Compressed Sensing

Felix Krahmer
Office hour Mo 1-2, office 104
Tel 10584, f.krahmer@math.uni-goettingen.de

~~Class when~~?

Bode

1. Introduction

We start by considering two motivating examples.

1.1 Imaging

Image compression schemes, such as JPG or JPG 2000, follow the general two-step strategy

1. Acquire the complete image in full resolution
2. Prosess stored image, throw away large part of information

→ Works for digital cameras, as sensors are cheap

→ Unnecessary cost for Infrared imaging (sensors are expensive)
Unnecessary time expense for Magnetic Resonance imaging. (subsequent measurements)

$\boxed{\text{Can we work with less sensors / measurements?}}$

Observation: Information in images is relatively small, but localized.

→ if we just choose a subset of the pixels, we will miss relevant parts and have more than we need of "noninteresting" parts.

Goal: Sensors should retrieve signal, which captures information from "everywhere".

→ MRI measurements: Fourier transform evaluation provides good model, does the job.

→ Infrared imaging: Use micromirror array, the sensor measures linear combination of subset of the pixels

Mirror points towards sensor → pixel is included in the sum
Mirror points away → pixel is not included

Reorienting the mirrors → different linear combination,
<u>one</u> sensor can make all the measurements

1.2 Machine Learning

Suppose you want to learn a function

$$f(t) = \sum_{\ell=1}^{N} x_\ell \psi_\ell(t) + \text{Noise}$$

from samples $f(t_j)$, where ψ_ℓ is a known representation system.

Need to find N coefficients, ⟶ need N samples (at least)
In practical applications, one only has access to less.

<u>Guiding paradigm in machine learning:</u>
If there are several possible solutions take the simplest one.

1.3 Sparsity and Compressibility

One way to make "simple" rigorous is that it should involve only few of the ψ_ℓ's.

Similarly, JPG / JPG 2000 are based on preserving only the largest coefficients in some representation system.

Notation $\|x\|_p := \left(\sum_{j=1}^{N} |x_j|^p \right)^{1/p}$ $0 < p < \infty$

$\|x\|_\infty := \sup_j |x_j|$

p-norm / quasi norm

$[N] = \{1, \dots, N\}$

Def 1.1 (Sparsity, Support)

For ~~an~~ a vector $x \in \mathbb{R}^N$ or $\mathbb{C}^N$ its support is defined as
$$\operatorname{supp} x = \{k : x_k \neq 0\}.$$
x is k-sparse if $|\operatorname{supp} x| \leq s$.

Furthermore, denote $\|x\|_0 := |\operatorname{supp} x|$; ~~we speak~~ with some abuse of notation we speak of the 0-norm. (as $\|x\|_p \xrightarrow[p \to 0]{} \|x\|_0$)

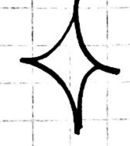
Even though keeping only few coefficients will provide a good approximation of the signal, the signal will not be exactly sparse. Rather, the following quantity will be small

Def 1.2 (error of best s-term approximation)

The error of best s-term approximation of a vector $x \in \mathbb{R}^N$ or $\mathbb{C}^N$ is
$$\sigma_s(x)_p := \inf\{\|x - z\|_p,\ z \text{ is } s\text{-sparse}\}$$

A vector x is called compressible if $\sigma_s(x)_p$ decays quickly ~~small~~ in s. Ideally, a meaningful compressed sensing result should have a variant for compressible, not just for sparse vectors.

The following proposition shows that a good model for compressible vectors is the ℓ_p-ball, $p < 1$.



Proposition 1.3 (Stechkin)

For any $q>p>0$ and any $x\in\mathbb{C}^N$,

$$\sigma_s(x)_q \le \frac{1}{s^{\frac{1}{p}-\frac{1}{q}}}\|x\|_p$$

The proof uses the ~~notion of~~ following definition

Definition 1.4 The nonincreasing rearrangement $x\in\mathbb{C}^N$ is the vector $x^*\in\mathbb{R}_+^N$ for which

$$x_1^*\ge x_2^*\ge\dots\ge x_N^*\ge 0$$

and there is a permutation $\pi:[N]\to[N]$ with $x_j^*=|x_{\pi(j)}|$ for all $j\in[N]$.

Proof of Prop. 1.3: Let x^* be the decreasing rearrangement of $x\in\mathbb{C}^N$, we have

$$\sigma_s(x)_q^q=\sum_{j=s+1}^N (x_j^*)^q\le (x_s^*)^{q-p}\sum_{j=s+1}^N (x_j^*)^p$$

$$\underset{x_s^*\le x_\ell^*,\ \ell\le s}{\le}\left(\frac{1}{s}\sum_{j=1}^s (x_j^*)^p\right)^{\frac{q-p}{p}}\left(\sum_{j=s+1}^N (x_j^*)^p\right)$$

$$\le\left(\frac{1}{s}\|x\|_p^p\right)^{\frac{q-p}{p}}\|x_p^p\|=\frac{1}{s^{q/p-1}}\|x_p\|^q$$

Take $\frac{1}{q}$-th power to obtain the proposition. $\square$

Remark 1.5: ~~Note that the~~ Using convex analysis, The constant can ~~in fact~~ be refined to obtain

$$\sigma_s(x)_q\le\frac{c_{p,q}}{s^{1/p-1/q}}\|x\|_p$$

where

$$c_{p,q}:=\left[\left(\frac{p}{q}\right)^{p/q}\left(1-\frac{p}{q}\right)^{1-p/q}\right]^{1/p}$$

For $p=1$, $q=2$ this yields

$$\sigma_s(x)_2\le\frac{1}{2\sqrt{s}}\|x\|_1$$

1.4 The compressed sensing problem

Note that the two examples yield problems of the same structure

$$\boxed{\text{Solve } A\cancel{x} = y \quad \text{knowing that } x \text{ is sparse (or compressible)} \qquad (P)}$$

Does this problem even have a unique solution?
Not always → choose A as a subset of rows of I.

~~sec~~ 1.6

Theorem ~~2.13~~ 1.6

Given $A \in \mathbb{C}^{m\times N}$, the following properties are equivalent

(a) every s-sparse vector $x \in \mathbb{C}^N$ is the unique s-sparse solution of $Az = Ax$, that is, if $A\cancel{x} = Az$ and both x and z are s-sparse, then $x = z$.

(b) The null space $\ker A$ does not contain any $2s$-sparse vector other than the zero vector, that is, $\ker A \cap \{z \in \mathbb{C}^N : \|z\|_0 \leq 2s\} = \{0\}$

(c) For every $S \subset [N]$ with $\operatorname{card}(S) \leq 2s$, the submatrix A_S is injective

(d) Every set of $2s$ columns of A is linearly independent

Proof: (a) ⟺ (b) Let x and z be s-sparse with $A\cancel{x} = Az$. Then $x - z$ is $2s$-sparse and $A(x-z) = 0$. If the kernel does not contain any $2s$-sparse vector different from the zero vector, then $x = z$.

Conversely, assume that for every s-sparse vector ~~different from the zero vector then x = z~~ $x \in \mathbb{C}^N$ we have $\{z \in \mathbb{C}^N : Az = Ax;\ \|z\|_0 \leq s\} = \{x\}$. Let $v \in \ker A$ be $2s$-sparse. We can write $v = x - z$ for s-sparse vectors x, z with $\operatorname{supp} x \cap \operatorname{supp} z = \emptyset$.

Then $Ax = Az$ and by assumption $x = z$.
Since the supports of x and z are disjoint
it follows that $x = z = 0$ and $v = 0$ ↯

For the equivalence of (b), (c), (d). ~~we observe~~
~~result for a 2s-sparse vector v with S = supp v~~
~~we have $Av = A_S v_S$~~ use linear algebra. ☐

Lecture II

Last time:

(i) Every s-sparse $x \in \mathbb{C}^N$ is the unique solution of $Ax = Az$ $\iff$ (iv) Every set of 2s columns is linearly independent.

(iv) $\Rightarrow m \geq 2s$ $\rightsquigarrow$ Need at least 2s measurements
in fact this number is also sufficient

Theorem 1.7 For any $N \geq 2s$, there exists a measurement matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ rows such that every s-sparse vector $x \in \mathbb{C}^N$ can be recovered from its measurement vector $y = Ax \in \mathbb{C}^m$ as ~~a solution of~~ the unique solution of minimal support.

Proof: Fix $t_N > \dots > t_2 > t_1 > 0$ and consider the matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ defined by

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_N \\ \vdots & \vdots & & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & & t_N^{2s-1} \end{pmatrix}$$

The square matrix $A_S \in \mathbb{C}^{2s \times 2s}$ obtained by restricting A to the columns in a set $S = \{j_1 < \dots < j_{2s}\}$ is a Vandermonde matrix.
As proved in the homework, $\det(A_S) = \prod_{k<\ell} (t_{j_\ell} - t_{j_k}) > 0$,
so A_S is invertible and hence injective. The result follows from Theorem 1.6. ☐

Note that in the proof of Theorem 1.7, the assumption that the t_i are positive is not necessary, all we need is that they are pairwise different.

In particular, consider the matrix consisting of the first $m = 2s$ rows of the discrete Fourier transform (DFT) matrix $M = \left(e^{2\pi i \frac{(\ell-1)(k-1)}{N}}\right)_{k,\ell=1}^{N}$, i.e.,

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{2\pi i \cdot \frac{1}{N}} & e^{2\pi i \frac{2}{N}} & \cdots & e^{2\pi i \frac{N-1}{N}} \\ 1 & e^{2\pi i \frac{2}{N}} & e^{2\pi i \frac{2 \cdot 2}{N}} & & \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & e^{2\pi i \frac{(2s-1)}{N}} & e^{2\pi i \frac{2(2s-1)}{N}} & \cdots & e^{2\pi i \frac{(2s)(N-1)}{N}} \end{pmatrix}$$

is suitable. Applying A to a vector $x \in \mathbb{C}^N$ is the same as computing the first $2s$ ~~to obtain~~ Fourier coefficients of the function ~~$x: [N] \to \mathbb{C}, x(k) = x_k$~~.

$x : [N] \to \mathbb{C}$, $x(k) = x_k$, that is

$$\hat{x}(j) := \sum_{k=0}^{N-1} x(k) e^{-2\pi i \frac{jk}{N}}, \quad 0 \le j \le N-1$$

We will now derive an efficient method to find the s-sparse solution x.

Define the polynomial of degree s

$$p(t) := \prod_{k \in S} \left(1 - e^{-i2\pi k/N} e^{i2\pi t/N}\right), \quad \text{where } S = \operatorname{supp} x.$$

Note that
$$\left.\begin{aligned} t \in S &\Rightarrow p(t) = 0 \\ t \notin S &\Rightarrow x(t) = 0 \end{aligned}\right\} \Longrightarrow p(t)x(t) = 0 \quad \forall t \in [N]$$

Thus $\forall\, 0 \le j \le N-1$

$$0 = \widehat{p(\cdot)x(\cdot)}[j] = \sum_{k=0}^{N-1} p(k)x(k)\, e^{-2\pi i \frac{jk}{N}}$$

To estimate this, we use the following Lemma

<u>Lemma 1.8</u>: Let $y : [N] \to \mathbb{C}$. Then $y(k) = \frac{1}{N} \sum_{\ell=0}^{N-1} \hat{y}(\ell) e^{+2\pi i \frac{k\ell}{N}}$

<u>Proof:</u>
$$\sum_{\ell=0}^{N-1} \hat{y}(\ell) e^{2\pi i \frac{k\ell}{N}} = \sum_{\ell=0}^{N-1} \sum_{k'=0}^{N-1} y(k')\, e^{2\pi i \frac{(k-k')\ell}{N}}$$
$$= \sum_{k'=0}^{N-1} y(k') \underbrace{\sum_{\ell=0}^{N-1} e^{2\pi i \frac{(k-k')\ell}{N}}}_{= \begin{cases} 0 & k \ne k' \\ N & k = k' \end{cases}} = N\, y(k) \qquad \square$$

Thus $$0 = \sum_{k=0}^{N-1} \left(\frac{1}{N} \sum_{\ell=0}^{N-1} \hat{p}(\ell) e^{2\pi i \frac{k\ell}{N}} \frac{1}{N} \sum_{m=0}^{N-1} \hat{x}(m) e^{2\pi i \frac{km}{N}} \right) e^{-2\pi i \frac{jk}{N}}$$

$$= \sum_{\ell,m=0}^{N-1} \frac{1}{N^2} \hat{p}(\ell) \hat{x}(m) \underbrace{\sum_{k=0}^{N-1} e^{2\pi i \frac{k(\ell+m-j)}{N}}}_{= \begin{cases} 0, & \ell \not\equiv j-m \mod N \\ N & \ell \equiv j-m \mod N \end{cases}}$$

$$= \frac{1}{N} \sum_{\ell=0}^{N-1} \hat{p}(\ell) \hat{x}(j-\ell) \qquad (*)$$

Note that $p(t)$ is a polynomial of degree s, thus

$$p(t) = 1 + \sum_{\ell=1}^{s} \alpha_\ell e^{2\pi i \frac{\ell t}{N}} \Rightarrow \hat{p}(0) = N, \quad \hat{p}(\ell) = 0 \text{ for } s < \ell < N$$

and $(*)$ becomes

$$\begin{pmatrix} \hat{x}(s-1) & \hat{x}(s-2) & \dots & \hat{x}(0) \\ \hat{x}(s) & \hat{x}(s-1) & \dots & \hat{x}(1) \\ \vdots & & & \\ \hat{x}(2s-2) & & \dots & \hat{x}(s-1) \end{pmatrix} \begin{pmatrix} \hat{p}(1) \\ \hat{p}(2) \\ \vdots \\ \hat{p}(s) \end{pmatrix} = - \begin{pmatrix} \hat{x}(s) \\ \hat{x}(s+1) \\ \vdots \\ \hat{x}(2s-1) \end{pmatrix} \qquad (**)$$

Recall: $\hat{x}(0), \dots, \hat{x}(2s-1)$ are known.

Can we solve for $(\hat{p}(1), \dots \hat{p}(s))$?

No, e.g., $x = [1, 0, \dots, 0]^T \Rightarrow$ All entries of matrix are 1
$\hat{x} = [1, \dots, 1]^T$

Hence we only find some solution $[\hat{q}(1), \dots, \hat{q}(s)]$

Recall: We also had $\hat{p}(0) = 1$, $\hat{p}(k) = 0$, $k > s$,

so extend $\hat{q}$ accordingly by $\hat{q}(0) = 1$, $\hat{q}(k) = 0$, $k > s$.

Note that with this extension, $(**)$ is equivalent to $(*)$ and $(*)$ is equivalent to $x(t)\, q(t) = 0$ for all $0 \le t < N$

Thus q vanishes on S. Since q is of degree s, there cannot be more zeros $\Rightarrow$ support of x can be found by solving for zeros of q. We obtain

Algorithm 1.9 (Prony-Method)

1. Find some solution $[\hat{q}(1), \dots \hat{q}(s)]^T$ of $(**)$
2. Find $\{$zeros of $q\} = \operatorname{supp} x$
3. Solve overdetermined system given by A's columns corresponding to the support.

Remark: Drawbacks of this method
- Sensitivity to noise
- Works for exactly sparse vectors, not for compressible.

Goal of this class:

- Find criteria for when a matrix allows for approximate recovery even when the ~~exact~~ solution is just compressible / there is noise
- Find matrices that have such properties.

Class II Side 11

2. ℓ_1-minimization and the Restricted Isometry Property

Motivation ⓐ: In a more general setup, no efficient algorithm like Prony's method is known. While uniqueness still holds under the above cond. there is no efficient algorithm to find the ℓ_0-minimizer $\operatorname{argmin}_{Ax=y} \|x\|_0$.

Idea: Use convex relaxation

$$\boxed{x^{\#} = \operatorname{argmin}_{Ax=y} \|x\|_1} \quad (P_1)$$

Motivation ⓑ: The condition above ensures invertibility of submatrices, but there may be bad conditioning.
Use same idea, but require submatrices of $2s$ colums to be well-conditioned rather than just invertible

Definition 2.1 (RIP)

A matrix $A \in \mathbb{C}^{m \times N}$ has the restricted isometry property (RIP) of order k and level δ if for all s-sparse $x \in \mathbb{C}^N$

$$(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2 .$$

The restricted isometry constant $\delta_k(A)$ is the smallest δ such that A has (k,δ)-RIP

Goal of this section: ~~Condition~~ Under the assumption of RIP, (P_1) can be used for recovery.

