \in \mathbb{R}$ is sub-exponential random variable if $\exists \beta, \kappa > 0$ such that

$$P\left[|X| \geq t\right] \leq \beta e^{-\kappa t}, \forall t > 0$$

We can understand this as saying that the random variable decays exponentially.

Definition 4.1.8 (Sub-gaussian Random Variable)**.** We say that $X \in \mathbb{R}$ is subgaussian random variable if $\exists \beta, \kappa > 0$ such that

$$P\left[|X| \geq t\right] \leq \beta e^{-\kappa t^2}, \forall t > 0$$

Again, we understand this as saying that the random variable decays faster than exponential, at a rate comparable to a Gaussian. What types of random variables fit these definitions? The following examples provide some indication as to the richness of the classification.

Example 4.1.9. If X is sub-gaussian, then X^2 is sub-exponential with the same β and κ:

$$\beta e^{-\kappa t^2} \geq P\left[|X| \geq t\right] = P\left[|X|^2 \geq t^2\right]$$

but after a relabeling of t we rewrite as

$$\beta e^{-\kappa t} \geq P\left[X^2 \geq t\right]$$

for all $t > 0$.

Example 4.1.10. All bounded random variables, e.g. Radamacher, Bernoulli, uniform on bounded interval, all discrete random variables, are subgaussian.

Example 4.1.11. A Gaussian random variable $X \sim \mathcal{N}(0,1)$ is sub-gaussian with $\beta = 1$ and $\kappa = 1/2$

Theorem 4.1.12 (Bernstein's Inequality for sub-expoential random variables)**.** *Let $X_1, \ldots, X_m$ be independent mean 0 sub-expoential random variables such that $\exists \beta, \kappa > 0$ where $P\left[|X_\ell| \geq t\right] \leq \beta e^{-\kappa t}$, $\forall t > 0, \forall \ell \in [m]$. Then*

$$P\left[\left|\sum_{\ell=1}^{m} X_\ell\right| \geq t\right] \leq 2\exp\left(\frac{-(\kappa t)^2/2}{2\beta m + \kappa t}\right)$$

Proof. By Lemma 4.1.6 we have

$$\begin{aligned}
\mathbb{E}\left[|X_\ell|^n\right] &= n\int_0^\infty P\left[|X| \geq t\right] t^{n-1} dt \\
&= \beta n \int_0^\infty e^{-\kappa t} t^{n-1} dt && \text{let } \kappa t = u \\
&= \beta n \kappa^{-n} \int_0^\infty e^{-u} u^{n-1} du && \text{Notice } \Gamma \text{ function definition} \\
&= \beta \kappa^{-n} n!
\end{aligned}$$

Now apply Bernstein's Inequality 4.1.5 with $R = \kappa^{-1}$ and $\sigma_\ell^2 = 2\beta\kappa^{-2}$ to obtain the desired bound. □

Homework 4.1.1. Prove that if X is a bounded random variable then $\exists t_0 \in \mathbb{R}^+$ such that X is sub-gaussian $\forall \beta, \kappa > 0$ satisfying $t_0 = \sqrt{\frac{\ln \beta}{\kappa}}$

Homework 4.1.2. Prove that if $X \sim \mathcal{N}(0,1)$ then

$$P\left[|X| \geq t\right] \leq e^{-t^2/2}, \forall t > 0$$

Homework 4.1.3. Let X be uniformly distributed on $[-1,1]$ show that

$$\mathbb{E}\left[|X|^2\right] = 1/3$$

and that

$$\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(\theta^2/6\right) = \exp\left(\theta^2 \mathbb{E}[|X|^2]/2\right)$$

4.2 Stability of Subgaussians as a Class of Random Variables

Lemma 4.2.1. *If X is a sub-gaussian random variable with parameters $\beta, \kappa > 0$ then*

$$\|X\|_p = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq \kappa^{-1}\beta^{1/2}p^{1/2}, \ \forall p \geq 1$$

Proof. By Lemma 4.1.6

$$\begin{aligned}
\mathbb{E}\left[|X_\ell|^p\right] &= p\int_0^\infty P\left[|X| \geq t\right] t^{p-1}dt \\
&= \frac{p}{(2\kappa)^{p/2}}\int_0^\infty P\left[|X| \geq \frac{u}{\sqrt{2\kappa}}\right] u^{p-1}du && \text{let } t = \frac{u}{\sqrt{2\kappa}} \\
&\leq \frac{p\beta}{(2\kappa)^{p/2}}\int_0^\infty e^{-u^2/2}u^{p-1}du && X \text{is sub-gaussian} \\
&= \frac{p\beta}{2\kappa^{p/2}}\Gamma\left(\frac{p}{2}\right) \\
&= \frac{p\beta}{\kappa^{p/2}}\sqrt{\frac{\pi}{2}}\left(\frac{p}{2}\right)^{p/2-1/2}e^{-p/2}e^{1/6p}\Gamma\left(\frac{p}{2}\right) \\
&= \kappa^{-p/2}\beta p^{p/2}\left[\frac{1}{2^{p/2}}\sqrt{p\pi}e^{-p/2+1/6p}\right]
\end{aligned}$$

$\left[\frac{1}{2^{p/2}}\sqrt{p\pi}e^{-p/2+1/6p}\right]$ is monotonically decreasing for $p \geq 1$ and is bounded by 1 from above, which then taking p-th root leads to the desired final bound. □

Lemma 4.2.2. *If X is sub-gaussian with parameters β, κ then $\exists c \in (0, \kappa)$ and $\tilde{c} \geq 1 + \frac{\beta c \kappa^{-1}}{1 - c\kappa^{-1}}$ such that $\mathbb{E}\left[\exp\left(cX^2\right)\right] \leq \tilde{c}$.*

Proof. By using Lemma 4.1.6 with $p \leftarrow 2n$ then

$$\mathbb{E}\left[|X|^{2n}\right] \leq \beta \kappa^{-n} n!$$

$$\begin{aligned}
\mathbb{E}\left[\exp\left(cX^2\right)\right] &= \int_0^\infty \sum_{n=0}^{\infty} \frac{c^n X^{2n}}{n!} p(x) dx \\
&= \sum_{n=0}^{\infty} \frac{c^n \mathbb{E}\left[X^{2n}\right]}{n!} && \text{Fubini's Theorem} \\
&\leq 1 + \sum_{n=1}^{\infty} c^n \beta \kappa^{-n} && \text{Series converges when } c \in (0, \kappa) \\
&= 1 + \frac{\beta c \kappa^{-1}}{1 - c\kappa^{-1}}
\end{aligned}$$

□

Theorem 4.2.3 (Alternative Characterization of Sub-gaussian Property)**.** *Let $X \in \mathbb{R}$ be a random variable*

1. *If X is sub-gaussian with $\mathbb{E}[X] = 0$ then $\forall c \in \mathbb{R}^+$ with $c > \max\left\{\frac{1}{2\kappa} + \frac{4e^2}{\kappa}\ln(1+\beta), \frac{\sqrt{2}\beta e^2}{\kappa\sqrt{\pi}}\right\}$ then*

   $$\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(c\theta^2\right), \forall \theta \in \mathbb{R}^+ \tag{4.1}$$

2. *If 4.1 holds for some $c \in \mathbb{R}^+$ then $\mathbb{E}[X] = 0$ and X is sub-gaussian with parameters $\beta = 2$ and $\kappa = \frac{1}{4c}$*

Proof. We begin with the second part of the statement of the theorem. Assume inequality 4.1 holds.

$$\begin{aligned} P[X \geq t] &= P\left[\exp\left(\theta X \geq e^{\theta t}\right)\right] \\ &\leq e^{-\theta t}\mathbb{E}\left[\exp\left(\theta X\right)\right] && \text{Markov's Inequality} \\ &\leq e^{c\theta^2 - \theta t} && \text{by hypothesis} \end{aligned}$$

Setting $\theta = \frac{t}{2c}$ minimizes the right hand side and we have $P[X \geq t] \leq e^{\frac{-t^2}{4c}}$. We can repeat the same argument to conclude that $P[-X \geq t] \leq e^{\frac{-t^2}{4c}}$ and so conclude by a union bound that $P[|X| \geq t] \leq 2e^{\frac{-t^2}{4c}}$, i.e. X is sub-gaussian with with parameters $\beta = 2$ and $\kappa = \frac{1}{4c}$.

To see that the random variable must have mean zero, recall the bounds $(1+x) \leq e^x$, $\forall x \in \mathbb{R}$. Use this bound, taking expectation with respect to the random variable X, and using the series definition of the exponential we have

$$\begin{aligned} 1 + \mathbb{E}\left[\theta X\right] &\leq \mathbb{E}\left[\exp\left(\theta X\right)\right] \\ \implies 1 + \theta\mathbb{E}\left[X\right] &\leq \exp\left(c\theta^2\right) \\ \implies \theta\mathbb{E}\left[X\right] &\leq \frac{1}{2}c\theta^2 + \mathcal{O}\left(\theta^4\right) \end{aligned}$$

So sending $\theta \to 0$ yields $\mathbb{E}[X] = 0$.

Now we consider the first part of the theorem. Assume that the random variable X is sub-gaussian with $c > \max\left\{\frac{1}{2\kappa} + \frac{4e^2}{\kappa}\ln\left(1+\beta\right), \frac{\sqrt{2}\beta e^2}{\kappa\sqrt{\pi}}\right\}$. For the moment, consider $|\theta| \leq \theta_0$ for some yet to be determined θ_0

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\theta X\right)\right] &= 1 + \theta\mathbb{E}\left[X\right] + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}\left[X^n\right]}{n!} && \text{Fubini, linearity of expectation} \\
&= 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}\left[X^n\right]}{n!} && \text{mean zero} \\
&\leq 1 + \sum_{n=2}^{\infty} \frac{|\theta|^n \kappa^{n/2} \beta n^{n/2}}{\sqrt{2\pi} n^n e^{-n}} && \text{Lemma 4.2.1. Sterling's formula} \\
&= 1 + \frac{\beta}{\sqrt{2\pi}} \frac{\theta^2 e^2}{\kappa} \sum_{n=0}^{\infty} \theta_0^n \kappa^{-n/2} e^n && \text{re-indexing, } |\theta| \leq \theta_0 \\
&= 1 + \theta^2 \frac{\beta}{\sqrt{2\pi}} \frac{\theta^2 e^2}{\kappa} \frac{1}{1 - \frac{1}{2}} && \text{set } \theta_0 = \frac{\sqrt{\kappa}}{2e} \\
&\leq \exp(c\theta^2) && \text{when } c > \frac{\sqrt{2}\beta e^2}{\kappa\sqrt{\pi}}
\end{aligned}$$

We now must consider the case when $|\theta| > \theta_0$. We wish to show that

$$\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(c\theta^2\right) \iff \mathbb{E}\left[\exp\left(\theta X - c\theta^2\right)\right] \leq 1$$

Notice that by completing the square, for any positive constant c

$$\theta X - c\theta^2 = -\left(\sqrt{c}\theta - \frac{X}{2\sqrt{c}} + \frac{X^2}{4c}\right) \leq \frac{X^2}{4c}$$

So then

$$\mathbb{E}\left[\exp\left(\theta X - c\theta^2\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{X^2}{4c}\right)\right]$$

In particular for constant $\frac{1}{2\kappa}$ then,

$$\mathbb{E}\left[\exp\left(\theta X - \frac{1}{2\kappa}\theta^2\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{\kappa X^2}{2}\right)\right]$$

So then in Lemma 4.2.2, where for $c = \frac{\kappa}{2}$ we have for $\tilde{c} > 1 + \frac{\beta c \kappa^{-1}}{1 - c\kappa^{-1}} = 1 + \beta$. Noting this then we have that

$$\mathbb{E}\left[\exp\left(\frac{\kappa}{2}X^2\right)\right] \leq 1 + \beta$$

So combining the two inequalities we have

$$\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(\frac{\theta^2}{2\kappa}\right)\left(1+\beta\right)$$

Now let $\rho = \ln\left(1+\beta\right)\theta_0^{-2}$

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\theta X\right)\right] &\leq \left(1+\beta\right)\exp\left(\frac{\theta^2}{2\kappa}\right) \\
&= \left(1+\beta\right)\exp\left(-\rho\theta^2\right)\exp\left(\frac{\theta^2}{2\kappa}\right)\exp\left(\rho\theta^2\right) \\
&\leq \left(1+\beta\right)\exp\left(-\rho\theta_0^2\right)\exp\left(\left(\frac{1}{2\kappa}+\rho\right)\theta^2\right) \\
&\leq \exp\left(\left(\frac{1}{2\kappa}+\rho\right)\theta^2\right)
\end{aligned}$$

Noting that $\theta_0 = \frac{\sqrt{\kappa}}{2e}$ and $\rho = \frac{4e^2}{\kappa}\ln\left(1+\beta\right)$ we have then the desired bound. □

Theorem 4.2.4 (Stability of Sub-gaussians)**.** *Let* $\mathbf{X} = X_1, \ldots, X_m$ *be independent mean zero sub-gaussian random variables such that* $\mathbb{E}\left[\exp\left(\theta X_\ell\right)\right] \leq \exp\left(c\theta^2\right)$ *for* $\ell \in [m], \theta \in \mathbb{R}^+$*. Let* $\mathbf{a} \in \mathbb{R}^m$ *and define* $z = \langle \mathbf{a}, X\rangle$*. Then* z *is sub-gaussian with*

1.

$$\mathbb{E}\left[\exp\left(\theta z\right)\right] \leq \exp\left(c\|\mathbf{a}\|_2^2\theta^2\right)$$

2.

$$P\left[|z| \geq t\right] \leq 2\exp\left(\frac{-t^2}{4c\|\mathbf{a}\|_2^2}\right), \forall t > 0$$

Proof. 1.

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\theta\sum_{\ell=1}^m a_\ell X_\ell\right)\right] &= \prod_{\ell=1}^m \left[\exp\left(\theta a_\ell X_\ell\right)\right] \qquad X_\ell \text{ independent} \\
&\leq \prod_{\ell=1}^m \exp\left(ca_\ell^2\theta^2\right) \\
&\leq \exp\left(c\|\mathbf{a}\|_2^2\theta^2\right)
\end{aligned}$$

2. This follows from part 2 of Theorem 4.2.3.

□

Definition 4.2.5. A sub-gaussian random variable X allows a parameter c if

$$\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(c\theta^2\right), \forall \theta \in \mathbb{R}^+$$

Lemma 4.2.6. *Let $\mathbf{Z} \in \mathbb{R}^N$ be a random vector with independent, mean zero, variance 1, sub-gaussian entries that all allow the same parameter $c \in \mathbb{R}^+$. Then*

1.

$$\mathbb{E}\left[\left|\langle \mathbf{Z}, \mathbf{x}\rangle\right|^2\right] = \|\mathbf{x}\|_2^2, \forall \mathbf{x} \in \mathbb{R}^N$$

2. $\langle \mathbf{Z}, \mathbf{x}/\|\mathbf{x}\|_2\rangle$ is sub-gaussian and also allows the parameter c

Proof. 1. We expand the square of the sum, use linearity of expectation, independence of variables and the mean zero and variance of one of all the random variables to obtain:

$$\begin{aligned}
\mathbb{E}\left[\left|\langle \mathbf{Z}, \mathbf{x}\rangle\right|^2\right] &= \mathbb{E}\left[\left(\sum_{\ell=1}^{N} Z_\ell x_\ell\right)^2\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{N}\sum_{k=1}^{N} Z_\ell Z_k x_\ell x_k\right] \\
&= \sum_{\ell=1}^{N}\sum_{k=1}^{N} \mathbb{E}\left[Z_\ell Z_k\right] x_\ell x_k \\
&= \sum_{\ell=1}^{N} \mathbb{E}\left[Z_\ell^2\right] x_\ell^2 + \sum_{\ell=1}^{N}\sum_{k\neq\ell} \mathbb{E}\left[Z_\ell\right]\mathbb{E}\left[Z_k\right] x_\ell x_k \\
&= \sum_{\ell=1}^{N} (1) x_\ell^2 \\
&= \|\mathbf{x}\|_2^2
\end{aligned}$$

2. Follows from part 1. of Theorem 4.2.4

□

Theorem 4.2.7 (Concentration Inequality for Sub-gaussian Random Variables)**.** *Let $\Phi \in \mathbb{R}^{m\times N}$ be a matrix with independent, mean zero, variance one sub-gaussian entries that all allow parameter c. Then $\forall \mathbf{x} \in \mathbb{R}^n$ and $t \in (0,1)$,*

$$P\left(\left|m^{-1}\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \geq t\|\mathbf{x}\|_2^2\right) \leq 2\exp\left(-\tilde{c}mt^2\right)$$

where $\tilde{c}$ depends only on c, $\tilde{c} = \frac{1}{8c(16c+1)}$

Proof. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \in \mathbb{R}^N$ be the rows of the matrix Φ. Define

$$Z_\ell = |\langle \mathbf{Y}_\ell, \mathbf{x}\rangle|^2 - \|\mathbf{x}\|_2^2,\ \ell \in [m].$$

By Lemma 4.2.6 we have that $\mathbb{E}[Z_\ell] = 0$. Furthermore, $\langle \mathbf{Y}_\ell, \mathbf{x}/\|\mathbf{x}\|_2\rangle$ is sub-gaussian with parameter c. Now using the characterization of sub-gaussian random variables seen in Theorem 4.2.3, we have that $\langle \mathbf{Y}_\ell, \mathbf{x}/\|\mathbf{x}\|_2\rangle$ works as a sub-gaussian random variable for $\beta = 2$ and $\kappa = 1/4c$ with mean 0.

Therefore

$$P\left[|\langle \mathbf{Y}_\ell, \mathbf{x}/\|\mathbf{x}\|_2\rangle| \geq r\right] \leq \beta e^{-\kappa r^2}$$

Squaring the random variable then gives us a sub-exponential random variable concentration result,

$$P\left[|\langle \mathbf{Y}_\ell, \mathbf{x}/\|\mathbf{x}\|_2\rangle|^2 \geq \tilde{r}\right] \leq \beta e^{-\kappa\tilde{r}}$$

where $\tilde{r} = r^2$. Note

$$\frac{1}{\|\mathbf{x}\|_2}\left(m^{-1}\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right) = \frac{1}{m}\sum_{\ell=1}^{m}\left(\frac{|\langle \mathbf{Y}_\ell, \mathbf{x}\rangle|^2 - \|\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2}\right)$$
$$= \frac{1}{m\|\mathbf{x}\|_2^2}\sum_{\ell=1}^{m} Z_\ell$$

We now have what we need to satisfy Bernstein's inequality for sub-exponential random variables. That is

$$P\left[\frac{1}{m\|\mathbf{x}\|_2^2}\left|\sum_{\ell=1}^{m} Z_\ell\right| \geq t\right] = P\left[\left|\sum_{\ell=1}^{m}\frac{Z_\ell}{\|\mathbf{x}\|_2^2}\right| \geq mt\right]$$
$$\leq 2\exp\left(\frac{-mt^2\kappa^2}{4\beta + 2\kappa t}\right)$$
$$\leq 2\exp\left(-mt^2\tilde{c}\right)$$

when $\beta = 2, \kappa = \frac{1}{4c}$ and $\tilde{c} = \frac{1}{8c(16c+1)}$ □

Theorem 4.2.8. *Let $S \subset \mathbb{R}^N$ be an arbitrary finite subset of $\mathbb{R}^N$. Let $p, \epsilon \in (0,1)$. Finally, let $\Phi \in \mathbb{R}^{m\times N}$ be a matrix with independent, mean zero, variance one, sub-gaussian entries all allowing the parameter c. Then*

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \|\frac{1}{\sqrt{m}}\Phi(\mathbf{x}-\mathbf{y})\|_2^2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2$$

will hold for all $\mathbf{x}, \mathbf{y} \in S$ with probability at least p, provided that $m \geq \frac{8c(16c+1)}{\epsilon^2}\ln\left(\frac{|S|^2}{1-p}\right)$.

Proof. The proof of this theorem is left as an exercise to the reader. □

Theorem 4.2.9. *Let $S \subset \mathbb{C}^N$ be an arbitrary finite subset of $\mathbb{C}^N$, $p, \epsilon \in (0,1)$. Then Φ as in Theorem 4.2.8 will satisfy*

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \|\frac{1}{\sqrt{m}}\Phi(\mathbf{x}-\mathbf{y})\|_2^2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2^2$$

will hold for all $\mathbf{x}, \mathbf{y} \in S$ *with probability at least* p*, provided that* $m \geq 4\frac{8c(16c+1)}{\epsilon^2} \ln\left(\frac{4|S|^2}{1-p}\right)$.

Proof. The proof of this theorem is left as an exercise to the reader. □

Theorem 4.2.10. *Let* $p, \epsilon \in (0,1)$ *and* Φ *as in Theorem 4.2.8. Choose* m *such that*

$$m \geq s\left(\frac{32c(16c+1)}{\epsilon^2}\right) \ln\left(\frac{\left(\frac{eN}{s}\right)\left(\frac{48}{\epsilon}\right)^2}{(1-p)^{1/s}}\right)$$

then $\frac{1}{\sqrt{m}}\Phi$ *will have the* (s,ϵ)*-RIP property (see Definition 3.5.1) with probability at least* p.

Proof. The proof of this theorem is left as an exercise to the reader. □

Homework 4.2.1 (Towards Φ being sparse)**.** 1. $f(x) = p\delta(x) + \frac{(1-p)^{3/2}}{\sqrt{2\pi}} \exp\left(-\frac{x^2(1-p)}{2}\right)$ for $p \in (0,1)$. Show that X with density f is mean 0, variance 1 and sub-gaussian with parameter $c = \frac{1}{2(1-p)}$

2. Let X have density

$$f(x) = p\delta(x) + \frac{(1-p)}{2}\left[\delta\left(x - \frac{1}{\sqrt{1-p}}\right) + \delta\left(x + \frac{1}{\sqrt{1-p}}\right)\right]$$

for $p \in (0,1)$. Show that X with density f is mean 0, variance 1 and sub-gaussian with parameter $c = \frac{1}{1-p}$

3. Consider Theorem 4.2.7 and how c has to scale in order to end up having fewer non-zero entries in Φ with the same probability decay.

Homework 4.2.2. Prove Theorem 4.2.8 (Hint: use union bound) and discuss why it also implies a proof of Theorem 3.1.3.

Homework 4.2.3. Prove that if $\Phi \in \mathbb{R}^{m\times N}$ is an ϵ-JL map of both $S \subset \mathbb{R}^N$ and $T \subset \mathbb{R}^N$ then Φ is an ϵ-JL map of any $R \subset \mathbb{C}^N$ with $\Re(R) \subset S$ and $\Im(R) \subset T$.

Homework 4.2.4. Prove Theorem 4.2.10. Hint: see Homework 3.2.2 and Theorem 3.3.1 as well as the following consequence of Sterling's approximation

$$\left(\frac{N}{s}\right)^s \leq \binom{N}{s} \leq \left(\frac{eN}{s}\right)^s, \forall N, s \in \mathbb{Z}^+, N \geq s$$

Homework 4.2.5 (optional)**.** Suppose X is a Radamacher random variable. Show that X allows a parameter $c = \frac{1}{2}$ (Definition 4.2.5), i.e. $\mathbb{E}\left[\exp\left(\theta X\right)\right] \leq \exp\left(\frac{\theta^2}{2}\right)$.

4.3 Bounded Orthonormal Systems and the RIP

Let $\mathcal{D} \subset \mathbb{R}^d$ and ν be a probability measure on $\mathcal{D}$.

Let $\mathcal{B} = \{\phi_1, \ldots, \phi_N\}$ be an orthonormal system of functions. That is $\phi_j : \mathcal{D} \to \mathbb{C}$, $\forall j \in [N]$ with

$$\int_{\mathcal{D}} \phi_j(t)\overline{\phi_k(t)}d\nu(t) = \delta_{jk}$$

Definition 4.3.1. We call an orthonormal system $\mathcal{B}$ a bounded orthonormal system (BOS) with constant K if

$$K = \max_{j\in[N]} \|\phi_j\|_\infty = \max_{j\in[N]} \sup_{\mathbf{t}\in\mathcal{D}} |\phi_j(\mathbf{t})| < \infty$$

Example 4.3.2 (Trigonometric Polynomials $K = 1$)**.** Let $\mathcal{D} = [0, 1]$ and let ν be the uniform (Lebesgue) measure on $[0, 1]$. For $\omega \in \mathbb{Z}$ define $\phi_\omega(t) = e^{2\pi i\omega t}$

Example 4.3.3 (Columns of Discrete Fourier Transform)**.** Let $\mathcal{D} = [N]$ and let ν be the counting measure on $[N]$. The DFT matrix is defined as

$$F_{n,k} = \frac{1}{\sqrt{N}} e^{-\frac{2\pi ink}{N}}, \forall k, n \in [N]$$

and so the discrete functions $\phi_k(n) = \sqrt{N}F_{t,k}$ are the normalized columns of F

Example 4.3.4 (Unitary Matrix)**.** Let $\mathcal{D} = [N]$ and let ν be the counting measure on $[N]$. Then any unitary matrix U, where we use the columns to form discrete

functions is a BONS where

$$K = \max_{n,k\in[N]} \left|\sqrt{N}U_{n,k}\right|$$

Example 4.3.5 (Chebyshev Polynomials of the First Kind)**.** Consider the following polynomials

$$T_0(x) = 1$$

$$T_1(x) = x$$

$$T_2(x) = 2x^2 - 1$$

$$\vdots$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

Let $\mathcal{D} = [-1, 1]$ and for $A \subset \mathcal{D}$ let

$$\nu(A) = \frac{1}{\pi}\int_A \frac{dx}{\sqrt{1-x^2}}$$

ν be the measure. The set $\mathcal{B} = \left\{\tilde{T}_0(x) = 1, \tilde{T}_1(x) = \sqrt{2}x, \ldots, \tilde{T}_n(x) = \sqrt{2}T_n\right\}$ is a bounded orthonormal system with respect to ν where $K = \max_{j\in[N]} \max_{x\in[-1,1]} \left|\tilde{T}_j\right| = \sqrt{2}$.

Another equivalent definition for the polynomials is $T_j(x) = \cos(j \arccos(x))$

We conclude this section with a theorem that describes how to construct an RIP matrix by sampling a BOS.

Theorem 4.3.6. *Let $\Phi \in \mathbb{C}^{m\times N}$ be a matrix formed by sampling m points $\mathbf{t}_1, \ldots, \mathbf{t}_m \in \mathcal{D}$ independently according to ν for the BOS $\mathcal{B} = \{\phi_1, \ldots, \phi_N\}$ and setting $\Phi_{\ell,k} = \phi_k(\mathbf{t}_\ell)$, $\forall \ell \in [m], k \in [N]$.*

If for $\epsilon \in (0, 1)$ we have

$$m \geq \frac{cK^2}{\epsilon^2} s \ln^4 N$$

then with probability at least $1 - N^{-\ln^3 N}$ *the matrix* $\frac{1}{\sqrt{m}}\Phi$ *with have the RIP of order* (s, ϵ). *Here* $c > 0$ *is an absolute constant independent of* s, K, ϵ, N

See proof of Theorem 12.32 in [12].

Homework 4.3.1. Prove that $K \geq 1$ must always hold for any BOS. Give an example of an orthonormal system for which $K = \infty$ i.e. that isn't bounded

Homework 4.3.2. Verify that $\{\phi_\omega(t)\}_{\omega \in \mathbb{Z}}$ is a BONS with $K = 1$.

Homework 4.3.3. Verify that $\{\phi_k(n)\}_{k \in [N]}$ is a BONS with $K = 1$.

4.4 Interpolation, Function Approximation from Randomly Sampled Data

We now proceed to show an application of Theorem 4.3.6. Our project will be to find a suitable approximation to a function which lies (nearly) in the span of some BOS of interest. We introduce some notation

Given our domain $\mathcal{D} \subseteq \mathbb{R}^D$ Suppose we have a function $f : \mathcal{D} \to \mathbb{C}$ where

$$f(\mathbf{t}) = \sum_{j=1}^{N} x_j \phi_j(\mathbf{t}) + \epsilon(\mathbf{t})$$

for $\mathbf{t} \in \mathcal{D}$. We will imagine that the function ϵ is some relatively small function with respect to some yet to be determined norm, and so we see that f lies nearly in the span of the BOS $\mathcal{B} = \{\phi_1, \ldots, \phi_N\}$.

Suppose we have samples $f(\mathbf{t}_1), \ldots, f(\mathbf{t}_m)$ for points $\mathbf{t}_1, \ldots, \mathbf{t}_m$ randomly sampled according to ν

Our goal then is to find $\tilde{f}$ based on the samples such that $\|f - \tilde{f}\|$ is small. We will argue that if $m \ll N$ then you can solve with instance optimal guarantees with s on the order $m/\log(N/s)$.

Consider the following system of equations

$$\begin{pmatrix} f(\mathbf{t}_1) \\ f(\mathbf{t}_2) \\ \vdots \\ f(\mathbf{t}_m) \end{pmatrix} = \begin{pmatrix} \phi_1(\mathbf{t}_1) & \phi_2(\mathbf{t}_1) & \dots & \phi_N(\mathbf{t}_1) \\ \phi_1(\mathbf{t}_2) & \phi_2(\mathbf{t}_2) & \dots & \phi_N(\mathbf{t}_2) \\ \vdots & & \ddots & \vdots \\ \phi_1(\mathbf{t}_m) & \phi_2(\mathbf{t}_m) & \dots & \phi_N(\mathbf{t}_m) \end{pmatrix} \begin{pmatrix} | \\ | \\ \mathbf{x} \\ | \end{pmatrix} + \begin{pmatrix} \epsilon(\mathbf{t}_1) \\ \epsilon(\mathbf{t}_2) \\ \vdots \\ \epsilon(\mathbf{t}_m) \end{pmatrix}$$

In this problem, our left hand side is given, the matrix Φ has the RIP property according to Theorem 4.3.6 and ℓ^1 minimization will allow us to approximate a sparse $\mathbf{x}$ which we then use directly to write $\tilde{f}$ in terms of the given BOS, $\tilde{f}(\mathbf{t}) = \sum_{j=1}^{N} = \tilde{x}_j \phi_j(\mathbf{t})$.

Theorem 4.4.1. *Choose $\delta > 0$ and $\eta > 0$. Let $U \in \mathbb{R}^{N \times N}$ be an orthonormal matrix obeying $U^*U = I$ and $\max_{i,j} |F_{i,j}| \leq K/\sqrt{N}$. Define a random sampling matrix $H \in \mathbb{R}^{m \times N}$ with rows chosen i.i.d. uniformly at random from the rows of U and set $\Phi = \frac{1}{\sqrt{m}} H$. Then with probability at least $1 - e^{-\eta}$*

$$\left| \|\Phi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \max\left(\delta, \delta^2\right) \|\mathbf{x}\|_2^2$$

will hold $\forall \mathbf{x} \in K_s$ provided

$$m \geq cK^2 \left\lceil \frac{s}{2\delta^2} \right\rceil \left(\ln^4 N + \eta\right)$$

where c is a universal constant.

Note that when $\delta < 1$, we have Theorem 12.32 from [12].

Lemma 4.4.2. *Let $\delta, s \in [1, \infty)$ be such that $\Phi \in \mathbb{C}^{m \times N}$ has the RIP of order $\left(2\left\lceil \frac{s}{\delta} \right\rceil, \frac{1}{2}\right)$. Then, Φ will also satisfy*

$$\max_{S \subset [N], |S| \leq s} \left\| \Phi_S^* \Phi_S - I \right\|_{2 \to 2} = \sup_{\mathbf{x} \in K_s \setminus \{0\}} \frac{\left| \|\Phi \mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right|}{\|\mathbf{x}\|_2^2} \leq \frac{\delta}{2} \tag{4.2}$$

Proof. Choose $\mathbf{x} \in K_s \setminus \{\mathbf{0}\}$.

$$T = \text{supp}(\mathbf{x}) = \{j_1, \ldots, j_s\} \subset [N]$$

Let $d = \gcd\left(s, \left\lceil \frac{s}{\delta} \right\rceil\right)$. Denote $n = \frac{s}{d}$.