Lecture III

We work in generalized setup:

(stability) (a) x is not assumed to be sparse, just compressible

(robustness) (b) the measurements are affected by noise: $y = Ax + e$ with $\|e\|_2 \leq \eta$.

Problem with (b): The solution we are looking for is not an exact solution of $Ax = y$ any more

Thus we consider the problem

$$x^\# = \underset{\|Ax-y\|_2 \leq \eta}{\operatorname{argmin}} \|x\|_1 \qquad (P_{1,\eta})$$

instead.

Important tool: Null-space property. As we will see, equivalent to the fact that (P_1) recovers compressible vectors to a satisfactory degree. $\Leftrightarrow$ links RIP and P_1

2.1 THE ~~NEAR~~ ROBUST NULL SPACE PROPERTY

Definition 2.2 (Robust Null-Space-Property)

The matrix $A \in \mathbb{C}^{m \times N}$ is said to ~~satisfy~~ have the robust null space property (with respect to the ℓ_2-norm) with constants $0 < \rho < 1$ and $\tau > 0$ relative to a set $S \subseteq [N]$ if

$$\|v_S\|_1 \leq \rho \|v_{S^c}\|_1 + \tau \|Av\|_2 \quad \text{for all } v \in \mathbb{C}^N$$

It is said to have the robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ if it satisfies the robust null space property with constants ρ, τ relative to any set S with $\operatorname{card}(S) \leq s$.

Remark: The name null space property stems from its implication $\|v_S\|_1 \leq \rho \|v_{\bar{S}}\|_1$ for vectors $v \in \ker A$.

Theorem 2.3 Suppose that a matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{C}^N$, a solution $x^{\#}$ of $(P_{1,\eta})$ with $y = Ax + e$ ~~and~~ and $\|e\|_2 \le \eta$ approximates the vector x with ℓ_1-error

$$\|x - x^{\#}\|_1 \le \frac{2(1+\rho)}{(1-\rho)} \sigma_s(x)_1 + \frac{4\tau}{1-\rho} \eta. \qquad (+)$$

Remark: This is exactly what we aimed for: When the noise is small, vectors x close to sparse vectors are approximately recovered.

In fact the robust NSP is necessary and sufficient to have stable and robust recovery in the sense of $(+)$; we have the following ~~equivalence~~ stronger result

Theorem 2.4 The matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S if and only if

$$\|z - x\|_1 \le \frac{1+\rho}{1-\rho}\left(\|z\|_1 - \|x\|_1 + 2\|x_{S^c}\|_1\right) + \frac{2\tau}{1-\rho}\|A(z-x)\|_2$$

for all vectors $x, z \in \mathbb{C}^N$.

For the proof, use

Lemma 2.5 Given a set $S \subset [N]$ and vectors $x, z \in \mathbb{C}^N$, $\|(x-z)_{S^c}\|_1 \le \|z\|_1 - \|x\|_1 + \|(x-z)_S\|_1 + 2\|x_{S^c}\|_1$

Proof: Write $\|x\|_1 = \|x_{S^c}\|_1 + \|x_S\|_1 \le \|x_{S^c}\|_1 + \|(x-z)_S\|_1 + \|z_S\|_1$

Triangle inequality

and bound

$$\|(x-z)_{S^c}\|_1 \le \|x_{S^c}\|_1 + \|z_{S^c}\|_1$$

Sum up the inequalities to obtain

$$\|x\|_1 + \|(x-z)_{S^c}\|_1 \le 2\|x_{S^c}\|_1 + \|(x-z)_S\|_1 + \|z\|_1,$$ as desired $\square$

Proof of Thm 2.4:

"$\Leftarrow$" Assume that the matrix A satisfies ~~(2.4)~~ ⓪ for all vectors $x, z \in \mathbb{C}^N$. Thus, for $v \in \mathbb{C}^N$, taking $x = -v_S$ and $z = v_{S^c}$ yields

$$\|v\|_1 \le \frac{1+\rho}{1-\rho}\left(\|v_{S^c}\|_1 - \|v_S\|_1\right) + \frac{2\tau}{1-\rho}\|Av\|_2$$

and thus, multiplying by $1-\rho$,

$$(1-\rho)\left(\|v_S\|_1 + \|v_{S^c}\|_1\right) \le (1+\rho)\left(\|v_{S^c}\|_1 - \|v_S\|_1\right) + 2\tau\|Av\|_2$$

Rearrange to obtain

$$2\|v_S\|_1 \le 2\rho\|v_{S^c}\|_1 + 2\tau\|Av\|_2 .$$

Divide by 2 to obtain the robust NSP, as desired.

"$\Rightarrow$" Assume that A has the robust NSP relative to S. For $x, z \in \mathbb{C}^N$, set $v := z - x$. Then ~~the robust NSP and Lemma 2.5 yield~~

$$\|v_S\|_1 \le \rho\|v_{S^c}\|_1 + \tau\|Av\| \qquad \text{(Robust NSP)}$$

$$\|v_{S^c}\|_1 \le \|z\|_1 - \|x\|_1 + \|v_S\|_1 + 2\|x_{S^c}\|_1 \qquad \text{(Lemma 2.5)}$$

Combining the two equations yields (~~adding~~ adding, dividing by $(1-\rho)$)

$$\|v_{S^c}\|_1 \le \frac{1}{1-\rho}\left(\|z\|_1 - \|x\|_1 + 2\|x_{S^c}\|_1 + \tau\|Av\|_2\right)$$

Now

$$\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \overset{\text{robust NSP}}{\le} (1+\rho)\|v_{S^c}\|_1 + \tau\|Av\|_1$$

$$\le \frac{1+\rho}{1-\rho}\left(\|z\|_1 + \|x\|_1 + 2\|x_{S^c}\|_1\right) + \frac{2\tau}{1-\rho}\|Av\|_2$$

as desired $\square$

Proof of Theorem 2.3: Choose $z = x^\#$ in Theorem 2.4, $S = \{\text{largest } s \text{ entries of } x\}$. As $x^\#$ is the ℓ^1-minimizer, $\|z\|_1 \le \|x\|_1$, so

$$\|x - x^\#\|_1 \le \frac{2(1+\rho)}{1-\rho}\, 2\cdot\underbrace{\|x_{S^c}\|_1}_{\sigma_s(x)_1} + \frac{2\tau}{1-\rho}\underbrace{\|Ax - Ax^\#\|_2}_{= \|(Ax-y) - (Ax^\# - y)\|_2 \le 2\eta} \qquad \square$$

2.2 Restricted isometries

Recall A has the (δ,s)-RIP if for all s-sparse $x \in \mathbb{C}^N$

$$(1-\delta)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2$$

Normalize $\|x\|_2 = 1$, then

$$-\delta \le \|Ax\|_2^2 - 1 \le \delta$$

or

$$\left|\, \|Ax\|_2^2 - 1 \right| \le \delta .$$

So

$$\delta_s(A) = \sup_{\substack{x \in S^{N-1} \\ x \text{ s-sparse}}} \left|\, \|Ax\|_2^2 - 1 \right|$$

$$= \sup_{x \in S^{N-1}} \sup_{|S| \le s} \left|\, \|A_S x_S\|_2^2 - 1 \right|$$

restriction to columns indexed by S

restriction to entries indexed by S

$$= \sup_{\substack{|S| = s \\ S \subset [N]}} \sup_{x \in S^{s-1}} \left|\, \|A_S x\|_2^2 - 1 \right|$$

$$= \sup_{|S| = s} \sup_{x \in S^{s-1}} x^*(A_S^* A_S - I)x$$

$$= \sup_{|S| = s} \|A_S^* A_S - I\|$$

Spectral norm
$\|A\| = \sup_{\|x\|_2 = 1} \|Ax\|_2$

Here, support set on the left and on the right are the same.

Restricted Isometry Property has also implications for different support sets.

<u>Proposition 2.6</u> Let $u, v \in \mathbb{C}^N$ be vectors with $\|u\|_0 \le s$ and $\|v\|_0 \le t$. If $\text{supp}(u) \cap \text{supp}(v) = \emptyset$, then

$$|\langle Au, Av\rangle| \le \delta_{s+t} \|u\|_2 \|v\|_2 .$$

<u>Proof:</u>

Proof: Let $S := \operatorname{supp}(u) \cup \operatorname{supp}(v)$ and let u_S, v_S be the restrictions of $u, v \in \mathbb{C}^N$ to S. Since u and v have disjoint supports, we have $\langle u_S, v_S \rangle = 0$.
It follows that

$$
\begin{aligned}
|\langle Au, Av \rangle| &= |\langle A_S u_S, A_S v_S \rangle - \langle u_S, v_S \rangle| \\
&= |\langle (A_S^* A_S - Id) u_S, v_S \rangle| \\
&\leq \|(A_S^* A_S - Id) u_S\|_2 \, \|v_S\|_2 \\
&\leq \underbrace{\|A_S^* A_S - Id\|}_{\leq \delta_{|S|} = \delta_{s+t}} \underbrace{\|u_S\|_2}_{=\|u\|_2} \underbrace{\|v_S\|_2}_{=\|v\|_2}
\end{aligned}
$$

$\square$

~~Theorem~~

Theorem 2.7: Suppose that the $2s$th restricted isometry constant of the matrix $A \in \mathbb{C}^{m \times N}$ satisfies

$$\delta_{2s} < \frac{1}{3}$$

Then for any $x \in \mathbb{C}^N$ and $y \in \mathbb{C}^m$ with $\|Ax - y\|_2 \leq \eta$ a solution $x^\#$ of

$$\underset{z \in \mathbb{C}^N}{\text{minimize}} \ \|z\|_1 \quad \text{subject to} \quad \|Az - y\|_2 \leq \eta$$

approximates x with errors

$$\|x - x^\#\|_1 \leq C \sigma_s(x)_1 + D\sqrt{s}\,\eta,$$

$$\|x - x^\#\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(x)_1 + D\eta.$$

where the constants C, D depend only on δ_{2s}

~~Proof: Next time~~

Remark: State-of-the-art result: Same holds when ~~$\delta_{2s} < 0.6246$~~ $\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246$

Sharp bound: $\delta_s < \frac{1}{3} \Rightarrow \ell_1$-minimization works

$\delta_s > \frac{1}{2} \Rightarrow \exists$ a matrix such that ℓ_1-minimization does not work

2.3 The ℓ^2-robust Null space property

Until now: error bound $\|x - x^\#\|_1 \le \dots$

We seek bounds of the form $\|x - x^\#\|_2 \le \dots$

Need modified null space property.

Definition 2.7 (ℓ_2-robust null-space property)

A matrix $A \in \mathbb{C}^{m \times N}$ is said to have the ℓ_2-robust null space property of order s with respect to the ℓ_2-norm with constants $0 < \rho < 1$ and $\tau > 0$ if, for any set $S \subset [N]$ with $\operatorname{card}(S) \le s$,

$$\|v_S\|_2 \le \frac{\rho}{\sqrt{s}} \|v_{S^c}\|_1 + \tau \|Av\|_2 \quad \text{for all } v \in \mathbb{C}^N.$$

Theorem 2.8 Suppose that $A \in \mathbb{C}^{m \times N}$ has the ℓ_2-robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{C}^N$, a solution $x^\#$ of $(P_{1,\eta})$ with $y = Ax + e$ and $\|e\|_2 \le \eta$ approximates x with ℓ_2-error

$$\|x - x^\#\|_2 \le \frac{C}{\sqrt{s}} \sigma_s(x)_1 + D \eta, \quad 1 \le p \le 2$$

for some constants $C, D > 0$ depending only on ρ and τ.

~~Remark: For $p = 2$, we obtain $\|x - x^\#\|_2 \le \frac{C}{\sqrt{s}} \sigma_s(x)_1 + D\eta$~~

We again prove a stronger statement. ~~showing an equivalence between NSP and recovery properties.~~

Theorem 2.8 follows in the same way as the implication "Theorem 2.4 $\Rightarrow$ Theorem 2.3"

Theorem 2.9 Suppose $A \in \mathbb{C}^{m \times N}$ has the ℓ_2-robust NSP of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $x, z \in \mathbb{C}^N$ and $1 \leq p \leq q$

$$\| z - x \|_{\cancel{p}\,2} \leq \frac{C}{s^{\frac{1}{2}}} \left(\|z\|_1 - \|x\|_1 + 2\,\sigma_s(x)_1 \right) + D\, \| A(z-x) \|_2$$

where $C := (1+\rho)^2/(1-\rho)$ and $D := (3+\rho)\tau/(1-\rho)$

Proof:

The ℓ_2-robust NSP implies the Robust NSP as defined in Definition 2.2 with modified constants:

$$\| v_S \|_1 \underset{\substack{\uparrow \\ \text{Cauchy} \\ \text{-Schwartz}}}{\leq} \sqrt{s}\, \| v_S \|_2 \underset{\substack{\uparrow \\ \ell^2\text{-robust} \\ \text{NSP}}}{\leq} \rho \| v_{S^c} \| + \tau \sqrt{s}\, \| A v \|_2$$

for all $v \in \mathbb{C}^N$ and all $S \subset [N]$ with $|S| \leq s$.

Hence we can apply Theorem 2.4 with S the set of largest s entries of x

$$\| z - x \|_1 \leq \frac{1+\rho}{1-\rho} \left(\|z\|_1 - \|x\|_1 + 2\, \underbrace{\sigma_s(x)_1}_{\|x_{S^c}\|_1} \right) + \frac{2\tau\sqrt{s}}{1-\rho} \| A(z-x) \|_2$$

for all vectors $x, z \in \mathbb{C}^N$.

Now set S to be the set of largest entries of $z - x$

$$\| z - x \|_2 \underset{\substack{\uparrow \\ \text{Triangle} \\ \text{Inequality}}}{\leq} \underbrace{\| (z-x)_{S^c} \|_2}_{\sigma_s(z-x)_2} + \| (z-x)_S \|_2$$

$$\underset{\substack{\uparrow \\ \text{Prop. 1.3}}}{\leq} \frac{1}{\sqrt{s}} \| z - x \|_1 + \| (z-x)_S \|_2$$

$$\underset{\substack{\uparrow \\ \text{Robust } \ell_2\text{-NSP}}}{\leq} \frac{1}{\sqrt{s}} \| z - x \|_1 + \frac{\rho}{\sqrt{s}} \| (z-x)_{S^c} \|_1 + \tau \| A(z-x) \|_2$$

$$\leq \frac{1+\rho}{\sqrt{s}} \| z - x \|_1 + \tau \| A(z-x) \|_2$$

$$\leq \frac{1+\delta}{\sqrt{s}} \cdot \frac{1+\delta}{1-\delta}\left(\|z\|_1 - \|x\|_1 + 2\sigma_s(x)_1\right) + \frac{1+\delta}{\cancel{\sqrt{s}}} \cdot \frac{2\sqrt{\delta}}{1-\delta}\|A(z-x)\|_2$$

$$+ \tau \|A(z-x)\|_2$$

$$\leq \frac{(1+\delta)^2}{1-\delta} \cdot \frac{1}{\sqrt{s}}\left(\|z\|_1 - \|x\|_1 + 2\sigma_s(x)_1\right) + \frac{(3+\delta)\tau}{1-\delta}\|Az-x\|_2$$

$\square$

~~2~~.4 Connecting RIP and NSP

__Theorem 2.10:__ Suppose that $A \in \mathbb{C}^{m\times N}$ has $(\delta, 2s)$-RIP for some $\delta < \frac{1}{3}$.

Then A has the l_2-robust NSP of order s with constants $0<\rho<1$ and $\tau > 0$ depending only on δ.