Define a cover of T by $S_1, S_2, \ldots S_n \subset T$ where

$$S_\ell = \left\{ j_{(\ell-1)d+1 \mod s}, \ldots, j_{(\ell-1)d+\left\lceil \frac{s}{\delta} \right\rceil \mod s} \right\}$$

Note that each $j_\ell \in T$ will belong to $k = \frac{\left\lceil \frac{s}{\delta} \right\rceil}{d}$ sets (So for example, if $s = 21, \delta = 4$, then we would have $n = 7$ sets length 6 each. Any element from the original set T of size $s = 21$ would appear in precisely 2 sets S_ℓ)

So $\mathbf{x} = \frac{1}{k} \sum_{\ell=1}^{n} \mathbf{x}_{S_\ell}$ and $\|\mathbf{x}\|_2^2 = \frac{1}{k} \sum_{j=1}^{n} \|\mathbf{x}_{S_\ell}\|_2^2$ Now consider the quantity

$$
\begin{aligned}
\left|\|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| &= |\langle\Phi\mathbf{x},\Phi\mathbf{x}\rangle - \langle\mathbf{x},\mathbf{x}\rangle| \\
&= \left|\langle\Phi\left(\frac{1}{k}\sum_{j=1}^{n}\mathbf{x}_{S_j}\right),\Phi\left(\frac{1}{k}\sum_{i=1}^{n}\mathbf{x}_{S_i}\right)\rangle - \langle\left(\frac{1}{k}\sum_{j=1}^{n}\mathbf{x}_{S_j}\right),\left(\frac{1}{k}\sum_{i=1}^{n}\mathbf{x}_{S_i}\right)\rangle\right| \\
&= \frac{1}{k^2}\left|\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\langle\Phi_{S_j\cup S_i}\mathbf{x}_{S_j},\Phi_{S_i\cup S_j}\mathbf{x}_{S_i}\rangle - \langle\mathbf{x}_{S_j},\mathbf{x}_{S_i}\rangle\right)\right| \\
&= \frac{1}{k^2}\left|\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\langle\Phi^*_{S_j\cup S_i}\Phi_{S_j\cup S_i}\mathbf{x}_{S_j},\mathbf{x}_{S_i}\rangle - \langle\mathbf{x}_{S_j},\mathbf{x}_{S_i}\rangle\right)\right| \\
&= \frac{1}{k^2}\left|\sum_{j=1}^{n}\sum_{i=1}^{n}\langle\left(\Phi^*_{S_j\cup S_i}\Phi_{S_j\cup S_i} - I\right)\mathbf{x}_{S_j},\mathbf{x}_{S_i}\rangle\right| \\
&\leq \frac{1}{k^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left|\langle\left(\Phi^*_{S_j\cup S_i}\Phi_{S_j\cup S_i} - I\right)\mathbf{x}_{S_j},\mathbf{x}_{S_i}\rangle\right| \\
&\leq \frac{1}{k^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\|\left(\Phi^*_{S_j\cup S_i}\Phi_{S_j\cup S_i} - I\right)\mathbf{x}_{S_j}\|_2\|\mathbf{x}_{S_i}\|_2 \\
&\leq \frac{1}{k^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\|\Phi^*_{S_j\cup S_i}\Phi_{S_j\cup S_i} - I\|_{2\to 2}\|\mathbf{x}_{S_j}\|_2\|\mathbf{x}_{S_i}\|_2 \\
&\leq \frac{1}{k^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{2}\|\mathbf{x}_{S_j}\|_2\|\mathbf{x}_{S_i}\|_2 \\
&\leq \frac{1}{4k^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\|\mathbf{x}_{S_j}\|_2^2 + \|\mathbf{x}_{S_i}\|_2^2\right) \\
&= \frac{1}{4k^2}\left(\sum_{j=1}^{n}n\|\mathbf{x}_{S_j}\|_2^2 + \sum_{i=1}^{n}n\|\mathbf{x}_{S_i}\|_2^2\right) \\
&= \frac{kn}{2k^2}\|\mathbf{x}\|_2^2 \\
&= \frac{d}{2\left\lceil\frac{s}{\delta}\right\rceil}\frac{s}{d}\|\mathbf{x}\|_2^2 \\
&\leq \frac{s}{2\left(\frac{s}{\delta}\right)}\|\mathbf{x}\|_2^2 \\
&= \frac{\delta}{2}\|\mathbf{x}\|_2^2
\end{aligned}
$$

Rearranging terms, and taking the supremum then produces the desired bound.

Equality between the quotient and operator norm is given in Lemma 3.5.6. □

Lemma 4.4.3. *Suppose that* $\Phi \in \mathbb{C}^{m\times N}$ *has the RIP of order* (s,ϵ)*. Then* $\forall \mathbf{x} \in \mathbb{C}^N$

1. $\|\Phi\mathbf{x}\|_2 \leq \sqrt{1+\epsilon}\left[\frac{\|\mathbf{x}\|_1}{\sqrt{s}} + \|\mathbf{x}\|_2\right]$
2. $\|\Phi\|_{2\to 2} = \sigma_1(\Phi) \leq \sqrt{1+\epsilon}\left(\sqrt{\frac{N}{s}}+1\right)$

Proof. Suppose $\{j_k\}_{k=0}^{N-1}$ is an ordering of the entries of $\mathbf{x}$ in descending magnitude, i.e. $|x_{j_k}| \geq |x_{j_{k+1}}|$. Now partition the indices $\{j_k\}_{k=1}^{N}$ into $\lfloor \frac{N}{s} \rfloor + 1$ blocks of size at most s, e.g. $S_0 = \{j_0,\ldots,j_{s-1}\}, S_1 = j_s,\ldots,j_{2s-1},\ldots$ Thus

$$\begin{aligned}
\|\Phi\mathbf{x}\|_2 &= \sqrt{\sum_{\ell=0}^{\lfloor \frac{N}{s} \rfloor} \|\Phi_{S_\ell}\mathbf{x}_{S_\ell}\|_2^2} \\
&\leq \sum_{\ell=0}^{\lfloor \frac{N}{s} \rfloor} \|\Phi_{S_\ell}\mathbf{x}_{S_\ell}\|_2 \\
&\leq \sqrt{1+\epsilon}\sum_{\ell=0}^{\lfloor \frac{N}{s} \rfloor} \|\mathbf{x}_{S_\ell}\|_2 && \Phi\text{has } (s,\epsilon)-\text{RIP} \\
&\leq \sqrt{1+\epsilon}\left(\|\mathbf{x}_{S_0}\|_2 + \frac{1}{\sqrt{s}}\sum_{\ell=0}^{\lfloor \frac{N}{s} \rfloor} \|\mathbf{x}_{S_\ell}\|_1\right) && \text{Lemma 3.6.1} \\
&\leq \sqrt{1+\epsilon}\left(\|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{s}}\right)
\end{aligned}$$

which corresponds to the first inequality. Now using Holder's inequality, we conclude that $\|\mathbf{x}\|_1 \leq \sqrt{N}\|\mathbf{x}\|_2$. Using this on the $\|\mathbf{x}\|_1$ term from the above inequality, and noting that the bound holds for the supremum over all $\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$, we obtain our second inequality. □

Next we study a theorem wish will allow us to construct JL maps from BOS based RIP matrices.

Theorem 4.4.4 (Krahmer-Ward)**.** *Let $S \subset \mathbb{R}^N$ have $|S| = M$. Suppose that $\Phi \in \mathbb{R}^{m\times N}$ has the RIP of order $(2s, \eta/4)$ for some $\eta, p \in (0,1)$ and $s \geq 16 \ln\left(\frac{4M}{p}\right)$. Let $\psi \in \{-1,1\}^N$ have i.i.d. uniform Radamacher entries. Then*

$$(1-\eta)\left\|\mathbf{x}\right\|_2^2 \leq \left\|\Phi Diag(\psi)\mathbf{x}\right\|_2^2 \leq (1+\eta)\left\|\mathbf{x}\right\|_2^2$$

$\forall \mathbf{x} \in S$ with probability at least $1-p$

Proof. Let $\mathbf{x} \in S$. We may assume without loss of generality by a scaling argument that $\|\mathbf{x}\|_2 = 1$.

Let $k = \left\lceil \frac{N}{s} \right\rceil$. Order the indices $[N]$ based on descending magnitude of the entries of $\mathbf{x}$ and then partition these indices into the minimal number of disjoint blocks $S_1, \ldots, S_k$ of at most size s. So $S_1 \subset [N]$ contains the s largest entries of $\mathbf{x}$ and so on.

Denote the diagonal matrix which has entries along its diagonal equal to the entries of the vector $\boldsymbol{\psi}$ as $D_{\boldsymbol{\psi}}$. For our choice of $\boldsymbol{\psi}$ then $D_{\boldsymbol{\psi}} \in \{-1,0,1\}^{N\times N}$. Now we consider the following decomposition of the norm

$$
\begin{aligned}
\|\Phi D_{\boldsymbol{\psi}}\mathbf{x}\|_2^2 &= \langle \Phi D_{\boldsymbol{\psi}}\left(\mathbf{x}_{S_1}+\mathbf{x}_{\bar{S}_1}\right), \Phi D_{\boldsymbol{\psi}}\left(\mathbf{x}_{S_1}+\mathbf{x}_{\bar{S}_1}\right)\rangle \\
&= \langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}\rangle + 2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle + \langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle \\
&= \|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}\|_2^2 + 2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle + \langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle \\
&= \|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}\|_2^2 + 2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle + \langle \Phi D_{\boldsymbol{\psi}}\left(\sum_{j=2}^{k}\mathbf{x}_{S_j}\right), \Phi D_{\boldsymbol{\psi}}\left(\sum_{i=2}^{k}\mathbf{x}_{S_i}\right)\rangle \\
&= \|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}\|_2^2 + 2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle + \sum_{j=2}^{k}\sum_{i=2}^{k}\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_i}\rangle \\
&= \|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}\|_2^2 + 2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle + \sum_{j=2}^{k}\sum_{i=2i\neq j}^{k}\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_i}\rangle + \sum_{j=2}^{k}\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}, \Phi D \\
&= \underbrace{\sum_{j=1}^{k}\|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}\|_2^2}_{\text{Term I}} + \underbrace{2\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle}_{\text{Term II}} + \underbrace{\sum_{j=2}^{k}\sum_{i=2i\neq j}^{k}\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_i}\rangle}_{\text{Term III}}
\end{aligned}
$$

We will proceed then to bound each of the three terms.

Term (I): Note that $D_{\boldsymbol{\psi}}$ is unitary and also does not change the sparsity of $\mathbf{x}_{S_j}$. Since Φ has the RIP of order $(2s, \eta/4)$ we have

$$
\sum_{j=1}^{k}\|\Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}\|_2^2 \leq \left(1+\frac{\eta}{4}\right)\sum_{j=1}^{k}\|D_{\boldsymbol{\psi}}\mathbf{x}_{S_j}\|_2^2 = \left(1+\frac{\eta}{4}\right)\sum_{j=1}^{k}\|\mathbf{x}_{S_j}\|_2^2 = \left(1+\frac{\eta}{4}\right)\|\mathbf{x}\|_2^2
$$

The lower bound is obtained in an analogous way, and so we have

$$
\left(1-\frac{\eta}{4}\right)\|\mathbf{x}\|_2^2 \leq \text{Term I} \leq \left(1+\frac{\eta}{4}\right)\|\mathbf{x}\|_2^2
$$

Term (II): Note that since $D_{\boldsymbol{\psi}}$ is a diagonal matrix, $D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1} = D_{\mathbf{x}_{\bar{S}_1}}\boldsymbol{\psi}$

$$
\begin{aligned}
\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle &= \langle \Phi_{S_1} D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\mathbf{x}_{S_1}, \Phi_{\bar{S}_1} D_{\mathbf{x}_{\bar{S}_1}}\boldsymbol{\psi}_{\bar{S}_1}\rangle \\
&= \langle D_{\mathbf{x}_{\bar{S}_1}}\Phi_{\bar{S}_1}^{*}\Phi_{S_1} D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\mathbf{x}_{S_1}, \boldsymbol{\psi}_{\bar{S}_1}\rangle
\end{aligned}
$$

Note that the left vector in the inner product is deterministic save for the term $D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}$ whereas the vector on the right depends on $\boldsymbol{\psi}_{\bar{S}_1}$; since the entries of $\boldsymbol{psi}$ are i.i.d. and the support of $\boldsymbol{\psi}_{\bar{S}_1}$ and $\boldsymbol{\psi}_{S_1}$ are of course disjoint, we have that the two vectors are independent of each other. Denote $\mathbf{a} = D_{\mathbf{x}_{\bar{S}_1}} \Phi^*_{\bar{S}_1} \Phi_{S_1} D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}} \mathbf{x}_{S_1}$, the inner product then becomes

$$\langle a, \boldsymbol{\psi}_{\bar{S}_1} \rangle = \sum_{j \in \bar{S}_1} a_j \psi_j$$

Note that by independence $\mathbb{E}[a_j \psi_j] = \mathbb{E}[a_j]\mathbb{E}[\psi_j]$ and furthermore that $\mathbb{E}[\psi_j] = 0$, since ψ_j is a Radamacher random variable. And since $|a_j \psi_j| = |a_j|$ if we can argue that a_j is bounded, then we will have satisfied the hypothesis of Hoeffdings' inequality, Theorem 4.1.2. Note that $\mathbf{a} \in \ell^2$ implies that a_j is bounded. Note that for a generic diagonal matrix D, and compatible generic matrix A that $\|DA\|_{2\to 2} \leq \|D\|_\infty \|A\|_{2\to 2}$. Of interest to us is

$$
\begin{aligned}
\|D_{\mathbf{x}_{\bar{S}_1}}\Phi^*_{\bar{S}_1}\Phi_{S_1}D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\mathbf{x}_{S_1}\|_2 &= \frac{\|\mathbf{a}\|_2^2}{\|\mathbf{a}\|_2} \\
&= \frac{\langle \mathbf{a}, \mathbf{a}\rangle}{\|\mathbf{a}\|_2} \\
&= \langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \mathbf{a}\rangle \\
&= \sup_{\|\mathbf{z}\|_2=1} \langle \mathbf{z}, \mathbf{a}\rangle \\
&= \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \langle \mathbf{z}_{S_j}, D_{\mathbf{x}_{\bar{S}_j}}\Phi^*_{\bar{S}_j}\Phi_{S_1}D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\mathbf{x}_{S_1}\rangle \\
&\leq \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2\|\mathbf{x}_{S_1}\|_2\|D_{\mathbf{x}_{\bar{S}_j}}\Phi^*_{\bar{S}_j}\Phi_{S_1}D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\|_{2\to 2} \\
&\leq \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2\|\mathbf{x}_{S_1}\|_2\|D_{\mathbf{x}_{\bar{S}_j}}\|_\infty\|\Phi^*_{\bar{S}_j}\Phi_{S_1}\|_{2\to 2}\|D_{\boldsymbol{\psi}_{\boldsymbol{S_1}}}\|_\infty \\
&\leq \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2\|\mathbf{x}_{S_1}\|_2\|\mathbf{x}_{\bar{S}_j}\|_\infty\|\Phi^*_{\bar{S}_j}\Phi_{S_1}\|_{2\to 2} \\
&\leq \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2 (1) \left(\frac{\|\mathbf{x}_{j-1}\|_2}{\sqrt{s}}\right)\left(\frac{\eta}{4}\right) \\
&\leq \frac{\eta}{4\sqrt{s}} \sup_{\|\mathbf{z}\|_2=1} \sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2\|\mathbf{x}_{j-1}\|_2 \\
&\leq \frac{\eta}{4\sqrt{s}} \sup_{\|\mathbf{z}\|_2=1} \frac{1}{2}\sum_{j\geq 2} \|\mathbf{z}_{S_j}\|_2^2 + \|\mathbf{x}_{j-1}\|_2^2 \\
&\leq \frac{\eta}{4\sqrt{s}}
\end{aligned}
$$

So the sum of mean zero, bounded random variables in Term II, $\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle$, satisfies Hoeffdings' inequality and thus

$$
P\left[|\langle \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{S_1}, \Phi D_{\boldsymbol{\psi}}\mathbf{x}_{\bar{S}_1}\rangle| \geq t\right] \leq 2\exp\left(\frac{-8st^2}{\eta^2}\right) \tag{4.3}
$$

Term (III): Denote $B = D_{\mathbf{x}_{S_j}} \Phi_{S_j}^* \Phi_{S_i} D_{\mathbf{x}_{S_i}}$

$$\sum_{j=2}^{k} \sum_{\substack{i=2 \\ i \neq j}}^{k} \langle \Phi D_{\boldsymbol{\psi}} \mathbf{x}_{S_j}, \Phi D_{\boldsymbol{\psi}} \mathbf{x}_{S_i} \rangle = \sum_{j=2}^{k} \sum_{\substack{i=2 \\ i \neq j}}^{k} \langle \boldsymbol{\psi}, D_{\mathbf{x}_{S_j}} \Phi_{S_j}^* \Phi_{S_i} D_{\mathbf{x}_{S_i}} \boldsymbol{\psi} \rangle$$

$$= \boldsymbol{\psi}^* B \boldsymbol{\psi}$$

Observe that

$$(B)_{k,\ell} = \begin{cases} x_k \Phi_k^* \Phi_\ell x_\ell & \text{if } k, \ell \in [N] \setminus S_1, k \neq \ell \\ 0 & \text{otherwise} \end{cases}$$

And so B is a symmetric matrix with zeros along its axis. In order to apply Lemma 4.4.5, which is stated and proved below, we need to show that the operator norm and Frobenius norm are bounded. Consider then the operator norm of B. We bound it by use of Cauchy-Schwarz, operator norm bounds and bounds similar to those seen in the Term II bound.

$$
\begin{aligned}
\|B\|_{2\to 2} &= \sup_{\|\mathbf{z}\|_2=1} \langle \mathbf{z}, B\mathbf{z}\rangle \\
&\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \langle \mathbf{z}_j, D_{\mathbf{x}_{S_j}}\Phi_{S_j}^*\Phi_{S_\ell} D_{\mathbf{x}_{S_\ell}}\mathbf{z}_\ell\rangle \\
&\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \|\mathbf{z}_j\|_2 \|D_{\mathbf{x}_{S_j}}\|\Phi_{S_j}^*\Phi_{S_\ell} D_{\mathbf{x}_{S_\ell}}\mathbf{z}_\ell\|_2 \\
&\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \|\mathbf{z}_j\|_2 \|\mathbf{z}_\ell\|_2 \|\mathbf{x}_{S_j}\|_\infty \|\Phi_{S_j}^*\Phi_{S_\ell}\|_{2\to 2}\|\mathbf{x}_{S_\ell}\|_\infty \\
&\le \sup_{\|\mathbf{z}\|_2\le 1} \sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \|\mathbf{z}_j\|_2 \|\mathbf{z}_\ell\|_2 \frac{\|\mathbf{x}_{S_{j-1}}\|_2}{\sqrt{s}}\frac{\eta}{4}\frac{\|\mathbf{x}_{S_{\ell-1}}\|_2}{\sqrt{s}} \\
&= \frac{\eta}{4s}\sup_{\|\mathbf{z}\|_2\le 1} \sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \left(\|\mathbf{z}_j\|_2\|\mathbf{x}_{S_{j-1}}\|_2\right)\left(\|\mathbf{z}_\ell\|_2\|\mathbf{x}_{S_{\ell-1}}\|_2\right) \\
&\le \frac{\eta}{4s}\sup_{\|\mathbf{z}\|_2\le 1} \frac{1}{4}\sum_{j=2}^{k}\sum_{\substack{\ell=2\\ \ell\neq j}}^{k} \left(\|\mathbf{z}_j\|_2^2 + \|\mathbf{x}_{S_{j-1}}\|_2^2\right)\left(\|\mathbf{z}_\ell\|_2^2 + \|\mathbf{x}_{S_{\ell-1}}\|_2^2\right) \\
&\le \frac{\eta}{4s}\sup_{\|\mathbf{z}\|_2\le 1} \frac{1}{2}\left(\|\mathbf{z}\|_2^2 + \|\mathbf{x}\|_2^2\right) \\
&= \frac{\eta}{4s}
\end{aligned}
$$

Using Homework **??**, $\|\mathbf{x}_{S_t}\|_\infty \le \|\mathbf{x}_{S_{t-1}}\|_2/\sqrt{s}$:

$$
\begin{aligned}
\|B\|_F &= \sum_{j=2}^{k}\sum_{\substack{t=2\\ t\neq j}}^{k}\sum_{i\in S_j}\sum_{\ell\in S_t}\left(x_i\Phi_i^*\Phi_\ell x_\ell\right)^2 \\
&\le \sum_{j=2}^{k}\sum_{\substack{t=2\\ t\neq j}}^{k}\sum_{i\in S_j} x_i^2\|D_{\mathbf{x}_{S_t}}\Phi_{S_t}^*\Phi_i\|_2^2 \\
&\le \sum_{j=2}^{k}\sum_{\substack{t=2\\ t\neq j}}^{k}\sum_{i\in S_j} x_i^2\|\mathbf{x}_{S_t}\|_\infty\|\Phi_{S_t}^*\Phi_i x_\ell\|_2^2 \quad \le \left(\frac{\eta}{4}\right)^2\sum_{j=2}^{k}\|\mathbf{x}_{S_j}\|_2^2\sum_{\substack{t=2\\ t\neq j}}^{k}\frac{\|\mathbf{x}_{S_{t-1}}\|_2^2}{s} \le \frac{\eta^2}{16s}
\end{aligned}
$$

So both the operator and Frobenius norm of B is bounded, and we can apply Lemma 4.4.5 to conclude that

$$P\left[|\textbf{Term III}| \geq r\right] \leq 2\exp\left(-2\min\left\{\frac{3r^2}{8\eta^2}, \frac{r}{8\eta}\right\}\right) \tag{4.4}$$

Setting $t = \eta/8$ and $r = \eta/2$ in Equations 4.3 and 4.4 we have that

$$\left(1 - \frac{7}{8}\eta\right)\|\mathbf{x}\|_2^2 \leq \|\Phi\text{Diag}(\psi)\mathbf{x}\|_2^2 \leq \left(1 + \frac{7}{8}\eta\right)\|\mathbf{x}\|_2^2 \tag{4.5}$$

fails to hold with at most probability

$$2\left[\exp\left(\frac{-s}{8}+\right) + \exp\left(\frac{-s}{16}+\right)\right]$$

by the union bound over the events $|\text{Term II}| \geq t$ and $|\text{Term III}| \geq r$. Note that using $s \geq 16\ln\left(\frac{4M}{p}\right)$ we can conclude that

$$2\left[\exp\left(\frac{-s}{8}+\right) + \exp\left(\frac{-s}{16}+\right)\right] \leq \frac{p}{M}$$

□

Union bounding over all M points in S then, we have that the norm inequality 4.5 fails with at most probability p for any $\mathbf{x} \in S$and so taking the complement we have that the inequality holds for $\forall \mathbf{x} \in S$ with probability at least $1 - p$.

Lemma 4.4.5 (Radamacher Chaos)**.** *Let $B \in \mathbb{R}^{N\times N}$ be symmetric with zeros on its diagonal, and let $\boldsymbol{\psi} \in \mathbb{R}^N$ be a Radamacher random vector. Then for $t > 0$*

$$\begin{aligned} P\left[|\boldsymbol{\psi}^* B\boldsymbol{\psi}| \geq t\right] &\leq 2\exp\left(-\min\left\{\frac{3t^2}{128\|B\|_F^2}, \frac{t}{32\|B\|_{2\to 2}}\right\}\right) \\ &= \begin{cases} 2\exp\left(\frac{-3t^2}{128\|B\|_F^2}\right) & 0 < t \leq \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}} \\ 2\exp\left(\frac{-t^2}{32\|B\|_{2\to 2}}\right) & t > \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}} \end{cases} \end{aligned}$$

Proof. We estimate $\mathbb{E}\left[\exp\left(\theta\boldsymbol{\psi}^* B\boldsymbol{\psi}\right)\right]$, the moment generating function of the random variable $\boldsymbol{\psi}^* B\boldsymbol{\psi}$.

$$\mathbb{E}\left[\exp\left(\theta\boldsymbol{\psi}^* B\boldsymbol{\psi}\right)\right] \leq \mathbb{E}\left[\exp\left(4\theta\boldsymbol{\psi}^* B\boldsymbol{\psi}'\right)\right]$$

where we have applied Lemma 4.4.6

$$= \mathbb{E}_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{\psi'}} \left[\exp\left(4\theta \sum_{k=1}^{N} \psi'_k \left(\sum_{j=1}^{N} \psi_j B_{jk} \right) \right) \right]$$

denote $\mathbf{a}$, where $a_k = \sum_{j=1}^{N} \psi_j B_{jk}$. So $\sum_{k=1}^{N} \psi'_k \left(\sum_{j=1}^{N} \psi_j B_{jk} \right) = \langle \boldsymbol{\psi'}, \mathbf{a} \rangle$.

$$= \mathbb{E}_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{\psi'}} \left[\exp\left(4\theta \langle \boldsymbol{\psi'}, \mathbf{a} \rangle \right) \right]$$

By Homework 4.2.5, the entries of ψ' are subguassian random variables which allow parameter $1/2$. So by Theorem 4.2.4, $4\langle \boldsymbol{\psi'}, \mathbf{a} \rangle$ is itself a subguassian random variable which allows the parameter $\frac{16\|\mathbf{a}\|_2^2}{2}$. That is $\mathbb{E}_{\boldsymbol{\psi'}} \left[\exp\left(\theta 4 \langle \boldsymbol{\psi'}, \mathbf{a} \rangle \right) \right] \leq \exp\left(\theta^2 \frac{16\|\mathbf{a}\|_2^2}{2} \right)$ by definition of subguassian random variable (see Theorem 4.2.3 and Definition 4.2.5). Thus we can eliminate the inner expectation over $\boldsymbol{\psi'}$ using the subgaussian bound

$$\leq \mathbb{E}_{\boldsymbol{\psi}} \left[\exp\left(\theta^2 \frac{16\|\mathbf{a}\|_2^2}{2} \right) \right]$$
$$= \mathbb{E}_{\boldsymbol{\psi}} \left[\exp\left(8\theta^2 \sum_{k=1}^{N} \left(\sum_{j=1}^{N} \psi_j B_{jk} \right)^2 \right) \right]$$

Note that $\sum_{k=1}^{N} \left(\sum_{j=1}^{N} \psi_j B_{jk} \right)^2 = \langle B\boldsymbol{\psi}, B\boldsymbol{\psi} \rangle$, where we have used symmetry of B. Furthermore $\langle B\boldsymbol{\psi}, B\boldsymbol{\psi} \rangle = (B\boldsymbol{\psi})^*(B\boldsymbol{\psi}) = \boldsymbol{\psi}^* B^* B \boldsymbol{\psi} = \boldsymbol{\psi}^* B^2 \boldsymbol{\psi}$. So we can rewrite the above inequality as

$$\mathbb{E}\left[\exp\left(\theta \boldsymbol{\psi}^* B \boldsymbol{\psi} \right) \right] \leq \mathbb{E}\left[\exp\left(8\theta^2 \boldsymbol{\psi}^* B^2 \boldsymbol{\psi} \right) \right] \tag{4.6}$$

Consider the following

$$\begin{aligned} \boldsymbol{\psi}^* B^2 \boldsymbol{\psi} &= \boldsymbol{\psi}^* \left(B^2 - \mathrm{Diag}(B^2) + \mathrm{Diag}(B^2) \right) \boldsymbol{\psi} \\ &= \boldsymbol{\psi}^* \left(\mathrm{Diag}(B^2) \right) \boldsymbol{\psi} + \boldsymbol{\psi}^* \left(B^2 - \mathrm{Diag}(B^2) + \right) \boldsymbol{\psi} \\ &= \mathrm{Trace}(B^2) + \boldsymbol{\psi}^* \left(B^2 - \mathrm{Diag}(B^2) \right) \boldsymbol{\psi} \\ &= \|B\|_F^2 + \boldsymbol{\psi}^* \left(B^2 - \mathrm{Diag}(B^2) \right) \boldsymbol{\psi} \end{aligned}$$

Observe that $\boldsymbol{\psi}^* \left(\text{Diag}(B^2)\right) \boldsymbol{\psi} = \text{Trace}(B^2)$ since it is equivalent to $\text{diag}(B^2)^*(\boldsymbol{\psi} * \boldsymbol{\psi})$ where $\text{diag}(B^2)$ is the vector formed by the diagonal entries of B^2, and $*$ denotes entry-wise multiplication; thus $(\boldsymbol{\psi} * \boldsymbol{\psi}) = \{1\}^N$. Also, $\|B\|_F^2 = \sum_{\ell=1}^N \|\mathbf{b}_\ell\|_2^2 = \sum_{\ell=1}^N \langle \mathbf{b}_\ell, \mathbf{b}_\ell \rangle = \text{Trace}(B^2)$. So estimating the right hand side of inequality 4.6 with these facts,

$$\begin{aligned}\mathbb{E}\left[\exp\left(8\theta^2 \boldsymbol{\psi}^* B^2 \boldsymbol{\psi}\right)\right] &= \mathbb{E}\left[\exp\left(8\theta^2\left[\|B\|_F^2 + \boldsymbol{\psi}^*\left(B^2 - \text{Diag}(B^2)\right)\boldsymbol{\psi}\right]\right)\right] \\ &= \exp\left(8\theta^2\|B\|_F^2\right)\mathbb{E}\left[\exp\left(8\theta^2 \boldsymbol{\psi}^*\left(B^2 - \text{Diag}(B^2)\right)\boldsymbol{\psi}\right)\right]\end{aligned}$$

Now, using Lemma 4.4.6 where $\mathbf{y} = \text{diag}(B^2)$ we obtain

$$\leq \exp\left(8\theta^2\|B\|_F^2\right)\mathbb{E}\left[\exp\left(32\theta^2 \boldsymbol{\psi}^* B^2 \boldsymbol{\psi}'\right)\right]$$

Again, we will employ the subguassian bound - in this instance $\tilde{\mathbf{a}}$ is a vector with entries $\tilde{a}_k = \sum_{j=1}^N \psi_j (B^2)_{jk}$. So $\sum_{k=1}^N \psi'_k \left(\sum_{j=1}^N \psi_j (B^2)_{jk}\right) = \langle \boldsymbol{\psi'}, \tilde{\mathbf{a}} \rangle$. We note that $32\langle \boldsymbol{\psi'}, \tilde{\mathbf{a}} \rangle$ is subgaussian random variable which allows parameter $\frac{1024\|\tilde{\mathbf{a}}\|_2^2}{2}$ and that $\|\tilde{\mathbf{a}}\|_2^2 = \sum_{k=1}^N \left(\sum_{j=1}^N \psi_j (B^2)_{jk}\right)^2 = \boldsymbol{\psi}^* B^4 \boldsymbol{\psi}$. We obtain the bound

$$\leq \exp\left(8\theta^2\|B\|_F^2\right)\mathbb{E}\left[\exp\left(512\theta^4 \boldsymbol{\psi}^* B^4 \boldsymbol{\psi}'\right)\right]$$

The matrix B^4 and B^2 are both positive definite; we can use this along with the the operator norm to arrive at the following bound for the argument of the exponential function above:

$$\begin{aligned}0 < \boldsymbol{\psi}^* B^4 \boldsymbol{\psi} &= \langle B^2\boldsymbol{\psi}, B^2\boldsymbol{\psi}\rangle \\ &= \|B^2\boldsymbol{\psi}\|_2^2 \\ &= \|B^2\boldsymbol{\psi}\|_2 \|B^2\boldsymbol{\psi}\|_2 \\ &\leq \|B\|_{2\to 2}^2 \boldsymbol{\psi}^* B^2 \boldsymbol{\psi}\end{aligned}$$

So then noting that the argument is positive on the right hand side, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(8\theta^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)\right] &\leq \exp\left(8\theta^2\|B\|_F^2\right)\mathbb{E}\left[\exp\left(512\theta^4\|B\|_{2\to 2}^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)\right] \\ &= \exp\left(8\theta^2\|B\|_F^2\right)\mathbb{E}\left[\exp\left(8\theta^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)^{64\theta^2\|B\|_{2\to 2}^2}\right]\end{aligned}$$

Consider the function $g(y) = y^{64\theta^2\|B\|_{2\to 2}^2}$, if $64\theta^2\|B\|_{2\to 2}^2 < 1$ then $g(y)$ is concave and so by Jensen's inequality $\mathbb{E}[g(y)] \leq g(\mathbb{E}[y])$. So in particular, for $y = \exp\left(8\theta^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)$ we obtain

$$\leq \exp\left(8\theta^2\|B\|_F^2\right)\left(\mathbb{E}\left[\exp\left(8\theta^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)\right]\right)^{64\theta^2\|B\|_{2\to 2}^2}$$

Now after a rearrangement of terms and noting the bound in 4.6 we have

$$\mathbb{E}\left[\exp\left(\theta\boldsymbol{\psi}^*B\boldsymbol{\psi}\right)\right] \leq \mathbb{E}\left[\exp\left(8\theta^2\boldsymbol{\psi}^*B^2\boldsymbol{\psi}\right)\right] \leq \exp\left(\frac{8\theta^2\|B\|_F^2}{1-64\theta^2\|B\|_{2\to 2}^2}\right) \tag{4.7}$$

when $\theta < 1/8\|B\|_{2\to 2}$. We can now use the established inequality on the moment generating function to bound the probability as stated in the hypothesis by way of Markov's inequality:

$$\begin{aligned}P\left[\boldsymbol{\psi}^*B\boldsymbol{\psi} \geq t\right] &= P\left[\exp\left(\boldsymbol{\psi}^*B\boldsymbol{\psi}\right) \geq e^{\theta t}\right] \\ &\leq e^{-\theta t}\mathbb{E}\left[\exp\left(\theta\boldsymbol{\psi}^*B\boldsymbol{\psi}\right)\right] \\ &\leq \exp\left(-\theta t + \frac{8\theta^2\|B\|_F^2}{1-64\theta^2\|B\|_{2\to 2}^2}\right)\end{aligned}$$

Now in the event that $0 < t <\leq \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}}$ we set $\theta = (16\|B\|_{2\to 2})^{-1}$ and obtain from 4.7 $2\exp\left(\frac{-3t^2}{128\|B\|_F^2}\right)$. When $t > \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}}$ set $\theta = \frac{3t}{64\|B\|_F^2}$. We have then our desired result

$$P\left[|\boldsymbol{\psi}^*B\boldsymbol{\psi}| \geq t\right] \leq \begin{cases} & 0 < t \leq \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}} \\ 2\exp\left(\frac{-t^2}{32\|B\|_{2\to 2}}\right) & t > \frac{4}{3}\frac{\|B\|_F^2}{\|B\|_{2\to 2}}\end{cases}$$

□

Lemma 4.4.6 (Decoupling)**.** *Let $B \in \mathbb{R}^{N\times N}$ be a symmetric matrix with zeros on its diagonal. Let $\boldsymbol{\psi} \in \mathbb{R}^N$ be a vector of independent, mean zero random variables. If $f : \mathbb{R} \to \mathbb{R}$ is a convex function then*

$$E = \mathbb{E}\left[f\left(\boldsymbol{\psi}^* B\boldsymbol{\psi}\right)\right] \leq \mathbb{E}\left[f\left(4\boldsymbol{\psi}^*\left(B + Diag(\mathbf{y})\right)\boldsymbol{\psi}'\right)\right]$$

where $\boldsymbol{\psi}'$ denotes an i.i.d vector with the same distribution as $\boldsymbol{\psi}$ and $\mathbf{y} \in \mathbb{R}^N$ is arbitrary.