__Proof:__ Let $v \in \mathbb{C}^N$ be given. Let S_0 be the set of the s largest (absolute) entries of v and partition the complement $S_0^c = S_1 \cup S_2 \ldots$

where S_i is the set of s largest (absolute) entries of v __not__ in S_j for any $j < i$

That is, S_1 = index set of largest absolute entries in S_0^c
S_1 = index set of largest absolute entries in $(S_0 \cup S_1)^c$
etc.

* ~~Insert from next page~~

By the RIP, we write

$$\|v_{S_0}\|_2^2 \leq \frac{1}{1-\delta_{2s}}\|A(v_{S_0})\|_2^2 = \frac{1}{1-\delta_{2s}}\left\langle A(v_{S_0}), \overbrace{Av - A(v_{S_1}) - A(v_{S_2}) - \ldots}^{v = v_{S_0} + v_{S_1} + \ldots}\right\rangle$$

$$= \frac{1}{1-\delta_{2s}}\langle \underset{\uparrow}{A(v_{S_0})}, Av\rangle + \frac{1}{1-\delta_{2s}}\sum_{k\geq 1}\langle \underset{\uparrow}{A(v_{S_0})}, A(-v_{S_k})\rangle$$

Cauchy Schwartz   Proposition 2.6

$$\leq \frac{1}{1-\delta_{2s}}\|A(v_{S_0})\|_2\|Av\|_2 + \frac{1}{1-\delta_{2s}}\sum_{k\geq 1}\delta_{2s}\|v_{S_0}\|_2\|v_{S_k}\|_2$$

(*) Observe that for $j \geq 1$

$$\|v_{S_j}\|_2 = \sqrt{\sum_{\ell \in S_j} v_\ell^2} \leq \sqrt{s \max_{\ell \in S_j} v_\ell^2}$$

$$\leq \sqrt{s \min_{\ell \in S_{j-1}} v_\ell^2} = \sqrt{s} \min_{\ell \in S_{j-1}} |v_\ell|$$

minimum is less than average

$$\leq \sqrt{s} \frac{1}{s} \sum_{\ell \in S_{j-1}} |v_\ell| = \frac{1}{\sqrt{s}} \|v_{S_{j-1}}\|_1$$

$$\underset{\substack{\uparrow \\ \text{RIP and observation}}}{\leq} \frac{1+\delta_s}{1-\delta_{2s}} \|v_{S_0}\|_2 \|Av\|_2 + \frac{\delta_{2s}}{1-\delta_{2s}} \|v_{S_0}\|_2 \sum_{k \geq 1} \frac{1}{\sqrt{s}} \|v_{S_{k-1}}\|_1$$

$$\leq \|v_{S_0}\|_2 \cdot \left(\frac{1+\delta_s}{1-\delta_{2s}} \|Av\|_2 + \frac{\delta_{2s}}{1-\delta_{2s}} \cdot \frac{1}{\sqrt{s}} \|v\|_1 \right)$$

Dividing by $\|v_{S_0}\|_2$, we obtain

$$\|v_{S_0}\|_2^2 \leq \underbrace{\frac{1+\delta_s}{1-\delta_{2s}}}_{J} \|Av\|_2 + \underbrace{\frac{\delta_{2s}}{1-\delta_{2s}}}_{\frac{\rho}{2}} \frac{1}{\sqrt{s}} \|v\|_1)$$

This is the desired ~~inequality~~ result provided that

$$\rho := \frac{2\delta_{2s}}{1-\delta_{2s}} < 1 \iff \delta_{2s} < \frac{1}{3}$$ $\square$

As a consequence we obtain the main result of this section

Theorem 2.11:
Suppose that $A \in \mathbb{C}^{m \times N}$ has $(\delta, 2s)$-RIP for some $\delta < \frac{1}{3}$.
Then for any $x \in \mathbb{C}^N$ and $y \in \mathbb{C}^m$ with $\|Ax - y\|_2 \leq \eta$ a solution $x^\#$ of

$$\underset{z \in \mathbb{C}^N}{\text{minimize}} \ \|z\|_1 \quad \text{subject to} \quad \|Az - y\|_2 \leq \eta$$

approximates x with error

$$\|x - x^\#\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(x)_1 + D\eta$$

where the constants C, D depend only on δ_{2s}.

Remarks:

- ~~One also obtains [illegible] under the same assumpti~~

  $$\cancel{\|x - x^\#\|_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\,\eta}$$

- ~~By Cauchy Schwarz this implies~~

- State-of-the-art result: Same holds when $\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246$

- Sharp bounds are known for δ_s:

  $\delta_s < \frac{1}{3} \Rightarrow$ ℓ_1-minimization recovers sparse vectors

  $\delta_s > \frac{1}{3} \Rightarrow$ $\exists$ a matrix and a sparse vector such that ℓ_1-minimization does not work

- Using similar techniques, one obtains under the same assumptions

$$\|x - x^\#\|_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\,\eta.$$

3 Basic tools from probability theory and functional analysis

Motivating questions: – Find examples for RIP matrices $\boxed{1}$
– ~~First~~ show that matrices of practical relevance have RIP $\boxed{2}$

~~...~~

Easy for ~~...~~ large m $\to$ for $m \geq N$ take ~~...~~ isometry

Smallest possible m : $m \approx C s \log \frac{N}{s}$

No deterministic matrix know that achieves this or anything close to this.

Way out: ~~"Generic" matrices satisfy~~

Randomized construction.

Answer to $\boxed{1}$ uses basic tools

answer to ② is considerably more involved

This section: Ingredients to prove $\boxed{1}$

3.1 Essentials from probability

Setup: $(\Omega, \Sigma, \mathbb{P})$

Ω: ~~...~~ sample space
Σ: σ-algebra (admissible events)
$\mathbb{P}$: probability measure ~~$(\mathbb{P}(\Omega)=1)$~~

Denote, for an event $B \in \Sigma$, the probability of the event by ~~$\mathbb{P}(B)$~~.

$$\mathbb{P}(B) = \int_B d\,\mathbb{P}(\omega) = \int_\Omega I_B(\omega)\, d\,\mathbb{P}(\omega)$$

Where the characteristic function $1_B(\omega) = \begin{cases} 1 & \text{if } \omega \in B \\ 0 & \text{else} \end{cases}$

~~...~~ If $\mathbb{P}(B^c) = 0$ ~~holds~~ ~~almost surely~~ we say that B holds almost surely (a.s.)

Lemma 3.1 (Union bound)

For a collection of event $B_\ell \in \Sigma$, $\ell = 1, \dots, n$, we have

$$\mathbb{P}\left(\bigcup_{\ell=1}^{n} B_\ell\right) \leq \sum_{\ell=1}^{n} \mathbb{P}(B_\ell)$$

Proof: By induction. Equality for $n=1$.

For induction step, assume

$$\mathbb{P}\left(\bigcup_{\ell=1}^{n-1} B_\ell\right) \leq \sum_{\ell=1}^{n-1} \mathbb{P}(B_\ell)$$

Then

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\ell=1}^{n} B_\ell\right) &= \mathbb{P}\left(\bigcup_{\ell=1}^{n-1} B_\ell \cup B_n\right) \\
&= \int_\Omega \chi_{\bigcup_{\ell=1}^{n-1} B_\ell \cup B_n} \, d\mathbb{P}(\omega) \\
&\leq \int_\Omega \left[\chi_{\bigcup_{\ell=1}^{n-1} B_\ell} + \chi_{B_n}\right] d\mathbb{P}(\omega) \\
&= \mathbb{P}\left(\bigcup_{\ell=1}^{n-1} B_\ell\right) + \mathbb{P}(B_n) \\
&\leq \sum_{\ell=1}^{n} \mathbb{P}(B_\ell)
\end{aligned}$$

$\square$

A <u>random variable X</u> is a real-valued <u>measurable</u> function on (Ω, Σ).

<u>Intuition:</u> Takes its value at random according to some distribution

<u>Examples:</u> (1) <u>Rademacher random variable</u> given by $\mathbb{P}(X=1) = \mathbb{P}(X=-1) = \frac{1}{2}$

(2) <u>Standard normal RV:</u> $\mathbb{P}(X < \alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$

distribution function

A random variable has a probability density function (pdf) $\phi : \mathbb{R} \to \mathbb{R}_+$ if

$$\mathbb{P}(a < X \leq b) = \int_a^b \phi(t)\, dt \quad \text{for all } a < b \in \mathbb{R}$$

Example (2) ~~does not~~ has a pdf, (1) does not.

The <u>expectation or mean</u> of a random variable will be denoted by

$$\mathbb{E} X = \int_\Omega X(\omega)\, d\mathbb{P}(\omega)$$

If X has density ϕ $\Rightarrow$ $\mathbb{E}\, g(X) = \int_{\mathbb{R}} g(t)\, \phi(t)\, dt$.

$\mathbb{E}\, X^p \to$ p-th moment of X

$\mathbb{E}\, |X|^p \to$ p-th absolute moment of X (sometimes just "moment")

Lemma 3.2 (Important ~~(in)equalities~~ facts)

Let X, Y be random variables on a common probability space.

(1) $|\mathbb{E}\, XY| \leq (\mathbb{E}\,|X|^p)^{1/p} (\mathbb{E}\,|X|^q)^{1/q}$ for $\frac{1}{p}+\frac{1}{q}=1$

(Hölder's inequality)

(2) $|\mathbb{E}\, XY| \leq \sqrt{\mathbb{E}\,|X|^2\, \mathbb{E}\,|Y|^2}$ (Cauchy-Schwarz-ineq.)

(3) $(\mathbb{E}\,|X+Y|^p)^{\frac{1}{p}} \leq (\mathbb{E}\,|X|^p)^{\frac{1}{p}} + (\mathbb{E}\,|Y|^p)^{\frac{1}{p}}$ (Minkowski's inequality)

(4) Fubini: If $\int_{A\times B} |f(x,y)|\, d(\nu\otimes\mu)(x,y) < \infty$ ..

Proof: Real analysis class (don't use that measure is prob) $\square$

Proposition 3.3 The absolute moments of a random variable X can be expressed as

$$\mathbb{E}\,|X|^p = p\int_0^\infty \mathbb{P}(|X|\geq t)\, t^{p-1}\, dt, \quad p>0$$

Proof: $$\mathbb{E}\,|X|^p = \int_\Omega |X|^p\, d\mathbb{P} = \int_\Omega \int_0^{|X|^p} 1\, dx\, d\mathbb{P}$$

$$= \int_\Omega \int_0^\infty \chi_{\{|X|^p \geq x\}}\, dx\, d\mathbb{P}$$

$$\overset{\text{Fubini}}{=} \int_0^\infty \int_\Omega \chi_{\{|X|^p\geq x\}}\, d\mathbb{P}\, dx = \int_0^\infty \mathbb{P}(|X|^p \geq x)\, dx$$

Change of vars $x = t^p$

$$= p\int_0^\infty \mathbb{P}(|X|^p \geq t^p)\, t^{p-1}\, dt = p\int_0^\infty \mathbb{P}(|X|\geq t)\, t^{p-1}\, dt$$

$\square$

The *tail* of a probability distribution is the function $t \mapsto \mathbb{P}(|X| \geq t)$.
The tail can be estimated as follows

Theorem 3.4 (Markov's Inequality)
Let X be a random variable. Then

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \quad \text{for all } t > 0.$$

Proof: Note that $\mathbb{P}(|X| \geq t) = \mathbb{E}\chi_{\{|X| \geq t\}}$
and $t\chi_{\{|X| \geq t\}} \leq |X|$. Thus

$$t\,\mathbb{P}(|X| \geq t) = \mathbb{E}\,t\,\chi_{\{|X| \geq t\}} \leq \mathbb{E}|X| \quad \square$$

~~Corollary 3.5~~: Note that

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(|X|^p \geq t^p) \leq t^{-p}\,\mathbb{E}|X|^p.$$

3.2 Moments and tails

We use the Γ-function as defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\,dt$$

Intuition: Interpolation of factorial to non-integers.
Indeed, $\Gamma(n) = (n-1)!$ for $0 < n \in \mathbb{Z}$.

Stirling's formula:

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x} \exp\left(\frac{\theta(x)}{12x}\right) \quad \text{for } x > 0 \text{ and } 0 \leq \theta(x) \leq 1$$

functional equation:

$$\Gamma(x+1) = x\,\Gamma(x)$$ (as for the factorial)

Proposition 3.5 Suppose that a random variable Z satisfies, for some $\gamma > 0$,

$$\mathbb{P}\left(|Z| \geq e^{1/\gamma} \alpha u\right) \leq \beta e^{-u^\gamma/\gamma} \quad \text{for all } u > 0.$$

Then for $p > 0$,

$$\mathbb{E}|Z|^p \leq \beta \alpha^p (e\gamma)^{p/\gamma} \Gamma\left(\frac{p}{\gamma} + 1\right) \qquad (*)$$

As a consequence, for $p \geq 1$,

$$\left(\mathbb{E}|Z|^p\right)^{1/p} \leq C_1 \alpha \left(C_{2,\gamma} \beta\right)^{1/p} p^{1/\gamma} \quad \text{for all } p \geq 1 \qquad (**)$$

where $C_1 = e^{1/(2e)} \approx 1.2$ and $C_{2,\gamma} = \sqrt{\frac{2\pi}{\gamma}}\, e^{\gamma/12}$.

Proof: Using Proposition 3.3, we obtain

$$\mathbb{E}|Z|^p = p \int_0^\infty \mathbb{P}(|Z| > t)\, t^{p-1}\, dt \underset{\substack{\uparrow \\ \text{change of variables} \\ t = e^{1/\gamma}\alpha u}}{=} p\, \alpha^p e^{p/\gamma} \int_0^\infty \mathbb{P}\left(|Z| \geq e^{1/\gamma} \alpha u\right) u^{p-1}\, du$$

$$\overset{\substack{\text{Hypothesis} \\ \downarrow}}{\leq} p\, \alpha^p e^{p/\gamma} \int_0^\infty \beta e^{-u^\gamma/\gamma} u^{p-1}\, du$$

$$\underset{\substack{\uparrow \\ \text{change of variables} \\ v = \frac{u^\gamma}{\gamma}}}{=} p\, \beta\, \alpha^p e^{p/\gamma} \int_0^\infty e^{-v} (\gamma v)^{p/\gamma - 1}\, dv$$

$$\text{Definition of } \gamma\text{-function} \longrightarrow = \beta \alpha^p (e\gamma)^{p/\gamma} \frac{p}{\gamma} \Gamma\left(\frac{p}{\gamma}\right) \overset{\substack{\text{functional equation} \\ \downarrow}}{=} \beta \alpha^p (e\gamma)^{p/\gamma} \Gamma\left(\frac{p}{\gamma} + 1\right)$$

This proves $(*)$. For $(**)$, use Stirling's formula:

$$\mathbb{E}|Z|^p \leq \beta \alpha^p \cdot (e\gamma)^{p/\gamma} \sqrt{2\pi} \left(\frac{p}{\gamma}\right)^{p/\gamma + 1/2} e^{-p/\gamma} e^{\gamma/(12p)}$$

$$= \sqrt{2\pi}\, \beta\, \alpha^p \cdot e^{\gamma/(12p)} p^{p/\gamma + \frac{1}{2}} \gamma^{-\frac{1}{2}}$$

Using $p \geq 1$, we conclude

$$\left(\mathbb{E}|Z|^p\right)^{1/p} \leq \left(\frac{\sqrt{2\pi}\, e^{\gamma/12}}{\sqrt{\gamma}} \beta\right)^{1/p} \alpha\, p^{\frac{1}{\gamma}} p^{1/(2p)}$$

Finally $p^{\frac{1}{2p}} = e^{\frac{1}{2p}\log p}$ yields maximum at $p = e \Rightarrow p^{\frac{1}{2p}} \leq e^{\frac{1}{2e}} \Rightarrow (**)$

3.3 ~~Rand~~ Collections of random variables and ~~vec~~ random vectors

A <u>random vector</u> $X = [X_1, \dots, X_n]^T \in \mathbb{R}^n$ is a collection of n random variables on a common probability space $(\Omega, \Sigma, \mathbb{P})$. Its expectation is the vector $\mathbb{E}X = [\mathbb{E}X_1, \dots, \mathbb{E}X_n]^T \in \mathbb{R}^n$.

~~Vets~~ X has a <u>joint probability ~~den~~ density</u> if there exists a function $\phi : \mathbb{R}^n \to \mathbb{R}_+$ s.t. for any measurable domain $D \subset \mathbb{R}^n$

$$\mathbb{P}(X \in D) = \int \phi(t_1, \dots, t_n)\, dt_1 \dots dt_n .$$

A collection of random variables is (stochastically) independent if, for all $t_1, \dots, t_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \le t_1, \dots X_n \le t_n) = \prod_{e=1}^{n} \mathbb{P}(X_e \le t_e)$$

For independent random variables, we have

$$\mathbb{E}\left[\prod_{e=1}^{n} X_e\right] = \prod_{e=1}^{n} \mathbb{E}[X_e]$$

if they have a joint probability density function ϕ, then

$$\phi(t_1, \dots, t_n) = \phi_1(t_1) \cdots \phi_n(t_n)$$

where ϕ_i is the pdf of X_i

Similarly, a collection of random vectors $X_1 \in \mathbb{R}^{n_1}, \dots, X_m \in \mathbb{R}^{n_m}$ are independent if for any collection of measurable sets $A_e \subset \mathbb{R}^{n_e}, i \in [m]$

$$\mathbb{P}(X_1 \in A_1, \dots, X_m \in A_m) = \prod_{e=1}^{m} \mathbb{P}(X_e \in A_e)$$

A collection of independent random vectors $X_1, \dots, X_m$ that have the same distribution are called independent identically distributed (i.i.d)

If X' is independent from X and has the same distribution, we say X' is an independent copy of X.

3.4 Subgaussian random variables and random vectors

Definition 3.6: A random variable X is called subgaussian if there exist constants $\beta, \kappa > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t^2} \quad \text{for all } t > 0$$

It is called subexponential if

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t} \quad \text{for all } t > 0$$

Remark: last semester equivalent definition, see Proposition 3.8 below

Lemma 3.7: X is subgaussian $\Leftrightarrow$ X^2 is subexponential

Proof: "$\Rightarrow$" $\mathbb{P}(|X^2| \geq t) = \mathbb{P}(|X| \geq \sqrt{t}) \overset{X \text{ is subgaussian}}{\leq} \beta e^{-\kappa(\sqrt{t})^2} = \beta e^{-\kappa t}$.

"$\Leftarrow$" $\mathbb{P}(|X| \geq t) = \mathbb{P}(|X^2| \geq t^2) \overset{X^2 \text{ is subexponential}}{\leq} \beta e^{-\kappa t^2}$ $\square$

Examples: Standard normal, Rademacher

Proposition 3.8: Let X be a random variable

(a) if X is subgaussian, then there exist constants $c > 0$, $C > 1$ such that $\mathbb{E}[\exp(cX^2)] \leq C$

(b) If $\mathbb{E}[\exp(cX^2)] \leq C$ for some constants $c, C > 0$ then X is subgaussian. More precisely, $\mathbb{P}(|X| > t) \leq C e^{-ct^2}$

(c) if X is subgaussian, with β, κ as in the def then $(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq \tilde{C} \kappa^{-\frac{1}{2}} \beta^{\frac{1}{p}} p^{\frac{1}{2}}$ for all $p \geq 1$, where $\tilde{C}$ is an absolute constant

Proof: (c) follows directly from ~~Theorem~~ Theorem 3.5.

~~To prove~~ (a) by setting $\gamma = 2$, $\alpha = \frac{1}{\sqrt{2e\kappa}}$.

For $p = 2n$, ~~the constant $\tilde{C}$ takes~~ we obtain the specific form

$$\mathbb{E}|X|^{2n} \leq \beta \kappa^{-n} n!$$

Thus

$$\mathbb{E}[\exp(cX^2)] \overset{\text{Taylor series}}{=} 1 + \sum_{n=1}^{\infty} \frac{c^n \mathbb{E}[X^{2n}]}{n!} \leq 1 + \beta \sum_{n=1}^{\infty} \frac{c^n \kappa^{-n} n!}{n!} \overset{\text{geometric series}}{=} \underbrace{1 + \frac{\beta c \kappa^{-1}}{1 - c\kappa^{-1}}}_{=: C}$$

provided $c < \kappa$, which proves (a)

To prove (b), we estimate using Markov's inequality, Theorem 3.4

$$\mathbb{P}(|X|\ge t) = \mathbb{P}(\exp(cX^2) \ge \exp ct^2) \le \mathbb{E}[\exp(cX^2)]e^{-ct^2} \le Ce^{-ct^2} \qquad \square$$

For centered subgaussian random variables, there is an equivalent characterization using the <u>Laplace transform</u>.

<u>Proposition 3.9</u> Let X be a random variable

(a) If X is subgaussian with $\mathbb{E}X = 0$ then there exists a constant c (depending only on β and K) such that

$$\mathbb{E}[\exp(\theta X)] \le \exp(c\theta^2) \quad \text{for all } \theta \in \mathbb{R} \quad (\ddagger)$$

(b) Conversely, if $(\ddagger)$ holds, then $\mathbb{E}X = 0$ and X is sub-gaussian with parameters $\beta = 2$ and $K = \frac{1}{4c}$

<u>Remark</u>: A constant c satisfying $(\ddagger)$ is called <u>subgaussian parameter</u> of X.

<u>Proof</u>: To prove (b), assume $(\ddagger)$ and estimate for $\theta, t > 0$:

$$\mathbb{P}(X \ge t) = \mathbb{P}[\exp(\theta X) \ge \exp(\theta t)] \overset{\text{Markov}}{\le} \mathbb{E}[\exp(\theta X)]e^{-\theta t} \overset{(\ddagger)}{\le} e^{c\theta^2 - \theta t}.$$

Choosing $\theta = \frac{t}{2c}$ (this is in fact optimal) yields

$$\mathbb{P}(X \ge t) \le e^{-\frac{t^2}{4c}}$$

~~Similarly, $\mathbb{P}(-X \ge t) \le \mathbb{P}(\exp((-\theta)(-X)) \ge \exp$~~

Similarly $\mathbb{P}(-X \ge t) \le \mathbb{E}[\exp(-\theta X)]e^{-\theta t} \overset{(\ddagger)\text{ for } -\theta}{\le} e^{c\theta^2 - \theta t}$

$$\Rightarrow \mathbb{P}(-X \ge t) \le e^{-\frac{t^2}{4c}}$$

and hence $\mathbb{P}(|X| \ge t) \le 2e^{-\frac{t^2}{4c}}$ by Lemma 3.1 (Union bound)

To show that $\mathbb{E}X = 0$, note that for $|\theta| < 1$ one has by Taylor's theorem for some $|\eta| < 1$

$$e^{\theta X} = 1 + \theta X + \frac{\eta^2 X^2}{2} \ge 1 + \theta X$$

$$\left.\begin{array}{l} \Rightarrow \mathbb{E}e^{\theta X} \ge 1 + \theta \mathbb{E}X \\ \mathbb{E}e^{\theta X} \le e^{c\theta^2} \overset{\text{Taylor's theorem}}{\le} 1 + c\theta^2 + \frac{\xi^2\theta^4}{2} \text{ for some } |\xi| < |c| \end{array}\right\} \Rightarrow \mathbb{E}X \le c\theta + \frac{\xi^2\theta^3}{2} \text{ for } \theta \text{ arbitrarily small} \Rightarrow \mathbb{E}X = 0$$

To prove (a), assume w.l.o.g. $\theta \geq 0$ (otherwise consider $-X$)

Expand

$$\mathbb{E}\exp(\theta X) = 1 + \theta \underbrace{\mathbb{E}(X)}_{\substack{=0 \\ \text{by assumption}}} + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E} X^n}{n!} \leq 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}|X|^n}{n!}$$

$$\overset{\boxed{\text{Prop } 3.8(c)}}{\leq} 1 + \beta \sum_{n=2}^{\infty} \frac{\theta^n \tilde{C}^n K^{-\frac{n}{2}} n^{\frac{n}{2}}}{n!}$$

Recall that by Stirling's formula, we have $n! \geq \sqrt{2\pi}\, n^n e^{-n}$

and thus

$$\mathbb{E}\exp(\theta X) \leq 1 + \frac{\beta}{\sqrt{2\pi}} \sum_{n=2}^{\infty} \frac{\theta^n \tilde{C}^n K^{-\frac{n}{2}} n^{\frac{n}{2}}}{n^n e^{-n}}$$

$$\leq 1 + \theta^2 \frac{\beta(\tilde{C}e)^2}{\sqrt{2\pi}\, K} \sum_{n=0}^{\infty} \left(\tilde{C} e \theta_0 K^{-\frac{1}{2}}\right)^n$$

provided $K_0 := \tilde{C} e \theta K^{-\frac{1}{2}} < 1$

$$\cancel{= 1 + \theta^2 \frac{\beta(\tilde{C}e)^2}{\sqrt{2\pi}\, K} \sum_{n=0}^{\infty} \left(\tilde{C} e \theta_0 K^{-1/2}\right)^n}$$

$$= 1 + \theta^2 \frac{\beta(\tilde{C}e)^2}{\sqrt{2\pi}\, K} \frac{1}{1 - \tilde{C} e \theta K^{-1/2}}$$

$$\text{Provided } K_0 \leq \tfrac{1}{2}: \quad \leq 1 + \theta^2 \underbrace{\frac{2\beta(\tilde{C}e)^2}{\sqrt{2\pi}\, K}}_{=: C_4}$$

The condition $\frac{1}{2} > K_0 = \tilde{C} e \theta K^{-\frac{1}{2}}$ translates to $\theta \leq \frac{K^{\frac{1}{2}}}{2e\tilde{C}}$

So it remains the case $\theta > \frac{K^{\frac{1}{2}}}{2e\tilde{C}} =: \theta_0$

What we seek to prove can be rewritten as

$$\mathbb{E}\exp\left[\theta X - c_2 \theta^2\right] \leq 1.$$

where we are free to choose $c_2 > 0$

complete the square

$$\theta X - c_2 \theta^2 = -\left(\sqrt{c_2}\,\theta - \frac{X}{2\sqrt{c_2}}\right)^2 + \frac{X^2}{4c_2} \leq \frac{X^2}{4c_2}.$$

So we have $\mathbb{E}\exp\left[\theta X - c_2\theta^2\right] \leq \mathbb{E}\exp\left[\frac{X^2}{4c_2}\right]$

By Theorem 3.8 (a) we can control this for $\frac{1}{4c_2} = \tilde{C}$

Hence choose $c_2 = \frac{1}{4\tilde{C}}$ and obtain

$$\mathbb{E}\exp\left[\theta X - c_2\theta^2\right] \leq \exp(\tilde{C}X^2) \leq \tilde{C}.$$

Define $\rho = \ln(\tilde{C})\,\theta_0^{-2}$ ~~to~~ obtain

$$\mathbb{E}\left[\exp \theta X\right] \le \tilde{C} \exp(c_2 \theta^2) = \tilde{C} \exp(-\rho\,\theta^2) \exp((c_2+\rho)\,\theta^2)$$

$$\overset{\text{case } \theta > \theta_0}{\le} \underbrace{\tilde{C} \exp(-\rho\,\theta_0^2)}_{=1}\, e^{(c_2+\rho)\theta^2} \le e^{(c_2+\rho)\theta^2}.$$

Setting $c = \max(c_1, c_2+\rho)$ completes the proof $\square$

The sum of independent centered subgaussian random variables is again subgaussian:

Theorem 3.10: Let $X_1, X_2, \dots, X_M$ be a sequence of independent mean zero subgaussian random variables with subgaussian parameter c in (†). Let $a \in \mathbb{R}^M$ be some vector.
Then $Z := \sum_{\ell=1}^{M} a_\ell X_\ell$ is subgaussian with parameter $c\|a\|_2^2$, i.e.,

$$\mathbb{E} \exp(\theta Z) \le \exp(c\,\|a\|_2^2\, \theta^2)$$

and thus

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} a_\ell X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{4c\|a\|_2^2}\right) \quad \text{for all } t > 0.$$

Proof: By independence we have

$$\mathbb{E} \exp\left(\theta \sum_{\ell=1}^{M} a_\ell X_\ell\right) = \mathbb{E} \prod_{\ell=1}^{M} \exp(\theta a_\ell X_\ell) = \prod_{\ell=1}^{M} \mathbb{E} \exp(\theta a_\ell X_\ell)$$

$$\underset{X_\ell \text{ are subgaussian}}{\le} \prod_{\ell=1}^{M} \exp(c\,\theta^2 a_\ell^2) = \exp(c\,\|a\|_2^2\,\theta^2).$$

This proves that Z is subgaussian with parameter $c\|a\|_2^2$.
The tail bound follows from Proposition 3.9. $\square$

Definition 3.11 Let Y be a random vector on $\mathbb{R}^N$.

(a) if $\mathbb{E}|\langle Y, x\rangle|^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^N$, then Y is called isotropic.

(b) If, for all $x \in \mathbb{R}^N$ with $\|x\|_2 = 1$, the random variable $\langle Y, x\rangle$ is subgaussian with parameter bounded by c independent of x, that is,

$$\mathbb{E}\left[\exp(\theta\langle Y, x\rangle)\right] \le \exp(c\,\theta^2) \quad \text{for all } \theta \in \mathbb{R},\ \|x\|_2 = 1$$

then Y is called a subgaussian random vector.

Again c is called the subgaussian parameter of Y.

Remark: Isotropic subgaussian vectors does not necessarily have independent entries.

However, if the entries of a random vector X are independent subgaussian random variables of ~~second moment~~ mean zero and second moment 1 then X is isotropic and subgaussian with ~~constant~~ parameter independent of the dimension, as given by the following lemma.

Lemma 3.12: Let $Y \in \mathbb{R}^N$ be a random vector with independent, mean zero, ~~const~~ subgaussian entries Y_ℓ with $\mathbb{E}\, Y_\ell^2 = 1$, and subgaussian parameter c. Then Y is an isotropic subgaussian vector with parameter c.

Proof: Let $x \in \mathbb{R}^N$ with $\|x\|_2 = 1$. Since the Y_ℓ are independent, mean zero and $\mathbb{E}\, Y_\ell^2 = 1$ we have

$$\mathbb{E}\, |\langle Y, x\rangle|^2 = \sum_{\ell,\ell'=1}^{N} x_\ell x_{\ell'} \underbrace{\mathbb{E}\, Y_\ell Y_{\ell'}}_{0 \text{ if } \ell \neq \ell'} = \sum_{\ell=1}^{N} x_\ell^2 \underbrace{\mathbb{E}\, Y_\ell^2}_{=1} = \|x\|_2^2$$

Thus Y is isotropic.

Furthermore, according to Theorem 3.10, $Z = \langle Y, x\rangle = \sum_{\ell=1}^{N} x_\ell Y_\ell$ is subgaussian with parameter $c \Rightarrow Y$ is subg. with parameter c.

$\square$

3.5 The covering argument

<u>Motivation</u>: Restricted isometry constant is supremum over infinite set of sparse vectors. Need discretization of that set.

<u>Definition 3.13</u>: Let T be a subset of a metric space (X,d). For $t>0$, the <u>covering number</u> $\mathcal{N}(T,d,t)$ is the smallest $N\in\mathbb{N}$ such that T can be covered with balls

$$B(x_\ell,t)=\{x\in X,\ d(x,x_\ell)\le t\},\ x_\ell\in T,\ \ell=1,\dots,N,\ \text{i.e.,}$$

$$T\subseteq\bigcup_{\ell=1}^{N}B(x_\ell,t)$$

The set of points $\{x_1,\dots,x_N\}$ is called a <u>t-covering</u>.

The <u>packing number</u> $\mathcal{P}(T,d,t)$, for $t>0$, is ~~defined~~ ~~as~~ the maximal $P\in\mathbb{Z}$ s.t. there exist $x_\ell\in T$, $\ell=1,\dots,P$, which are <u>t-separated</u>, i.e., $d(x_\ell,x_k)>t$ for all $k,\ell=1,\dots,P$, $k\neq\ell$.

If $(X,\|\cdot\|)$ is a vector space we also write $\mathcal{N}(T,\|\cdot\|,t)$ and $\mathcal{P}(T,\|\cdot\|,t)$.