Proof. Let $\boldsymbol{\delta} \in [0,1]^N$ be a Bernouli random vector where the entries are i.i.d and equal to 1 or 0 with equal probability. Note that $\mathbb{E}\left[\delta_k(1-\delta_j)\right] = 1/4$ for $j \neq k$

$$\begin{aligned}
E &= \mathbb{E}\left[f\left(\boldsymbol{\psi}^* B\boldsymbol{\psi}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\psi}}\left[f\left(\sum_{\substack{k,j \\ k\neq j}}^{N} \psi_j B_{jk}\psi_k\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\psi}}\left[f\left(4\sum_{\substack{k,j \\ k\neq j}}^{N} \mathbb{E}_{\boldsymbol{\delta}}\left[\delta_k(1-\delta_j)\right]\psi_j B_{jk}\psi_k\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\psi}}\mathbb{E}_{\boldsymbol{\delta}}\left[f\left(4\sum_{\substack{k,j \\ k\neq j}}^{N} \delta_k(1-\delta_j)\psi_j B_{jk}\psi_k\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\delta}}\mathbb{E}_{\boldsymbol{\psi}}\left[f\left(4\sum_{\substack{k,j \\ k\neq j}}^{N} \delta_k(1-\delta_j)\psi_j B_{jk}\psi_k\right)\right]
\end{aligned}$$

where we have used Jensen's inequality and Fubini's Theorem. Denote the set $\sigma(\boldsymbol{\delta}) = \{j \in [N] | \delta_j = 1\}$, i.e. the set of indices where the Bernouli random vector is equal to one.

$$\leq \mathbb{E}_{\boldsymbol{\delta}}\mathbb{E}_{\boldsymbol{\psi}|_{\sigma(\boldsymbol{\delta})}}\mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma(\boldsymbol{\delta})}}}\left[f\left(4\sum_{j\in\sigma(\boldsymbol{\delta})}\sum_{k\in\overline{\sigma(\boldsymbol{\delta})}} \psi_j B_{jk}\psi_k\right)\right]$$

Notice that ψ_j and ψ_k are independent given $j \in \sigma(\boldsymbol{\delta})$ and $k \in \overline{\sigma(\boldsymbol{\delta})}$, so we are able to replace ψ_k with ψ'_k with no change in expectation.

$$\leq \mathbb{E}_{\boldsymbol{\delta}} \mathbb{E}_{\boldsymbol{\psi}|_{\sigma(\boldsymbol{\delta})}} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma(\boldsymbol{\delta})}}} \left[f\left(4 \sum_{j \in \sigma(\boldsymbol{\delta})} \sum_{k \in \overline{\sigma(\boldsymbol{\delta})}} \psi_j B_{jk} \psi'_k \right) \right]$$

Note that $E \leq E_{\boldsymbol{\delta}} \mathbb{E}_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{\psi}'} \left[f\left(4 \sum_{j \in \sigma} \sum_{k \in \bar{\sigma}} \psi'_k B_{kj} \psi_j \right) \right]$ implies that there exists at least one particular Bernouli vector δ^* such that the inequality holds, i.e. if not, then the inequality in expectation over δ could not hold. Let us denote $\sigma = \sigma(\delta^*)$; the indices of the entries of the now fixed vector $\boldsymbol{\delta^*}$ where the value is one.

$$\leq \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(4 \sum_{j \in \sigma} \sum_{k \in \bar{\sigma}} \psi_j B_{jk} \psi'_k \right) \right]$$

Since $\mathbb{E}_{\boldsymbol{\psi}'|_\sigma} [\psi'_k] = 0 = \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} [\psi_j]$ we can include additional summands in the following way

$$\leq \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(4 \sum_{j \in \sigma} \left\{ \sum_{k \in \bar{\sigma}} \psi_j B_{jk} \psi'_k + \sum_{k \in \sigma} \psi_j B_{jk} \mathbb{E}_{\boldsymbol{\psi}'|_\sigma} [\psi'_k] \right\} + 4 \sum_{j \in \overline{\sigma}} \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} [\psi_j] \sum_{k \in \overline{\sigma}} B_{jk} \psi'_k \right) \right]$$

Now denote $\tilde{B} = B + \operatorname{diag}(\mathbf{y})$; and use this matrix in the previously introduced summands to obtain

$$\leq \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(4 \sum_{j \in \sigma} \left\{ \sum_{k \in \bar{\sigma}} \psi_j B_{jk} \psi'_k + \sum_{k \in \sigma} \psi_j \tilde{B}_{jk} \mathbb{E}_{\boldsymbol{\psi}'|_\sigma} [\psi'_k] \right\} + 4 \sum_{j \in \overline{\sigma}} \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} [\psi_j] \sum_{k \in \overline{\sigma}} \tilde{B}_{jk} \psi'_k \right) \right]$$

Now by linearity of expectation we have

$$= \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(\mathbb{E}_{\boldsymbol{\psi}'|_\sigma} \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} \left[4 \sum_{j \in \sigma} \left\{ \sum_{k \in \bar{\sigma}} \psi_j B_{jk} \psi'_k + \sum_{k \in \sigma} \psi_j \tilde{B}_{jk} \psi'_k \right\} + 4 \sum_{j \in \overline{\sigma}} \psi_j \sum_{k \in \overline{\sigma}} \tilde{B}_{jk} \psi'_k \right] \right) \right]$$

$$= \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(\mathbb{E}_{\boldsymbol{\psi}'|_\sigma} \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} \left[4 \sum_{j \in \sigma} \sum_{k} \psi_j \tilde{B}_{jk} \psi'_k + 4 \sum_{j \in \overline{\sigma}} \sum_{k \in \overline{\sigma}} \psi_j \tilde{B}_{jk} \psi'_k \right] \right) \right]$$

$$= \mathbb{E}_{\boldsymbol{\psi}|_\sigma} \mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}} \left[f\left(\mathbb{E}_{\boldsymbol{\psi}'|_\sigma} \mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}} \left[4 \sum_{j,k} \psi_j \tilde{B}_{jk} \psi'_k \right] \right) \right]$$

Apply Jensen's inequality on the inner expectations to write

$$\leq \mathbb{E}_{\boldsymbol{\psi}|_\sigma}\mathbb{E}_{\boldsymbol{\psi}'|_{\overline{\sigma}}}\mathbb{E}_{\boldsymbol{\psi}'|_\sigma}\mathbb{E}_{\boldsymbol{\psi}|_{\overline{\sigma}}}\left[f\left(4\sum_{j,k}\psi_j\tilde{B}_{jk}\psi'_k\right)\right]$$

$$= \mathbb{E}_{\boldsymbol{\psi}}\mathbb{E}_{\boldsymbol{\psi}'}\left[f\left(4\sum_{j,k}\psi_j\tilde{B}_{jk}\psi'_k\right)\right]$$

rewriting the sum as vector-matrix multiplication and noting that we have the expectation over each random vector we have

$$= \mathbb{E}\left[f\left(4\boldsymbol{\psi}^*\tilde{B}\boldsymbol{\psi}'\right)\right]$$

□

Homework 4.4.1. Show that $\|\Phi^*_{S_t}\Phi_i\|_2^2 \leq \left(\frac{\eta}{4}\right)^2$ when $i \notin S_t$ and Φ has the RIP of order $(2s, \eta/4)$

Homework 4.4.2. Show that $B \in \mathbb{C}^{m\times N}$ is ϵ-JL map of $S \subset \mathbb{R}^N$ if and only if $C \in \mathbb{R}^{2m\times N}$ where

$$C = \begin{pmatrix} \Re(B) \\ \Im(B) \end{pmatrix}$$

is an ϵ-JL map of $S \in \mathbb{R}^N$. Use this to show that Theorem 4.4.4 holds as stated if $\Phi \in \mathbb{C}^{m\times N}$. **Hint:** Use the first part of the exercise to argue that a real matrix has the desired Krahmer-Ward property, and again using the exercise to argue that the real matrix having the Krahmer-Ward property implies the complex matrix must as well.

Homework 4.4.3. Show that $B \in \mathbb{C}^{m\times N}$ is ϵ-JL map of $S = \left\{\mathbf{x}_j + i\mathbf{y}_j | \mathbf{x}_j, \mathbf{y}_j \in \mathbb{R}^b\right\} \subset \mathbb{C}^N$ if and only if

$$A = \left(\begin{array}{c|c} \Re(B) & -\Im(B) \\ \hline \Im(B) & \Re(B) \end{array}\right) \in \mathbb{R}^{2m\times 2N}$$

is an ϵ-JL map for $\tilde{S}$ where

$$\tilde{S} = \begin{pmatrix} \mathbf{x}_i \\ \\ \mathbf{y}_i \end{pmatrix}$$

Homework 4.4.4. Use the previous homework problems to prove the following variant of Theorem 4.4.4: Let $S \subset \mathbb{C}^N$ with $|S| = M$. Suppose that $\Phi \in \mathbb{C}^{m \times N}$ is $(2s, \eta/4)$-RIP matrix for $\eta, \delta \in (0,1)$ and $s \geq 16 \ln(4M/\delta)$. Let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \in \mathbb{R}^N$ have i.i.d Radamacher entries then

$$(1-\eta)\|\mathbf{x}\|_2^2 \leq \|\Phi \mathrm{Diag}(\boldsymbol{\psi}_1)\Re(\mathbf{x}) + i\mathrm{Diag}(\boldsymbol{\psi}_2)\Im(\mathbf{x})\|_2^2 \leq (1+\eta)\|\mathbf{x}\|_2^2$$

4.5 General Metric Space Embeddings

Definition 4.5.1. A metric space is a pair (S, ρ) where S is a set and $\rho : S \times S \to \mathbb{R}^+$ satisfies

1. $\rho(x,y) = 0 \iff x, y \in S$
2. $\rho(x,y) = \rho(y,x) \forall x, y \in S$
3. $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$

Definition 4.5.2. A finite metric space is a metric space where $|S|$ is finite

Definition 4.5.3. A compact metric space is a metric space (S, ρ) where $|S|$ may be infinite and has the property that every open cover of S contains a finite subcover. That is, for all collections of points $\{x_i\}_{i \times \mathcal{I}} \subset S$ and radii $\{r_i\} \in \mathbb{R}^+ \setminus \{0\}$ where $S = \bigcup_{i \in \mathcal{I}} \{y \in S | \rho(x_i, y) < r_i\}$ there exists a finite cover, $\mathcal{F} \subseteq \mathcal{I}$ such that $S = \bigcup_{i \in \mathcal{F}} \{y \in S | \rho(x_i, y) < r_i\}$. Equivalently, a compact metric space has the property that for any $\epsilon > 0$, the ϵ-covering number with respect to ρ is finite.

Example 4.5.4 (Norm Induced Metric)**.** Let $S \subset \mathbb{R}^N$ and let $\rho : S \times S \to \mathbb{R}^+$ be $\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ where $\|\cdot\|$ is any norm on $\mathbb{R}^N$. Then (S, ρ) is a metric space.

Example 4.5.5 (Weighted Graph Metric Space)**.** Let (V, E) be an undirected, connected weighted graph with no loops or multiple edges. V is the set of vertices and $E \subset V \times V$ is the set of edges, we may define $w : E \to (0, \infty)$ as the weight function, where $w(v, v) = 0$ for $v \in V$. A path from vertex p to q where $p, q \in V$ is a set of edges $e_1, \ldots, e_d \in E$ such that $e_j = (u, v)$ is in the path if and only if $e_{j-1} = (\cdot, u)$ and $e_{j+1} = (v, \cdot)$ for $j = 2, \ldots, d-1$ and $e_1 = (p, \cdot)$, $e_d = (\cdot, q)$. That is, a string of edges that connects p and q.

The shortest path distance from p to q for $p, q \in V$

$$\rho(p, q) = \inf_{\mathcal{E} \in (p,q)-\text{paths}} \sum_{e_j \in \mathcal{E}} w(e_j)$$

defines a metric for V

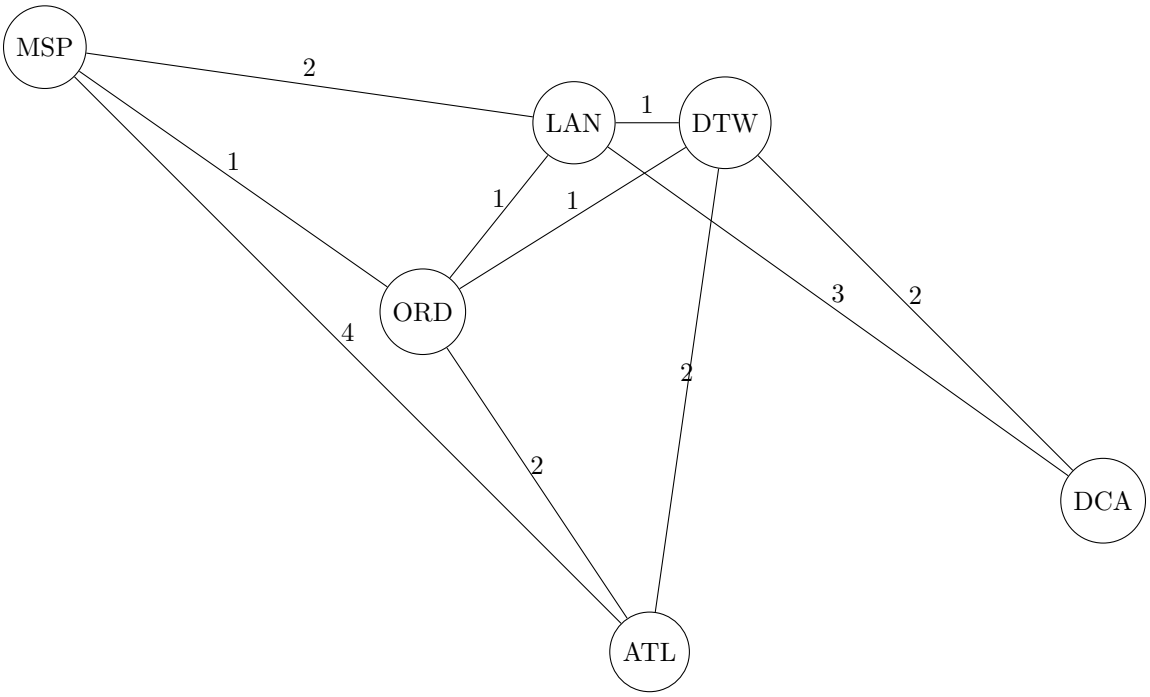


Figure 4.1: Example of a weighted graph using airport codes as the nodes. The edges could represent available flights over a certain period of time, and the weights could be travel time in hours

Example 4.5.6 (Edit Distance over Words)**.** Let A be some alphabet, a finite set of elements such as the letters in the English language. Let $V = \bigcup_{j=1}^{N} A^j =$

$\bigcup_{j=1}^{N} \underbrace{A \times A \times \cdots \times A}_{\text{Cartesian product of } A \, j\text{-times with itself}}$. That is, V is the set of tuples of A with length less than or equal to where N is the length of the longest word. We will build an edge set in the following way. For any pair $(x, y) \in V^2$ such that $x = (z, u)$ for some $z, u \in \emptyset \times \bigcup_{j=1}^{N-1} A^j$ with

$$y \in \{(a, z, u), (z, a, u), (a, u, a)\}$$

for some $a \in A$. That is, (x, y) is a pair of words, where y differs from x by an insertion of a single letter. We assign the weight function w whereeach edge in the edge set has weight one. So for example $w(\texttt{at}, \texttt{cat}) = 1$ and $w(\texttt{rat}, \texttt{cat}) = 2$ if we consider the alphabet A to be the 26 letters of the English alphabet.

Example 4.5.7 (Manifold)**.** Let $\mathcal{M} \subset \mathbb{R}^N$ be a compact, smooth, Riemannian manifold of $\mathbb{R}^N$. Then $(\mathcal{M}, \text{geodesic distance})$ is a compact metric space.

A central question then for our study of metric spaces and embeddings is as follows: given a large, potentially complicated finite metric space, such as the graph example given in 4.5.6. Can we compute approximate distance in a different metric more quickly? A related problem is then can we compress a table of distances (e.g. for a graph (V, E) then the distance table has $\binom{|V|}{2}$ entries to store) such that the table take less storage while still approximately preserving distance. As is typical in our applications we wish to understand how to trade accuracy for time and space in some optimal way.

Definition 4.5.8 ((α, β)-Distortion Embedding)**.** Let $\alpha, \beta \in (0, \infty)$ with $\alpha \leq \beta$. An (α, β)-distortion embedding of a metric space (X, ρ) into (Y, η) is a function $f : X \to Y$ such that

$$\alpha\eta\left(f(\mathbf{x}, f(\mathbf{y}))\right) \leq \rho(\mathbf{x}, \mathbf{y}) \leq \beta\eta\left(f(\mathbf{x}, \mathbf{y})\right)$$

Where our interest lies in finding functions f and η which are able to be computed quickly (as compared to $\rho(\cdot,\cdot)$). Or, finding metric spaces Y which are easier to store in comparison to X.

Definition 4.5.9. If $\alpha = \beta = 1$ then f in Definition 4.5.8 is called an isometric embedding.

Example 4.5.10. Let $\epsilon \in (0,1)$, $S \subset \mathbb{R}^N$ be finite, and define $S \times S \to \mathbb{R}^+$ to be $d_S^N(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$. The JL lemma ensures that there exists a $\left((1+\epsilon)^{-1/2},(1-\epsilon)^{-1/2}\right)$-distortion embedding of (S, d_S^N) into $\left(S' \subset \mathbb{R}^m, d_{S'}^N\right)$ where $m = \mathcal{O}\left(\frac{\log|S|}{\epsilon^2}\right)$. Furthermore the embeddings are easy to generate as random matrices.

We are wish to find embeddings into $\mathbb{R}^N$ with an ℓ^p-norm defined metric on the space $Y \subset \mathbb{R}^N$ - this is desirable since ℓ^p-norms are easier to implement and compute than metrics in the original metric space.

Definition 4.5.11. We say that (X,ρ) is an (α,β)-embeddable into $\ell^p(\mathbb{R}^N)$ if there exists (α,β)-distortion embedding of (X,ρ) into $S \subset \mathbb{R}^N, \|\mathbf{x}-\mathbf{y}\|_p$ with $|X| = |S|$. That is, when (X,ρ) is (α,β)-embeddable into $\ell^p(\mathbb{R}^N)$ there exists $f : X \to S \subset \mathbb{R}^N$ such that

$$\alpha\|\mathbf{f}(x)-\mathbf{f}(y)\|_p \le \rho(x,y) \le \beta\|\mathbf{f}(x)-\mathbf{f}(y)\|_p,\ \forall x,y \in X$$

Definition 4.5.12. We say that (X,ρ) embeds isometrically into ℓ_N^p if there exists $(1,1)$-distortion embedding of (X,ρ) into $\ell^p(\mathbb{R}^N)$

Theorem 4.5.13. *Every finite metric space (X,ρ) embeds isometrically into $\ell_{|X|}^\infty$*

Proof. Define $f : X \to \mathbb{R}^{|X|}$. Consider each element of the metric space $X = \left\{x_1,\ldots,x_{|X|}\right\}$. Our function then returns a vector of length $|X|$ such that $f(x_j) = \left(\rho(x_1,x_i),\ldots,\rho(x_{|X|},x_i)\right)$.

Consider then the infinity norm of the vectors in the metric space Y

$$\begin{aligned}\|\mathbf{f}(x) - \mathbf{f}(x_j)\|_\infty &\geq \left|\left(f(x_i) - f(x_j)\right)_j\right| \\ &= |\rho(x_j, x_i) - \rho(x_j, x_j)| \\ &= \rho(x_i, x_j)\end{aligned}$$

which holds for any choice of $i, j \in [|X|]$. We have then that $\beta = 1$ We now consider the lower bound of $\rho(x_i, x_j)$

$$\begin{aligned}\left|\left(\mathbf{f}(x_i) - \mathbf{f}(x_j)\right)_k\right| &= |\rho(x_k, x_i) - \rho(x_k, x_j)| \\ &\leq \rho(x_i, x_j)\end{aligned}$$

where we have used the reverse triangle inequality. Since this holds for any choice of $k \in [|X|]$, it will hold for the maximum over k.

$$\|\mathbf{f}(x) - \mathbf{f}(x_j)\|_\infty = \max_{k \in [|X|]} \left|\left(f(x_i) - f(x_j)\right)_k\right|$$

and so we have then that $\alpha = 1$. Thus f defines an isometric embedding into $\ell^\infty(\mathbb{R}^{|X|})$ □

Computationally, in the proof above, we are simply storing all distances between pairs of elements of X through the use of f.

$$\begin{array}{c} \\ \\ \\ \\ \end{array} \overset{f(x_i)}{\overset{\downarrow}{\boxed{\begin{array}{c} \\ \\ \\ \\ \end{array}}}} \left(\begin{array}{ccccc} \rho(x_1, x_1) & \dots & \rho(x_1, x_i) & \dots & \rho(x_1, x_{|X|}) \\ \rho(x_2, x_1) & \dots & \rho(x_2, x_i) & \dots & \rho(x_2, x_{|X|}) \\ \vdots & & \vdots & & \vdots \\ \rho(x_{|X|}, x_1) & \dots & \rho(x_{|X|}, x_i) & \dots & \rho(x_{|X|}, x_{|X|}) \end{array}\right)$$

Definition 4.5.14. A subset $\mathcal{C} \subseteq X$ is called an ϵ-cover of (X, ρ) if $\forall x \in X$, $\exists y \in \mathcal{C}$ such that $\rho(x, y) < \epsilon$. The ϵ-covering number of (X, ρ) is the cardinality of a minimal ϵ-cover of (X, ρ).

As can be seen in Homework 4.5.4, for approximation purposes, it is sufficient to embed a cover of a compact metric space to preserve approximately all metric distances in X, even when the set has an infinite number of points, and that distance computations will required N ρ-distance computations. If (X, ρ) has low dimension structure this can be done efficiently using for example a cover-tree construction as shown in [3]

Homework 4.5.1. Show that every finite metric space is a compact metric space

Homework 4.5.2. Show that a metric space (S, ρ) from Example 4.5.4 is a compact metric space if and only if S is closed and bounded with respect to Euclidean distance. **Hint:** Recall that any two norms $\|\cdot\|_*$, $\|\cdot\|_\dagger$ on $\mathbb{R}^N$, there exists positive constants D, C such that

$$D\|\mathbf{x}\|_* \leq \|\mathbf{x}\|_\dagger \leq C\|\mathbf{x}\|_*, \forall \mathbf{x} \in \mathbb{R}^N$$

Homework 4.5.3. Show that (α, β)-distortion embedding is always one-to-one such that $|X| \leq |Y|$ when one exists.

Homework 4.5.4. Let $C_\epsilon^\rho = N$ denote the ϵ covering number of (X, ρ) (see Definition 3.2.2) then

1. N is finite if (X, ρ) is a compact metric space.

2. Prove there exists a map $f : X \to \ell^\infty(\mathbb{R}^N)$ such that

$$|d_N^\infty(f(x), f(y)) - \rho(x, y)| < 2\epsilon, \forall x, y \in X$$

Application 4.5.15 (Computing the diameter of a set)**.** We now consider a particular, more useful, embedding into $\ell^\infty(\mathbb{R}^N)$ that will reveal for $S \subseteq \mathbb{R}^n$, its diameter $\operatorname{diam}_p(S) = \max_{\mathbf{x},\mathbf{y}\in S} \|\mathbf{x}-\mathbf{y}\|_p$.

Similar to the naive nearest neighbor problem the computation, the diameter of a set takes $\mathcal{O}(n|S|^2)$ to compute since it requires a distance calculation between all pairs of points.

Lemma 4.5.16. *If $p > q \geq 1$ then $\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_q$, $\forall \mathbf{x} \in \mathbb{R}^n$*

Proof. Consider

$$\begin{aligned}\|\mathbf{x}\|_p^p &= \sum_{j=1}^n |x_j|^p \\ &= \sum_{j=1}^n |x_j|^q |x_j|^{p-q} \\ &\leq \max_j |x_j|^{p-q} \left(\sum_{j=1}^n |x_j|^q\right)\end{aligned}$$

Note that

$$\max_{j\in[n]} |x_j|^q \leq \sum_{j=1}^n |x_j|^q \iff \max_{j\in[n]} |x_j|^{p-q} \leq \left(\sum_{j=1}^n |x_j|^q\right)^{\frac{p-q}{q}} = \left(\|\mathbf{x}\|_q^q\right)^{\frac{p-q}{q}}$$

therefore

$$\begin{aligned}\|\mathbf{x}\|_p^p &\leq \left(\|\mathbf{x}\|_q^q\right)^{\frac{p-q}{q}} \|\mathbf{x}\|_q^q \\ &= \|\mathbf{x}\|_q^p\end{aligned}$$

which implies the result after taking p-th root □

Lemma 4.5.17. *Let $p > 1$. Then*

$$\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_p n^{1-1/p}$$

holds $\forall \mathbf{x} \in \mathbb{R}^n$

Proof. Apply lemma 4.5.16 and Holder's Inequality. □

Lemma 4.5.18. *If (S, d_1^n) is a finite $\ell^1(\mathbb{R}^n)$ metric space then it can be isometrically embedded into $\ell^\infty(\mathbb{R}^{2^n})$ using a linear embedding $f : S \to \mathbb{R}^{2^n}$.*

Proof. Let $\tilde{S} \subset \mathbb{R}^{2^n}$ be the set $\{-1, 1\}^n$, all possible vertices of the n-cube. Noting

$$\operatorname{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

we have the following identity between the ℓ^1-norm and maximum of inner products over $\tilde{S}$. Observe, on one hand:

$$\|\mathbf{x}\|_1 = \sum_{j=1}^{n} |x_j| = \sum_{j=1}^{n} \operatorname{sgn}(x_j) x_j \leq \max_{\mathbf{y} \in \tilde{S}} \langle \mathbf{y}, \mathbf{x} \rangle$$

on the other hand, using Holder's inequality

$$\max_{\mathbf{y} \in \tilde{S}} \langle \mathbf{y}, \mathbf{x} \rangle \leq \|\mathbf{y}\|_\infty \|\mathbf{x}\|_1 = \|\mathbf{x}\|_1$$

and so $\|\mathbf{x}\|_1 = \max_{\mathbf{y} \in \tilde{S}} \langle \mathbf{y}, \mathbf{x} \rangle$. We define our embedding $f : \mathbb{R}^n \to \mathbb{R}^{2^n}$ as

$$f(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{y} \rangle \right)_{\mathbf{y} \in \tilde{S}}$$

As seen above, $\|\mathbf{x}\|_1 = \|f(\mathbf{x})\|_\infty$, and so we see that f is the embedding of S into $\ell^\infty(\mathbb{R}^{2^n})$ □

Note that $\mathbf{diam}_\infty(S)$ can be computed in time linear in $|S|$. We can see this by

considering the following calculation

$$\begin{aligned}\operatorname{diam}_\infty(S) &= \max_{\mathbf{x},\mathbf{y}\in S} \|\mathbf{x}-\mathbf{y}\|_\infty \\ &= \max_{\mathbf{x},\mathbf{y}\in S} \max_{j\in[2^N]} |x_j - y_j| \\ &= \max_{j\in[2^N]} \max_{\mathbf{x},\mathbf{y}\in S} |x_j - y_j| \\ &= \max_{j\in[2^N]} \left| \max_{\mathbf{x}\in S} x_j - \min_{\mathbf{y}\in S} y_j \right|\end{aligned}$$

but for any coordinate j, $\max_{\mathbf{x}\in S} x_j$ requires $|S|$ comparisons. Likewise for $\min_{\mathbf{y}\in S} y_j$. Since there are 2^N coordinates then, this calculation takes $\mathcal{O}(2^N|S|)$ time to compute.

Algorithm 4.5.1 Diameter in ℓ^p-norm of S

Input: $S \subseteq \mathbb{R}^N, p \in [1,\infty)$
Output: Estimate for $\operatorname{diam}_p(S)$
for $\mathbf{x}$ in S **do**
  Compute $\mathbf{x}' = f(\mathbf{x}) \in \mathbb{R}^{2^N}$ where f defined as in Lemma 4.5.18
end for
for each $j \in [2^n]$ **do**
  $M_j \leftarrow \max_{\mathbf{x}'\in S'} x_j$
  $m_j \leftarrow \min_{\mathbf{x}'\in S'} x_j$
  $\ell_j \leftarrow M_j - m_j$
end for
return $\max_{j\in[2^n]} \ell_j$

Recall that $f(\mathbf{x})$ is computed by taking 2^N inner products of vectors of length n. Since this is computed for each element of $\mathbf{x} \in S$ the runtime of the first loop of Algorithm 4.5.1 is $\mathcal{O}(|S|N2^N)$. The second loop requires comparing $|S|$ values for each of the 2^N coordinates for a runtime of $\mathcal{O}|S|2^N$. Thus the overall runtime of the algorithm is $\mathcal{O}(|S|N2^N)$.

The brute force method requires computing the ℓ^p-norm of the difference of all possible pairs of vectors. Each norm computation takes on order N operations and there are on order $|S|^2$ such pairs to compute for an overall runtime of $\mathcal{O}(|S|^2N)$. So when $|S| \geq 2^N$ then the Algorithm 4.5.1 will save time.

Theorem 4.5.19. *Algorithm 4.5.1 outputs an estimate E that satisfies*

$$diam_p(S) \leq E \leq n^{1-\frac{1}{p}} diam_p(S)$$

Proof. Denote the output of the algorithm as E. Note that by Lemma 4.5.18, we have that $\ell^1(\mathbb{R}^N)$ is embedded isometrically into $\ell^\infty(\mathbb{R}^{2^N})$ by f. That is

$$E = \max_{\mathbf{x},\mathbf{y}} \|f(\mathbf{x}) - f(\mathbf{y})\|_\infty = \max_{\mathbf{x},\mathbf{y}} \|\mathbf{x} - \mathbf{y}\|_1$$

On the other hand, by Lemma 4.5.16, $\|\mathbf{x} - \mathbf{y}\|_p \leq \|\mathbf{x} - \mathbf{y}\|_1$ and also $\|\mathbf{x} - \mathbf{y}\|_1 \leq n^{1-\frac{1}{p}}\|\mathbf{x} - \mathbf{y}\|_p$. Combining then these results we see

$$\mathbf{diam}_p(S) \leq \mathbf{diam}_1(S) = E \leq n^{1-\frac{1}{p}} \mathrm{diam}_(S)$$

□

Our algorithm and theorem then demonstrate that we can productively use the fact that all finite metric spaces embed into ℓ^∞. This depended greatly

4.6 Frechet Embedding Methods for Finite Metric Spaces

Definition 4.6.1 (Frechet Embedding)**.** Let (X, ρ) be a finite metric space. Let $S_j \subset X$ for $j \in [r]$ be r non-empty subsets. Let $\alpha_1, \ldots, \alpha_r \in \mathbb{R}$. The Frechet embedding f of (X, ρ) into $\ell^p(\mathbb{R}^r)$ is given by

$$f(\mathbf{x})_j = \alpha_j \min_{\mathbf{z} \in S_j} \rho(\mathbf{z}, \mathbf{x}),\ \forall x \in X, j \in [r] \tag{4.8}$$

Lemma 4.6.2. *Let (X, ρ) be a finite metric space. Let $f : X \to \mathbb{R}^r$ be a Frechet embedding as in Definition 4.8 where $\alpha_1 = \alpha_2 = \cdots = \alpha_r$. Then*

$$\|f(\mathbf{x}) - f(\mathbf{y})\|_p \leq \alpha r^{1/p} \rho(\mathbf{x}, \mathbf{y})$$

$\forall p \geq 1$, $\forall \mathbf{x}, \mathbf{y} \in X$

Proof. Fix $\mathbf{x}, \mathbf{y} \in X$. We will show that for any nonempty subset $S \subseteq X$

$$|\rho(\mathbf{x}, S) - \rho(y, S)| \leq \rho(\mathbf{x}, \mathbf{y}) \tag{4.9}$$

where $\rho(\mathbf{x}, S) = \min_{\mathbf{z} \in S} \rho(\mathbf{x}, \mathbf{z})$. If 4.9 holds then

$$\left(\sum_{j=1}^{r} |\alpha\rho(\mathbf{x}, S_j) - \alpha\rho(\mathbf{y}, S_j)|^p \right)^{1/p} \leq \alpha \left(\sum_{j=1}^{r} |\rho(\mathbf{x}, \mathbf{y})|^p \right)^{1/p} = \alpha r^{1/p} \rho(\mathbf{x}, \mathbf{y})$$

To see that 4.9 holds, if S is nonempty, $\rho(\mathbf{x}, S) \leq \rho(\mathbf{x}, \mathbf{w})$, $\forall \mathbf{x} \in X, \forall \mathbf{w} \in S$ by definition. Now suppose $\tilde{\mathbf{w}} \in S$ is the minimizer for $\rho(\mathbf{y}, S) = \rho(\mathbf{y}, \tilde{\mathbf{w}})$. Then

$$\begin{aligned} \rho(\mathbf{x}, S) - \rho(\mathbf{y}, S) &\leq \rho(\mathbf{x}, \tilde{\mathbf{w}}) - \rho(\mathbf{y}, \tilde{\mathbf{w}}) \\ &\leq \rho(\mathbf{x}, \mathbf{y}) \end{aligned}$$

Similarly, $\rho(\mathbf{y}, S) - \rho(\mathbf{x}, S)$ holds when $\tilde{\mathbf{w}}$ is the minimizer for the distance $\rho(\mathbf{x}, S)$. □

Lemma 6.4 indicates that any constant Frechet embedding will satisfy some (γ, β)-distortion criteria with $\beta = \alpha r^{1/p}$. So the main difficulty is to find a lower bound for the distortion. We would ideally want a bound that is proportional to $r^{1/p}$ i.e. some small $\tilde{c}$ so that $\gamma = r^{1/p}\tilde{c}$

An embedding into $\ell^2(\mathbb{R}^{\mathcal{O}(\log_2 |X|)})$ for any finite metric space (X, ρ)

1. Let $j = 1, \ldots, \lfloor \log_2 |X| \rfloor + 1$
2. For each j above, construct m random subsets $A_{i,j} \subseteq X$, $\forall i \in [m]$ by drawing 2^j entries from X i.i.d uniformly at random.

Theorem 4.6.3 (Bourgain)**.** *Let $f : X \to \mathbb{R}^{qm}$ be defined by*

$$\tilde{f}(\mathbf{y})_{mi:m(i+1)} = \left(\frac{1}{m} \min_{\mathbf{z} \in A_{i,1}}, \ldots, \frac{1}{m} \min_{\mathbf{z} \in A_{i,m}}\right)$$

where q denotes $\lfloor \log_2 |X| \rfloor + 1$. There exists universal constants $c, \tilde{c} \in \mathbb{R}^+$ such that

$$\tilde{c}\rho(\mathbf{x}, \mathbf{y}) \leq \|\tilde{f}(\mathbf{x}) - \tilde{f}(\mathbf{y})\|_2 \leq q\rho(\mathbf{x}, \mathbf{y})$$

holds $\forall \mathbf{x}, \mathbf{y} \in X$ *with high probability provided that*

$$m \geq c \log_2 |X|$$

The theorem asserts that there exists $(\tilde{c}, \mathcal{O}(\log_2 |X|))$-distortion embedding of (X, ρ) into $\ell^2(\mathcal{O}(\log_2^2 |X|))$. We can further improve the compression by use of an ϵ-JL map. Schematically

$$(X, \rho) \xrightarrow{\text{Bourgain}} (\tilde{f}(X'), \ell^2(\mathbb{R}^{\mathcal{O}(\log^2 |X|)})) \xrightarrow{\epsilon\text{-JL}} (\Phi(\tilde{f}(X)), \ell^2(\mathbb{R}^{\mathcal{O}(\log_2 |X|)}))$$

Corollary 4.6.4. *There exists* $(1, \mathcal{O}(\log |X|))$*-distortion embeddings of* (X, ρ) *into* $\ell^2(\mathbb{R}^{\mathcal{O}(\log |X|)})$

What then is achieved in terms of compression for such an embedding? Note that storing all pairwise distances in the original metric requires $\mathcal{O}(|X|^2)$-space. Embedding via the Bourgain Theorem and a JL map would result in $\mathcal{O}(|X| \log |X|)$-space where we have approximations of quality $\mathcal{O}(\log |X|)$ distortion. There are known examples of metric spaces where the upperbound of the distortion is achieved by this approach; that state of affairs is not satisfactory for large $|X|$ and so our next result will address this deficiency by allowing for a range of space-accuracy trade-offs.

Goal. *We desire an* (α, β)*-distortion embedding* (X, ρ) *into* $\ell^\infty(\mathbb{R}^{\tilde{n}})$ *where* $\tilde{n} \leq |X|$ *with a distortion ratio* $\beta/\alpha = D$*, where* D *is any odd value we choose (i.e. we will be interested in the case* $D \geq 3$ *and* $D \leq \log |X|$*).*

Towards this goal, let $p = |X|^{-1/q}$ where $q = \frac{D+1}{2} > 1$.

Now let $j = 1, \ldots, q$, define $p_j = \min(1/2, p^j)$. Let $m \geq 24\gamma n^{1/q} \ln |X|$ for any $\gamma \geq 1$ we may choose.

Construct random subsets of X in the following way: for each $j = 1, \ldots, q$ and $i = 1, \ldots, m$ choose a subset $A_{i,j} \subseteq X$ where each element of X is included in $A_{i,j}$ with probability p_j.

Define $f' : X \to \mathbb{R}^{mq}$ by

$$(f'(\mathbf{v}))_{i,j} = \rho(\mathbf{v}, A_{i,j}),\ \forall j \in [q], i \in [m], \mathbf{v} \in X \tag{4.10}$$

Note that based on the definitions, $mq = \mathcal{O}(q\gamma|X|^{1/q} \ln |X|)$

Theorem 4.6.5. *The random embedding f' into $\ell^\infty(\mathbb{R}^{mq})$ as seen in 4.10 will satisfy*

$$\frac{1}{D}\rho(\mathbf{u}, \mathbf{v}) \leq \|f'(\mathbf{u}) - f'(\mathbf{v})\|_\infty \leq \rho(\mathbf{u}, \mathbf{v})$$

$\forall \mathbf{u}, \mathbf{v} \in X$ with provability at least $1 - |X|^{2(1-\gamma)}$ where $D = 2q - 1$

How does this compare in terms of storage? Recall that an isometric embedding of (X, ρ) into ℓ^∞ requires $\mathcal{O}(|X|^2)$ space. The embedding f' on the other hand requires $\mathcal{O}(D\gamma|X|^{\frac{2}{D+1}+1} \ln |X|)$ space.

Chapter V

LJL Embeddings of Arbitrary Subsets of $\mathcal{R}^D$, Manifold Models, Manifold Learning, and Dimensionality Reduction (MTH 994 Lectures 8 – 10)

5.1 Gaussian Widths and Applications (MTH 994 Lecture 8)

Definition 5.1.1. The Gaussian width of a set $T \subset \mathbb{R}^N$ is denoted $w(T)$ and is defined as

$$w(T) = \mathbb{E}_{\mathbf{g}} \left[\sup_{\mathbf{x} \in T} \langle \mathbf{g}, \mathbf{x} \rangle \right]$$

where $\mathbf{g} \in \mathbb{R}^N$ is a Gaussian random vector with i.i.d entries $g_j \sim \mathcal{N}(0,1)$. We define $w(\emptyset) = 0$.

As we shall see shortly, the Gaussian width is most useful when we consider normalized vectors, i.e. we consider $w(\tilde{T})$ where $\tilde{T} = \{\mathbf{t}/\|\mathbf{t}\|_2 | \mathbf{t} \in T\}$. Additionally, the width of a set is unchanged if we consider its closure $w(T) = w(\overline{T})$

Lemma 5.1.2. $\forall T \subseteq \mathbb{R}^N$, $w(T) \geq 0$

Proof. The proof is left as an exercise. □

Lemma 5.1.3. *If $S \subseteq \overline{T}$ then $w(S) \leq w(T)$*

Proof. The proof is left as an exercise. □

Lemma 5.1.4. *Let $U \in \mathbb{R}^{N \times N}$ be unitary, $\mathbf{y} \in \mathbb{R}^N$ and $T \subseteq \mathbb{R}^N$. Then $w(U(T) + \mathbf{y}) = w(T)$*

Proof.

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathbf{x}\in U(T)+\mathbf{y}}\langle \mathbf{g},\mathbf{x}\rangle\right] &= \mathbb{E}\left[\sup_{\mathbf{z}\in T}\langle \mathbf{g}, U\mathbf{z}+\mathbf{y}\rangle\right] \\
&= \mathbb{E}\left[\sup_{\mathbf{z}\in T}\langle \mathbf{g}, U\mathbf{z}\rangle + \langle\right] + \mathbb{E}\left[\mathbf{g},\mathbf{y}\rangle\right] && \text{linearity of } \mathbb{E} \\
&= \mathbb{E}\left[\sup_{\mathbf{z}\in T}\langle U^*\mathbf{g}, \mathbf{z}\rangle\right] && \text{Lemma 2.5.1} \\
&= \sqrt{\frac{1}{(2\pi)^N}}\int_{\mathbb{R}^N} f_T\left(U^*\mathbf{g}\right) e^{-\frac{\|\mathbf{g}\|_2^2}{2}} d\mathbf{g} && f_T\left(U^*\mathbf{g}\right) = \sup_{\mathbf{z}\in T}\langle \mathbf{g}, U\mathbf{z}\rangle + \langle \\
&= \sqrt{\frac{1}{(2\pi)^N}}\int_{U^*\mathbb{R}^N} f_T\left(\mathbf{h}\right) e^{-\frac{\|U\mathbf{h}\|_2^2}{2}} |\det U^*| d\mathbf{h} && \text{change of variable } \mathbf{h} = U^*\mathbf{g} \\
&= \sqrt{\frac{1}{(2\pi)^N}}\int_{\mathbb{R}^N} f_T\left(\mathbf{h}\right) e^{-\frac{\|\mathbf{h}\|_2^2}{2}} d\mathbf{h} && |\det(U)| = 1, \|U\mathbf{h}\|_2 = \|\mathbf{h}\|_2 \\
&= w(T)
\end{aligned}$$

□

Example 5.1.5. Consider $B_N^{\ell^2} = B_N^{\ell^2}(\mathbf{0}, 1)$, the unit ℓ^2-ball in $\mathbb{R}^N$. Then

$$\begin{aligned}
w\left(B_N^{\ell^2}(\mathbf{0},1)\right) &= \mathbb{E}\left[\sup_{\mathbf{x}\in B_N^{\ell^2}}\langle \mathbf{g},\mathbf{x}\rangle\right] \\
&= \mathbb{E}\left[\|\mathbf{g}\|_2\right] && \mathbf{g}/\|\mathbf{g}\|_2 \in B_N^{\ell^2} \\
&\leq \left(\mathbb{E}\left[\|\mathbf{g}\|_2^2\right]\right)^{1/2} && \text{Jensen's inequality} \\
&= \left(\mathbb{E}\left[\sum_{j=1}^N g_j^2\right]\right)^{1/2} \\
&= \sqrt{N}
\end{aligned}$$

A lower bound in terms of $\sqrt{N}$ can likewise be found and so we see $w(T) \approx \sqrt{N}$.

Example 5.1.6. Consider $T = B^N(\mathbf{0}, 1) \cap \mathcal{L}_{\mathcal{B}}^k$, the unit ℓ^2-ball intersected with a k-dimensional subspace. Note that there exists U unitary such that

$$U\left(\left\{(x_1,\ldots,x_k,0,\ldots,0) | x_1^2+\cdots+x_k^2 = 1\right\}\right) = \mathcal{B}$$

We can use the previous example then to show that $w(T) \leq \sqrt{k}$

Lemma 5.1.7. *If $a \in \mathbb{R}$, $T, S \subseteq \mathbb{R}^N$ then*

1. $w(T+S) = w(T) + w(S)$

2. $w(aT) = |a| w(T)$

Proof. The proof is left as an exercise. □

Lemma 5.1.8. *Let $T \subseteq \mathbb{R}^N$ then*

$$w(T-T) = 2w(T)$$

Proof. Immediate consequence of Lemma 5.1.7. □

Lemma 5.1.9. *Let $T \subseteq \mathbb{R}^N$ then*

$$\frac{1}{\sqrt{2\pi}} diam(T) \leq w(T) \leq \frac{\sqrt{N}}{2} diam(T)$$

where $diam(T) = \sup\{\|\mathbf{x}-\mathbf{y}\|_2 | \mathbf{x}, \mathbf{y} \in T\}$

Proof. Fix $\mathbf{x}, \mathbf{y} \in T$

$$\begin{aligned} w(T) &= \frac{1}{2} w(T-T) && \text{Lemma 5.1.8} \\ &= \frac{1}{2} \mathbb{E}\left[\sup_{\mathbf{u}-\mathbf{v} \in T-T} \langle \mathbf{u}-\mathbf{v}, \mathbf{g} \rangle\right] \\ &\geq \frac{1}{2} \mathbb{E}\left[\max\left(\langle \mathbf{x}-\mathbf{y}, \mathbf{g} \rangle, \langle \mathbf{y}-\mathbf{x}, \mathbf{g} \rangle\right)\right] \\ &= \frac{1}{2} \mathbb{E}\left[|\langle \mathbf{x}-\mathbf{y}, \mathbf{g} \rangle|\right] \end{aligned}$$

But $\langle \mathbf{x}-\mathbf{y}, \mathbf{g} \rangle \sim \mathcal{N}(0, \|\mathbf{x}-\mathbf{y}\|_2)$ so $|\langle \mathbf{x}-\mathbf{y}, \mathbf{g} \rangle|$ is a folded normal distribution, which has an expectation of $\|\mathbf{x}-\mathbf{y}\|_2 \sqrt{\frac{2}{\pi}}$ and so taking the supremum over all distances $\mathbf{x}-\mathbf{y}$ we have the lower bound:

$$\frac{1}{2} \|\mathbf{x}-\mathbf{y}\|_2 \sqrt{\frac{2}{\pi}} = \frac{1}{\sqrt{2\pi}} \text{diam}(T) \leq w(T)$$

Now consider

$$\begin{aligned} w(T) &= \frac{1}{2}w(T-T) && \text{Lemma 5.1.8} \\ &= \frac{1}{2}\mathbb{E}\left[\sup_{\mathbf{u}-\mathbf{v}\in T-T}\langle \mathbf{u}-\mathbf{v}, \mathbf{g}\rangle\right] && \text{Cauchy-Schwarz} \\ &\leq \frac{1}{2}\mathbb{E}\left[\sup_{\mathbf{u}-\mathbf{v}}\|\mathbf{u}-\mathbf{v}\|_2\|\mathbf{g}\|_2\right] \\ &\leq \frac{\sqrt{N}}{2}\text{diam}(T) \end{aligned}$$

which corresponds to our desired upper bound □

Lemma 5.1.10. *$w(T)$ is finite if and only if T is bounded.*

Proof. The proof is left as an exercise. □

The following two lemmas will be used to show that the gaussian width of a set and the set's convex hull are equal.

Lemma 5.1.11. *If $\mathcal{C} \subseteq T \subseteq \mathbb{R}^N$ is an ϵ-cover of T then conv($\mathcal{C}$) is a 2ϵ-cover of $\overline{conv(T)}$.*

Proof. Let $\mathbf{x} \in \overline{\text{conv}(T)}$ and $\mathbf{y} \in \text{conv}(T)$ be such that $\|\mathbf{x}-\mathbf{y}\|_2 < \epsilon$. By Caratheodory's Theorem then $\exists \mathbf{t}, \ldots, \mathbf{t}_{\tilde{N}} \in T$ such that $\sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{t}_\ell$ where $\tilde{N} \leq N+1, \alpha_e ll > 0 \forall \ell \in [\tilde{N}]$. For every $\mathbf{t}_\ell \in T$ let $\mathbf{z}_\ell \in \mathcal{C}$ be such that $\|\mathbf{t}_\ell - \mathbf{z}_\ell\|_2 \leq \epsilon$. So $\mathbf{z} = \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \mathbf{z}_\ell \in \text{conv}(\mathcal{C})$ and

$$\begin{aligned} \|\mathbf{z}-\mathbf{x}\|_2 &\leq \|\mathbf{x}-\mathbf{y}\|_2 + \|\mathbf{y}-\mathbf{z}\|_2 \\ &\leq \epsilon + \|\sum_{\ell=1}^{\tilde{N}} \alpha_\ell(\mathbf{t}_\ell - \mathbf{z}_\ell)\|_2 \\ &\leq \epsilon + \sum_{\ell=1}^{\tilde{N}} \alpha_\ell \|\mathbf{t}_\ell - \mathbf{z}_\ell\|_2 \\ &\leq 2\epsilon \end{aligned}$$

□

Lemma 5.1.12. *If $\mathcal{C} \subseteq \mathbb{R}^N$ is a finite set then $w(\mathcal{C}) = w(conv(\mathcal{C})$*

Proof. Consider $\mathcal{C} = \left\{\mathbf{c}_1, \ldots, \mathbf{c}_{|\mathcal{C}|}\right\} \subset \mathbb{R}^N$. By definition of convex hull, we have

$$\begin{aligned} w(\text{conv}(\mathcal{C})) &= \mathbb{E}\left[\sup_{\mathbf{z}\in\text{conv}(\mathcal{C})} \langle \mathbf{g}, \mathbf{z}\rangle\right] \\ &= \mathbb{E}\left[\sup_{\left\{\alpha_1,\ldots,\alpha_{|\mathcal{C}|}, \sum\alpha_j=1\right\}} \langle \mathbf{g}, \sum_{j=1}^{|\mathcal{C}|} \alpha_j \mathbf{c}_j\rangle\right] \\ &= \mathbb{E}\left[\sup_{\left\{\alpha_1,\ldots,\alpha_{|\mathcal{C}|}, \sum\alpha_j=1\right\}} \alpha_j \langle \mathbf{g}, \mathbf{c}_j\rangle\right] \\ &= \mathbb{E}\left[\sup_{j\in[|\mathcal{C}|]} \langle \mathbf{g}, \mathbf{c}_j\rangle\right] \\ &= w(\mathcal{C}) \end{aligned}$$

Since the sum is maximized then when $\alpha_j = 1$ for $\arg\min_j \langle \mathbf{g}, \mathbf{c}_j\rangle$ □

Theorem 5.1.13. *Let $T \subseteq \mathbb{R}^N$ be bounded. Then $w(T) = w(conv(T))$*

Proof. Note that $T \subset \text{conv}(T)$ and so by Lemma 5.1.3 $w(T) \leq w(\text{conv}(T))$.

Now assume for eventual contradiction that $w(T)$ is not greater than $w(\text{conv}(T))$. So then $\exists\epsilon'$ such that $\epsilon' = w(\text{conv}(T)) - w(T) > 0$. Now let $\epsilon = \epsilon'\sqrt{N}/4 > 0$ and let $\mathcal{C} \subseteq T$ be an ϵ-cover of T. Thus

$$\begin{aligned} w(\text{conv}(T)) &= \mathbb{E}\left[\sup_{\mathbf{x}\in\text{conv}(T)} \langle \mathbf{g}, \mathbf{x}\rangle\right] \\ &= \mathbb{E}\left[\sup_{\mathbf{x}\in\overline{\text{conv}(T)}} \langle \mathbf{g}, \mathbf{x}\rangle\right] \\ &= \mathbb{E}\left[\langle \mathbf{g}, \mathbf{x}_\mathbf{g}\rangle\right] \end{aligned}$$

since $f_\mathbf{g}(\mathbf{x}) = \langle \mathbf{g}, \mathbf{x}\rangle$ is a continuous function over a compact set, the set will contain its maximizers; denote such a maximizer by $\mathbf{x}_\mathbf{g}$. Since $\mathcal{C}$ is an ϵ-cover of T, by Lemma

5.1.11 we have that $\text{conv}(\mathcal{C})$ is an 2ϵ-cover of $\text{conv}(T)$ and therefore $\exists \mathbf{x}'_{\mathbf{g}} \in \text{conv}(\mathcal{C})$ such that $\mathbf{x}_{\mathbf{g}} = \mathbf{x}'_{\mathbf{g}} + \boldsymbol{\eta}$ where $\|\boldsymbol{\eta}\|_2 \leq 2\epsilon$ Thus

$$\begin{aligned}
w(\text{conv}(T)) &= \mathbb{E}\left[\langle \mathbf{g}, \mathbf{x}_{\mathbf{g}} \rangle\right] \\
&\leq \mathbb{E}\left[\sup_{\substack{\mathbf{x}\in\text{conv}(\mathcal{C}) \\ \mathbf{y}\in B(0,2\epsilon)}} \langle \mathbf{g}, \mathbf{x}+\mathbf{y} \rangle\right] \\
&= \mathbb{E}\left[\sup_{\mathbf{x}\in\text{conv}(\mathcal{C})} \langle \mathbf{g}, \mathbf{x} \rangle\right] + \mathbb{E}\left[\sup_{\mathbf{y}\in 2\epsilon B(0,1)} \langle \mathbf{g}, \mathbf{y} \rangle\right] \\
&= w(\text{conv}(\mathcal{C})) + w(2\epsilon B(0,1)) \\
&= w(\mathcal{C}) + 2\epsilon w(B(0,1)) \\
&\leq w(\mathcal{C}) + 2\epsilon\sqrt{N}
\end{aligned}$$

Where we have used Lemma 5.1.7, and Theorem 5.1.13 and calculation for width of unit ball seen in Example 5.1.6

$$\begin{aligned}
&= w(T) + \frac{\epsilon'}{2} \\
&< w(T) + \epsilon' \\
&= w(\text{conv}(T))
\end{aligned}$$

which is a contradiction, we negate the hypothesis and conclude that $w(\text{conv}(T)) \leq w(T) \geq w(\text{conv}(T))$ i.e. $w(T) = w(\text{conv}(T))$ □

Lemma 5.1.14. *Let $g \sim \mathcal{N}(0,1)$. Then for $\theta \in \mathbb{R}$, $\mathbb{E}\left[\exp(\theta g)\right] = \exp\left(\frac{\theta^2}{2}\right)$.*

Proof.

$$\begin{aligned}\mathbb{E}\left[\exp(\theta g)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\theta x}e^{-x^2/2}dx \\ &= e^{\frac{\theta^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2\theta x+\theta^2)}dx \\ &= e^{\frac{\theta^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-\theta)^2}dx \\ &= e^{\frac{\theta^2}{2}}\end{aligned}$$

□

Note that we can apply Theorem 4.2.4 and Lemma 5.1.14 to conclude that

$$\mathbb{E}\left[\exp\left(\theta\langle \mathbf{g}, \mathbf{x}\rangle\right)\right] \leq \exp\left(\frac{1}{2}\|\mathbf{x}\|_2^2\theta^2\right)$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I)$

Theorem 5.1.15. *Let $X_0, \ldots, X_{M-1}$ be a sequence of mean zero, subgaussian random variables (not necessarily independent) satisfying $\mathbb{E}\left[\exp(\theta X_\ell)\right] \leq \exp\left(c_\ell\theta^2\right), \forall \ell \in [M]$. Let $c = \max_{\ell\in[M]} c_\ell$. Then*

$$\begin{aligned}\mathbb{E}\left[\max_{\ell\in[M]} X_\ell\right] &\leq 2\sqrt{c\ln M} \\ \mathbb{E}\left[\max_{\ell\in[M]} |X_\ell|\right] &\leq 2\sqrt{c\ln(2M)}\end{aligned}$$

Proof. The inequalities hold when $M = 1$ since $\mathbb{E}[X_\ell] = 0 = \ln 1$. So we may assume

$M \geq 2$. Consider for $\beta \in \mathbb{R}^+$,

$$\begin{aligned}
\beta\mathbb{E}\left[\max_{\ell\in[M]} X_\ell\right] &= \mathbb{E}\left[\max_{\ell\in[M]} \beta X_\ell\right] \\
&= \mathbb{E}\left[\max_{\ell\in[M]} \ln\exp\left(\beta X_\ell\right)\right] \\
&= \mathbb{E}\left[\ln\max_{\ell\in[M]} \exp\left(\beta X_\ell\right)\right] \\
&\leq \mathbb{E}\left[\ln\sum_{\ell\in[M]} \exp\left(\beta X_\ell\right)\right] \\
&\leq \ln\left(\mathbb{E}\left[\sum_{\ell\in[M]} \exp\left(\beta X_\ell\right)\right]\right)
\end{aligned}$$

where we have used Jensen's inequality stated in terms of concave function $\ln(Y)$, $\mathbb{E}[\ln Y] \leq \ln E[Y]$

$$\begin{aligned}
&= \ln\left(\sum_{\ell\in[M]} \mathbb{E}\left[\exp\left(\beta X_\ell\right)\right]\right) \\
&\leq \ln\left(M\exp\left(c\beta^2\right)\right) \\
&= \ln\left(M\right) + c\beta^2
\end{aligned}$$

where we have used linearity of expectation and uniform subgaussian bound on the moment generating functions of random variables X_ℓ

Now after a rearrangement of terms and setting $\beta = \sqrt{\frac{\ln M}{c}}$ we obtain

$$\mathbb{E}\left[\max_{\ell\in[M]} X_\ell\right] \leq 2\sqrt{c\ln M}$$

To obtain the second inequality in the theorem statement, we note that $\max_{\ell\in[M]}\{|X_\ell|\} = \max_{j\in[2M]}\{Y_j\}$ where $Y_{2\ell} = X_\ell$, $Y_{2\ell+1} = -X_\ell$. I.e., consider the set of random variables and their opposites, which will have the same maximum as the absolute value.

We can then apply the first inequality on this doubled set to obtain.

$$\mathbb{E}\left[\max_{\ell\in[M]}|X_\ell|\right] \leq 2\sqrt{c\ln 2M}$$

□

Example 5.1.16. [Upper bound for Gaussian Width of Finite Set] Let $T \subset \mathbb{R}^N$ be a finite set of cardinality M. Note that for $\mathbf{z} = \mathbf{x} - \mathbf{y}$, $\mathbf{x}, \mathbf{y} \in T$, $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I)$ the random variable $X = \langle \mathbf{g}, \mathbf{z}\rangle$ is distributed normally with variance equal to the square of ℓ^2-norm of $\mathbf{z}$, i.e. $X \sim \mathcal{N}(\mathbf{0}, \|\mathbf{x}-\mathbf{y}\|_2^2)$ by Lemma 2.5.1 and X is subgaussian with parameter $\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ by Theorem 4.2.4 and Lemma 5.1.14. Observe

$$\begin{aligned}
w(T) &= \frac{1}{2}w(T-T) \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{\mathbf{x},\mathbf{y}\in T}\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{\mathbf{x},\mathbf{y}\in T}\langle \mathbf{g}, \mathbf{x}-\mathbf{y}\rangle\right] \\
&= \frac{1}{2}\mathbb{E}\left[\max_{\ell\in[|T-T|]} X_\ell\right] \\
&\leq \frac{1}{2}\left(2\sqrt{\left(\frac{1}{2}\max_{\mathbf{x},\mathbf{y}\in T}\|\mathbf{x}-\mathbf{y}\|_2^2\right)\ln M^2}\right) \\
&= \operatorname{diam}(T)\sqrt{\ln M}
\end{aligned}$$

where we have applied Theorem 5.1.15.

Example 5.1.17 (Gaussian Width of Standard Basis)**.** Let $T = \{\pm\mathbf{e}_j\}_{j=1}^N \subset \mathbb{R}^N$ be the set of standard basis vectors and their opposites in $\mathbb{R}^N$. Observe

$$\begin{aligned}
w(T) &= \mathbb{E}\left[\max_{j\in[N]}\max\left(\langle \mathbf{g}, \mathbf{e_j}\rangle, \langle \mathbf{g}, -\mathbf{e_j}\rangle\right)\right] \\
&= \mathbb{E}\left[\max_{j\in[N]}|g_j|\right]
\end{aligned}$$

Theorem 5.1.18. *Let* $\mathbf{g} \sim \mathcal{N}(0, I_{N\times N})$ *for* $N \geq 2$. *Then* $\mathbb{E}[\|\mathbf{g}\|_\infty] \geq 0.265\sqrt{\ln N}$.