<u>Lemma 3.14</u>: ~~For S, T ⊂ X, ...~~ (Properties)

Let (X,d) be a metric space, $S,T\subset X$, $\alpha>0$.

Then (a) $\mathcal{N}(S\cup T,d,t)\le\mathcal{N}(S,d,t)+\mathcal{N}(T,d,t)$

(b) $\mathcal{N}(T,\alpha d,t)=\mathcal{N}(T,d,\frac{t}{\alpha})$

(c) If $X=\mathbb{R}^n$ and d is induced by a norm $\|\cdot\|$ then $\mathcal{N}(\alpha T,d,t)=\mathcal{N}(T,d,\alpha^{-1}t)$

(d) If d' is another metric that satisfies $d'(x,y)\le d(x,y)$ for all $x,y\in T$ then $\mathcal{N}(T,d',t)\le\mathcal{N}(T,d,t)$

The same relations hold for $\mathcal{P}(\cdots)$ instead of $\mathcal{N}(\cdots)$.

<u>Proof</u>: (a) The union of the optimal coverings of S and T form a covering of $S\cup T$ (not necessarily optimal, hence $\le$)

(b) $\alpha d(x,y)\le t \iff d(x,y)\le\frac{t}{\alpha} \implies B(x_\ell,t)$ w.r.t. αd is the same as $B(x_\ell,\frac{t}{\alpha})$ w.r.t d.

(c) $\|\alpha(x-y)\|\le t\iff\|x-y\|\le\frac{t}{\alpha}$, similar conclusion as in b.

(d) $B(x_\ell,t)$ w.r.t d $\subseteq$ $B(x_\ell,t)$ w.r.t d'

Same proof for packing numbers $\square$

Lemma 3.15 Let T be a subset of a metric space (X, d) and $t>0$. Then $\mathcal{P}(T, d, 2t) \leq \mathcal{N}(T, d, t) \leq \mathcal{P}(T, d, t)$

Proof: Let $\{x_1, \dots, x_p\}$ be a $2t$-separated set and $\{x'_1, \dots, x'_N\}$ be a t-covering. ~~Then~~ ~~Because~~ By definition of a t-covering, for each x_{ℓ_1}, x_{ℓ_2} there exists x'_{j_1}, x'_{j_2} such that $d(x_{\ell_i}, x'_{j_i}) \leq t$
~~On the other hand~~ Now for $\ell_1 \neq \ell_2$ one has ~~otherwise~~ $j_1 \neq j_2$ as otherwise $d(x_{\ell_1}, x_{\ell_2}) \leq d(x_{\ell_1}, x_{j_1}) + d(x_{j_2}, x_{\ell_2}) \leq 2t$, a contradiction to the separatedness assumption.
So there is ~~not~~ at least as many x'_j s as x_ℓ s, and hence $\mathcal{P}(T, d, 2t) \leq \mathcal{N}(T, d, t)$.

~~On the other hand, t~~
For the other direction, let $\{x_1, \dots, x_p\}$ be a ~~maximal~~ ~~set of~~ t-separated set of maximal size. Then adding an arbitrary point would yield a set that is not t-separated; In other words, every point in T has distance $\leq t$ to some x_ℓ and $\{x_1, \dots, x_p\}$ is also a t-covering.
This proves the second inequality. $\square$

Proposition 3.16: ~~Volume~~ Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let U be a subset of the unit ball $B = \{x \in \mathbb{R}^n, \|x\| \leq 1\}$. Then the packing and covering numbers satisfy
$$\mathcal{N}(U, \|\cdot\|, t) \leq \mathcal{P}(U, \|\cdot\|, t) \leq \left(1 + \frac{2}{t}\right)^n.$$

Proof Lemma 3.15 proves the first inequality. For the second inequality, we use a volume argument.
Let $\{x_1, \dots, x_N\}$ be a maximal ~~with~~ t-separated set
That is, the $B(x_j, \frac{t}{2})$ are disjoint. Furthermore, they are all contained in $B(0, 1+\frac{t}{2}) = (1+\frac{t}{2})B$
Thus
$$\mathrm{vol}\left(B\left(0, 1+\tfrac{t}{2}\right)\right) = \left(1+\tfrac{t}{2}\right)^n \mathrm{vol}(B) \geq \mathrm{vol}\left(\bigcup_{j=1}^N B\left(x_j, \tfrac{t}{2}\right)\right) = \sum_{j=1}^N \mathrm{vol}\left(B\left(0, \tfrac{t}{2}\right)\right)$$
$$= \underbrace{\sum_{j=1}^N \left(\tfrac{t}{2}\right)^n \mathrm{vol}(B)}_{= N \cdots} \implies N \leq \left(1 + \tfrac{2}{t}\right)^n \qquad \square$$

Theorem 3.17 ((covering argument):

$\mathcal{N} = \mathcal{N}(S^{n-1}, \|\cdot\|_2, t)$

Let B be an $N \times n$-matrix and let $\mathcal{N}$ be an t-covering of S^{n-1} for some $t \in [0,1)$.

(a) $\max_{x \in \mathcal{N}_t} \|Bx\|_2 \leq \|B\| \leq (1-t)^{-1} \max_{x \in \mathcal{N}_t} \|Bx\|_2$

(b) If furthermore $N = n$ and A is symmetric

$$\max_{x \in \mathcal{N}_t} |\langle Bx, x\rangle| \leq \|B\| \leq (1-2t)^{-1} \max_{x \in \mathcal{N}} |\langle Bx, x\rangle|$$

Proof: Recall that by definition, $\|B\| = \sup_{x \in S^{n-1}} \|Bx\|_2 \geq \max_{x \in \mathcal{N}} \|Bx\|_2$

~~Furthermore if B is symmetric, then $\|B\|$ is the larg~~

~~$\|B\| = \sup_{x \in S^{n-1}} \langle Bx, x\rangle$~~

Claim: If B is symmetric, then $\|B\|$ is the largest in absolute eigenvalue of B.

Proof of Claim: Expand any y in eigenbasis $y = \sum \alpha_i v_i$

$$\|Ay\|_2 = \|\sum \alpha_i \lambda_i v_i\| = \sqrt{\sum |\alpha_i|^2 |\lambda_i|^2} \leq |\lambda_{max}| \sum |\alpha_i|^2 = |\lambda_{max}|$$

Equality for associated eigenvector.

$$\|B\| = \|B\underbrace{x}_{\text{eigenvector}}\| = |\lambda_{max}| = |\lambda_{max}| |\langle \underbrace{x}_{\text{eigenvector}}, x\rangle| = |\langle Bx, x\rangle| = \sup_{x \in S^{n-1}} |\langle Bx, x\rangle|$$

This establishes the first inequalities in (a) and (b)

To prove the second inequalities in (a) and (b) fix $x \in S^{n-1}$ for which $\|B\| = \|Bx\|_2$ or $\|B\| = \langle Bx, x\rangle$, respectively, and choose $y \in \mathcal{N}$ s.t. $\|x - y\|_2 \leq t$.

For (a) we have by the triangle inequality

$\|Bx - By\|_2 \leq \|B\| \|x - y\|_2 \leq t\|B\|$ and thus

$$\|By\|_2 \geq \|Bx\|_2 - \|Bx - By\|_2 \geq \|B\| - t\|B\| = (1-t)\|B\|$$

For (b), again by the triangle inequality

$$|\langle Bx, x\rangle - \langle By, y\rangle| = |\langle Bx, x-y\rangle + \langle B(x-y), y\rangle|$$
$$\leq \|B\| \|x\|_2 \|x-y\|_2 + \|A\| \|x-y\|_2 \|y\|_2 \leq 2t\|A\|$$

It follows like for (i) that

$$|\langle \mathcal{B}y, y\rangle| \geq |\langle \mathcal{B}x, x\rangle| - 2\varepsilon\|\mathcal{B}\| = (1-2\varepsilon)\|\mathcal{B}\|.$$

In both cases, taking the maximum over $y \in \mathcal{N}$ proves the result. $\square$

3.6 Bernstein's inequality (following R. Vershynin "HDP")

Bernstein's inequality concerns sums of subexponential random variables

Recall: X is subexponential if $\boxed{\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t}} \quad (*)$ for all $t > 0$

Theorem 3.17 (Bernstein's inequality)

Let $X_1, \dots, X_n$ be independent centered subexponential random variables with (uniform) parameter (β, κ) as in $(*)$. Then for all $t > 0$ and an absolute constant C

$$\mathbb{P}\left(\left| \sum_{\ell=1}^{n} X_\ell \right| \geq t \right) \leq 2 \exp\left(- \cancel{C} \frac{\kappa^2 t^2 / 4}{4\beta n + \kappa t} \right)$$

~~where~~

Proof:

Step 1. Estimating the moments

$$\mathbb{E} |X_\ell|^n = n \int_0^\infty \mathbb{P}(|X_\ell| \geq t)\, t^{n-1}\, dt \leq \beta n \int_0^\infty e^{-\kappa t} t^{n-1}\, dt$$

$$= \beta n \kappa^{-n} \underbrace{\int_0^\infty e^{-u} u^{n-1}\, du}_{= \Gamma(n) = (n-1)!} = \beta n!\, \kappa^{-n}$$ ~~[illegible]~~

Step 2: Estimating the moment generating function

$$\mathbb{E} \exp(t X_\ell) = 1 + t \underbrace{\mathbb{E} X_\ell}_{=0} + \sum_{p=2}^{\infty} \frac{t^p\, \mathbb{E} X^p}{p!}$$

$$\leq 1 + \sum_{p=2}^{\infty} \frac{\beta p!\, \kappa^{-p} t^p}{p!} = 1 + \beta \sum_{p=2}^{\infty} \left(\frac{t}{\kappa}\right)^p$$

$$\underset{\substack{\text{for } t < \kappa, \\ \text{geometric} \\ \text{series}}}{=} 1 + \beta \frac{t^2}{\kappa^2} \frac{1}{1 - \frac{t}{\kappa}} \underset{\text{for } t < \frac{\kappa}{2}}{\leq} 1 + 2\beta \frac{t^2}{\kappa^2} \leq \exp 2\beta \frac{t^2}{\kappa^2}$$

Step 3 (Exponential Markov inequality)

Denote $S = \sum_{i=1}^{M} X_e$. Then for $\lambda > 0$

$$\mathbb{P}(S > t) = \mathbb{P}\left(e^{\lambda K S} \geq e^{\lambda K t}\right)$$

$$\underset{\substack{\uparrow \\ \text{Markov} \\ \text{inequality}}}{\leq} e^{-\lambda K t}\, \mathbb{E}\, e^{\lambda K S}$$

$$\underset{\substack{\uparrow \\ \text{independence}}}{=} e^{-\lambda K t} \prod_{i=1}^{M} \mathbb{E} \exp(\lambda K X_i)$$

$$\underset{\substack{\uparrow \\ \text{Step 2,} \\ \text{provided } \lambda < \frac{1}{2}}}{\leq} e^{-\lambda K t} \prod_{i=1}^{M} \exp 2\beta \lambda^2 = e^{-\lambda K t + 2M\beta \lambda^2}$$

~~Choosing~~ $\cancel{\lambda = \frac{(Kt)^2}{2} \quad \frac{1}{2} \frac{(Kt)^2}{(Kt+2M\beta)^2}}$

Choosing $\lambda = \frac{Kt}{2(Kt+2M\beta)} < \frac{1}{2}$ yields

$$\mathbb{P}(S>t) \leq \exp\left(-\frac{(Kt)^2}{2(Kt+2M\beta)} + \underbrace{\frac{2M\beta\,(Kt)^2}{4(Kt+2M\beta)^2}}_{\leq \frac{(Kt+2M\beta)(Kt)^2}{4(Kt+2M\beta)^2} = \frac{(Kt)^2}{4(Kt+2M\beta)}}\right)$$

$$\leq \exp\left(-\frac{(K^2t^2)/4}{2\beta M + Kt}\right)$$

Step 4: Absolute values

The same estimate for $\tilde{X}_e = -X_e$ yields

$$\mathbb{P}(S < -t) \leq \exp\left(-\frac{(K^2t^2)/4}{2\beta M + Kt}\right)$$

and hence

$$\mathbb{P}(|S| > t) \underset{\substack{\text{union bound} \\ \downarrow}}{\leq} \mathbb{P}(S>t) + \mathbb{P}(S<-t) \leq 2\exp\left(-\frac{(K^2t^2)/4}{2\beta M + Kt}\right)$$

□

4 Examples of RIP matrices: Subgaussian Matrices

A <u>random matrix</u> is a matrix having random variables as its entries.

<u>Definition 4.1</u>: Let A be an $m \times N$ random matrix

(a) If the entries of A are independent Rademacher variables, i.e., $\mathbb{P}(X=+1) = \mathbb{P}(X=-1) = \frac{1}{2}$, then A is called a <u>Bernoulli random matrix</u>

(b) If the entries of A are independent standard normal random variables then A is called a <u>Gaussian random matrix</u>.

(c) If all entries of A are independent mean-zero subgaussian random variables of variance 1 with the same constants β, θ, i.e.,

$$\mathbb{P}(|A_{j,k}| \geq t) \leq \beta e^{-\kappa t^2} \text{ for all } t>0,\ j \in [m], k \in [N],$$

then A is called a <u>subgaussian random matrix</u>.

<u>Note</u>: In c) the entries are required to be <u>independent</u>, but not identically distributed.

The main result of this section is the following

<u>Theorem 4.2</u> Let A be an $m \times N$ subgaussian random matrix. Then there exists a constant $C>0$ (depending only on the subgaussian parameters β, K) such that the restricted isometry constant of $\frac{1}{\sqrt{m}} A$ satisfies $\delta_s \leq \delta$ with probability at least $1-\varepsilon$ provided

$$m \geq C \delta^{-2} (s \ln(eN/s) + \ln(2\varepsilon^{-1}))$$

<u>Remarks</u>: (I) Setting $\varepsilon = \exp(-\delta^2 m/(2C))$ yields a ~~stiless probability~~ simple condition $\boxed{m \geq 2C\delta^{-2} s \ln\left(\frac{eN}{s}\right)}$ with corresponding success probability $\geq 1 - 2\exp(-\delta^2 m/(2C))$

(II) The restricted isometry property for any s entails that ~~the Euclidean~~ norms of 1-sparse vectors must be preserved. This means that each column of ~~A~~ the RIP matrix must have norm approximately 1, which is why the normalization is $\frac{1}{\sqrt{m}}$.

Proof idea: Covering argument + union bound over points in the covering.

<u>Need</u>: Probability estimate for individual matrix in the covering "concentration inequality"

<u>Lemma 4.3</u> Let $A \left(= \begin{pmatrix} Y_1 \\ \vdots \end{pmatrix}\right)$ be an $m\times N$ random matrix whose rows are independent, isotropic, ~~and~~ <u>subgaussian random</u> vectors with the same subgaussian parameter c, i.e., $\mathbb{E}[\exp(\lambda \langle \underset{i\text{th row}}{Y_i}, x\rangle)] \le \exp(c\lambda^2) \quad \forall \lambda\in\mathbb{R},\ \|x\|_2 = 1.$

Then, for all $x\in\mathbb{R}^N$ and every $t\in(0,1)$

$$\mathbb{P}\left(\left|m^{-1}\|Ax\|_2^2 - \|x\|_2^2\right| \ge t\|x\|_2^2\right) \le 2\exp(-\tilde{c}t^2 m)$$

where $\tilde{c}$ depends only on c.

<u>Proof</u>: Normalize so that $\|x\|_2 = 1$.

Consider random variables $Z_\ell = |\langle Y_\ell, x\rangle|^2 - \|x\|_2^2$, $\ell\in[m]$

Since Y_ℓ is isotropic, this entails $\mathbb{E}Z_\ell = 0$.

By Lemma 3.7, Z_ℓ is subexponential with parameters β, κ depending only on c, $\mathbb{P}(|Z_\ell| > t) \le \beta\exp(-\kappa t)$

Now

$$m^{-1}\|Ax\|_2^2 - \|x\|_2^2 = \frac{1}{m}\sum_{\ell=1}^m \left(|\langle Y_\ell, x\rangle|^2 - \|x\|_2^2\right) = \frac{1}{m}\sum_{\ell=1}^m Z_\ell.$$

~~By~~ As the Z_ℓ are independent, we have by Bernstein's inequality, Theorem 3.16,

$$\mathbb{P}\left(\left|m^{-1}\sum_{\ell=1}^m Z_\ell\right| \ge t\right) = \mathbb{P}\left(\left|\sum_{\ell=1}^m Z_\ell\right| \ge tm\right) \le 2\exp\left(-\frac{\kappa^2 m^2 t^2/4}{2\beta m + \underbrace{\kappa t m}_{\le \kappa m}}\right)$$

$$\le 2\exp\left(-\frac{\kappa^2 t^2}{8\beta + 4\kappa} m\right) =: 2\exp(-\tilde{c}t^2 m) \qquad \square$$

Theorem 4.4 Suppose that an $m \times N$ random matrix A is drawn according to a probability distribution for which the concentration inequality

$$\mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| > t\|x\|_2^2\right) \le 2\exp(-\tilde{c}t^2 m)$$

holds for all $t \in (0,1)$ and $x \in \mathbb{R}^N$. If for $\delta, \varepsilon \in (0,1)$

$$m \geq C\delta^{-2}\left[s\log\frac{N}{s} + \ln(2\varepsilon^{-1})\right]$$

where C only depends on $\tilde{c}$; then with probability at least $1-\varepsilon$ the restricted isometry constants δ_s of A satisfies $\delta_s \leq \delta$

Proof: For fixed support $S \subset [N]$, $|S| = s$, the set $\mathcal{S}_S = \{x \in \mathbb{R}^N, \operatorname{supp} x \subset S, \|x\|_2 = 1\}$ is nothing but the unit sphere in the s-dimensional space $\mathbb{R}^s$.

Hence, by Prop. 3.16,

$$\mathcal{N}(\mathcal{S}_S, \|\cdot\|_2, \rho) < \left(1 + \frac{2}{\rho}\right)^s,$$

i.e., there exists a set $\mathcal{U}$ with $|\mathcal{U}| \leq \left(1+\frac{2}{\rho}\right)^s$ such that $\min_{u \in \mathcal{U}} \|z-u\|_2 \leq \rho$ for all $z \in \mathcal{S}_S$

By the covering argument, Theorem 3.17, we have

$$\|A_S^* A_S - Id_S\| \leq \max_{x \in \mathcal{U}} (1-2\rho)^{-1} \left|\langle x, (A_S^* A_S - Id_S)x\rangle\right|$$

$$\leq \max_{x \in \mathcal{U}} (1-2\rho)^{-1} \left|\|Ax\|_2^2 - \|x\|_2^2\right|$$

So choosing $\rho = \frac{1}{4}$

$$\mathbb{P}(\|A_S^* A_S - Id\| > \delta) \leq \mathbb{P}\left(\max_{x \in \mathcal{U}} (1-2\tfrac{1}{4})^{-1}\left[\|Ax\|_2^2 - 1\right] > \delta\right)$$

$$= \mathbb{P}\left(\exists x \in \mathcal{U} : \left|\|Ax\|_2^2 - \|x\|_2^2\right| > \cancel{(1-2\rho)}\tfrac{1}{2}\delta\right)$$

$$\underset{\text{union bound}}{\leq} \left(1+\frac{2}{\frac{1}{4}}\right)^s \mathbb{P}\left(\left|\|Ax\|_2^2 - \|x\|^2\right| > \tfrac{1}{2}\delta\|x\|_2^2\right)$$

$$\leq (1+8)^s \cdot 2\exp\left(-\tilde{c}\tfrac{\delta^2}{4} m\right)$$

$$= \exp\left(s \log 9 - \frac{\tilde{c}}{4}\delta^2 m\right) \qquad (\ddagger)$$

We proceed using another union bound over the possible support sets S. As each of them corresponds to selecting s of the N components, there are $\binom{N}{s}$ such sets.

Thus

$$\mathbb{P}(\delta_s > \delta) = \mathbb{P}\left(\sup_{\substack{S \subset [N] \\ |S| = s}} \|A_S^* A_S - \mathrm{Id}\| > \delta\right)$$

$$= \mathbb{P}\left(\exists S \subset [N], |S| = s : \|A_S^* A_S - \mathrm{Id}\| > \delta\right)$$

$$\overset{\text{union bound}}{\leq} \sum_{S \subset [N]} \mathbb{P}\left(\|A_S^* A_S - \mathrm{Id}\| > \delta\right)$$

$$\overset{(\ddagger)}{\leq} \sum_{S \subset [N]} \exp\left(s \log 9 - \frac{\tilde{c}}{4}\delta^2 m\right)$$

$$= \binom{N}{s} \exp\left(s \log 9 - \frac{\tilde{c}}{4}\delta^2 m\right)$$

$\binom{N}{s} \leq \frac{N^s}{s!} \leq \frac{N^s}{\left(\frac{s}{e}\right)^s}$

$$\leq \left(\frac{eN}{s}\right)^s \exp\left(s \log 9 - \frac{\tilde{c}}{4}\delta^2 m\right)$$

$$= \exp\left(s\left[\log(9e) + \log\left(\frac{N}{s}\right)\right] - \frac{\tilde{c}}{4}\delta^2 m\right)$$

$$\leq \exp\left(s \cdot 2\log(9e) \cdot \log\left(\frac{N}{s}\right) - \frac{\tilde{c}}{4}\delta^2 m\right)$$

Choosing $m \geq \underbrace{\frac{4}{\tilde{c}} \cdot 2\log(9e)}_{=: C}\left[\delta^{-2} s \log\left(\frac{N}{s}\right) + \ln(\varepsilon^{-1})\right]$

$$\geq \frac{4}{\tilde{c}}\left[2\log(9e)\,\delta^{-2} s \log\left(\frac{N}{s}\right) + \ln(\varepsilon^{-1})\right]$$

ensures $\mathbb{P}(\delta_s > \delta) < \exp\left(s \cdot 2\log(9e) \cdot \log\left(\frac{N}{s}\right) - s\,2\log(9e)\log\left(\frac{N}{s}\right) + \log\varepsilon\right)$

$= \varepsilon$, as desired $\square$

5 Advanced probabilistic tools

5.1 Rademacher sums and Symmetrization

Recall ε is a Rademacher variable if $\mathbb{P}(\varepsilon = \pm 1) = \frac{1}{2}$.

A Rademacher sum is of the form $\sum_{\ell=1}^{M} \varepsilon_\ell x_\ell$

where ε_ℓ are independent Rademacher variables. (x_ℓ scalar, vector, matrix)

Proof strategy: (1) ~~Replace~~ Estimate sum of independent random variables via a Rademacher sum

(2) Estimate Rademacher sum using standard tools.

Both these steps will be discussed in the following

(1) Introducing Rademacher variables into the sum

Lemma 5.1 (Symmetrization)

Assume that $\xi = (\xi_\ell)_{\ell=1}^{M}$ is a sequence of independent random vectors in a finite-dimensional vector space V with norm $\|\cdot\|$. Let $F: V \to \mathbb{R}$ be a convex function. Then

$$\mathbb{E}\, F\Big(\sum_{\ell=1}^{M} (\xi_\ell - \mathbb{E}[\xi_\ell])\Big) \leq \mathbb{E}\, F\Big(2 \sum_{\ell=1}^{M} \varepsilon_\ell \xi_\ell\Big),$$

Where $\varepsilon = (\varepsilon_\ell)_{\ell=1}^{M}$ is a Rademacher sequence independent of ξ. In particular, for $1 \leq p < \infty$,

$$\Big(\mathbb{E} \Big\| \sum_{\ell=1}^{M} (\xi_\ell - \mathbb{E}[\xi_\ell])\Big\|^p\Big)^{1/p} \leq 2 \Big(\mathbb{E} \Big\| \sum_{\ell=1}^{M} \varepsilon_\ell \xi_\ell \Big\|^p\Big)^{\frac{1}{p}}$$

Proof: Let $\xi' = (\xi_1', \dots, \xi_M')$ denote an independent copy of the sequence of random vectors $(\xi_1, \dots, \xi_M)$, also independent of the Rademacher sequence ε. Then

$$E := \mathbb{E}\, F\Big(\sum_{\ell=1}^{M} (\xi_\ell - \mathbb{E}[\xi_\ell])\Big) = \mathbb{E}\, F\Big(\sum_{\ell=1}^{M} (\xi_\ell - \mathbb{E}[\xi_\ell'])\Big)$$

$$\underset{\text{Jensen}}{\leq} \mathbb{E}\, F\Big(\sum_{\ell=1}^{M} (\xi_\ell - \xi_\ell')\Big)$$

Now $(g_e - g_e')_e$ has the same distribution of $(g_e' - g_e)_e$ and hence also as $(\varepsilon_e(g_e - g_e'))_e$.

So $$E \le \mathbb{E}\, F\Big(\sum_{e=1}^{M} \varepsilon_e (g_e - g_e')\Big) \underset{\text{convexity}}{\le} \mathbb{E}\Big(\tfrac12 F\Big(2\sum_{e=1}^{M} \varepsilon_e g_e\Big) + \tfrac12 F\Big(2\sum_{e=1}^{M} (-\varepsilon_e) g_e'\Big)\Big)$$

$$\underset{\substack{g_e' \sim g_e \\ \varepsilon_e \sim -\varepsilon_e}}{=} \mathbb{E}\Big(F\Big(2\sum_{e=1}^{M} \varepsilon_e g_e\Big)\Big)$$

The inequality for the moments follows by observing that for $p \in [1,\infty)$, $F(x) = \|x\|^p$ is a convex function. $\boxed{}$

② Estimating moments and tails of Rademacher sums

The moments of Rademacher sums can be estimated using Khintchine's inequality

Theorem 8.2 Let $a \in \mathbb{C}^M$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_M)$ be a Rademacher sequence. Then, for all $n \in \mathbb{N}$,
$$\mathbb{E}\Big|\sum_{e=1}^{M} \varepsilon_e a_e\Big|^{2n} \le \frac{(2n)!}{2^n n!} \|a\|_2^{2n}$$

The proof uses the following Lemma from calculus.

Lemma 8.3 (Multinomial theorem)

For $m, n \in \mathbb{N}$ and $x_1, \dots, x_m \in \mathbb{C}$, one has
$$\Big(\sum_{e=1}^{m} x_e\Big)^n = \sum_{k_1+k_2+\dots+k_m=n} \frac{n!}{k_1!\,k_2!\cdots k_m!} \prod_{j=1}^{m} x_j^{k_j}.$$

HW

Proof: By induction. For $n=0$, one must have $k_j = 0 \ \forall j \Rightarrow$ Both sides equal 1.

For $n>0$, assume the result has been established for $n-1$.

Then $$\Big(\sum_{e=1}^{m} x_e\Big)^n = \Big(\sum_{l=1}^{m} x_e\Big)\Big[\sum_{k_1+k_2+\dots+k_m=n-1} \frac{(n-1)!}{k_1!\cdots k_n!} \prod_{j=1}^{m} x_j^{k_j}\Big]$$

$$= \sum_{e=1}^{m} \sum_{k_1+\dots+k_m=n-1} \frac{(n-1)!}{k_1!\cdots k_n!} \Big(\prod_{\substack{j=1 \\ j\ne e}}^{m} x_j^{k_j}\Big) x_e^{k_e+1}$$

$$= \sum_{\tilde k_1+\dots+\tilde k_m=n} \sum_{e:\, \tilde k_e>0} \frac{(n-1)! \cdot \tilde k_e}{\tilde k_1! \cdots \tilde k_n!} \Big(\prod_{j=1}^{m} x_j^{\tilde k_j}\Big)$$

$\sum_{e:\, \tilde k_e>0} \tilde k_e = n$

$$= \sum_{\tilde k_1+\dots+\tilde k_m=n} \Big(\prod_{j=1}^{m} x_j^{\tilde k_j}\Big)$$ $\square$

Proof of Theorem 8.2:

- For a_ℓ real valued, we obtain

$$E: \mathbb{E}\Big|\sum_{\ell=1}^{M} \varepsilon_\ell a_\ell\Big|^{2n} \underset{\substack{\text{Lemma 8.3,}\\ \text{independence}}}{=} \sum_{k_1+k_2+\dots+k_M=2n} \frac{(2n)!}{k_1!\cdots k_M!}\, \mathbb{E}\,\varepsilon_1^{k_1}\cdots \mathbb{E}\,\varepsilon_M^{k_M}$$

$$\underset{\substack{\mathbb{E}\,\varepsilon_\ell^{k}=0\\ \text{for } k \text{ odd}}}{=} \sum_{j_1+j_2+\dots+j_M=n} \frac{(2n)!}{(2j_1)!\cdots(2j_M)!}\, \underbrace{\mathbb{E}\,\varepsilon_1^{2j_1}}_{=1}\cdots \underbrace{\mathbb{E}\,\varepsilon_M^{2j_M}}_{=1}$$

$$= \sum_{j_1+j_2+\dots+j_M=n} \frac{(2n)!}{(2j_1)!\cdots(2j_M)!}$$

- Now for integers satisfying $j_1+\dots+j_M=n$, one has

$$(2j_1)!\cdots(2j_M)! \geq 2^{j_1} j_1! \cdots 2^{j_M} j_M! = 2^{j_1+\dots+j_M} j_1!\cdots j_M! = 2^n j_1!\cdots j_M!$$

and hence

$$E \leq \frac{(2n!)}{2^n\, n!} \sum_{\substack{j_1+\dots+j_M=n\\ j_i\geq 0}} \frac{n!}{j_1!\cdots j_n!} |a_1|^{2j_1}\cdots |a_M|^{2j_M}$$

$$= \frac{(2n)!}{2^n n!} \Big(\sum_{j=1}^{M} |a_j|^2\Big)^n = \frac{(2n)!}{2^n n!} \|a\|_2^{2n}.$$

- For the complex case, we write

$$\Big(\mathbb{E}\Big|\sum_{\ell=1}^{M}\varepsilon_\ell a_\ell\Big|^{2n}\Big)^{1/2n} = \Big(\mathbb{E}\Big|\sum_{\ell=1}^{M} \varepsilon_\ell\,(\mathrm{Re}(a_\ell)+i\,\mathrm{Im}(a_\ell))\Big|^{2n}\Big)^{1/2n}$$

$$= \Big(\mathbb{E}\Big[\Big|\sum_{\ell=1}^{M}\varepsilon_\ell\,\mathrm{Re}(a_\ell)\Big|^2 + \Big|\sum_{\ell=1}^{M}\varepsilon_\ell\,\mathrm{Im}(a_\ell)\Big|^2\Big]^n\Big)^{1/2n}$$

$$\underset{\substack{\text{Triangle}\\ \text{inequality}}}{\leq} \Big(\Big(\mathbb{E}\Big|\sum_{\ell=1}^{M}\varepsilon_\ell\,\mathrm{Re}(a_\ell)\Big|^{2n}\Big)^{\frac{1}{n}} + \Big(\mathbb{E}\Big|\sum_{\ell=1}^{M}\varepsilon_\ell\,\mathrm{Im}(a_\ell)\Big|^{2n}\Big)^{\frac{1}{n}}\Big)^{\frac{1}{2}}$$

$$\underset{\substack{\text{Real}\\ \text{case}}}{\leq} \Big(\Big(\frac{(2n)!}{2^n n!}\Big)^{\frac{1}{n}} \big(\|\mathrm{Re}(a)\|_2^2 + \|\mathrm{Im}(a)\|_2^2\big)\Big)^{\frac{1}{2}} = \Big(\frac{(2n)!}{2^n n!}\Big)\, \|a\|_2.$$

$\square$

Corollary 5.4 (Hoeffding's inequality)

Let $a \in \mathbb{C}^M$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_M)$ be a Rademacher sequence. Then exists a constant C such that $\forall u > 0$

$$\mathbb{P}\left(\left| \sum_{\ell=1}^{M} \varepsilon_\ell a_\ell \right| \geq \|a\|_2 u \right) \leq 2\exp(-C u^2)$$

Proof: Choosing $C \leq \frac{1}{4}\log 2$ yields the result for all $u \leq \sqrt{2}$, as the resulting upper bound is 1, hence trivially satisfied.