.2.6

Proof. Consider the following upper bound on the probability that the absolute value of a gaussian random variable is less than some positive number δ

$$\begin{aligned}
P[|g| < \delta] &= 1 - P[|g| \geq \delta] \\
&= 1 - P[g < -\delta \cup g > \delta] \\
&= 1 - 2P[g > \delta] \\
&= 1 - \sqrt{\frac{2}{\pi}} \int_\delta^\infty e^{-\frac{t^2}{2}} dt \\
&\leq 1 - \sqrt{\frac{2}{\pi}} \int_\delta^{2\delta} e^{-\frac{t^2}{2}} dt \\
&\leq 1 - \sqrt{\frac{2}{\pi}} \int_\delta^{2\delta} e^{-\frac{(2\delta)^2}{2}} dt \\
&\leq 1 - \sqrt{\frac{2}{\pi}} \delta e^{-2\delta^2}
\end{aligned}$$

Thus

$$P[|g| < \delta] \leq 1 - \sqrt{\frac{2}{\pi}} \delta e^{-2\delta^2} \tag{5.1}$$

Now observe the following calculation for the expectation of the infinity norm of a Gaussian vector

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}\|_\infty] &= \mathbb{E}\left[\max_{j\in[N]} |g_j|\right] \\
&= \int_0^\infty P\left[\max_{j\in[N]} |g_j| \geq t\right] dt \\
&= \int_0^\infty \left(1 - P\left[\max_{j\in[N]} |g_j| < t\right]\right) dt \\
&= \int_0^\infty \left(1 - P\left[\bigcap_{j\in[N]} |g_j| < t\right]\right) dt
\end{aligned}$$

We have that the entries are i.i.d guassian and so we can apply the product property

$$
\begin{aligned}
&= \int_0^{\infty} \left(1 - \left(P\left[|g| < t\right]\right)^N\right) dt \\
&\geq \int_0^{\delta} \left(1 - \left(P\left[|g| < t\right]\right)^N\right) dt \\
&\geq \int_0^{\delta} \left(1 - \left(P\left[|g| < \delta\right]\right)^N\right) dt \\
&= \delta \left(1 - \left(P\left[|g| < \delta\right]\right)^N\right)
\end{aligned}
$$

applying 5.1 from above we obtain

$$
\geq \delta \left(1 - \left(1 - \sqrt{\frac{2}{\pi}}\delta e^{-2\delta^2}\right)^N\right)
$$

Now let $\delta = \sqrt{\frac{\ln N}{2}}$

$$
\begin{aligned}
&= \sqrt{\frac{\ln N}{2}} \left(1 - \left(1 - \sqrt{\frac{\ln N}{\pi}}\frac{1}{N}\right)^N\right) \\
&\geq \sqrt{\frac{\ln N}{2}} \left(1 - \exp\left(-\sqrt{\frac{\ln N}{\pi}}\right)\right) \\
&\geq \frac{1 - \exp\left(-\sqrt{\frac{\ln 2}{\pi}}\right)}{\sqrt{2}} \ln N \\
&\geq 0.265 \ln N
\end{aligned}
$$

□

Homework 5.1.1. Prove that $f_T(\mathbf{z}) = \sup_{\mathbf{x}\in T}\langle \mathbf{z}, \mathbf{x}\rangle$ is continuous whenever $T \subseteq \mathbb{R}^N$ is bounded.

Homework 5.1.2. Prove that $\exists C \in \mathbb{R}^+$ such that $|f(\mathbf{z})| \leq C\|\mathbf{z}\|_2$, $\forall \mathbf{z} \in \mathbb{R}^N$ whenever T is bounded.

Homework 5.1.3. Prove Lemma 5.1.2

Homework 5.1.4. Prove Lemma 5.1.3

Homework 5.1.5. Prove Lemma 5.1.7

Homework 5.1.6. Prove Lemma 5.1.10

Homework 5.1.7. Let $B_N^{\ell^1}$ be the ℓ^1 unit ball in $\mathbb{R}^N$.

1. Show that $w(B_N^{\ell^1}) = \mathbb{E}\left[\|\mathbf{g}\|_\infty\right], \mathbf{g} \sim \mathcal{N}(0, I_{N\times N})$
2. $0.265\sqrt{\ln N} \leq w(B_N^{\ell^1}) \leq 2\sqrt{\ln 2N}$
3. Show that $\exists T \subset \mathbb{R}^N$ such that upper bound in Example 5.1.16 is tight up to a constant.

5.2 Covering Number and Gaussian Widths (MTH 994 Lecture 9)

Definition 5.2.1. A mean 0 Gaussian process, $(X_t)_{\mathbf{t}\in T}$ is a collection of random variables with the property that for all finite subsets $T_0 \subset T \subset \mathbb{R}^N$ there exists $\sigma \in \mathbb{R}^+$ where

$$\sum_{t\in T_0} a_t X_t \sim \mathcal{N}(0, \sigma)$$

Example 5.2.2. The collection $(\langle \mathbf{g}, \mathbf{t}\rangle)\, \mathbf{t} \in T$ where $\mathbf{g} \sim \mathcal{N}(0, I_{N\times N})$ is a mean zero guassian process $\forall T \subseteq \mathbb{R}^N$. To see why, consider

1. $\langle \mathbf{g}, \mathbf{t}\rangle \sim \mathcal{N}(0, \|\mathbf{t}\|_2)$
2. $a\langle \mathbf{g}, \mathbf{t}\rangle \sim \mathcal{N}(0, a\|\mathbf{t}\|_2)$
3. $\sum_{\mathbf{t}\in T_0} a_{\mathbf{t}}\langle \mathbf{g}, \mathbf{t}\rangle \sim \mathcal{N}(0, \|\sum t \in T_0 a_{\mathbf{t}}\mathbf{t}\|_2)$

Note that $\langle \mathbf{g}, \mathbf{t}\rangle$ and $\langle \mathbf{g}, \mathbf{s}\rangle$ for $\mathbf{t}, \mathbf{s} \in T_0$ are not independent unless $\text{supp}(\mathbf{t}) \cap \text{supp}(\mathbf{s}) = \emptyset$. Of course our definition of Gaussian width is the supremum of this collection of

random variables.

$$\mathbb{E}\left[\sup_{\mathbf{t}\in T}\langle \mathbf{g}, \mathbf{t}\rangle\right]$$

Example 5.2.3. Given $T \subseteq \mathbb{R}^N$ let $X_{\mathbf{t}} \sim \mathcal{N}(0, \sigma)$ where $X_{\mathbf{t}}$ and $X_{\mathbf{s}}$ are independent when $\mathbf{t} \neq \mathbf{s}$ then

$$\sum_{t\in T_0} a_{\mathbf{t}} X_{\mathbf{t}} \sim \mathcal{N}(0, \sigma\sqrt{\sum_{\mathbf{t}\in T_0} a_{\mathbf{t}}^2})$$

Theorem 5.2.4 (Sudahov-Fernique Inequality)**.** *Let* $(X)_{\mathbf{t}\in T}$ *and* $(Y)_{\mathbf{t}\in T}$ *be mean zero Gaussian processes. Assume that*

$$\mathbb{E}\left[(X_{\mathbf{t}} - X_{\mathbf{s}})^2\right] \leq \mathbb{E}\left[(Y_{\mathbf{t}} - Y_{\mathbf{s}})^2\right]$$

then

$$\mathbb{E}\left[\sup_{\mathbf{t}\in T} X_{\mathbf{t}}\right] \leq \mathbb{E}\left[\sup_{\mathbf{t}\in T} Y_{\mathbf{t}}\right]$$

Proof. In light of Theorem 5.2.4, consider two random variables, $X_{\mathbf{t}}, Y_{\mathbf{t}}$ of the form seen in Example 5.2.2, $X_{\mathbf{t}} = \langle \mathbf{g}_1, \mathbf{t}\rangle$, $Y_{\mathbf{t}} = \langle \mathbf{g}_2, \mathbf{t}\rangle$ for all $\mathbf{t} \in T \subseteq \mathbb{R}^N$ where $\mathbf{g}_i \sim \mathcal{N}(0, \sigma_i)$, $i = 1, 2$ then $\sigma_1 \leq \sigma_2$ if and only if

$$\begin{aligned}\mathbb{E}\left[(X_{\mathbf{t}} - X_{\mathbf{s}})^2\right] &= \sigma_1^2\|\mathbf{t} - \mathbf{s}\|_2^2 \\ &\leq \sigma_2^2\|\mathbf{t} - \mathbf{s}\|_2^2 \\ &= \mathbb{E}\left[(Y_{\mathbf{t}} - Y_{\mathbf{s}})^2\right]\end{aligned}$$

for all $\mathbf{t}, \mathbf{s} \in T$. Note then

$$\begin{aligned}\mathbb{E}\left[|X_t|\right] &= \int_0^\infty P\left[|X_t| \geq a\right] da \\ &\leq \int_0^\infty P\left[|Y_t| \geq a\right] da \qquad\qquad = \mathbb{E}\left[|Y)t|\right]\end{aligned}$$

and so it is a reasonable suggestion that the supremum then would also hold, i.e.

$$\mathbb{E}\left[\sup_{\mathbf{t}\in T} X_{\mathbf{t}}\right] \leq \mathbb{E}\left[\sup_{\mathbf{t}\in T} Y_{\mathbf{t}}\right]$$

See section 7.2.3 in **??** for a full proof of this result. □

Lemma 5.2.5. *Let $A \in \mathbb{R}^{m\times N}$, $T \subset \mathbb{R}^N$ then*

$$\begin{aligned} w(AT) &\leq \|A\|_{T-T,2\to 2} w(T) \\ &\leq \|A\|_{2\to 2} w(T) \end{aligned}$$

Proof. Naturally, $\|A\|_{T-T,2\to 2} \leq \|A\|_{2\to 2}$ since $T - T \subset \mathbb{R}^N$. We will apply Theorem 5.2.4 on a well chosen $X_{\mathbf{t}}$ and $Y_{\mathbf{t}}$ to achieve the first bound on $w(AT)$). Consider $Y_{\mathbf{t}} = \|A\|_{T-T,2\to 2}\langle \mathbf{g}, \mathbf{t}\rangle$ and $X_{\mathbf{t}} = \langle \mathbf{g}, A\mathbf{t}\rangle$

$$\begin{aligned} \mathbb{E}\left[(Y_{\mathbf{t}} - Y_{\mathbf{s}})^2\right] &= \mathbb{E}\left[(\langle \mathbf{g}, \|A\|_{T-T,2\to 2}(\mathbf{t}-\mathbf{s})\rangle)^2\right] \\ &= \left(\|A\|_{T-T,2\to 2}\|_2\right)^2 \|\mathbf{t}-\mathbf{s}\|_2^2 \\ &\geq \|A(\mathbf{t}-\mathbf{s})\|_2^2 \\ &= \mathbb{E}\left[(\langle \mathbf{g}, A(\mathbf{t}-\mathbf{s})\rangle)^2\right] \\ &= \mathbb{E}\left[(X_{\mathbf{t}} - X_{\mathbf{s}})^2\right] \end{aligned}$$

and thus we can apply Theorem 5.2.4 and conclude

$$w(AT) = E\left[\sup_{\mathbf{t}\in T}\langle \mathbf{g}, A\mathbf{t}\rangle\right] \leq E\left[\sup_{\mathbf{t}\in T}\langle \mathbf{g}, \|A\|_{T-T,2\to 2}\mathbf{t}\rangle\right] = \|A\|_{T-T,2\to 2} w(T)$$

□

Theorem 5.2.6 (Subakov's Minorization for Gaussian Widths)**.** *Let $T \subset \mathbb{R}^N$ be bounded. Then $\forall \epsilon \geq 0$ we have*

$$w(T) \geq C\epsilon\sqrt{\log C_\epsilon^{\ell^2}(T)}$$

where $C \in \mathbb{R}^+$ is an absolute universal constant independent of both ϵ and T bounded below by $\frac{0.265}{\sqrt{2}}$

Proof. Let P_ϵ be a maximal ϵ-packing of T, and its cardinality (packng number) denoted $P_\epsilon^{\ell^2}$ is finite by Lemma 3.2.6 and by Lemma 3.2.5 we have $P_\epsilon^{\ell^2} \leq C_\epsilon^{\ell^2}$.

Since $P_\epsilon \subseteq T$ we have that $w(T) \leq w(P_\epsilon)$ by Lemma 5.1.3. Now consider the Gaussian process, $(Y_{\mathbf{t}})_{\mathbf{t}\in P_\epsilon} = \{\langle \mathbf{g}, \mathbf{t}\rangle\}_{\mathbf{t}\in P_\epsilon}$. We will compare that to a new Gaussian process $(X_{\mathbf{t}})_{\mathbf{t}\in P_\epsilon} = \left\{\frac{\epsilon}{\sqrt{2}} g_{\mathbf{t}}\right\}_{\mathbf{t}\in P_\epsilon}$ where $g_{\mathbf{t}} \sim \mathcal{N}(0,1)$ are i.i.d for each distinct $\mathbf{t} \in P_\epsilon$ (i.e. white noise). Observe on one hand, since $Y_{\mathbf{t}} - Y_{\mathbf{s}} \sim \mathcal{N}(\mathbf{0}, \|\mathbf{t}-\mathbf{s}\|_2)$ for $\mathbf{t} \neq \mathbf{s}$:

$$\begin{aligned}\mathbb{E}\left[(\langle \mathbf{g}, \mathbf{t}-\mathbf{s}\rangle)\right] &= \mathrm{Var}\left[Y_t - Y s\right] \\ &= \|\mathbf{t}-\mathbf{s}\|_2^2 \\ &\geq \epsilon^2\end{aligned}$$

since $\mathbf{t}, \mathbf{s}$ are distinct points in P_ϵ. On the other hand, using independence we see:

$$\begin{aligned}\mathbb{E}\left[(X_{\mathbf{t}-X_{\mathbf{s}}})^2\right] &= \mathbb{E}\left[X_{\mathbf{t}}^2\right] - 2\mathbb{E}\left[X_{\mathbf{t}}\right]\mathbb{E}\left[X_{\mathbf{s}}\right] + \mathbb{E}\left[X_{\mathbf{s}}^2\right] \\ &= \epsilon^2\end{aligned}$$

And so

$$\mathbb{E}\left[(X_{\mathbf{t}} - X_{\mathbf{s}})^2\right] \leq \mathbb{E}\left[(Y_{\mathbf{t}} - Y_{\mathbf{s}})^2\right]$$

and thus using the definition of Gaussian width and Theorem 5.2.4, and denoting $\mathbf{g} \in \mathbb{R}^{P_\epsilon^{\ell^2}}$ the Gaussian random vector with entries $(\mathbf{g})_{\mathbf{t}} = g_{\mathbf{t}}$

$$\begin{aligned}w(P_\epsilon) &= \mathbb{E}\left[\sup_{\mathbf{t}\in T} Y_{\mathbf{t}}\right] \\ &\geq \mathbb{E}\left[\sup_{\mathbf{t}\in T} X_{\mathbf{t}}\right] \\ &= \frac{\epsilon}{\sqrt{2}}\mathbb{E}\left[\|\mathbf{g}\|_\infty\right] \\ &\geq \frac{\epsilon}{\sqrt{2}}(0.265)\sqrt{\ln N}\end{aligned}$$

by an appeal to Theorem 5.1.18 □

We desire an upper bound of the Gaussian width of a set in terms of the covering number. First however we will need a lemma:

Lemma 5.2.7. *Let $T \subset \mathbb{R}^N$ be bounded. If $w(T_0) \leq C$ for all finite $T_0 \subseteq T$ then $w(T) \leq C$*

Proof. Choose $\epsilon > 0$ and let $C_\epsilon \subset T$ be a finite ϵ-cover of T. Then $\overline{T} \subset C_\epsilon + \epsilon\overline{B^N_{\ell^2}(0,1)}$ and

$$\begin{aligned} w(T) &= w(\overline{T}) \\ &\leq w(C_\epsilon + \epsilon\overline{B^N_{\ell^2}(0,1)}) \\ &= w(C_\epsilon) + w(\epsilon\overline{B^N_{\ell^2}(0,1)}) \\ &\leq C + \epsilon\sqrt{N} \end{aligned}$$

since this holds $\forall \epsilon > 0$ we get the result in the limit. □

Theorem 5.2.8 (Dudley Inequality for Gaussian Widths)**.** *Let $T \subset \mathbb{R}^N$ be bounded, and denote $\Delta = \sup_{\mathbf{x}\in T} \|\mathbf{x}\|_2 < \infty$ then*

1. $w(T) \leq 4\sqrt{2}\int_0^{\Delta/2} \sqrt{\ln\left(C_\epsilon^{\ell^2}(T)\right)}d\epsilon$

2. $\mathbb{E}\left[\sup_{\mathbf{t}\in T} |\langle \mathbf{g}, \mathbf{t}\rangle|\right] \leq 4\sqrt{2}\int_0^{\Delta/2} \sqrt{\ln\left(2C_\epsilon^{\ell^2}(T)\right)}d\epsilon$

Proof. Denote $\Delta = \sup_{\mathbf{x}\in T} \|\mathbf{x}\|_2$. We proceed in three parts.

Chaining definitions Pick any arbitrary finite subset $T_0 \subset T$. We will define a tree on T_0. The tree will provide a means to find a path from a root node $C_0 = \{\mathbf{0}\}$ to any (other) point in $\mathbf{x} \in T_0$ through a set of cover sets C_j with increasingly small covering radii. Towards this:

Let $\epsilon_j = \Delta 2^{-j}$. Naturally, by our definitions $\epsilon_0 \geq \|\mathbf{x} - \mathbf{0}\|_2 = \|\mathbf{x}\|_2$.

Let $C_j \subseteq T_0$ be a minimal ϵ_j-cover of T_0.

Since $|T_0|$ is finite, there exists some smallest natural number n where $C_n = T_0$ (e.g. $\arg\min_{n\in\mathbb{N}} \epsilon_n \leq \min_{\mathbf{x},\mathbf{y}\in T_0} \|\mathbf{x}-\mathbf{y}\|_2$)

Let $f_j : C_j \to C_{j-1}$ be a function that assigns $\mathbf{x} \in C_j$ to some point in C_{j-1} that is within ϵ_{j-1} distance, i.e.

$$\|\mathbf{x} - f_j(\mathbf{x})\|_2 \leq \epsilon_{j-1},\ \forall \mathbf{x} \in C_j$$

This is well defined since C_{j-1} defines a ϵ_j-cover of T_0.

Note then that we can create a path then from any point $\mathbf{x} \in T_0$ to $\mathbf{0}$ in the following manner then:

$$\mathbf{x}, f_n(\mathbf{x}), f_{n-1}(f_n(\mathbf{x})), \ldots, \bigcirc_{\ell=0}^{n} f_\ell(\mathbf{x}) = \mathbf{0}$$

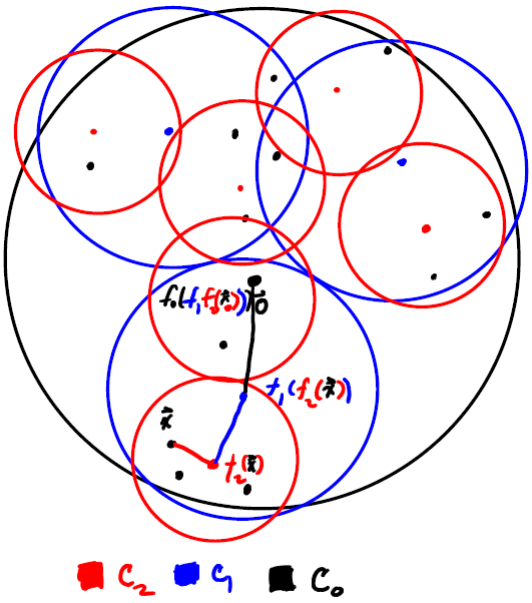


Figure 5.1: Schematic of chaining scheme and path

Induction on Gaussian Widths of Covers Consider the following hypothesis

$$w(C_j) \leq 2\sqrt{2\ln|C_1|}\epsilon_1 + \sum_{\ell=2}^{j} 4\sqrt{2\ln|C_\ell|}\,(\epsilon_\ell - \epsilon_{\ell+1})\ \forall j = 1,\ldots,n \tag{5.2}$$

We consider the base case, $j = 1$. First note $\langle \mathbf{g}, \mathbf{x} \rangle \sim \mathcal{N}(0, \|\mathbf{x}\|_2)$ and that $\mathbb{E}[\exp(\theta\langle \mathbf{g}, \mathbf{x}\rangle)] \leq \exp\left(\frac{\|\mathbf{x}\|_2^2}{2}\theta\right)$, for $\mathbf{g} \sim \mathcal{N}(0, I_{n\times n})$ - that is, we have Gaussian random variables that admit the subgaussian parameter $c = \frac{\|\mathbf{x}\|_2^2}{2}$ (see Lemma 4.2.6)

By definition of Gaussian width and Theorem 5.1.15

$$\begin{aligned} w(C_1) &= \mathbb{E}\left[\max_{\mathbf{x}\in C_1}\langle \mathbf{g}, \mathbf{x}\rangle\right] \\ &\leq \sqrt{4\left(\max_{\mathbf{x}\in C_1}\frac{\|\mathbf{x}\|_2^2}{2}\right)\ln|C_1|} \\ &\leq \Delta\sqrt{2\ln|C_1|} \\ &\leq 2\sqrt{2\ln|C_1|}\epsilon_1 \end{aligned}$$

Now assume that the hypothesis 5.2 holds for j and prove it holds for $j+1$. Consider

$$\begin{aligned} w(C_{j+1}) &= \mathbb{E}\left[\max_{\mathbf{x}\in C_{j+1}}\langle \mathbf{g}, \mathbf{x}\rangle\right] \\ &= \mathbb{E}\left[\max_{\mathbf{x}\in C_{j+1}}\langle \mathbf{g}, \mathbf{x} - f_{j+1}(\mathbf{x}) + f_{j+1}(\mathbf{x})\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\mathbf{x}\in C_{j+1}}\langle \mathbf{g}, \mathbf{x} - f_{j+1}(\mathbf{x})\rangle\right] + \mathbb{E}\left[\max_{\mathbf{x}\in C_{j+1}}\langle \mathbf{g}, f_{j+1}(\mathbf{x})\rangle\right] \end{aligned}$$

Now apply Theorem 5.1.15 to the random variables $\langle \mathbf{g}, \mathbf{x} - f_{j+1}(\mathbf{x})\rangle$ and note that $f_{j+1}(C_{j+1}) \subset C_j$ to obtain

$$\begin{aligned} &\leq \sqrt{4\left(\frac{1}{2}\max_{\mathbf{x}\in C_{j+1}}\|\mathbf{x} - f_{j+1}(\mathbf{x})\|_2^2\right)\ln|C_{j+1}|} + \mathbb{E}\left[\max_{\mathbf{x}\in C_j}\langle \mathbf{g}, \mathbf{x}\rangle\right] \\ &\leq \sqrt{4\left(\frac{1}{2}\max_{\mathbf{x}\in C_{j+1}}\|\mathbf{x} - f_{j+1}(\mathbf{x})\|_2^2\right)\ln|C_{j+1}|} + w(C_j) \end{aligned}$$

Apply now the induction hypothesis on $w(C_j)$

$$\leq \sqrt{4\left(\frac{1}{2}\max_{\mathbf{x}\in C_{j+1}}\|\mathbf{x} - f_{j+1}(\mathbf{x})\|_2^2\right)\ln|C_{j+1}|} + \sum_{\ell=2}^{j} 4\sqrt{2\ln|C_\ell|}\left(\epsilon_\ell - \epsilon_{\ell+2}\right) + 2\sqrt{2\ln|C_1|}\epsilon_1$$

Now note that

$$\epsilon_j = \Delta 2^{-j} = 4\Delta 2^{-j}\left(2^{-1} - 2^{-2}\right) = 4\left(\Delta 2^{-(j+1)} - \Delta 2^{-(j+2)}\right) = 4(\epsilon_{j+1} - \epsilon_{j+2})$$

So after a rearrangement of terms, we conclude

$$\begin{aligned} w(C_{j+1}) &\leq 2\sqrt{2\ln|C_1|}\epsilon_1 + \sum_{\ell=2}^{j+1} 4\sqrt{2\ln|C_\ell|}\,(\epsilon_\ell - \epsilon_{\ell+1}) \\ &= 4\sqrt{2\ln|C_1|}\,(\epsilon_1 - \epsilon_2) + \sum_{\ell=2}^{j+1} 4\sqrt{2\ln|C_\ell|}\,(\epsilon_\ell - \epsilon_{\ell+1}) \\ &= \sum_{\ell=1}^{j+1} 4\sqrt{2\ln|C_\ell|}\,(\epsilon_\ell - \epsilon_{\ell+1}) \end{aligned}$$

Integrating and bounding with T Using our claim from the previous section

$$\begin{aligned} w(C_n) &= w(T_0) \\ &\leq \sum_{\ell=1}^{j+1} 4\sqrt{2\ln|C_\ell|}\,(\epsilon_\ell - \epsilon_{\ell+1}) \\ &= 4\sqrt{2}\sum_{\ell=1}^{j+1}\int_{\epsilon_{\ell+1}}^{\epsilon_\ell} \sqrt{\ln|C_\ell|}d\epsilon \end{aligned}$$

Notice that $|C_\ell| = C_{\epsilon_\ell}^{\ell^2}(T_0) \leq C_\epsilon^{\ell^2}(T_0)$, $\epsilon \in [\epsilon_{\ell+1}, \epsilon_\ell]$

$$\begin{aligned} &\leq 4\sqrt{2}\sum_{\ell=1}^{j+1}\int_{\epsilon_{\ell+1}}^{\epsilon_\ell} \sqrt{\ln C_\epsilon^{\ell^2}(T_0)}d\epsilon \\ &= 4\sqrt{2}\int_{\epsilon_{n+1}}^{\Delta/2} \sqrt{\ln C_\epsilon^{\ell^2}(T_0)}d\epsilon \\ &\leq 4\sqrt{2}\int_{0}^{\Delta/2} \sqrt{\ln C_\epsilon^{\ell^2}(T_0)}d\epsilon \end{aligned}$$

$C_\epsilon^{\ell^2}(T_0) \leq C_\epsilon^{\ell^2}(T)$

$$\leq 4\sqrt{2}\int_{0}^{\Delta/2} \sqrt{\ln C_\epsilon^{\ell^2}(T)}d\epsilon$$

That is, we have a uniform bound on $w(T_0)$ for all $|T_0| < \infty$ and so by Lemma 5.2.7 we conclude that

$$w(T) \leq 4\sqrt{2}\int_{0}^{\Delta/2} \sqrt{\ln C_\epsilon^{\ell^2}(T)}d\epsilon$$

The proof of part 2 of the theorem statement is left as an exercise. □

Theorem 5.2.9. *Let $T \subset \mathbb{R}^N$ and $\Phi \in \mathbb{R}^{m\times N}$ have independent, mean zero, variance one subgaussian entries. Then with high probability we have*

$$\|\mathbf{x}-\mathbf{y}\|_2 - \delta \leq \|\frac{1}{\sqrt{m}}\Phi(\mathbf{x}-\mathbf{y})\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2 + \delta$$

holds $\forall \mathbf{x}, \mathbf{y} \in T$ where $\delta = \frac{Cw(T)}{\sqrt{m}}$ for an absolute constant $C > 0$ that only depends on the subgaussian parameters of the entries of Φ.

Proof. See Proposition 9.3.2 [32] □

The constant C can be understood in terms of the subgaussian parameters of Φ by looking at Lemma 4.2.1 and Propositions 2.5.2 and Lemma 3.4.2 in [32].

Corollary 5.2.10. *Let $S \subset \mathbb{R}^N$ be a finite set and $\Phi \in \mathbb{R}^{m\times N}$ have independent mean zero, variance one, subgaussian entries. Then with a fixed probability we have that*

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|\frac{1}{\sqrt{m}}\Phi(\mathbf{x}-\mathbf{y})\|_2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2$$

holds $\forall \mathbf{x}, \mathbf{y} \in S$ whenever

$$m \geq \frac{c' \ln(1+|S|)}{\epsilon^2}$$

for an absolute constant $c' > 0$ that is independent of the set S

Proof. Apply Theorem 5.2.9 to the set

$$T = \left\{ \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2} | \mathbf{x}, \mathbf{y} \in S, \mathbf{x} \neq \mathbf{y} \right\} \cup \{\mathbf{0}\}$$

and so we have with high probability that

$$\left| \|\frac{1}{\sqrt{m}}\Phi(\mathbf{u}-\mathbf{v})\|_2 - \|\mathbf{u}-\mathbf{b}\|_2 \right| \leq \delta$$

holds $\forall \mathbf{u}, \mathbf{v} \in T$. In particular, when $\mathbf{u} = \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2}$ and $\mathbf{v} = \mathbf{0}$ which yields

$$\left| \|\frac{1}{\sqrt{m}}\Phi(\mathbf{x}-\mathbf{y})\|_2 - \|\mathbf{x}-\mathbf{y}\|_2 \right| \leq \delta\|\mathbf{x}-\mathbf{y}\|_2$$

To see that the lower bound on the dimension m holds we use Example 5.1.16 to see that

$$\delta = \frac{cw(T)}{\sqrt{m}} \leq \epsilon\frac{\tilde{c}}{\sqrt{c'}}\frac{\operatorname{diam}(T)\ln(|T|)}{\sqrt{\ln(1+|S|)}}$$

but $T \subseteq B^{\ell^2}(0,1)$ and $|T| \leq |S|^2$ thus for large enough c' then $\delta \leq \epsilon$ □

Definition 5.2.11. We say that $\Phi \in \mathbb{C}^{m\times N}$ has the generalized Restricted Isometry Property of order (s,δ) for $\delta > 0$ and sparsity $0 < s < N$ if

$$\left| \|\Phi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \leq \max\left(\delta, \delta^2\right)\|\mathbf{x}\|_2^2$$

holds $\forall \mathbf{x} \in \mathbb{C}^N$ with $\|\mathbf{x}\|_0 \leq s$

Recall that if Φ has RIP of order $(2\left\lceil \frac{s}{\delta} \right\rceil, \frac{1}{2})$ then it has the generalized RIP of order (s,δ) for all $\delta \geq 1$ by Lemma 4.4.2 - so matrices with the property defined in 5.2.11 exist for any choice of parameters (s,δ) by applying constructions for $\epsilon < 1$ as stated in Theorem 4.2.10.

Additionally, Theorem 4.4.1 asserts that $\sqrt{\frac{N}{m}}RFD$ matrices have the generalized RIP of order (s,δ) with probability at least $1-e^{-\eta}$ whenever

$$m \geq c\left\lceil \frac{s}{2\delta^2} \right\rceil \left(\ln^4 N + \eta\right)$$

Definition 5.2.12. Let $L = \lceil \log_2 N \rceil$, $\delta > 0$ and $s \in [N] \setminus \{0\}$. For $\ell \in [L+1]$ let $(s_\ell, \delta_\ell) = (2^\ell s, 2^{\ell/2}\delta)$ be a sequence of sparsity and distortion levels. We will say that $\Phi \in \mathbb{R}^{m\times N}$ satisfies the multi-resolution RIP (MRIP) of order (s,δ) if it satisfies the generalized RIP of order $(2^\ell s, 2^{\ell/2}\delta)$, $\forall \ell \in [L+1]$.

Note that Theorem 4.4.1 implies an $\sqrt{\frac{N}{m}}RF$ matrix will have the MRIP of order (s,δ) with probability $1-e^{-\eta}$ provided

$$m \geq c\left\lceil \frac{s}{2\delta^2} \right\rceil (1+\eta)\log^4 N, \forall N > 1$$

Theorem 5.2.13. *Let $T \subset \mathbb{R}^N$ have $r(T) = \sup_{\mathbf{v}\in T}\|\mathbf{v}\|_2 \leq \infty$. Suppose that Φ has MRIP of order $\left(200(1+\eta), \frac{\delta r(T)}{c\max(r(T),w(T))}\right)$ where $c>0$ is an absolute constant. Then for $\psi \in \{-1,1\}^N$, a vector with i.i.d. uniform Radamacher entries.*

$$\sup_{\mathbf{x}\in T}\left|\|\Phi Diag(\psi)\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \max(\delta,\delta^2)r(T)^2$$

will hold with probability $1-e^{-\eta}$.

Note that if we want to state Theorem 5.2.13 with the usual multiplicative error guarantees for some arbitrary set $S \subset \mathbb{R}^N$, apply the theorem to the set

$$T = \left\{ \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2} | \mathbf{x},\mathbf{y} \in S, \mathbf{x} \neq \mathbf{y} \right\} \cup \{\mathbf{0}\}$$

Note then that $r(T) = 1$ and we have the bound

$$\left(1-\max(\delta,\delta^2)\right)\|\mathbf{x}-\mathbf{y}\|_2^2 \leq \sup_{\mathbf{x}\in T}\|\Phi \text{Diag}(\psi)(\mathbf{x}-\mathbf{y})\|_2^2 \leq \left(1+\max(\delta,\delta^2)\right)\|\mathbf{x}-\mathbf{y}\|_2^2$$

Homework 5.2.1. Show that

1. $C_\epsilon^{\ell^2}\left(B_{\ell^2}^N(0,1)\right) \leq e^{\frac{N}{c^2\epsilon^2}}$ for any $\epsilon > 0$. Is this a better bound than the bound found in Corollary 3.2.7 for any ϵ?
2. Use part 1.) to bound $C_\epsilon^{\ell^2}\left(B_{\ell^1}^N(0,1)\right)$
3. Let $T \subseteq \mathbb{R}^N$ be finite. Prove that $w(T) \geq \frac{0.265}{\sqrt{2}}\left(\min_{\mathbf{x},\mathbf{y}\in T}\|\mathbf{x}-\mathbf{y}\|_2\right)\sqrt{\ln|T|}$

Homework 5.2.2. Prove part 2 of Theorem 5.2.8.

Homework 5.2.3. Prove that the Gaussian complexity, $\gamma(T) = \mathbb{E}\left[\sup_{\mathbf{x}\in T} |\langle \mathbf{g}, \mathbf{x}\rangle|\right]$ satisfies

1. $w(T) = \frac{1}{2}w(T-T) = \frac{1}{2}\gamma(T-T)$
2. Recall that $\mathbb{E}\left[|\langle \mathbf{g}, \mathbf{z}\rangle|\right]$ where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, I) = \sqrt{\frac{2}{\pi}}\|\mathbf{z}\|_2$ (see Definition 2.1.7). show that

$$\gamma(T) \leq 2w(T) + \sqrt{\frac{2}{\pi}}\|\mathbf{y}\|_2$$

holds $\forall \mathbf{y} \in T$

Homework 5.2.4. Show that Theorem 4.4.1 implies an $\sqrt{\frac{N}{m}}RF$ matrix will have the MRIP of order (s, δ) with probability $1 - e^{-\eta}$ provided

$$m \geq c\left\lceil \frac{s}{2\delta^2} \right\rceil (1+\eta)\log^4 N, \forall N > 1$$

Chapter VI

Sublinear-Time Compressive Sensing, Sparse Fourier Transforms, and the Fast Approximation of Functions of Many Variables (MTH 994 Lectures 11 – ∞, Partially Transcribed by Craig Gross)

6.1 "Slow" combinatorial compressive sensing using binary low coherence matrices

We will now discuss how coherence propertries can be combined with deterministic constructions to achieve compressive sensing with quadratic dependence on sparsity s (compared with linear dependence for probabilistic approaches seen so far) but which do offer faster recovery and also guarantee error bounds (i.e. no chance of failure)

Definition 6.1.1 (Coherence)**.** Let $\Phi \in \mathbb{C}^{m \times N}$ be a matrix with ℓ^2-normalized columns $\boldsymbol{\varphi}_0, \dots, \boldsymbol{\varphi}_{N-1}$ having $\|\phi_j\| = 1 \, \forall j \in [N]$. The coherence $\mu(\Phi) = \max_{i \neq j} |\langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i|$

Note that if U is orthonormal, then $\mu(U) = 0$. If U contains two identical columns then $\mu(U) = 1$. So we see that $\mu(\Phi) \in [0, 1]$. Compared to other matrix properties of interest for compressive sensing settings, coherence is easy to compute.

We will show how low coherence matrices have the RIP property, and in turn can be used as JL maps. In order to do this, we will need the following classic theorem from linear algebra.

Theorem 6.1.2 (Gerschgorin Disc)**.** *Let λ be an eigenvalue of a square matrix $A \in \mathbb{C}^{N\times N}$. Then there exists an index $j \in [N]$ such that*

$$|\lambda - A_{jj}| \leq \sum_{\ell\in[N]\setminus\{j\}} |A_{j\ell}|$$

Proof. Let $(\lambda, \mathbf{u})$ be an eigenpair of the matrix. Let j be the index corresponding to the largest entry of the eigenvector, i.e. $|u_j| = \|\mathbf{u}\|_\infty$. Then $\sum_{\ell\in[N]} A_{j\ell}u_\ell = \lambda u_j$. That is since $A\mathbf{u} = \lambda\mathbf{u}$, the j-th entry of the vector $A\mathbf{u}$ is the inner product of the j-th row of A with $\mathbf{u}$ scaled by λ.

Now, moving the term A_jju_j to the other side, we obtain

$$\sum_{\ell\in[N]\setminus\{j\}} A_{j\ell}u_\ell = (\lambda - A_jj)u_j$$

Using the triangle inequality and bounding with the infinity norm, we have

$$|\lambda - A_{jj}||u_j| \leq \sum_{\ell\in[N]\setminus\{j\}} |A_{j\ell}u_\ell| \implies |\lambda - A_{jj}||u_j| \leq u_j \sum_{\ell\in[N]\setminus\{j\}} |A_{j\ell}|$$

Dividing each side then by $|u_j|$ yields the desired result. Note that every eigenpair may have a different center and radius, depending on which entry of the eigenvector is of largest magnitude. □

Corollary 6.1.3. *Every eigenvalue of A lies in at least one of the N circular disks in the complex plane with centers A_{jj} and radii $\sum_{i\neq j} |A_{ij}|$. Moreover if m of these disks form a connected domain that is disjoint from the other $N-m$ disks, then there are m eigenvalues of A within the domain.*

Theorem 6.1.4. *Let $\Phi \in \mathbb{C}^{m\times N}$ be a matrix with ℓ^2-normalized columns, take $s \in [N]$. Then $\forall s$-sparse vectors $\mathbf{x} \in \mathbb{C}^N$*

$$(1 - (s-1)\mu(\Phi))\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2^2 \leq (1 + (s-1)\mu(\Phi))\|\mathbf{x}\|_2^2$$

Proof. Let $S \subset [N]$, $|S| \leq s$. Then the matrix $\Phi_S^* \Phi_S \in \mathbb{C}^{s \times s}$ formed by omitting all columns of Φ not in S is positive semi-definite. Denote its largest and smallest eigenvalues as $\lambda_{\max}, \lambda_{\min}$. If $\mathbf{x}$ is s-sparse and $S = \text{supp}(\mathbf{x})$ then

$$\begin{aligned} \|\Phi \mathbf{x}\|_2^2 &= \|\Phi_S \mathbf{x}_S\|_2^2 \\ &= \langle \Phi_S \mathbf{x}_S, \Phi_S \mathbf{x}_S \rangle \\ &= \langle \Phi_S^* \Phi \mathbf{x}_S, \mathbf{x}_S \rangle \\ \leq \lambda_{\max} \|\mathbf{x}_S\|_2^2 & \\ \leq \lambda_{\max} \|\mathbf{x}\|_2^2 & \end{aligned}$$

In a similar fashion we can show that $\|\Phi \mathbf{x}\|_2^2 \geq \lambda_{\min} \|\mathbf{x}\|_2^2$. That is we have

$$\lambda_{\min} \|\mathbf{x}\|_2^2 \leq \|\Phi \mathbf{x}\|_2^2 \leq \lambda_{\max} \|\mathbf{x}\|_2^2$$

Gerschgorin's Disc theorem implies that there exists some index $j \in S$ such that

$$\left| \lambda - (\Phi_S^* \Phi_S)_{jj} \right| \leq \sum_{i \neq j}^{s} \left| (\Phi_S^* \Phi_S)_{ij} \right|$$

However, we know that $(\Phi_S^* \Phi_S)_{ij} = \langle \boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j \rangle$, so

$$|\lambda - 1| \leq \sum_{i \neq j}^{s} |\langle \boldsymbol{\varphi}_i, \boldsymbol{\varphi}_j \rangle| \leq (s-1)\mu(\Phi)$$

□

Theorem 6.1.4 immediately implies the following

1. Φ has the RIP of order $(s, (s-1)\mu(\Phi))$.
2. For $\Phi_S^* \Phi_S \in \mathbb{C}^{N \times N}$, all of its non-zero eigenvalues are contained in the interval $[1 - (s-1)\mu(\Phi_S^* \Phi_S), 1 + (s-1)\mu(\Phi_S^* \Phi_S)]$

Definition 6.1.5. Let $K, \alpha \in [N] := \{0, \ldots, N-1\}$. A matrix $A \in \{0,1\}^{m \times N}$ is (K, α)-*coherent* if the following conditions hold:

1. Every column of A contains at least K ones, and
2. For every $j, \ell \in [N]$, $j \neq \ell$, the inner product of the columns $\mathbf{a}_j$ and $\mathbf{a}_\ell$ satisfies $\langle \mathbf{a}_j, \mathbf{a}_\ell \rangle \leq \alpha$.

Homework 6.1.1. Fix $\omega \in [N] \setminus \{0\}$ and let $X_\ell = \exp\left(\frac{2\pi i u_\ell \omega}{N}\right)$ where u_ℓ are i.i.d. uniformly in $[N]$ random variables $\forall \ell \in [m]$.

1. Prove that $0 = \mathbb{E}\left[\frac{1}{m}\Re(X_\ell)\right] = \mathbb{E}\left[\frac{1}{m}\Im(X_\ell)\right]$
2. Use Theorem 4.1.2 twice to show that
$$P\left[\frac{1}{m}\left|\sum_{\ell=1}^{m} X_\ell\right| \geq t\right] \leq \frac{p}{N-1}$$
for any choice of $p \in (0,1)$ provided $m \geq \frac{4}{t^2}\ln\frac{4(n-1)}{p}$
3. Let $A \in \mathbb{C}^{m\times N}$ be given by
$$A_{\ell,\omega} = \frac{1}{\sqrt{m}}\exp\left(\frac{2\pi i u_\ell \omega}{N}\right),\ \ell \in [m],\ \omega \in [N]$$
Show that the columns of A are ℓ^2-normalized and that the coherence of $\mu(A) < \epsilon$ with probability greater than $1-p$ provided
$$m \geq \frac{4}{\epsilon^2}\ln\left(\frac{4(N-1)}{p}\right)$$
4. Show that A has the RIP of order (s,ϵ) for A with high probability when
$$m \geq C\frac{s^2}{\epsilon^2}\ln\left(\frac{4(N-1)}{p}\right)$$

Goal. *Keep α small while making K large.*

Proposition 6.1.6 (Welch bound)**.** *For a matrix $A \in \{0,1\}^{m\times N}$, the coherence satisfies*
$$\max_{1\leq j\neq \ell\leq N}\left|\left\langle \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|_2}, \frac{\mathbf{a}_\ell}{\|\mathbf{a}_\ell\|_2}\right\rangle\right| \geq \sqrt{\frac{N-m}{m(N-1)}}.$$

The Welch bound then gives a lower bound on the number of rows for a (K, α)-coherent matrix:

$$\frac{\alpha}{K} \geq \sqrt{\frac{N-m}{m(N-1)}} \implies m \geq \frac{K^2}{\alpha^2}\frac{N-m}{N-1}.$$

When, for example, $m \leq N/2$ (which we henceforth assume), we must then have $m = \Omega(K^2/\alpha^2)$.

Example 1 ([18], Theorem 2)**.** *Fix some probability threshold $\sigma \in [0, 1)$, and generate $M \in \{0, 1\}^{m \times N}$ where each entry is i.i.d. Bernoulli, i.e.,*

$$m_{i,j} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases}$$

where

$$p = \frac{\log_{4/\mathrm{e}}\left(\frac{3N^2}{1-\sigma}\right)}{K\,(1+o(1))}$$

for some $K \geq \alpha \geq 2\log_{4/\mathrm{e}}\left(\frac{3N^2}{1-\sigma}\right)$. Then M will be (K, α)-coherent with probability at least σ provided that $m \geq cK^2/\alpha$.

6.1.1 Deterministic block constructions

Example 2 ([10] and [21])**.**

1. *Choose a prime $p \in [N]$.*

2. *For each $j \in [N]$ we will consider the representation of j in base p, denoted*

$$j = j_0 + j_1 p + j_2 p^2 + \cdots + j_{\lceil \log_p N \rceil - 1} p^{\lceil \log_p N \rceil - 1},$$

   where $j_0, \ldots, j_{\lceil \log_p N \rceil - 1} \in [p]$.

3. *Now, we map every $j \in [N]$ to the polynomial*

$$Q_j(x) := j_0 + j_1 x + \cdots + j_{\lceil \log_p N \rceil - 1} x^{\lceil \log_p N \rceil - 1},$$

   over the finite field $\mathbb{Z}_p$.

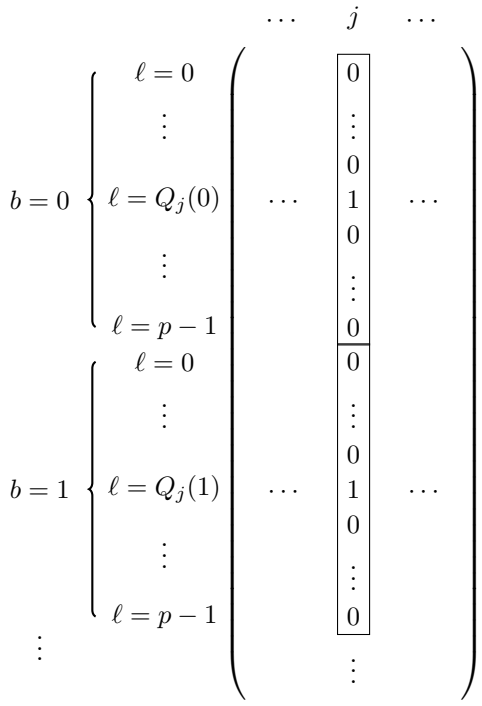


Figure 6.1: Deterministically constructed low coherence matrix M by the process described in 2.

4. Define $M = (\mathbf{m}_0, \ldots, \mathbf{m}_{N-1}) \in \{0,1\}^{p^2 \times N}$ *by*

$$m_{\ell+bp,j} = \begin{cases} 1 & \text{if } Q_j(b) = \ell, \\ 0 & \text{otherwise,} \end{cases}$$

of the form depicted in 6.1.

We now consider the coherence of M*,*

$$\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle = |\{b \in [p] : Q_j(b) = Q_\ell(b) \iff (Q_j - Q_\ell)(b) = 0\}|.$$

Since $Q_j - Q_\ell$ *is a polynomial of degree at most* $\lceil \log_p N \rceil - 1$*, it can have at most* $\lceil \log_p N \rceil$ *zeros. Thus,* M *is a* $\left(p, \lceil \log_p N \rceil\right)$*-coherent matrix with* p^2 *rows separated into* p *blocks.*

Example 3 (Construction by error-correcting codewords)**.** *We can view the previous construction as a special case of the construction of a* (K, α)*-coherent matrix where*

we view the columns as binary error-correcting codewords with Hamming weight K by specifying a lower bound on the Hamming distance.

Indeed, writing $M = (\mathbf{m}_0, \ldots, \mathbf{m}_{N-1}) \in \{0,1\}^{m \times N}$ where each codeword $\mathbf{m}_j \in \{0,1\}^m$ has Hamming weight K (that is, K nonzero entries), we calculate $\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle$ in terms of the Hamming distance $\Delta(\mathbf{m}_j, \mathbf{m}_\ell) := |\{i \in [m] : (\mathbf{m}_j)_i \neq (\mathbf{m}_\ell)_i\}|$. Let $i_k \in [m]$ be an index of $\mathbf{m}_j$ such that $(\mathbf{m}_j)_{i_k} = 1$ The corresponding entry of $\mathbf{m}_\ell$ will either satisfy

1. *$(\mathbf{m}_\ell)_{i_k} = 1$, and therefore this index increases $\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle$ by one or,*
2. *$(\mathbf{m}_\ell)_{i_k} = 0$, and therefore this index increases $\Delta(\mathbf{m}_j, \mathbf{m}_\ell)$ by one. Additionally, this "mismatched one" in $\mathbf{m}_j$ must have a corresponding "mismatched one" somewhere in $\mathbf{m}_\ell$ (since both codewords have the same Hamming weight) which again increases $\Delta(\mathbf{m}_j, \mathbf{m}_\ell)$ by one.*

Thus, after iterating through all K ones in $\mathbf{m}_j$, we account for $\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle$ and exactly half of $\Delta(\mathbf{m}_j, \mathbf{m}_\ell)$, that is

$$\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle = K - \frac{\Delta(\mathbf{m}_j, \mathbf{m}_\ell)}{2}.$$

A lower bound for the Hamming distance of $2(K - \alpha)$ will then ensure that M is (K, α)-coherent.

In 2, each column has a Hamming weight of exactly p when viewed as an error-correcting codeword. The Hamming distance $\Delta(\mathbf{m}_j, \mathbf{m}_\ell)$ is twice the number of block indices b which are not zeros of $Q_j - Q_\ell$. By same argument as before, this "number of non-zeros" is at least by $p - \lceil \log_p N \rceil$, giving

$$\Delta(\mathbf{m}_j, \mathbf{m}_\ell) \geq 2\left(p - \lceil \log_p N \rceil\right).$$

Thus, in the context of error-correcting codewords, we have again shown that the matrix constructed in 2 is $\left(p, \lceil \log_p N \rceil\right)$-coherent.

$$
\begin{array}{rl}
 & j = \begin{array}{cccccc} 0 & 1 & 2 & 3 & \cdots & \end{array} \\
p_1 = 2 \begin{cases} [j \equiv 0 \bmod 2] \\ [j \equiv 1 \bmod 2] \end{cases} & \left(\begin{array}{cccc|c|c} 1 & 0 & 1 & 0 & I_2 & \cdots \\ 0 & 1 & 0 & 1 & & \end{array} \right. \\
p_2 = 3 \begin{cases} [j \equiv 0 \bmod 3] \\ [j \equiv 1 \bmod 3] \\ [j \equiv 2 \bmod 3] \end{cases} & \left(\begin{array}{ccc|c|c} 1 & 0 & 0 & I_3 & \cdots \\ 0 & 1 & 0 & & \\ 0 & 0 & 1 & & \end{array} \right. \\
\vdots & \left(\begin{array}{ccc|c} & \vdots & & \ddots \end{array} \right)
\end{array}
$$

Figure 6.2: Deterministically constructed low coherence matrix M by the process described in 4 for $q = 1$.

Example 4 (A Fourier friendly construction [19, 20])**.** *Let*

$$p_0 = 1,\ p_1 = 2,\ p_2 = 3,\ p_3 = 5,\ \ldots,\ p_\ell = \textit{the } \ell\textit{th prime.}$$

1. For some starting index $q \in \mathbb{N}$, fix the K sequential primes $p_q, \ldots, p_{q+K-1}$.

2. Define $M \in \{0,1\}^{\left(\sum_{\ell=0}^{K-1} p_{q+\ell}\right) \times N}$ with rows indexed by

$$(\ell, h) \in ([q, q+K-1] \cap \mathbb{N}) \times [p_\ell]$$

by

$$m_{(\ell,h),j} = \begin{cases} 1 & \textit{if } j \equiv h \bmod p_{q+\ell}, \\ 0 & \textit{otherwise.} \end{cases}$$

6.2 gives an example of this constructed matrix for starting index $q = 1$. We now obtain K blocks of rows (one for each prime) where each column contains exactly one 1 in each block.

Considering the coherence of M, we calculate

$$\langle \mathbf{m}_j, \mathbf{m}_\ell \rangle = |\{p \in \{p_q, \ldots p_{q+K-1}\} : \ell \equiv j \bmod p\}| =: |R|.$$

By the Chinese Remainder Theorem, the product of all primes in R must divide $|\ell - j| < N$. This restricts R to sets of primes whose product is strictly less than N. Thus, for $j \neq \ell$,

$$\alpha := \min\{m \in [K] : p_q p_{q+1} \cdots p_{q+m} \geq N\}$$

provides an upper bound on the cardinality of R and therefore the coherence of M. Then,

$$p_q^\alpha \leq p_q p_{q+1} \cdots p_{q+\alpha-1} < N,$$

giving that $\alpha \leq \left\lfloor \log_{p_q} N \right\rfloor$, and therefore M is $\left(K, \left\lfloor \log_{p_q} N \right\rfloor\right)$-coherent. Additionally, if q is chosen so that $p_{q-1} < K \leq p_q$, bounds on q [11] and the prime number theorem give that $p_{K+q-1} = \mathcal{O}(K \log K)$, and therefore M has

$$m = \sum_{\ell=0}^{K-1} p_{q+\ell} = \mathcal{O}(K^2 \log K)$$

columns.

Theorem 6.1.7. *For all $\tilde{x} \in \mathbb{C}^N$ with $\|\tilde{x}\|_0 \leq s$, if we set $K = s \lfloor \log_s N \rfloor / \varepsilon$, then*

$$(1-\varepsilon)\|\tilde{x}\|_2^2 \leq \|M\tilde{x}\|_2^2 \leq (1+\varepsilon)\|\tilde{x}\|_2^2.$$

Proof. The coherence of $M/\sqrt{K}$ is bounded by ε/s which implies the restricted isometry property (RIP) of order s by standard arguments, e.g., [13, Theorem 5.3]. □

Now how is this matrix Fourier friendly? Let $\tilde{N} = \prod_{\ell=q}^{q+K-1} p_\ell$ and $\tilde{M} \in \{0,1\}^{m \times \tilde{N}}$ be as above with, e.g., $K \leq p_q < 2K$ (which is possible by Bertrand's postulate) where $K = s \lfloor \log_s N \rfloor / \varepsilon$. Recall that these assumptions imply $s \ll N \ll \tilde{N}$. Additionally let $\hat{\mathbf{f}} \in \mathbb{C}^{\tilde{N}}$ be such that $\sum_{j>N} |\hat{f}_j|$ is small (e.g., zero) and suppose that $\{\hat{f}_j\}_{j\in[N]}$ has a good s-sparse approximation (e.g., because it's s-sparse). In this case, we can use compressive sensing methods to recover $\hat{\mathbf{f}}$ using only the values of $M\hat{\mathbf{f}}$. Moreover, we can compute these values quickly by the following observations.

First, we see that

$$\tilde{M}\hat{\mathbf{f}} = \tilde{M} F_{\tilde{N}\times\tilde{N}} \left(F_{\tilde{N}\times\tilde{N}}^{-1} \hat{\mathbf{f}} \right) =: \tilde{M} F_{\tilde{N}\times\tilde{N}} \mathbf{f},$$

where $\mathbf{f} := \left\{ f\left(\frac{2\pi j}{\tilde{N}}\right) \right\}_{j\in[\tilde{N}]}$ *is the vector of* $\tilde{N}$ *equally spaced samples from* $f(x) := \sum_{j\in[\tilde{N}]} \hat{f}_j \mathbb{e}^{\mathbb{i}jx}$. *But note that*

$$\tilde{M}F_{\tilde{N}\times\tilde{N}} = \left(\begin{array}{c|c|c|c} I_{p_q} & I_{p_q} & I_{p_q} & \cdots \\ \hline I_{p_{q+1}} & I_{p_{q+1}} & \cdots & \\ \hline \vdots & & \ddots & \end{array}\right) F_{\tilde{N}\times\tilde{N}}$$

and therefore each row in the product is the product of a discrete spike train with $F_{\tilde{N}\times\tilde{N}}$. *For example, following the same indexing scheme as* $\tilde{M}$,

$$\left(\tilde{M}F_{\tilde{N}\times\tilde{N}}\right)_{(\ell,h),k} = \frac{1}{\tilde{N}} \sum_{j=0}^{\frac{\tilde{N}}{p_{q+\ell}}-1} \mathbb{e}^{\frac{-2\pi\mathbb{i}(h+jp_{q+\ell})k}{\tilde{N}}} = \begin{cases} \frac{1}{p_{q+\ell}} \mathbb{e}^{\frac{-2\pi\mathbb{i}hk}{\tilde{N}}} & \text{if } k \equiv 0 \bmod \frac{\tilde{N}}{p_{q+\ell}}, \\ 0 & \text{otherwise.} \end{cases}$$

Note then that this product is extremely sparse, with each block corresponding to $p_{q+\ell}$ *having at most* $p_{q+\ell}$ *nonzero columns, which makes for at most* m *nonzero columns. Additionally, the resulting structure of the product allows one to compute*

$$\tilde{M}\hat{\mathbf{f}} = \tilde{M}F_{\tilde{N}\times\tilde{N}}\mathbf{f} = \begin{pmatrix} F_{p_q\times p_q}\left(f\left(\frac{2\pi j}{p_q}\right)\right)_{j\in[p_q]} \\ F_{p_{q+1}\times p_{q+1}}\left(f\left(\frac{2\pi j}{p_{q+1}}\right)\right)_{j\in[p_{q+1}]} \\ \vdots \\ F_{p_{q+K-1}\times p_{q+K-1}}\left(f\left(\frac{2\pi j}{p_{q+K-1}}\right)\right)_{j\in[p_{q+K-1}]} \end{pmatrix}$$

via K *fast Fourier transforms of size at most* $P_{q+K-1} = \mathcal{O}(K\log K)$. *Thus,* $\tilde{M}\hat{\mathbf{f}}$ *can be computed in* $\mathcal{O}(K^2\log^2 K\log\log K)$ *time if we have sampling access to* f. *By our assumption that* $K = s\lfloor\log_s N\rfloor/\varepsilon$ *for compressive sensing, we effectively compute* $M\hat{\mathbf{f}}$ *in*

$$\mathcal{O}\left(\frac{s^2}{\log^2 s}\log^2 N\left(\log s + \log\log N\right)^2 \log\left(\log s + \log N\right)\right) = \mathcal{O}\left(s^2\log^{2+\varepsilon} N\right) = \mathcal{O}\left(s^2\log^3 N\right)$$

time, which is sublinear in N when $s \ll N$.

However, while we can compute the samples of $M\hat{\mathbf{f}}$ in sublinear time, standard compressive sensing algorithms are all $\mathcal{O}(N)$-time, so the entire recovery process will not be sublinear. The solution which we now pursue will be to avoid these standard RIP based recovery algorithms.

6.2 Toward sublinear-time compressive sensing using low coherence matrices

6.2.1 Majority δ-reconstructing matrices

We begin with some definitions and examples toward an alternative to the RIP.

Running Example. *The matrix $A \in \{0,1\}^{4\times 6}$ given by*

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

is $(2,1)$-coherent.

Definition 6.2.1. If M has at least K ones in every column, then $M(n) \in \{0,1\}^{K\times N}$ for $n \in [N]$ is the $K \times N$ submatrix of M created by selecting the first K rows of M with nonzero entries in the nth column.

Running Example. *For $n = 2$,*

$$A = \begin{pmatrix} 1 & 1 & \boxed{1} & 0 & 0 & 0 \\ 1 & 0 & \boxed{0} & 1 & 1 & 0 \\ 0 & 1 & \boxed{0} & 1 & 0 & 1 \\ 0 & 0 & \boxed{1} & 0 & 1 & 1 \end{pmatrix} \mapsto A(2) = \begin{pmatrix} 1 & 1 & \boxed{1} & 0 & 0 & 0 \\ 0 & 0 & \boxed{1} & 0 & 1 & 1 \end{pmatrix}$$

column 2 ↓ *column 2 is all ones* ↓

and for $n = 5$,

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & \boxed{0} \\ 1 & 0 & 0 & 1 & 1 & \boxed{0} \\ 0 & 1 & 0 & 1 & 0 & \boxed{1} \\ 0 & 0 & 1 & 0 & 1 & \boxed{1} \end{pmatrix} \overset{\text{column 5}}{\downarrow} \mapsto A(5) = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & \boxed{1} \\ 0 & 0 & 1 & 0 & 1 & \boxed{1} \end{pmatrix} \overset{\text{column 5 is all ones}}{\downarrow}.$$

Definition 6.2.2. If M has at least K ones in every column, then $M'(n) \in \{0,1\}^{K \times N-1}$ is the $K \times (N-1)$ submatrix of $M(n)$ created by removing its nth column.

Running Example. *For* $n = 2$ *and* $n = 5$ *as above,*

$$A'(2) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \quad \textit{and} \quad A'(5) = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Definition 6.2.3. For $\mathbf{x} \in \mathbb{C}^N$, after ordering its entries by magnitude

$$|x_{j_1}| \geq |x_{j_2}| \geq \ldots \geq |x_{j_N}|$$

(where ties are broken for $|x_{j_i}| = \left|x_{j_{i+1}}\right|$ so that $j_i \leq j_{i+1}$), we define the index sets

$$S_{k,1} := \{j_1, \ldots, j_k\},\ S_{k,2} := \{j_{k+1}, \ldots, j_{2k}\},\ \ldots,\ S_{k,r} := \{j_{(r-1)k+1}, \ldots, j_N\},$$

for $r = \left\lfloor \frac{N-1}{k} \right\rfloor + 1$.

Definition 6.2.4. For $\mathbf{x} \in \mathbb{C}^N$ and any index set $S \subseteq [N]$ we define *the restriction of* $\mathbf{x}$ *to* S denoted $\mathbf{x}|_S \in \mathbb{C}^N$ to be the vector with entries

$$(\mathbf{x}|_S)_j := \begin{cases} x_j & \text{if } j \in S \\ 0 & \text{otherwise.} \end{cases}$$

Definition 6.2.5. Given $\mathbf{x} \in \mathbb{C}^N$, the s-sparse vector $\mathbf{x}_s \in \mathbb{C}^N$ is defined as $\mathbf{x}_s := \mathbf{x}|_{S_{s,1}}$.

We can now present our alternative to the RIP.

Definition 6.2.6 (Majority δ-reconstructing)**.** Let $\delta \in (0,1)$, and $M \in \{0,1\}^{m\times N}$ have at least K ones in every column. We will say that M is *majority δ-reconstructing for* $\mathbf{x} \in \mathbb{C}^N$ if the set

$$B_n := \left\{ j : \left| (M(n)\mathbf{x})_j - x_n \right| \leq \delta \left\| \mathbf{x} - \mathbf{x}_{\lfloor 1/\delta \rfloor} \right\|_1 \right\} \subseteq [K] \tag{6.1}$$

has cardinality $|B_n| > K/2$ for all $n \in [N]$.

Note. *More generally, one can change the fraction involved in the cardinality lower-bound of B_n in Definition 6.2.6 from $K/2$ to $\frac{c-2}{c}K$ for any $c \geq 4$. Doing so allows for modified reconstruction procedures in the next section. Furthermore, Theorem 6.3.3 guarantees that such matrices can be constructed for any desired $c \geq 4$.*

6.2.2 Reconstruction algorithm

For measurements of a vector taken with a δ-reconstructing matrix, we provide 6.2.1 to rapidly construct an approximation. It was first used in the sublinear-time Fourier setting in [19] and with (K,α)-coherent matrices in [2]. In the algorithm, $\mathbf{n} \in \mathbb{C}^m$ represents arbitrary additive errors on our measurements of $\mathbf{x}$ given by $\mathbf{y} = M\mathbf{x} + \mathbf{n}$, and $\mathbf{n}|_{M(n)} \in \mathbb{C}^K$ contains the K entries of $\mathbf{n} \in \mathbb{C}^m$ associated with the K rows of $M(n)$ in M.

Algorithm 6.2.1 Median recovery, $\mathrm{MR} : \mathbb{C}^m \times \{0,1\}^{m\times N} \times \mathcal{P}([N]) \times \mathbb{N} \to \mathbb{C}^N$.

INPUT: $\mathbf{y} = M\mathbf{x} + \mathbf{n}$, M, $S \subseteq [N]$, s. **OUTPUT:** $\mathbf{z}|_{\tilde{S}} \in (\mathbb{C} \times [N])^{\min(|S|,2s)}$ (interpreted as vector in $\mathbb{C}^N$). **for** $n \in S$ Let $\Re(z_n) \leftarrow$ median of $\Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)$ entries. Let $\Im(z_n) \leftarrow$ median of $\Im\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)$ entries. Sort $\{z_n\}_{n\in S}$ by magnitude so that $|z_{n_1}| \geq \ldots \geq |z_{n_{|S|}}|$. $\tilde{S} := \{n_1, \ldots, n_{\min(2s,|S|)}\}$. Output $\mathbf{z}|_{\tilde{S}}$ with

$$\left(\mathbf{z}|_{\tilde{S}}\right)_j := \begin{cases} z_j & \text{if } j \in \tilde{S} \\ 0 & \text{otherwise} \end{cases}$$

as an approximation to $\mathbf{x}$.