For $u \geq \sqrt{2}$, one can find an integer n such that

$$2n \leq u^2 \leq 4n.$$

Thus by Markov's inequality

$$\mathbb{P}\left(\left| \sum_{\ell=1}^{M} \varepsilon_\ell a_\ell \right| \geq \|a\|_2 u \right) = \mathbb{P}\left(\left| \sum_{\ell=1}^{M} \varepsilon_\ell a_\ell \right|^{2n} \geq \|a\|_2^{2n} u^{2n} \right)$$

$$\overset{\text{Markov}}{\leq} \frac{\mathbb{E}\left| \sum_{\ell=1}^{M} \varepsilon_\ell a_\ell \right|^{2n}}{\|a\|_2^{2n} u^{2n}} \underset{\text{Theorem 5.2}}{\leq} \frac{(2n)!}{2^n n!} \cdot \frac{\|a\|_2^{2n}}{\|a\|_2^{2n} u^{2n}}$$

$$\leq \frac{(2n)^{2n}}{2^n n! \, (2n)^{2n}} \leq 2^{-n} \underset{u^2 \leq 4n}{\leq} e^{-\frac{1}{4}\log 2 \cdot u^2}$$

So for $C = \frac{1}{4}\log 2$, we obtain the result for all u.

Remark: Using a more refined analysis (Stirling's formula), one obtains the optimal constant $C = \frac{1}{2}$.

5.2 Dudley's inequality

<u>Goal:</u> Estimate suprema of stochastic processes

<u>Stochastic process</u>: Collection $(X_t)_{t\in T}$ for some set T
of random variables

Intuition (for following section): T = set of sparse vectors
$X_t = |\langle t, (A^*A - I)t\rangle|$ for some random matrix
$\Rightarrow \sup X_t$ = restricted isometry constant

1. Estimate $\mathbb{E} \sup X_t$
2. Estimate deviation $\mathbb{P}(|\sup X_t - \mathbb{E}\sup X_t| > t)$

<u>Step 1:</u> ~~This class~~ Based on Dudley's inequality

<u>Step 2:</u> Requires technical tools based on concentration of measure

Hence focus on Step 1, gives general idea about behaviour of $\sup X_t$.

<u>Technical assumptions</u>.

- To avoid measurability issues, consider the <u>lattice supremum</u>

$$\mathbb{E} \sup_{t\in T} X_t := \sup \left\{ \mathbb{E} \sup_{t\in F} X_t,\ F\subset T,\ F \text{ finite} \right\}$$

This agrees with the ~~usual~~ supremum for
  - countable sets T
  - continuous dependence on the parameter (as above)
  <u>Proof:</u> consider countable dense subsets.

- Assume the process is centered, $\mathbb{E} X_t = 0$ for all $t\in T$.

- Associated to the process X_t, $t\in T$, we define the pseudo-metric

$$d(s,t) := \left(\mathbb{E}|X_s - X_t|^2\right)^{1/2} \qquad s,t\in T. \qquad (X)$$

[<u>Pseudometric:</u> (1) $d(x,y) = d(y,x)$
(2) $d(x,z) \le d(x,y) + d(y,z)$.
Third condition in the definition of a metric (*) $d(x,y) = 0 \Leftrightarrow x = y$
does not hold here as one could have $X_t = X_s$ for $s \neq t$]

Def. 5.5 A centered stochastic process X_t, $t \in T$, is called subgaussian with parameter $c > 0$ if

$$\mathbb{E} \exp(\theta (X_s - X_t)) \le \exp(\theta^2 c\, d(s,t)^2), \quad s,t \in T, \ \theta > 0$$

with d being the pseudometric defined in $(\times)$.

Examples:

1. A process X_t is called a centered Gaussian process if for every finite collection $t_1, \dots, t_n \in T$, the random vector $(X_{t_1}, \dots, X_{t_n})$ is a mean-zero Gaussian random vector.

Then
$$X_t - X_s \text{ is a univariate Gaussian}$$
with $\mathbb{E}\, X_t - X_s = 0$ and, by definition, variance $\mathbb{E}|X_s - X_t|^2 = d(s,t)^2$

Thus $\frac{X_t - X_s}{d(s,t)}$ is standard normal

and hence subgaussian, and we have by Proposition 3.9.

$$\mathbb{E}\left[\exp\left(\tilde{\theta}\, \frac{X_t - X_s}{d(s,t)}\right)\right] \le \exp(c\, \tilde{\theta}^2) \quad \text{for all } \theta \in \mathbb{R},$$

~~and~~ which, for $\tilde{\theta} = \theta\, d(s,t)$, implies that X_t is a subgaussian process.

Examples of Gaussian processes: Brownian motion, $X_t = \sum_{j=1}^{M} g_j f_j(t)$ for functions $f_j(t)$.

2. A Rademacher process has the form
$$X_t = \sum_{j=1}^{M} \varepsilon_j x_j(t),$$
where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_M)$ is a Rademacher sequence.

and $x_j : T \to \mathbb{R}$, $j \in [M]$, are arbitrary functions.

X_t is centered as $\mathbb{E}\, \varepsilon_j = 0$ for all j.

The pseudo-metric $(\times)$ for such a process becomes

$$d(s,t) = \sqrt{\mathbb{E}|X_t - X_s|^2} = \left(\mathbb{E}\left|\sum_{j=1}^{M} \varepsilon_j (x_j(t) - x_j(s))\right|^2\right)^{\frac{1}{2}} = \left(\sum_{j=1}^{M} (x_j(t) - x_j(s))^2\right)^{\frac{1}{2}} = \|x(t) - x(s)\|_2$$

On the other hand, Hoeffding's inequality, Corollary 5.4, implies that

$$\mathbb{P}\left(\left| \sum_{e=1}^{M} \varepsilon_e \left(x_e(t) - x_e(s) \right) \right| > \| x(t) - x(s) \|_2 \, u \right) \leq 2\exp(-cu^2)$$

Thus the subgaussian parameter of $X_t - X_s$ is bounded by $\tilde{c}\, \| x(t) - x(s) \|_2 = \tilde{c}\, d(s,t)$ where $\tilde{c}$ is an absolute constant, (again by Proposition 3.9), which implies that X_t is a subgaussian process.

By Proposition 3.9, an equivalent way to define a subgaussian process is that X_t satisfies the tail estimate

$$(*) \quad \mathbb{P}\left(|X_s - X_t| \geq u \, d(s,t) \right) \leq 2 \exp\left(-\frac{u^2}{c}\right).$$

~~To estimate sup X_t for a subgaussian process X_t use following intuition~~

How to estimate $\sup_{t \in T} X_t$ for subgaussian process X_t?

First approach: Covering argument

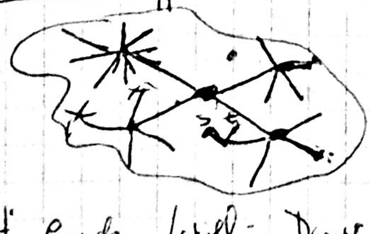
$d(s,t)$ small

Problem: X_s and X_t are estimated completely indepently.

If we have estimated X_s, it is much better to estimate $|X_s - X_t|$ using $(*)$, then X_t using the triangle inequality

Result: Suboptimal bound, as dependencies are not exploited

Second approach: Chaining



Start with one point, in the next step consider only differences to that point (using $(*)$). Result: 4 points. At third step consider the differences to the closest of these four pts etc.

At each level: Dense distribution of points → covering numbers

Theorem 5.6 (Dudley's inequality)

Let X_t, $t \in T$, be a centred real-valued subgaussian process with associated pseudo-metric d and subgaussian parameter C. Then, for any $t_0 \in T$,

$$\text{(a)} \quad \mathbb{E} \sup_{t \in T} X_t \leq 8\sqrt{C} \int_0^{\Delta(T)/2} \sqrt{\ln(\mathcal{N}(T,d,u))}\, du$$

$$\text{(b)} \quad \mathbb{E} \sup_{t \in T} |X_t| \leq 8\sqrt{C} \int_0^{\Delta(T)/2} \sqrt{\ln(2\mathcal{N}(T,d,u))}\, du$$

where $\Delta(T) = \sup_{t \in T} \sqrt{\mathbb{E}|X_t|^2}$ is the radius of T w.r.t. d.

Remark: If $(X_t)_{t \in T}$ is a Gaussian process, T is a subset of a finite dimensional space and d is induced by a norm, these bounds are known to be sharp up to logarithmic factor in the dimension.

Proof of Theorem 5.6

We write Δ shorthand for $\Delta(T)$. If $\Delta = \infty$, both sides of (a) and (b) are infinite, so the result holds. Hence from now on, assume $\Delta < \infty$.

Fix $F \subset T$ finite. At the very end we will take the supremum over all F to obtain the lattice supremum.

For each $n \in \mathbb{N}$, set $\varepsilon_n := 2^{-n}\Delta$ and $N_n = \mathcal{N}(T, d, \varepsilon_n)$.

By the definition of the covering number, we can find subsets $T_n \subset T$ of cardinality at most N_n such that for all $t \in T$ (and hence, in particular for all $t \in F$), there exists $s \in T_n$ such that $d(t,s) \leq \varepsilon_n$.

Write ~~s/z~~ $\phi_n(t)$ for this particular s.

To bound $\mathbb{E} \max_{t \in T_n} X_t$, we estimate

$$(0) \quad \max_{t \in T_n} X_t = \max_{t \in T_n} \left(X_t - X_{\phi_{n-1}(t)} + X_{\phi_{n-1}(t)} \right) \leq \underbrace{\max_{t \in T_n} \left(X_t - X_{\phi_{n-1}(t)} \right)}_{(8)} + \max_{t \in T_{n-1}} X_t$$

To estimate $\otimes$, ~~...~~ let $\beta > 0$ (to be fixed later) and bound

$$\beta\, \mathbb{E} \max_{t \in T_n} \left(X_t - X_{\phi_{n-1}(t)}\right) \overset{\text{log is monotone}}{=} \mathbb{E} \log \max_{t \in T_n} \exp\left(\beta\left(X_t - X_{\phi_{n-1}(t)}\right)\right)$$

$$\underset{\max \leq \Sigma}{\leq} \mathbb{E} \log\left(\sum_{t \in T_n} \exp\left(\beta\left(X_t - X_{\phi_{n-1}(t)}\right)\right)\right)$$

$$\overset{\text{Jensen}}{\leq} \log\left(\sum_{t \in T_n} \mathbb{E} \exp\left(\beta\left(X_t - X_{\phi_{n-1}(t)}\right)\right)\right)$$

$$\overset{X_t \text{ is subgaussian}}{\leq} \log\left(\sum_{t \in T_n} \exp \beta^2 c\, d(t, \phi_{n-1}(t))^2\right)$$

$$\leq \log\left(N_n \exp \beta^2 c\, \varepsilon_{n-1}^2\right)$$

$$= \log(N_n) + \beta^2 c\, \varepsilon_{n-1}^2$$

Choosing $\beta = \varepsilon_{n-1}^{-1} c^{-\frac{1}{2}} \cdot \sqrt{\log(N_n)}$ yields

$$\varepsilon_{n-1}^{-1} c^{-\frac{1}{2}} \sqrt{\log N_n} \cdot \mathbb{E} \max_{t \in T_n} \left(X_t - X_{\phi_{n-1}(t)}\right) \leq 2 \log N_n$$

$$\Rightarrow \mathbb{E} \max_{t \in T_n} \left(X_t - X_{\phi_{n-1}(t)}\right) \leq 2 \varepsilon_{n-1} \sqrt{c \cdot \log N_n} \qquad (i)$$

Furthermore note that for $n = 1$, repeating the exact same argument with 0 instead of $X_{\phi_{n-1}}$ yields

$$\mathbb{E} \max_{t \in T_0} X_t \leq 2 \Delta \sqrt{c \log N_n} \qquad (ii)$$

(as $d(t, 0) \leq \Delta$, Δ takes the role of ε_{n-1})

Thirdly, the same argument also yields

$$\mathbb{E} \sup_{t \in F} \left(X_t - X_{\phi_n(t)}\right) \leq 2 \varepsilon_n \sqrt{c \log |F|} \qquad (iii)$$

Now note that ~~...~~ by the monotonicity of the covering number

$$N_n = \mathcal{N}(T, d, \varepsilon_n) < \mathcal{N}(T, d, u) \;\; \forall u < \varepsilon_n \overset{\text{integrate over } u}{\Rightarrow} \underbrace{\sqrt{\log N_n}}_{(\varepsilon_n - \varepsilon_{n+1})} \leq \int_{\varepsilon_{n+1}}^{\varepsilon_n} \sqrt{\log \mathcal{N}(T, d, u)}\, du \qquad (iv)$$

Hence for $n \in \mathbb{N}$ $\quad \mathbb{E} \max_{t\in F} X_t \le \mathbb{E} \max_{t\in T_n} X_t + \mathbb{E} \max_{t\in F} (X_t - X_{\phi_n(t)})$

$$\overset{(iii)}{\le} \mathbb{E} \max_{t\in T_n} X_t + 2\varepsilon_n \sqrt{c \log |F|}$$

$$\overset{(0)}{\le} \mathbb{E} \max_{t\in T_{n-1}} X_t + \mathbb{E} \max_{t\in T_n} (X_t - X_{\phi_{n-1}(t)}) + 2\varepsilon_n \sqrt{c\log|F|}$$

Repeated application of (0)

$$\le \mathbb{E} \max_{t\in T_1} X_t + \sum_{j=2}^{n} \mathbb{E} \max_{t\in T_j} (X_t - X_{\phi_{j-1}(t)}) + 2\varepsilon_n \sqrt{c\log|F|}$$

$$\overset{(i),(ii)}{\le} 2\underbrace{\Delta}_{=4(\varepsilon_1-\varepsilon_2)} \sqrt{c\log N_1} + \sum_{j=2}^{n} 2\underbrace{\varepsilon_{j-1}}_{=4(\varepsilon_{j}-\varepsilon_{j+1})}\sqrt{c\log N_j} + 2\varepsilon_n\sqrt{c\log|F|}$$

$$= \sum_{j=1}^{n} 8(\varepsilon_j - \varepsilon_{j+1})\sqrt{c\log N_j} + 2\varepsilon_n \sqrt{c\log|F|}$$

$$\overset{(iv)}{\le} \sum_{j=1}^{n} \int_{\varepsilon_{j+1}}^{\varepsilon_j} 8\sqrt{c\log N(T,d,u)}\,du + 2\varepsilon_n\sqrt{c\log|F|}$$

$$= \int_{\varepsilon_n}^{\varepsilon_1} 8\sqrt{c\log N(T,d,u)}\,du + 2\varepsilon_n \sqrt{c\log|F|}$$

$$\le \int_0^{\Delta/2} 8\sqrt{c\log N(T,d,u)}\,du + 2\varepsilon_n\sqrt{c\log|F|}$$

This bound is valid for all $n\in\mathbb{N}$. Letting $n\to\infty$, we obtain

$$\mathbb{E}\max_{t\in F} X_t \le \int_0^{\Delta/2} 8\sqrt{c\log N(T,d,u)}\,du$$

and hence for the lattice supremum

$$\mathbb{E}\sup_{t\in T} X_t := \sup_{\substack{F\subset T\\ F \text{ finite}}} \mathbb{E}\max_{t\in F} X_t \le \int_0^{\Delta/2} 8\sqrt{c\log N(T,d,u)}\,du$$

For the bound regarding $|X_t|$, define $\tilde{T} = T\times\{0,1\}$ and define the stochastic process Y_t, $t\in\tilde{T}$ via $Y_{(t,0)} = X_t$, $Y_{(t,1)} = -X_t$. Then $\sup_{t\in T}|X_t| = \sup_{t\in\tilde{T}} Y_t$. Furthermore as $\tilde{T} = T\times\{1\} \cup T\times\{0\}$, $\mathcal{N}(\tilde{T},d,u) \le 2\mathcal{N}(T,d,u)$, which gives rise to the additional factor 2 in the log!