Complexity analysis

- 6.2.1–6.2.1 can be performed in $\mathcal{O}(|S| \cdot K)$-time using fast median algorithms [5] assuming that the rows of $M(n)$ can be identified in $\mathcal{O}(K)$-time for any $n \in [N]$. Note that this is indeed the case for our Fourier-friendly matrices in 4 by simply computing $n \bmod p_{q+\ell}$ for K values of ℓ.
- 6.2.1 can be performed in $\mathcal{O}(|S| \log |S|)$ time using, e.g., merge sort [24].
- 6.2.1–6.2.1 output $\mathbf{z}|_{\tilde{S}}$ in a compressed format in $\mathcal{O}(s)$-time and space.

Therefore, the total runtime/flop count of 6.2.1 is $\mathcal{O}\left(|S| \cdot \max\left(K, \log |S|\right)\right)$. Thus, the algorithm is fast if $|S|$ and K are small. In 6.2.2, we will analyze when it is also accurate.

Note. *6.2.1 is trivially parallelizable, since the z_n values can be computed in parallel, and efficient parallel sorting methods exist (see e.g. [27] for a comparison of several standard parallelized sorting algorithms).*

Note. *If one changes Definition 6.2.6 to require that $|B_n| > \frac{c-2}{c}K$ for any even $c \geq 4$ then one may use a median-of-means strategy in 6.2.1–6.2.1 as opposed to simply taking medians. More explicitly, this can be done by, e.g., modifying Line 6.2.1 to instead compute $\Re(z_n)$ by first sorting the entries of $\Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)$,*

$$\Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)_{j_1} \geq \Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)_{j_2} \geq \cdots \geq \Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)_{j_K},$$

and then setting

$$\Re(z_n) \leftarrow \text{Mean}\left(\Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)_{j_{\left\lfloor \frac{2(K-1)}{c} \right\rfloor + 1}}, \ldots, \Re\left(M(n)\mathbf{x} + \mathbf{n}|_{M(n)}\right)_{j_{K - \left\lfloor \frac{2(K-1)}{c} \right\rfloor}}\right).$$

If the noise vector $\mathbf{n}$ has, e.g., mean 0 i.i.d. random entries, then such median-of-means strategies can help to reduce the effect of the noise on the z_n estimates.

However, herein we will generally assume that $\mathbf{n}$ *represents an arbitrary (and deterministic) set of measurement errors.*

Theoretical analysis

The following theorem provides approximation guarantees for the output of 6.2.1.

Theorem 6.2.7 (Modified from [2])**.** *Let* $\beta \in [1, \infty)$*,* $s \in [N]$*, and* $\mathbf{x}, \mathbf{n} \in \mathbb{C}^N$*. Suppose that* M *is majority* δ*-reconstructing for* $\mathbf{x} \in \mathbb{C}^N$ *with* $\delta \leq \frac{1}{s}$ *, and that* $S \subseteq [N]$ *contains the set*

$$\mathcal{C}_{s.\beta} := \left\{ n \in [N] : |x_n| > \beta \left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty \right) \right\}.$$

Then,

$$\|\mathbf{x} - \mathrm{MR}(M\mathbf{x} + \mathbf{n}, M, S, s)\|_2 \leq \|\mathbf{x} - \mathbf{x}_{2s}\|_2 + C_\beta \left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{\sqrt{s}} + \sqrt{s}\|\mathbf{n}\|_\infty \right)$$

for an absolute constant $C_\beta \in \mathbb{R}^+$ *depending only on* β*. Furthermore,* $C_\beta \leq 6 + 2\sqrt{2}\beta$*.*

We will prove 6.2.7 with the help of a couple of lemmas.

Lemma 6.2.8. *Every* z_n *estimate produced in 6.2.1–6.2.1 of 6.2.1 satisfies*

$$|z_n - x_n| \leq \sqrt{2} \left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty \right).$$

Proof. This follows directly from the majority δ-reconstructing property of M for $\mathbf{x} \in \mathbb{C}^N$. Indeed, since the set B_n in the definition of M being majority δ-reconstructing for $\mathbf{x} \in C^N$ has cardinality $|B_n| > K/2$, reordering the vector $\Re(M(n)\mathbf{x} + \mathbf{n}|_{M(n)})_{j \in [K]}$ by magnitude ensures that there exists indices in B_n of elements in the reordered vector which lay on either side of the median $\Re(z_n)$. Thus, there exists some $j \in B_n$ such that

$$\begin{aligned} |\Re(z_n) - \Re(x_n)| &\leq \left| \Re(M(n)\mathbf{x} + \mathbf{n}_{M(n)})_j - \Re(x_n) \right| \leq |(M(n)\mathbf{x})_j + n_j - x_n| \\ &\leq \frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty =: \delta' \end{aligned} \tag{6.2}$$

where the first inequality holds by the previous argument, and the third inequality holds by the definition of B_n. Similarly, $|\Im(z_n) - \Im(x_n)| \leq \delta'$. Thus,

$$|z_n - x_n| \leq \sqrt{(\delta')^2 + (\delta')^2} = \sqrt{2}\delta'. \qquad \square$$

Lemma 6.2.9. *If $n \in \mathcal{C}_{s,\beta} \setminus \tilde{S}$ for $\tilde{S}$ in 6.2.1 of 6.2.1, then*

$$|x_n| \leq \left(\beta + 2\sqrt{2}\right)\left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty\right).$$

Proof. Note first that under the reordering $|x_{j_1}| \geq \ldots \geq |x_{j_N}|$, $j_\ell \in \mathcal{C}_{s,\beta}$ implies that $j_k \in \mathcal{C}_{s,\beta}$ for all $1 \leq k \leq \ell$. When $N > 2s$,

$$\|\mathbf{x} - \mathbf{x}_s\|_1 \geq \sum_{\ell=s+1}^{2s} |x_{j_\ell}| \geq s|x_{j_{2s+1}}|.$$

Thus, $\mathcal{C}_{s,\beta} \subseteq S_{2s,1} \cap S$ for all $\beta \geq 1$, giving $|\mathcal{C}_{s,\beta}| \leq \min(2s, |S|) = |\tilde{S}|$. Therefore if $n \in \mathcal{C}_{s,\beta} \setminus \tilde{S}$, there must exist some $j \in \tilde{S} \setminus \mathcal{C}_{s,\beta}$ in particular satisfying

$$|x_j| \leq \beta\left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty\right) \quad \text{and} \quad |z_j| \geq |z_n|.$$

Hence, for δ' as in (6.2),

$$\beta\delta' + \sqrt{2}\delta' \geq |x_j| + \sqrt{2}\delta' \geq |z_j| \geq |z_n| \geq |x_n| - \sqrt{2}\delta'$$

where the second and fourth inequalities hold by 6.2.8. $\square$

Lemma 6.2.10. *The distance between $\mathbf{x}$ and $\mathbf{x}|_{\mathcal{C}_{s,\beta}}$ in ℓ_2 can be bounded by*

$$\left\|\mathbf{x} - \mathbf{x}|_{\mathcal{C}_{s,\beta}}\right\|_2 \leq \|\mathbf{x} - \mathbf{x}_{2s}\|_2 + \sqrt{2s}\beta\left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty\right).$$

Proof. Since $\mathcal{C}_{s,\beta} \subseteq S_{2s,1}$,

$$\left\|\mathbf{x} - \mathbf{x}|_{\mathcal{C}_{s,\beta}}\right\|_2^2 = \|\mathbf{x} - \mathbf{x}_{2s}\|_2^2 + \sum_{n \in S_{2s,1} \setminus \mathcal{C}_{s,\beta}} |x_n|^2 \leq \|\mathbf{x} - \mathbf{x}_{2s}\|_2^2 + 2s\beta^2\left(\frac{\|\mathbf{x} - \mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty\right)^2.$$

$\square$

We are now prepared to prove our theorem concerning the output of the Median Recovery algorithm.

Proof of 6.2.7. Let $\delta' := \frac{\|\mathbf{x}-\mathbf{x}_s\|_1}{s} + \|\mathbf{n}\|_\infty$ as in (6.2). Then,

$$\begin{aligned}\|\mathbf{x} - \mathrm{MR}(M\mathbf{x}+\mathbf{n}, M, S, s)\|_2 &\leq \|\mathbf{x} - \mathbf{x}|_{\tilde{S}}\|_2 + \|\mathbf{x}|_{\tilde{S}} - \mathbf{z}|_{\tilde{S}}\|_2 \\ &\leq \|\mathbf{x} - \mathbf{x}|_{\tilde{S}}\|_2 + \sqrt{2s}(\sqrt{2}\delta'),\end{aligned}$$

by 6.2.8. Splitting $\tilde{S} = \mathcal{C}_{s,\beta} \sqcup (\tilde{S} \setminus \mathcal{C}_s)$ followed by applications of 6.2.10 and 6.2.9 on these pieces respectively, we bound

$$\begin{aligned}\|\mathbf{x} - \mathrm{MR}(M\mathbf{x}+\mathbf{n}, M, S, s)\|_2 &\leq \left\|\mathbf{x} - \mathbf{x}|_{\mathcal{C}_{s,\beta}}\right\|_2 + \sqrt{\sum_{n \in \mathcal{C}_{s,\beta} \setminus \tilde{S}} |x_n|^2} + 2\sqrt{s}\delta' \\ &\leq \|\mathbf{x} - \mathbf{x}_{2s}\| + \sqrt{\sum_{n \in \mathcal{C}_{s,\beta} \setminus \tilde{S}} |x_n|^2} + (2+\sqrt{2}\beta)\sqrt{s}\delta' \\ &\leq \|\mathbf{x} - \mathbf{x}_{2s}\| + \sqrt{2s\left((\beta+2\sqrt{2})\delta'\right)^2} + (2+\sqrt{2}\beta)\sqrt{s}\delta' \\ &= \|\mathbf{x} - \mathbf{x}_{2s}\|_2 + \left(\sqrt{2}(\beta+2\sqrt{2}) + 2 + \sqrt{2}\beta\right)\sqrt{s}\delta'\end{aligned}$$

as desired. □

Corollary 6.2.11. *Under the assumptions of 6.2.7, we will also have*

$$\|\mathbf{x} - \mathrm{MR}(M\mathbf{x}, M, S, s)\|_2 \leq C'_\beta \frac{\|\mathbf{x}-\mathbf{x}_s\|_1}{\sqrt{s}}$$

when $\mathbf{n} = 0$. *Here,* $C'_\beta \in \mathbb{R}^+$ *is an absolute constant with* $C'_\beta \leq C_\beta + 1$ *for* C_β *the constant in 6.2.7.*

Proof. By [13, Proposition 2.3], we have

$$\|\mathbf{x} - \mathbf{x}_{2s}\|_2 \leq \frac{\|\mathbf{x}-\mathbf{x}_s\|_1}{\sqrt{s}},$$

finishing the proof. □

Recap. *Up to now, we have seen that there are some nice (K, α)-coherent matrix constructions, including one that allows for sublinear-time evaluations of $M\hat{\mathbf{x}}$ whenever one has access to $\mathbf{x} \in \mathbb{C}^{\tilde{N}}$ (or sampling access to the trigonometric polynomial $f(x) = \sum_{\omega \in [N]} \hat{x}_\omega e^{\mathrm{i}\omega x}$). Additionally, we have seen that the majority δ-reconstructing property allows for standard compressive sensing error guarantees to be obtained.*

Next. *We will*

1. *relate (K, α)-coherent matrices to the majority δ-reconstructing property by showing how to construct matrices with this property using any (K, α)-coherent matrix (including, crucially, our Fourier friendly matrix), and*
2. *then consider fast algorithms for* rapidly *finding small sets $S \supseteq \mathcal{C}_{s,\beta}$. This will allow the Median Recovery algorithm to run in sublinear-time while still having good error bounds.*

6.3 Constructions of majority δ-reconstructing matrices

We will give two sets of majority δ-reconstructing matrix constructions using (K, α)-coherent matrices. Both sets will have meaningful representations in the Fourier context.

1. The first type construction will hold for all $\mathbf{x} \in \mathbb{C}^N$deterministically, but will have a suboptimal number of rows.
2. The second type of construction will have a near-optimal number of rows, but will only hold for a given (but a priori unknown) single vector $\mathbf{x} \in \mathbb{C}^N$ with high probability.

We begin with the following lemmas to help with the deterministic construction.

Lemma 6.3.1. *Suppose $M \in \{0,1\}^{m\times N}$ is (K,α)-coherent. Let $n \in [N]$, $s \in [1, K/\alpha] \cap \mathbb{N}$, and $\mathbf{x} \in \mathbb{C}^{N-1}$. Then at most $s\alpha$ of the entries in $M'(n)\mathbf{x} \in \mathbb{C}^K$ will have magnitude greater than or equal to $\|\mathbf{x}\|_1/s$.*

Proof. By Markov's inequality, we have

$$\begin{aligned}\left|\left\{j : |(M'(n)\mathbf{x})_j| \geq \frac{\|\mathbf{x}\|_1}{s}\right\}\right| &\leq \frac{s}{\|\mathbf{x}\|_1}\|M'(n)\mathbf{x}\|_1 \\ &\leq s\|M'(n)\|_{1\to 1}.\end{aligned}$$

Furthermore, if we denote the columns $M'(n) = (\mathbf{m}'_k)_{k\in[N-1]}$ and $M(n) = (\mathbf{m}_\ell)_{\ell\in[N]}$, we calculate

$$\begin{aligned}\|M'(n)\|_{1\to 1} &= \max_{k\in[N-1]} \|\mathbf{m}'_k\|_1 \\ &= \max_{\ell\in[N]\setminus\{n\}} \langle \mathbf{m}_\ell, \mathbf{m}_n\rangle \\ &\leq \alpha,\end{aligned}$$

finishing the proof. □

Lemma 6.3.2. *Suppose $M \in \{0,1\}^{m\times N}$ is a (K,α)-coherent matrix. Let $n \in [N]$, $s \in [1, K/\alpha] \cap \mathbb{N}$, $S \subset [N]$ with $|S| \leq s$, and $\mathbf{x} \in \mathbb{C}^{N-1}$. Then $M'(n)\mathbf{x}$ and $M'(n)\left(\mathbf{x} - \mathbf{x}|_S\right)$ will differ in at most $s\alpha$ of their K entries.*

Proof. Let $\mathbb{1} \in \mathbb{C}^{N-1}$ be the vector of all ones. We have for $\mathcal{B} := \{j : (M'(n)\mathbf{x})_j \neq (M'(n)(\mathbf{x} - \mathbf{x}|_S))_j\}$ that

$$|\mathcal{B}| = \left|\left\{j : (M'(n)\mathbf{x}|_S)_j \neq 0\right\}\right| \leq \left|\left\{j : (M'(n)\mathbb{1}|_S)_j \geq 1\right\}\right| \tag{6.3}$$

since the nonzero entries of $M'(n)$ are all ones. Now, we may apply 6.3.1 to 6.3 with $M'(n)$ applied to $\mathbb{1}|_S$, which has $\|\mathbb{1}|_S\|_1 = |S| \leq s$ to learn that $|\mathcal{B}| \leq \alpha s$. □

We are now prepared to provide our first majority δ-reconstructing matrix construction.

Theorem 6.3.3 (Modified from [2])**.** *Suppose M is (K, α)-coherent. Let $s := 1/\delta \in [1, K/\alpha] \cap \mathbb{N}$ and $c \in [4, \infty) \cap \mathbb{N}$. If $K > c\alpha/\delta$, then M will be majority δ-reconstructing for all $\mathbf{x} \in \mathbb{C}^N$. In particular, the cardinality of B_n (6.1) will be such that $|B_n| > \left(\frac{c-2}{c}\right) K$ for all $n \in [N]$ and $\mathbf{x} \in \mathbb{C}^N$.*

Proof. Let $n \in [N]$ and $\mathbf{x} \in \mathbb{C}^N$. Furthermore, let $\mathbf{y} \in \mathbb{C}^{N-1}$ be $\mathbf{x}$ with x_n removed so that

$$y_j = \begin{cases} x_j & \text{if } j < n \\ x_{j+1} & \text{if } j \geq n. \end{cases}$$

Finally, let $\mathbf{m}_0, \ldots \mathbf{m}_{N-1} \in \{0,1\}^K$ be the columns of $M(n)$.

We have that

$$M(n)\mathbf{x} = x_n \mathbf{m}_n + M'(n)\mathbf{y} = x_n \mathbb{1} + M'(n)\mathbf{y}.$$

6.3.2 tells us that at most $s\alpha = \alpha/\delta$ entries of $M'(n)\mathbf{y}$ differ from those in $M'(n)\left(\mathbf{y} - \mathbf{y}_s\right)$. Of the remaining entries of $M'(n)\mathbf{y}$ (of which there are at least $K - s\alpha$), at most $s\alpha$ of them have magnitudes greater than or equal to $\delta\|\mathbf{y} - \mathbf{y}_s\|_1$ by 6.3.1 (since removing the at most $s\alpha$ rows from M for which $M'(n)\left(\mathbf{y} - \mathbf{y}_s\right) \neq M'(n)\mathbf{y}$ will leave us with another $(K - s\alpha, \alpha)$-coherent matrix and $s \in [1, \frac{K-s\alpha}{\alpha}]$ as $K > c\alpha s > 2\alpha s$). Hence, at least

$$\begin{aligned} K - 2s\alpha &= K - \frac{2\alpha}{\delta} \\ &= K - \frac{2}{c}\frac{c\alpha}{\delta} \\ &> K - \frac{2}{c}K = \left(\frac{c-2}{c}\right) K \end{aligned}$$

entries of $M'(n)\mathbf{y}$ will have magnitudes bounded above by

$$\frac{1}{\delta}\|\mathbf{y} - \mathbf{y}_s\|_2 \leq \frac{1}{\delta}\|\mathbf{x} - \mathbf{x}_s\|_1.$$

The result follows after noting that $\frac{c-2}{c} \geq \frac{1}{2}$ for all $c \in [4, \infty)$. □

Example 5. *Recalling the Fourier-friendly matrices from 4, we note that setting* $K = 4s \lfloor \log_s N \rfloor$ *for* $p_{q-1} < K \leq p_q$*, will yield a majority* $(\delta = 1/s)$*-reconstructing matrix for any* $s \leq N$*. It will have at most*

$$m = \mathcal{O}\left(s^2 \log_s^2 N \log(s \log_s N)\right)$$

rows.

If we use this Fourier-friendly matrix with the Median Recovery algorithm, 6.2.1, we can approximate $\hat{\mathbf{x}}$ *for any* $\mathbf{x} \in \mathbb{C}^{\tilde{N}}$ *using only* $\mathcal{O}(m)$ *samples from* $\mathbf{x}$*, and the algorithm will run in* $\mathcal{O}(Ns \log N)$*-time if* $S = [N]$ *is used.*

Next. *We still have two concerns to take care of.*

1. *How can we quickly find a smaller* S *containing* $C_{s,\beta}$*?*
2. *How can we reduce the number of rows in the Fourier-friendly construction to scale linearly in* s*?*

Appendices

Appendix A

Partial Examples of the .tex Style I Like (SVD CMSE Lec 5 Review Content – SVD Section)

Classical Building Blocks: Definitions, FFT, SVD, Linear and Semidefinite Programming

$[N] = \{0, \ldots, N-1\}$.

In this chapter we briefly introduce the reader to some of the fundamental algorithms and techniques often used as building blocks to in order to construct of methods for the analysis of massive data sets.

A.1 Some Notation

NEED TO REWRITE BELOW WITH RESPECTABLE NOTATION!!!!

INDEXES ALWAYS START FROM 0!!!!

:= for definitions, = for equalities that logically follow from those definitions.

n and N will always denote dimension sizes...

vectors (bolded), matrices, unitary, ENTRIES of matrix/vectors notation, $\cdot$ always denotes scalar multiplication (never a dot product).

A^* = conj. transpose, range(A)=column span of A = image of A when considered as a linear map.

ℓ_p-norms of vectors

A ball in $\mathbb{R}^N$ around a is B(a,radius)

vectors are considered equivalent to $1 \times N$ size matrices

computational complexity, operation counts (not time except for hueristic). Little o and big O, theta, omega.

A.2 The Discrete Fourier Transform (DFT)

In later chapters a particular orthonormal basis of $\mathbb{C}^N$, known as the discrete Fourier transform basis, will become important for computational reasons. In this section we will define the discrete Fourier basis and then present an algorithm, called the Fast Fourier Transform (FFT), for quickly expressing any given vector in terms of the Fourier basis. The speed of the FFT in combination with the many interesting and useful properties of the Fourier basis will allow for several developments later on in Chapters **??** and **??**. The reader should feel free to skip this section until just before then.

Define

$$f := \mathrm{e}^{\frac{-2\pi \mathrm{i}}{N}} \tag{A.1}$$

and let $F \in \mathbb{C}^{N \times N}$ be the $N \times N$ matrix defined by

$$F_{\omega,j} := \frac{f^{\omega \cdot j}}{\sqrt{N}} \tag{A.2}$$

for $\omega, j \in [N]$. The matrix F is called the **Discrete Fourier Transform (DFT) Matrix** of size N. It is not difficult to show that the columns of F form an orthonormal basis of $\mathbb{C}^N$ (i.e., one can easily show that F is a unitary matrix – see problem 1 below).

The **Discrete Fourier Transform (DFT)** of a vector $\mathbf{v} \in \mathbb{C}^N$ is simply

$$\hat{\mathbf{v}} := F\mathbf{v} \tag{A.3}$$

with entries given by $\hat{v}_\omega$ for all $\omega \in [N] = \{0, \ldots, N-1\} \subset \mathbb{N}$. Similarly, the **Inverse Discrete Fourier Transform (IDFT)** of a vector $\mathbf{v} \in \mathbb{C}^N$ is

$$\widehat{\mathbf{v}}^{-1} := F^{-1}\mathbf{v} = F^*\mathbf{v}. \tag{A.4}$$

As we shall see, the DFT walks hand in hand with our next definition.

The **discrete convolution** of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$, denoted by $\mathbf{u} \star \mathbf{v} \in \mathbb{C}^N$, is defined entrywise with its k^{th} entry given by

$$(\mathbf{u} \star \mathbf{v})_k := \sum_{j=0}^{N-1} u_j \cdot v_{(k-j) \bmod N}.^1 \tag{A.5}$$

The discrete convolution of two vectors has the following useful relationship to the two vectors' Discrete Fourier Transforms.

Theorem A.2.1. *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$. *Then*

$$\widehat{(\boldsymbol{u} \star \boldsymbol{v})}_\omega = \sqrt{N} \cdot \hat{u}_\omega \hat{v}_\omega \tag{A.6}$$

holds for all $\omega \in [N]$.

[1] Recall that "$x \bmod N$" denotes the unique integer $w \in [N]$ satisfying $x = w + k \cdot N$ for some $k \in \mathbb{Z}$.

Proof: To obtain (A.6) we compute

$$(\widehat{\mathbf{u} \star \mathbf{v}})_\omega = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} (\mathbf{u} \star \mathbf{v})_k \, f^{\omega \cdot k} = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \left(\sum_{j=0}^{N-1} u_j \cdot v_{(k-j) \bmod N} \right) f^{\omega \cdot k}. \tag{A.7}$$

Rearranging the final double sum we obtain

$$(\widehat{\mathbf{u} \star \mathbf{v}})_\omega = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} u_j \, f^{\omega \cdot j} \left(\sum_{k=0}^{N-1} v_{(k-j) \bmod N} \, f^{\omega \cdot (k-j)} \right) = \sqrt{N} \cdot \hat{u}_\omega \hat{v}_\omega. \tag{A.8}$$

Here we have used the fact that $f^{n \cdot N} = 1$ for all $n \in \mathbb{Z}$ so that $f^{\omega \cdot (k-j)} = f^{\omega \cdot ((k-j) \bmod N)}$ always holds. □

Theorem A.2.1 tells us that the DFT of the convolution of two vectors is equal to the entrywise product of the DFTs of the two vectors. Using this relationship we can compute the discrete convolution of $\mathbf{u}$ and $\mathbf{v}$ using their DFTs. Let $\mathbf{u} \odot \mathbf{v} \in \mathbb{C}^N$ denote the entrywise (or Hadamard) product of the two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$. That is, let

$$(\mathbf{u} \odot \mathbf{v})_j := u_j v_j \tag{A.9}$$

for all $j \in [N]$. Theorem A.2.1 now directly implies that

$$\mathbf{u} \star \mathbf{v} = \sqrt{N} \cdot \widehat{\hat{\mathbf{u}} \odot \hat{\mathbf{v}}}^{-1} = \sqrt{N} \cdot F^* \left(F\mathbf{u} \odot F\mathbf{v} \right). \tag{A.10}$$

Note that the last expression of (A.10) could be computed quickly if we could find a way to quickly calculate both $F\mathbf{u}$ and $F^*\mathbf{u}$ for any given $\mathbf{u}$. As it turns out this is in fact possible!

A.2.1 The Fast Fourier Transform (FFT)

As seen above, computing the DFT of a vector $\mathbf{u} \in \mathbb{C}^N$ requires the computation of $F\mathbf{u}$. Computing $F\mathbf{u}$ directly via a generic matrix-vector multiply requires $\mathcal{O}(N^2)$

operations. The Fast Fourier Transform (FFT) allows us to reduce this computational cost from $\mathcal{O}(N^2)$ operations to $\mathcal{O}(N \ln N)$ operations. In what follows we will outline the recursive construction of the FFT algorithm via sum splitting techniques.

Let $\mathbf{u} \in \mathbb{C}^N$, and suppose that its dimension, N, has the prime factorization

$$N = p_1 \cdot p_2 \cdots p_m, \text{ where } p_1 \leq p_2 \leq \cdots \leq p_m \text{ are } N\text{'s prime factors.}$$

Choose $\omega \in [N]$. It's not too difficult to see that

$$\hat{u}_\omega = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} u_j \, f^{\omega \cdot j}. \tag{A.11}$$

By splitting the sum (A.11) for $\hat{u}_\omega$ into p_1 smaller subsums, one for each possible residue modulo p_1, we can see that

$$\hat{u}_\omega = \frac{1}{\sqrt{N}} \sum_{k=0}^{p_1-1} f^{\omega \cdot k} \left(\sum_{j=0}^{\frac{N}{p_1}-1} u_{k+p_1 \cdot j} \, f^{\omega \cdot p_1 \cdot j} \right). \tag{A.12}$$

Let's now rewrite the internal sum of (A.12) in order to realize some progress.

Given $k \in [p_1]$, define $\mathbf{u}^{(k,p_1)} \in \mathbb{C}^{N/p_1}$ to be the vector whose entries are the entries of $\mathbf{u}$ having indexes that are congruent to k modulo p_1,

$$\left(\mathbf{u}^{(k,p_1)}\right)_j := u_{k+j \cdot p_1} \tag{A.13}$$

for all $j \in [N/p_1]$.[2] Our equation (A.12) for $\hat{u}_\omega$ now becomes

$$\hat{u}_\omega = \frac{1}{\sqrt{p_1}} \sum_{k=0}^{p_1-1} f^{\omega \cdot k} \left(\frac{1}{\sqrt{N/p_1}} \sum_{j=0}^{\frac{N}{p_1}-1} \left(\mathbf{u}^{(k,p_1)}\right)_j \, f^{p_1 \cdot \omega \cdot j} \right) \tag{A.14}$$

$$= \frac{1}{\sqrt{p_1}} \left(\sum_{k=0}^{p_1-1} f^{\omega \cdot k} \left(\widehat{\mathbf{u}^{(k,p_1)}} \right)_{\omega \bmod \frac{N}{p_1}} \right). \tag{A.15}$$

For the sake of clarity we emphasize that the vector $\widehat{\mathbf{u}^{(k,p_1)}} \in \mathbb{C}^{N/p_1}$ in (A.15) is exactly $F\mathbf{u}^{(k,p_1)}$, where $F \in \mathbb{C}^{\frac{N}{p_1} \times \frac{N}{p_1}}$ is now the DFT Matrix of size N/p_1. We

[2] Note that we used an integer divisor of N, i.e. p_1, exactly to ensure that $\frac{N}{p_1} \in \mathbb{N}$.

strongly recommend that you verify the equality of (A.14) and (A.15) for yourself before reading further.

At this point it's useful to ask ourselves what we've managed to accomplish by reformulating (A.11) into (A.15). Mainly, we can now compute $\hat{\mathbf{u}} \in \mathbb{C}^N$ with fewer operations than before by computing it in two steps. First, we compute $\widehat{\mathbf{u}^{(k,p_1)}} \in \mathbb{C}^{\frac{N}{p_1}}$ for all $k \in [p_1]$. Next, we use the vectors $\widehat{\mathbf{u}^{(0,p_1)}}, \ldots, \widehat{\mathbf{u}^{(p_1-1,p_1)}}$ computed in the first step in order to compute each entry of $\hat{\mathbf{u}}$ via (A.15). The first step can be accomplished with p_1 matrix-vector multiplications, each of which can be computed using $\mathcal{O}(N^2/p_1^2)$ operations (recall that $\widehat{\mathbf{u}^{(k,p_1)}} = F\mathbf{u}^{(k,p_1)}$, where F is the DFT Matrix of size N/p_1). Hence, the first step can be completed using $\mathcal{O}(N^2/p_1)$ total operations. Step two only requires $\mathcal{O}(p_1 N)$ total operations in order to finish calculating $\hat{\mathbf{u}}$, $\mathcal{O}(p_1)$-operations for each $\hat{u}_\omega$. Putting it all together, we can see that (A.15) allows us compute $\hat{\mathbf{u}} \in \mathbb{C}^N$ using a grand total of $\mathcal{O}(p_1 N + N^2/p_1)$ operations, as opposed to computing it directly via (A.3) using $\Theta(N^2)$ operations.

Although the computational gain obtained from (A.15) is modest, it is important to note that the sum-splitting technique used to obtain it can now be employed again in order to compute each $\widehat{\mathbf{u}^{(k,p_1)}}$, $k \in [0, p_1)$, more quickly. That is, we may split up the sum for $(\widehat{\mathbf{u}^{(k,p_1)}})_\omega$ into p_2 additional sums, etc.. Repeatedly sum-splitting in this fashion leads to the recursive **Fast Fourier Transform (FFT)** shown in Algorithm A.2.1. Analogous sum-splitting leads to the **Inverse Fast Fourier Transform (IFFT)** which can be obtained from Algorithm A.2.1 by replacing line 10's $f^{k\omega}$ by $f^{-k\omega}$ and replacing each $\hat{\mathbf{u}}$ by a $\hat{\mathbf{u}}^{-1}$.

We are now ready to analyze the computational complexity of the FFT. Let T_N be the number of operations used by Algorithm A.2.1 to compute $\hat{\mathbf{u}} \in \mathbb{C}^N$. In order to determine an equation for T_N we note that lines 6 – 8 require $p_1 \cdot T_{\frac{N}{p_1}}$ operations

Algorithm A.2.1 FAST FOURIER TRANSFORM (FFT)

1: **Input: Vector** $\mathbf{u} \in \mathbb{C}^N$**, Dimension** N**, Dimension's prime factorization** $p_1 \leq \cdots \leq p_m$
2: **Output:** $\hat{\mathbf{u}} \in \mathbb{C}^N$
3: **if** $N == 1$ **then**
4: Return $\mathbf{u}$
5: **end if**
6: **for** k from 0 to $p_1 - 1$ **do**
7: $\widehat{\mathbf{u}^{(k,p_1)}} \leftarrow \text{FFT}\left(\mathbf{u}^{(k,p_1)}, \frac{N}{p_1}, p_2 \leq p_3 \leq \cdots \leq p_m\right)$
8: **end for**
9: **for** ω from 0 to $N - 1$ **do**
10: $\hat{u}_\omega \leftarrow \frac{1}{\sqrt{p_1}} \left(\sum_{k=0}^{p_1-1} f^{k\omega} \left(\widehat{\mathbf{u}^{(k,p_1)}}\right)_{\omega \bmod \frac{N}{p_1}} \right)$
11: **end for**
12: Return $\hat{\mathbf{u}}$

while lines 9 – 11 use $\mathcal{O}(p_1 N)$ operations. Therefore we have

$$T_N = \mathcal{O}(p_1 N) + p_1 \cdot T_{\frac{N}{p_1}}. \tag{A.16}$$

However, Algorithm A.2.1 is recursively invoked again to compute $\widehat{\mathbf{u}^{(0,p_1)}}, \ldots, \widehat{\mathbf{u}^{(p_1-1,p_1)}}$ by sum-splitting in line 7. Taking this into account we can see that

$$T_{\frac{N}{p_1}} = \mathcal{O}\left(\frac{p_2 N}{p_1}\right) + p_2 \cdot T_{\frac{N}{p_1 p_2}}. \tag{A.17}$$

We now have

$$T_N = \mathcal{O}(p_1 N) + p_1 \cdot \left(\mathcal{O}\left(\frac{p_2 N}{p_1}\right) + p_2 \cdot T_{\frac{N}{p_1 p_2}} \right) = \mathcal{O}\left(N(p_1 + p_2)\right) + p_1 p_2 \cdot T_{\frac{N}{p_1 p_2}}. \tag{A.18}$$

Repeating this recursive sum-splitting $n \leq m$ times shows us that

$$T_N = \mathcal{O}\left(N \cdot \sum_{l=1}^{n} p_l \right) + \prod_{l=1}^{n} p_l \cdot T_{\frac{N}{p_1 \cdots p_n}}. \tag{A.19}$$

Using that $T_1 = \mathcal{O}(1)$ (see Algorithm A.2.1's lines 3 – 5) we have

$$T_N = \mathcal{O}\left(N \cdot \sum_{l=1}^{m} p_l \right) + \mathcal{O}(N) = \mathcal{O}(m \cdot p_m \cdot N). \tag{A.20}$$

Note that $m \leq \log_2 N$ while p_m is N's largest prime factor. We have proven the following theorem in the course of the subsequent discussion.

Theorem A.2.2. *Let* $\mathbf{u} \in \mathbb{C}^N$ *and suppose that* N *has the prime factorization* $N = p_1 \cdots p_m$*, where* $p_1 \leq p_2 \leq \cdots \leq p_m$ *are the prime factors of* N *ordered from smallest to largest. Then, we may compute* $\hat{\mathbf{u}} = F\mathbf{u}$ *using* $\mathcal{O}\left(N \cdot \sum_{l=1}^{m} p_l\right)$ *operations.*

Theorem A.2.2 tells us that the FFT can significantly speed up computation of the DFT. For example, if N is a power of 2 we'll have $m = \log_2 N$ and $p_m = 2$ leaving Algorithm A.2.1 with an $\mathcal{O}(N \ln N)$ operation count. This is a clear improvement over the $\Theta(N^2)$ operations required to in order to compute (A.3) directly. However, if N has large prime factors the improvement is less impressive. In the worst case, when N is prime, we have $m = 1$ and $p_1 = N$. This leaves Algorithm A.2.1 with a $\mathcal{O}(N^2)$ runtime which, in practice, is slower than the direct method (A.3).

The inability of Algorithm A.2.1 to handle vectors with sizes containing large prime factors isn't a setback when one may dictate, with little or no repercussions, the dimension of the vectors they work with. A popular choice is to simply force N to be a power of 2. However, sometimes one simply needs to compute the DFT of a vector whose size contains (or may contain) large prime factors. In the next subsection we discuss how to do this efficiently.

A.2.2 The FFT for Vectors of Arbitrary Size

As discussed in the previous subsection, Algorithm A.2.1 may not be a very efficient means of computing $\hat{\mathbf{u}} \in \mathbb{C}^N$ when N contains large prime factors. One way of addressing this issue is to rewrite $\hat{\mathbf{u}}$ as a discrete convolution of two vectors of a slightly larger dimension, $\tilde{N}$, that contains only small prime factors. This discrete convolution can then be computed quickly by Algorithm A.2.1 which will be efficient for vectors of dimension $\tilde{N}$.

Let $\omega \in [N]$. We may rewrite $\hat{u}_\omega$ as

$$\hat{u}_\omega = f^{-\frac{\omega^2}{2}} f^{\frac{\omega^2}{2}} \cdot \hat{u}_\omega = \frac{f^{\frac{\omega^2}{2}}}{\sqrt{N}} \cdot \sum_{j=0}^{N-1} u_j \, f^{\omega \cdot j - \frac{\omega^2}{2}} = \frac{f^{\frac{\omega^2}{2}}}{\sqrt{N}} \cdot \sum_{j=0}^{N-1} u_j \, f^{\frac{-(\omega-j)^2}{2}} f^{\frac{j^2}{2}}. \tag{A.21}$$

Note that the last sum in (A.21) resembles a convolution. In order to make the resemblance more concrete we will define two new vectors.

Let $\tilde{N} = 2^{\lceil \log_2 N \rceil + 1}$. Now define $\tilde{\mathbf{u}} \in \mathbb{C}^{\tilde{N}}$ by

$$\tilde{u}_j = \begin{cases} u_j \cdot f^{\frac{j^2}{2}} & \text{if } 0 \leq j < N \\ 0 & \text{if } N \leq j < \tilde{N} \end{cases}, \tag{A.22}$$

and define $\mathbf{v} \in \mathbb{C}^{\tilde{N}}$ by

$$v_h = \begin{cases} f^{\frac{-h^2}{2}} & \text{if } 0 \leq h < N \\ 0 & \text{if } N \leq h \leq \tilde{N} - N \\ f^{\frac{-(h-\tilde{N})^2}{2}} & \text{if } \tilde{N} - N < h < \tilde{N} \end{cases}. \tag{A.23}$$

Equation (A.21) now becomes

$$\hat{u}_\omega = \frac{f^{\frac{\omega^2}{2}}}{\sqrt{N}} \cdot \sum_{j=0}^{\tilde{N}-1} \tilde{u}_j \cdot v_{(\omega-j) \bmod \tilde{N}} = \frac{f^{\frac{\omega^2}{2}}}{\sqrt{N}} \cdot (\tilde{\mathbf{u}} \star \mathbf{v})_\omega. \tag{A.24}$$

Note that this final convolution, $\tilde{\mathbf{u}} \star \mathbf{v} \in \mathbb{C}^{\tilde{N}}$, can be computed efficiently by the FFT and IFFT using (A.10) since $\tilde{N}$ is a power of two. We have now established the following theorem.

Theorem A.2.3. *Let $\boldsymbol{u} \in \mathbb{C}^N$. Then, $\hat{\boldsymbol{u}} \in \mathbb{C}^N$ can be calculated using $\mathcal{O}(N \ln N)$ operations.*

Theorem A.2.3 generalizes Theorem A.2.2 to handle all values of N efficiently. We are now in the position to declare that the DFT of any vector in $\mathbb{C}^N$ can be calculated using only $\mathcal{O}(N \ln N)$ operations!

A.2.3 References

The FFT was first published and analyzed as a computer algorithm by Cooley and Tukey in 1965 [7], despite similar techniques being utilized much earlier (e.g., by Gauss and many others [16]). Cooley and Tukey's algorithm is particularly efficient for vector dimensions, N, whose prime factorizations contain only small prime factors. Later variants of the FFT [4, 29] allowed the FFT to also be utilized effectively for vector sizes whose prime factorizations contain larger primes. This section has primarily followed these three papers. For more information on Fourier methods and algorithms we recommend that the interested reader consult the relevant chapters of [28], [22], [8], or [6]. For a fast FFT implementation we recommend FFTW [14].

A.2.4 Exercises

1. Prove that the DFT matrix, F, is unitary.

2. Prove that $\|\hat{\mathbf{v}}\|_2^2 = \|\mathbf{v}\|_2^2$ holds for all $\mathbf{v} \in \mathbb{C}^N$. This equality is sometimes referred to as Parseval's equality in the context of the discrete Fourier basis.

3. Suppose $p, N \in \mathbb{N}$ are such that $N/p \in \mathbb{N}$ (i.e., suppose that p divides N). Given $\mathbf{u} \in \mathbb{C}^p$, let $\mathbf{v} \in \mathbb{C}^N$ be a longer vector with entries given by

$$v_j = \begin{cases} u_{pj/N} & \text{if } j \equiv 0 \bmod (N/p) \\ 0 & \text{else} \end{cases},$$

   and let $\mathbf{w} \in \mathbb{C}^N$ be another longer vector with entries given by $w_j = u_{j \bmod p}$. Compute the N-length DFTs $\hat{\mathbf{v}}, \hat{\mathbf{w}} \in \mathbb{C}^N$ in terms of the p-length DFT of $\mathbf{u}$.

4. Let $a, b, c \in [N]$ be such that a is invertible modulo N.[3] Furthermore, suppose that $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ satisfy

$$v_j = \mathrm{e}^{\frac{2\pi \mathrm{i} cj}{N}} u_{aj+b}$$

[3] A value $a \in [N]$ is invertible modulo N if there exists an $h \in [N]$ such that $(a\ h) \equiv 1 \bmod N$. Any $a \in [N]$ that is relatively prime to N will be invertible modulo N by the Fermat-Euler Theorem.

for all $j \in [N]$. Calculate $\hat{\mathbf{v}}$ in terms of $\hat{\mathbf{u}}$. How does a affect $\hat{\mathbf{v}}$ when $c = b = 0$? How does b affect $\hat{\mathbf{v}}$ when $a = 1$ and $c = 0$? How does c affect $\hat{\mathbf{v}}$ when $a = 1$ and $b = 0$?

5. Let $q(x) = \sum_{j=0}^{N-1} q_j x^j$ and $r(x) = \sum_{j=0}^{N-1} r_j x^j$ be two polynomials of degree at most $N-1$. Define $t(x) = q(x) \cdot r(x)$ to be their product. We know that $t(x)$ is a polynomial of degree at most $2N-2$ which can be written as $t(x) = \sum_{j=0}^{2N-2} t_j x^j$. Show that $t_0, \ldots, t_{2N-2}$ can be computed explicitly using only $\mathcal{O}(N \ln N)$ operations.

6. Consider the matrix $D_2 \in \mathbb{N}^{N \times N}$ whose entries are given by

$$(D_2)_{i,j} = \begin{cases} -2 & \text{if } i = j \\ 1 & \text{if } (i-j) \equiv 1 \bmod N \\ 1 & \text{if } (i-j) \equiv N-1 \bmod N \\ 0 & \text{otherwise} \end{cases} . \tag{A.25}$$

This is an example of a *circulant matrix* (see problem 8). Show that $FD_2 = EF$, where $E \in \mathbb{R}^{N \times N}$ is a diagonal matrix with entires given by

$$(E)_{i,j} = \begin{cases} 2\cos(2\pi j/N) - 2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} . \tag{A.26}$$

7. Let $D_{2r} \in \mathbb{N}^{N \times N}$ be defined by $D_{2r} := D_2^r$. Use the previous exercise to show that $FD_{2r} = E^r F$ for all $r \in \mathbb{Z}^+$.

8. Given a vector $\mathbf{u} \in \mathbb{C}^N$ we can define a *circulant matrix* $\mathrm{Circ}(\mathbf{u}) \in \mathbb{C}^{N \times N}$ with entries given by

$$(\mathrm{Circ}(\mathbf{u}))_{i,j} = u_{(j-i) \bmod N}.$$

Show that $\mathrm{Circ}(\mathbf{u})F = \sqrt{N}FD$, where D is a diagonal matrix with entries given

by

$$(D)_{i,j} = \begin{cases} (F\mathbf{u})_j & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

In other words, show that any given circulant matrix will have the columns of F as its eigenvectors.

9. Let $g : [0,1] \to \mathbb{R}$ be a twice continuously differentiable and periodic function. Any such g will have a *Fourier series expansion* of the form

$$g(x) = \sum_{\omega \in \mathbb{Z}} c_\omega \mathrm{e}^{2\pi \mathrm{i} \omega x} \ \forall x \in [0,1], \tag{A.27}$$

where the *Fourier series coefficients* $c_\omega \in \mathbb{C}$ satisfy (i) $c_\omega = \overline{c_{-\omega}}$ for all $\omega \in \mathbb{Z}$, and (ii) $\sum_{\omega \in \mathbb{Z}} |c_\omega| < \infty$. Let $\mathbf{u} \in \mathbb{R}^N$ be a vector whose entries are given by $u_j = g(j/N)$ for all $j \in [N]$. Show that the vector $F\mathbf{u}$ has entries

$$(F\mathbf{u})_j = \sqrt{N} \sum_{\omega \equiv j \bmod N} c_\omega.$$

10. Let $\mathbf{u} \in \mathbb{R}^N$ be a vector whose entries are given by a twice continuously differentiable and periodic function $g : [0,1] :\to \mathbb{R}$ as follows: $u_j = g(j/N)$ for all $j \in [N]$. Use Taylor's Theorem to explain why $N^2 D_2 \mathbf{u}$ should approximate values from the second derivative of g quite well for large N, where D_2 is the matrix defined by (A.25). Next, use the Fourier series expansion of g'' to argue that the matrix E from (A.26) should have $(E)_{\omega,\omega} \approx -(2\pi\omega/N)^2$ for all $\omega \in [N]$.[4] Finally, verify that the entries of E do indeed have this property for large N.

11. **COMPUTATIONAL EXERCISE:** Implement both the FFT and the IFFT for vectors of size 2^n, $n \in \mathbb{N}$. Produce a plot showing that each is indeed faster than the corresponding naive method for directly computing the (I)DFT of an arbitrary vector.

[4] The author grants you special permission to ignore c_ω for all $\omega \notin [N]$ when you make your argument.

A.3 The Singular Value Decomposition (SVD)

The Singular Value Decomposition (SVD) is arguably the most useful fact of Linear Algebra, which is itself arguably the most useful and ubiquitous of mathematical subjects (with respect to computation in particular). The SVD's utility in data analysis is underscored by the fact that it has been (re)discovered at least three times in different scientific communities. In this section we will review the SVD of a given matrix $A \in \mathbb{C}^{n\times N}$. Most sections of the book hereafter will use the SVD repeatedly and often – it is well worth refreshing yourself here, and familiarizing yourself with our notation, before moving on.

The following theorem defines, and guarantees the existence of, the SVD for an arbitrary rectangular matrix $A \in \mathbb{C}^{n\times N}$.

Theorem A.3.1. *Every $A \in \mathbb{C}^{n\times N}$ can be decomposed into $A = U\Sigma V^*$, where:*

1. $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{N\times N}$ are both unitary, and

2. $\Sigma \in [0,\infty)^{n\times N}$ is a unique diagonal matrix,

$$\Sigma_{i,j} = \begin{cases} \sigma_j & \text{if } i = j < \text{the rank of } A \\ 0 & \text{otherwise} \end{cases},$$

with $\sigma_0 \geq \sigma_2 \geq \cdots \geq \sigma_{r-1} \geq 0$, where $r =$ the rank of $A \leq \min(n, N)$.

The j^{th}-largest diagonal entry of Σ, $\sigma_j \in [0,\infty)$, is called the j^{th} **singular value** *of A.*

The proof of Theorem A.3.1 follows from the next 2 lemmas which will also help us establish some notation.

Lemma A.3.2. *Let $\{\mathbf{w}_0, \cdots, \mathbf{w}_{N-1}\}$ be an orthonormal basis for $\mathbb{C}^N$. Define $s_j :=$*

$\|A\mathbf{w}_j\|_2$, *and*

$$\mathbf{h}_j := \begin{cases} \mathbf{0} & \text{if } s_j = 0 \\ \frac{1}{s_j} A\mathbf{w}_j \in \mathbb{C}^n & \text{if } s_j \neq 0 \end{cases}. \tag{A.28}$$

Finally, let W *to be the unitary matrix* $(\mathbf{w}_0 \cdots \mathbf{w}_{N-1}) \in \mathbb{C}^{N\times N}$. *Then,*

$$A = (\mathbf{h}_0 \cdots \mathbf{h}_{N-1}) \begin{pmatrix} s_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{N-1} \end{pmatrix} W^*.$$

Proof: We have

$$AW = (\mathbf{h}_0 \cdots \mathbf{h}_{N-1}) \begin{pmatrix} s_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{N-1} \end{pmatrix}, \tag{A.29}$$

and $W^{-1} = W^*$. Note that we are free to reorder the columns of W in order to achieve $s_0 \geq s_1 \geq \cdots \geq s_{N-1} \geq 0$. The resulting sequence of ordered s_j-values will be uniquely determined by W. The number of nonzero s_j-values will equal the rank of A if $\{\mathbf{w}_0, \cdots, \mathbf{w}_{N-1}\}$ contains an orthonormal basis for the null space of A as a subset. □

Lemma A.3.2 already yields a large family of decompositions for A with several of the structural properties promised by Theorem A.3.1. The next lemma tells us how to choose W in order to ensure that the $\mathbf{h}_j$-vectors from (A.28) also yield a matrix $U \in \mathbb{C}^{n\times n}$ that is unitary. As a happy coincidence, our choice of W will also both contain an orthonormal basis for the null space of A as subset of its columns, and guarantee the uniqueness of the ordered s_j-values subsequently induced by Lemma A.3.2. Theorem A.3.1 will be proven as a result.

Lemma A.3.3. *The matrix* $U := (\mathbf{h}_0 \cdots \mathbf{h}_{N-1})$ *from* (A.29) *will be essentially unitary if and only if* $\mathbf{w}_0, \cdots, \mathbf{w}_{N-1}$ *from* (A.28) *are an orthonormal set of eigenvectors of* $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{N\times N}$.

Proof: Let λ_l be the eigenvalue of A^*A associated with an eigenvector $\mathbf{w}_l$ for all $0 \leq l < N$. Considering the inner product of two nonzero $\mathbf{h}_j$-vectors from (A.28) we have that

$$\langle \mathbf{h}_j, \mathbf{h}_l \rangle = \frac{1}{s_j s_l}\langle A\mathbf{w}_j, A\mathbf{w}_l\rangle = \frac{1}{s_j s_l}(A\mathbf{w}_j)^* A\mathbf{w}_l = \frac{1}{s_j s_l}\mathbf{w}_j^*\left(A^*A\mathbf{w}_l\right) = \frac{\lambda_l}{s_j s_l}\mathbf{w}_j^*\mathbf{w}_l = 0$$

whenever $j \neq l$. Similarly, $\langle \mathbf{h}_j, \mathbf{h}_j\rangle = 1$ will also hold for all nonzero $\mathbf{h}_j$ by the definition of s_j.

Now suppose that U is unitary. Having both

$$1 = \langle \mathbf{h}_j, \mathbf{h}_j\rangle = \frac{1}{s_j^2}\langle A\mathbf{w}_j, A\mathbf{w}_j\rangle = \frac{1}{s_j^2}\langle \mathbf{w}_j, A^*A\mathbf{w}_j\rangle$$

and

$$0 = \langle \mathbf{h}_l, \mathbf{h}_j\rangle = \frac{1}{s_l s_j}\langle A\mathbf{w}_l, A\mathbf{w}_j\rangle = \frac{1}{s_l s_j}\langle \mathbf{w}_l, A^*A\mathbf{w}_j\rangle$$

hold whenever $l \neq j$ necessitates that $A^*A\mathbf{w}_j$ be a scalar multiple of $\mathbf{w}_j$. Thus, we must choose $\mathbf{w}_0, \cdots, \mathbf{w}_{N-1} \in \mathbb{C}^N$ to be an orthonormal set of eigenvectors of $\mathbf{A}^*\mathbf{A}$ in order to obtain a unitary $U = (\mathbf{h}_0 \cdots \mathbf{h}_{N-1})$ in (A.29).

Finally, if the columns for U do not collectively span $\mathbb{C}^n$ we may use Gram-Schmidt orthogonalization to add additional columns to U until it does (being careful to associate them with zero s_j-values). Then, we may discard any zero columns of U until it is $n \times n$ (again, being careful to remove the associated zero rows of the diagonal matrix of s_j-values in (A.29)). □

The singular value decomposition of a matrix reveals many of its most important properties. The following theorem summarizes some of the most fundamental of these.

Theorem A.3.4. *Consider the singular value decomposition of $A \in \mathbb{C}^{n\times N}$, $A = U\Sigma V^*$. Let r be the rank of A. The following statements hold:*

1. *The r nonzero singular values of A are exactly the square roots of the positive eigenvalues of A^*A or AA^*.*
2. *The first r columns of U form an orthonormal basis for the column space of A (i.e., for the range, or image, of A).*
3. *The last $n - r$ columns of U form an orthonormal basis for the null space, or kernel, of A^*.*
4. *The first r columns of V form an orthonormal basis for the row space of $\overline{A}$ (i.e., for the range, or image, of A^*).*
5. *The last $N - r$ columns of V form an orthonormal basis for the null space, or kernel, of A.*
6. *If $n = N$ and A is Hermitian, then A will have λ as an eigenvalue if and only if there exists a $j \in [r]$ such that*
   - *$|\lambda|$ is the j^{th} singular value of A (i.e., $\sigma_j = |\lambda|$),*
   - *the j^{th} column of V, $\mathbf{v}_j \in \mathbb{C}^N$, is an eigenvector of A associated with λ, and*
   - *the j^{th} column of $U = \mathrm{sign}(\lambda)\mathbf{v}_j$.*

Proof: See Exercise 2. □

We now turn to a theorem and some corollaries that can be used to quickly simplify bounds on singular values for the sum or products of matrices.

Theorem A.3.5. *Let $A, B \in \mathbb{C}^{n\times N}$ and $q = \min(n, N)$. The following inequalities hold for the singular values of $A, B, A + B$, and AB^**

1. $\sigma_{i+j-1}(A+B) \leq \sigma_i(A) + \sigma_j(B)$

2. $\sigma_{i+j-1}(AB^*) \leq \sigma_i(A)\sigma_j(B)$

for $1 \leq i, j \leq q$, $i + j \leq q + 1$.

Proof. The theorem and proof can be found in [17]. □

Corollary A.3.6. *1.* $|\sigma_i(A+B) - \sigma_i(A)| \leq \sigma_1(B)$

2. $\sigma_{i+j-1}(AB^*) \leq \sigma_i(A)\sigma_1(B)$

Proof. The proof of the corollary is left an exercise. □

We are now prepared to concentrate on several other useful properties of the singular values of a given matrix.

Homework A.3.1. Prove Corollary A.3.6 using Theorem A.3.5

Homework A.3.2. Suppose every entry of matrix $A \in \mathbb{C}^{n\times N}$ is corrupted with additive error in magnitude less than or equal to ϵ. How much can the singular value $\sigma_i(A)$ change for any $i = 1, \ldots, \min(n, M)$?

A.3.1 Singular Values and Matrix Norms

Now discuss how singular values relate to the operator norm and Frobenius norm of a matrix.

Definition A.3.7 (Frobenius Norm)**.** The Froebenius norm of $A \in \mathbb{C}^{n\times N}$ is

$$\|A\|_F = \sqrt{\sum_{\ell,j} |A_{\ell,j}|^2} = \sqrt{\text{Trace}(A^*A)}$$

The equivalence of the two quantities in the definition can be seen by considering that the diagonal entries of the matrix A^*A are in fact the ℓ^2-norm of the columns of A. That is$(A^*A)_{j,j} = \langle \mathbf{a}_j, \mathbf{a}_j \rangle = \|\mathbf{a}_j\|_2^2$. Noting that the trace is the sum of the

diagonal entries then we then have the equivalence of the two quantities seen in A.3.7. For the same reason, the Froebenius norm can also be understood as equivalent to the ℓ^2-norm of a vector with nM entries equal to the entries of A. To see how the singular values relate to this norm, consider

$$\begin{aligned}
\|A\|_F &= \sqrt{\text{Trace}(A^*A)} \\
&= \sqrt{\text{Trace}(V\Sigma^*U^*U\Sigma V^*)} \\
&= \sqrt{\text{Trace}(V\Sigma^*\Sigma V^*)} \\
&= \sqrt{\text{Trace}(V^*V\Sigma^*\Sigma)} \\
&= \sqrt{\text{Trace}(\Sigma^*\Sigma)} \\
&= \sqrt{\sum_{j=1}^{\min(n,N)} (\sigma_j(A))^2}
\end{aligned}$$

where we have used the cyclic property of the trace. Thus the $\|A\|_F$ is equivalent to the ℓ^2-norm of a vector formed by the singular values of A

We recall Definition 3.1.9, the operator norm of $A \in \mathbb{C}^{n\times N}$ and will show how the singular values relate to this norm.

The ℓ^2-operator norm of $A \in \mathbb{C}^{n\times N}$ is

$$\|A\|_{2\to 2} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|_2=1} \|U\Sigma V^*\mathbf{x}\|_2$$

Note that V is a unitary matrix, and so its columns form an orthonormal basis; so for $\mathbf{x} \in \mathbb{C}^N$ we have that $\exists \alpha \in \mathbb{C}^N$ such that

$$\mathbf{x} = \sum_{j=1}^{N} \alpha_j \mathbf{v}_j,\ \|\alpha\|_2 = \|\mathbf{x}\|_2 = 1$$

Using this expansion of $\mathbf{x}$ in the basis defined by V along with the SVD of A, we have

$$\begin{aligned}
\|A\|_{2\to 2} &= \sup_{\|\alpha\|_2=1} \|U\Sigma V^* \left(\sum_{j=1}^{N} \alpha_j \mathbf{v}_j\right) \|_2 \\
&= \sup_{\|\alpha\|_2=1} \|U\Sigma \left(\sum_{j=1}^{N} \alpha_j V^* \mathbf{v}_j\right) \|_2 \\
&= \sup_{\|\alpha\|_2=1} \|U\Sigma \left(\sum_{j=1}^{N} \alpha_j \mathbf{e}_j\right) \|_2 \\
&= \sup_{\|\alpha\|_2=1} \|U \left(\sum_{j=1}^{N} \alpha_j \sigma_j(A) \mathbf{e}_j\right) \|_2 \\
&= \sup_{\|\alpha\|_2=1} \sqrt{\sum_{j=1}^{N} \left(\alpha_j \sigma_j(A)\right)^2}
\end{aligned}$$

Since the singular values are non-negative and in descending order, the sum above is maximized for $\alpha = e_1$ i.e. $\alpha_1 = 1$, $\alpha_j = 0 \forall j \neq 1$. Thus $\|A\|_{2\to 2} = \sigma_1(A)$.

A.3.2 Exercises

1. Use the singular value decomposition to help you construct a matrix $A \in \mathbb{R}^{2\times 2}$ with no eigenvectors whose square, $A^2 \in \mathbb{R}^{2\times 2}$, has two eigenvectors. How many eigenvectors will A^{709} have?

2. Prove Theorem A.3.4.

3. Use the singular value decomposition of $A \in \mathbb{C}^{n\times N}$ to help you find a formula for a matrix $A^\dagger \in \mathbb{C}^{N\times n}$ which inverts A on its range. More specifically, write a formula for a matrix $A^\dagger$ that satisfies both $AA^\dagger A = A$ and $A^\dagger A A^\dagger = A^\dagger$. When will $A^\dagger = A^{-1}$ hold?

4. Let $\alpha, \beta \in \mathbb{Z} \setminus \{0\}$. The $\frac{\alpha}{\beta}$-power of a full rank matrix $A \in \mathbb{C}^{N\times N}$ is a matrix $B \in \mathbb{C}^{N\times N}$ with the property that $B^\beta = A^\alpha$. Prove that there always exists a unitary matrix $W \in \mathbb{C}^{N\times N}$ such that the $\frac{\alpha}{\beta}$-power of AW exists. When can W

simply be the identity? How can one compute such a B and W for any given $A \in \mathbb{C}^{N \times N}$?

–Operator form, sum of rank 1 matrices –matrix norms –Best low rank approximation + regression (fitting a pointset...) – Computation section – Weyl's bounds and stability...

A.3.3 Computing the SVD of a Matrix

Notice that

$$\begin{aligned}\mathbf{A}^*\mathbf{A} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^*(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*) \\ &= \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \\ &= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^*.\end{aligned}$$

Then, $\mathbf{V}$ contains the eigenvectors of $\mathbf{A}^*\mathbf{A}$ as columns, and $\sigma_1, \sigma_2, \cdots, \sigma_q$ are the squared eigenvalues of $\mathbf{A}^*\mathbf{A}$.

Numerically, we can use, e.g., the QR algorithm to find the eigenvalues of $\mathbf{A}^*\mathbf{A}$ to get the singular values of $\mathbf{A}$. The shifted inverse power method, e.g., can be used to calculate $\mathbf{V}$. Similarly, from $\mathbf{A}\mathbf{A}^*$ we can find $\mathbf{U}$.

A.4 A Brief Introduction to Linear and Semidefinite Programming?

A.5 A brief review of computational complexity and asymptotic notation?

Bibliography

Bibliography

[1] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468, 2006.

[2] J. Bailey, M. A. Iwen, and C. V. Spencer. On the design of deterministic matrices for fast recovery of fourier compressible functions. *SIAM Journal on Matrix Analysis and Applications*, 33(1):263–289, 2012.

[3] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 97104, New York, NY, USA, 2006. Association for Computing Machinery.

[4] L. I. Bluestein. A Linear Filtering Approach to the Computation of Discrete Fourier Transform. *IEEE Transactions on Audio and Electroacoustics*, 18:451–455, 1970.

[5] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan. Time bounds for selection. *Journal of Computer and System Sciences*, 7(4):448 – 461, 1973.

[6] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, Inc., 2001.

[7] J. Cooley and J. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19:297–301, 1965.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to algorithms. *2nd Edition*, 2001.

[9] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, SCG '04, page 253262, New York, NY, USA, 2004. Association for Computing Machinery.

[10] R. A. DeVore. Deterministic constructions of compressed sensing matrices. *Journal of Complexity*, 23(4):918 – 925, 2007.

[11] P. Dusart. The kth prime is greater than $k(\ln k + \ln \ln k - 1)$ for $k \geq 2$. *Math. Comp.*, 68(225):411–415, 1999.

[12] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, 2013.

[13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[14] M. Frigo and S. Johnson. The design and implementation of fftw3. *Proceedings of IEEE 93 (2)*, pages 216–231, 2005.

[15] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[16] M. Heideman, D. Johnson, and C. Burrus. Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, 1(4):14–21, 1984.

[17] R. Horn, R. Horn, and C. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

[18] M. Iwen. Compressed sensing with sparse binary matrices: Instance optimal error guarantees in near-optimal time. *Journal of Complexity*, 30(1):1 – 15, 2014.

[19] M. A. Iwen. A deterministic sub-linear time sparse fourier algorithm via non-adaptive compressed sensing methods. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA 08, pages 20–29, USA, 2008. Society for Industrial and Applied Mathematics.

[20] M. A. Iwen. Simple deterministically constructible rip matrices with sublinear fourier sampling requirements. In *2009 43rd Annual Conference on Information Sciences and Systems*, pages 870–875, March 2009.

[21] B. S. Kašin. The diameters of octahedra. *Uspehi Mat. Nauk*, 30(4(184)):251–252, 1975.

[22] D. R. Kincaid and E. W. Cheney. *Numerical analysis: mathematics of scientific computing*, volume 2. Brooks/Cole Pacific Grove, CA, 2002.

[23] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[24] M. A. Kronrod. An optimal ordering algorithm without a field of operation. *Dokl. Akad. Nauk SSSR*, 186:1256–1258, 1969.

[25] K. G. Larsen and J. Nelson. Optimality of the johnson-lindenstrauss lemma. *CoRR*, abs/1609.02094, 2016.

[26] R. Motwani, A. Naor, and R. Panigrahi. Lower bounds on locality sensitive hashing. In *Proceedings of the Twenty-Second Annual Symposium on Computational Geometry*, SCG '06, page 154157, New York, NY, USA, 2006. Association for Computing Machinery.

[27] D. Pasetto and A. Akhriev. A comparative study of parallel sort algorithms. In *Proceedings of the ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion*, OOPSLA 11, page 203204, New York, NY, USA, 2011. Association for Computing Machinery.

[28] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier Analysis*. Springer, 2018.

[29] L. Rabiner, R. Schafer, and C. Rader. The Chirp z-Transform Algorithm. *IEEE Transactions on Audio and Electroacoustics*, AU-17(2):86–92, June 1969.

[30] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36):13212–13217, 2008.

[31] L. Trefethen and D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1997.

[32] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. 09 2018.