6 Random Sampling in Bounded Orthonormal Systems

Recall main motivation:

Recover images from undersampled linear measurements

Prime example: Magnetic Resonance Imaging (MRI)

Model: $f \in L^2([0,1]^2)$ image

$f(s,t) = \sum_{k,\ell} \underbrace{\hat{f}(k,\ell)}_{} \cdot e^{2\pi i\, ks+\ell t}$ Fourier series

MRI measurements correspond to Fourier coefficients

$$\boxed{\hat{f}(k,\ell) = \int f(s,t)\, e^{-2\pi i (ks+\ell t)}\, ds\, dt}$$

Discretization: Discrete ~~...~~ Image: $f \in \mathbb{C}^{N \times N}$

Discrete Fourier transform (DFT) ~~...~~

$$\mathcal{F} = \frac{1}{N} \left(e^{2\pi i\, k\omega + \ell\nu} \right)_{(\omega,\nu),(k,\ell) \in [0,N]^2}$$

DFT measurement: $$(\mathcal{F} f)_{(\omega,\nu)} = \frac{1}{N} \sum_{(k,\ell)\in[0,N]^2} e^{2\pi i\, k\omega+\ell\nu} f(k,\ell)$$

provides ~~...~~

Discrete approximation of MRI measurement

→ in some sense suboptimal, but commonly used in practice.

Allows for recovery: $\mathcal{F}^* \mathcal{F} = I \Rightarrow$ Recovery by applying adjoint.

Question: Can we recover approximately sparse signals from a small subset of the Fourier DFT measurements?

Resulting matrix: Partial Fourier → Select random rows of the DFT matrix

Accomplished so far

(1) $2s$ measurements allow for reconstruction (uniqueness), no efficient algorithms

(2) If ~~a~~ a matrix A has the Restricted Isometry Property of ~~the~~ order s and level ~~δ~~ $\delta < \frac{1}{3}$, then s-sparse vectors can be reconstructed via l_1-minimization

(3) Subgaussian matrices have the RIP with high probability for embedding dimensions $m = O(\delta^{-2} s \log(\frac{N}{s}))$

Need: (3') Partial Fourier matrices have the RIP ~~for embedding dim~~ for embedding dimensions $\geq$?

Again, use randomness (here random selection of frequencies).
Drawing ~~a~~ Partial random Fourier matrix is a special case of sampling a Bounded Orthonormal System

6.1 Bounded Orthonormal Systems

Definition 6.1 Let $D \subset \mathbb{R}^d$ be endowed with a probability measure ν. Further, let $\Phi = \{\phi_1, \dots, \phi_N\}$ be an orthonormal system of complex-valued functions on D, that is, for $j,k \in [N]$

$$\int_D \phi_j(t)\overline{\phi_k(t)}\, d\nu(t) = \delta_{jk} = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

Then Φ is called a bounded orthonormal system (BOS) with constant K if

$$\|\phi_j\|_\infty := \sup_{t \in D} |\phi_j(t)| \leq K \quad \text{for all } j \in N \qquad (\times)$$

~~A random ... for a finite BOS,~~

The random matrix of embedding dimension m associated to a BOS consists of m rows ~~drawn~~ ($l = 1 \dots m$) ~~independently~~ $(\phi_1(t_l), \dots, \phi_N(t_l))$ where the t_l are independently drawn according to ν

Remark: The smallest value that K can take is 1.

Indeed, $$1 \underset{\substack{\uparrow \\ \text{orthogonality}}}{=} \int_{\mathcal{D}} |\phi_1(t)|^2 \, d\nu(t) \underset{(*)}{\leq} \sup_{t \in \mathcal{D}} |\phi_1(t)|^2 \int_{\mathcal{D}} d\nu(t) \leq K$$

The inequality $(*)$ can only be an equality if $|\phi_1(t)| = \sup_{t \in \mathcal{D}} |\phi_1(t)|$ ν-a.s.

Hence $K = 1$ is only possible if $|\phi_1(t)| = 1$.

Thus $(\times)$ is a strong condition if $K = O(1)$

Interpretation: The ϕ_i distribute evenly over $\mathcal{D}$.

(~~Large~~ ~~constants~~ Large constants, say $K = O(N)$ do not suffice (not very easy to satisfy).

Examples:

1. Trigonometric Polynomials: Let $\mathcal{D} = [0,1]$ and set, for $k \in \mathbb{Z}$,
$$\phi_k(t) = e^{2\pi i k t}.$$
The probability measure ν is the Lebesgue measure on $[0,1]$.
Then, for all $j, k \in \mathbb{Z}$,
$$\int_0^1 \phi_k(t) \overline{\phi_j(t)} \, dt = \delta_{j,k}$$
The constant is $K = 1$.

Note: For any $\Gamma \subset \mathbb{Z}$, $\{\phi_k, k \in \Gamma\}$ is a BOS.
Common choice: $\Gamma = \{-q, -q+1, \dots, q-1, q\}$

Interpretation: Signal $f(t) = \sum_{\ell \in \Gamma} x_\ell \phi_\ell(t)$ & x_ℓ Fourier coefficients

~~Applying random matrix~~

Sampling f at random points $t_1, \dots, t_m \sim \nu$	$\Longleftrightarrow$	Applying random matrix of embedding dimension m associated to BOS to the vector $(x_1, \dots, x_N)$ of Fourier coefficients

Same idea works for Real Trigonometric Polynomials

2. Discrete Orthonormal Systems

Let $U \in \mathbb{C}^{N \times N}$ be a unitary matrix.

The renormalized columns $\sqrt{N}\, u_k \in \mathbb{C}^N$, $k \in [N]$ form an orthonormal system with respect to the discrete uniform measure on $[N]$ given by $\nu(B) = \frac{|B|}{N}$ for $B \subset [N]$.

Indeed,
$$\sum_{t=1}^{N} \sqrt{N}\, u_k(t)\, \overline{\sqrt{N}\, u_\ell(t)}\, \underbrace{\nu(t)}_{=\frac{1}{N}} = \langle u_k, u_\ell \rangle = \delta_{k\ell}$$

Boundedness translates to
$$\sqrt{N} \max_{k,t \in [N]} |U_{t,k}| = \max_{k,t \in [N]} |\sqrt{N}\, u_k(t)| \leq K$$

Revisiting our prime example

- The DFT matrix is unitary
- Dimension is N^2 (image)
- $|U_{t,k}| = \frac{1}{N} \Rightarrow \sqrt{N^2} \max_{k,t \in [N]} |U_{t,k}| = 1 \Rightarrow K = 1$

~~More examples~~

Random matrix associated to BOS is Partial random Fourier matrix.

Remainder of this class

Prove that Random matrix associated to BOS with embedding dimension $m \geq C\, \delta^{-2}\, s \log^4 N$ has (s,δ)-RIP ~~with high probability~~ in expectation.

6.2 Restricted Isometry Property for Bounded Orthonormal Systems

~~The~~ We prove the following slightly stronger bound

Theorem 6.2 Let $\mathcal{D} \subset \mathbb{R}^d$ and let $\{\phi_j\}_{j=1}^N$ be a BOS on $\mathcal{D}$ with associated probability measure ν. and constant $K \geq 1$. Let $\delta \in (0,1)$. Then exists $C > 0$ ~~the m ≥ N~~ such that if m satisfies

$$\frac{m}{\log(9m)} \geq C \delta^{-2} K^2 s \log^2(4s) \log(8N)$$

then the normalized random matrix associated to the BOS

$$\tilde{A} = \frac{1}{\sqrt{m}} \left(\phi_k(t_\ell) \right)_{k=1, \ell=1}^{N, m}, \quad t_\ell \sim \nu \text{ i.i.d.}$$

satisfies $\mathbb{E}\, \delta_s(\tilde{A}) \leq \delta$.

Proof: ~~Recall~~

$$\cancel{\delta_s = \max_{S \subseteq N, |S| \leq s} \|\tilde{A}_S^* \tilde{A}_S - Id\|}$$

Recall

$$\delta_s = \sup_{z \in D_{s,N}} |\langle z, (\tilde{A}_s^* \tilde{A}_s - Id) z \rangle|$$

where $D_{s,N} := \{ z \in \mathbb{C}^N : \|z\|_2 \leq 1, \|z\|_0 \leq s \} = \bigcup_{S \subset [N]} B_S$

where B_S denotes the unit sphere in $\mathbb{C}^S$ with respect to the ℓ_2-norm. ~~Then~~

Defining $|||B|||_s = \sup_{z \in D_{s,N}} |\langle z, Bz \rangle|$

(a norm for self-adjoint matrices, seminorm on all of $\mathbb{C}^{N \times N}$)

this fact reads

$$\delta_s = |||\tilde{A}^* \tilde{A} - Id|||_s .$$

We will now analyse $\tilde{A}^*\tilde{A}$.

Denote the (random) ℓ-th column of A^* by X_ℓ, that is, $X_\ell = \left(\overline{\phi_j(t_\ell)}\right)_{j=1}^N$

That is $\tilde{A}^* = \frac{1}{\sqrt{m}} \sum_{\ell=1}^m X_\ell e_\ell^*$, $\tilde{A} = \frac{1}{\sqrt{m}} \sum_{k=1}^m e_k X_k^*$

and $$\tilde{A}^*\tilde{A} = \frac{1}{m} \sum_{k,\ell=1}^m X_\ell \underbrace{e_\ell^* e_k}_{\delta_{k\ell}} X_k^* = \frac{1}{m} \sum_{\ell=1}^m X_\ell X_\ell^*$$

Now by the orthogonality $$\mathbb{E}\, X_\ell X_\ell^* = \left(\mathbb{E}\, \overline{\phi_j(t_\ell)}\, \phi_k(t_\ell)\right)_{j,k=1}^N = \left(\delta_{jk}\right)_{j,k=1}^N = Id$$

So $$\delta_s = \left|\left|\left| \frac{1}{m} \sum_{\ell=1}^m X_\ell X_\ell^* - Id \right|\right|\right|_s = \frac{1}{m} \left|\left|\left| \sum_{\ell=1}^m (X_\ell X_\ell^* - \mathbb{E}\, X_\ell X_\ell^*) \right|\right|\right|_s$$

~~Sum of mean zero~~ ~~We seek~~ We apply symmetrization, Lemma 5.1

$$\mathbb{E}\, \delta_s = \mathbb{E} \left|\left|\left| \sum_{\ell=1}^m (X_\ell X_\ell^* - \mathbb{E}\, X_\ell X_\ell^*) \right|\right|\right|_s \leq \mathbb{E} \left|\left|\left| \sum_{\ell=1}^m \varepsilon_\ell X_\ell X_\ell^* \right|\right|\right|_s$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)$ is a Rademacher sequence Independent of the random sampling points t_ℓ.

<u>Strategy:</u> Condition on t_ℓ, only consider randomness in ε_ℓ. Use the following lemma.

<u>Lemma 6.3</u> (Noncommutative Rudelson's lemma)

Let $x_1, \dots, x_m \in \mathbb{C}^N$ with $\|x_\ell\|_\infty \leq K$ for all $\ell \in [m]$

Then, for $s \leq m$

$$\mathbb{E} \left|\left|\left| \sum_{\ell=1}^m \varepsilon_\ell x_\ell x_\ell^* \right|\right|\right|_s \leq C_1 K \sqrt{s}\, \ln(4s) \sqrt{\ln(8N)\ln(9m)} \sqrt{\left|\left|\left| \sum_{\ell=1}^m x_\ell x_\ell^* \right|\right|\right|_s}$$

where C_1 is an absolute constant.

<u>Proof:</u> $$E := \mathbb{E} \left|\left|\left| \sum_{\ell=1}^m \varepsilon_\ell x_\ell x_\ell^* \right|\right|\right|_s = \mathbb{E} \sup_{u \in D_{s,N}} \left| \sum_{\ell=1}^m \varepsilon_\ell |\langle x_\ell, u\rangle|^2 \right|.$$

This is the supremum of a Rademacher process $X_u = \sum_{\ell=1}^m \varepsilon_\ell |\langle x_\ell, u\rangle|^2$, which has associated pseudo-metric

$$d(u,v) = \left(\mathbb{E}\, |X_u - X_v|^2\right)^{1/2} = \sqrt{\sum_{\ell=1}^m \left(|\langle x_\ell, u\rangle|^2 - |\langle x_\ell, v\rangle|^2\right)^2}$$

Seek to apply Dudley's inequality.

$\to$ metric is too complicated to allow for covering number estimates.

Hence we bound

$$d(u,v) = \left(\sum_{\ell=1}^{m} \left(|\langle x_\ell, u\rangle| - |\langle x_\ell, v\rangle| \right)^2 \left(|\langle x_\ell, u\rangle| + |\langle x_\ell, v\rangle| \right)^2 \right)^{\frac{1}{2}}$$

$$\le \max_{\ell\in[m]} \Big| |\langle x_\ell, u\rangle| - |\langle x_\ell, v\rangle| \Big| \sup_{u,v\in D_{s,N}} \sqrt{\sum_{\ell=1}^{m} \left(|\langle x_\ell, u\rangle| + |\langle x_\ell, v\rangle| \right)^2}$$

$$\le 2R \max_{\ell\in[m]} |\langle x_\ell, u-v\rangle| \qquad (\ddagger)$$

where $R = \sup_{u\in D_{s,N}} \sqrt{\sum_{\ell=1}^{m} |\langle x_\ell, u\rangle|^2} = \sqrt{\left\| \sum_{\ell=1}^{m} x_\ell x_\ell^* \right\|_s}$

So we reduced the problem to considering the seminorm

$$\|u\|_X := \max_{\ell\in[m]} |\langle x_\ell, u\rangle| \quad , \quad u \in \mathbb{C}^N$$

More precisely $(\ddagger)$ implies that

$$\left(\mathbb{E} \left| \frac{X_u}{2R} - \frac{X_v}{2R} \right|^2 \right)^{1/2} \le \|u-v\|_X$$

By Dudley's inequality, Theorem 5.6, there is an absolute constant C_2 such that

$$\boxed{\begin{aligned} E &= \mathbb{E} \sup_{u\in D_{s,N}} |X_u| = 2R\, \mathbb{E} \sup_{u\in D_{s,N}} \left| \frac{X_u}{2R} \right| \\ &\overset{\text{Thm 5.6}}{\le} 2C_2 R \int_0^{\Delta(D_{s,N}, \|\cdot\|_X)/2} \sqrt{\log 2\mathcal{N}(D_{s,N}, \|\cdot\|_X, t)}\, dt \end{aligned}}$$

(b) Estimate radius

For $u \in D_{s,N}$: $\|u\|_X = \max_{\ell\in[m]} |\langle x_\ell, u\rangle| \overset{\text{Hölder}}{\le} \|u\|_1 \max_{\ell\in[m]} \|x_\ell\|_\infty \overset{\text{Cauchy-Schwarz, } u \text{ s-sparse}}{\le} K\sqrt{s}\, \|u\|_2 \le K\sqrt{s}$

$$\Rightarrow \Delta(D_{s,N}, \|\cdot\|_X) \le \sup_{u\in D_{s,N}} \|u\|_X \le K\sqrt{s}$$

(•) Covering number estimate for large u

• ~~[illegible]~~ Define the norm

$$\|z\|_1^* = \sum_{j=1}^{N} \left(|\mathrm{Re}(z_j)| + |\mathrm{Im}(z_j)| \right), \quad z \in \mathbb{C}^N$$

For $z \in D_{s,N}$ we have

$$\|z\|_1^* = \|\mathrm{Re}\, z\|_1 + \|\mathrm{Im}\, z\|_1 \overset{\text{Cauchy-Schwarz}}{\leq} \sqrt{s}\,\|\mathrm{Re}\, z\|_2 + \sqrt{s}\,\|\mathrm{Im}\, z\|_2$$

$$\leq \sqrt{2s}\,\|z\|_2$$

$$\Rightarrow D_{s,N} \subset \sqrt{2s}\, B^N_{\|\cdot\|_1^*} = \{ x \in \mathbb{C}^N ; \|x\|_1^* \leq \sqrt{2s} \}$$

Idea: We will cover this superset of $D_{s,N}$, which induces a covering for $D_{s,N}$.

• <u>Lemma 6.4:</u> Let $U \in B^N_{\|\cdot\|_1^*}$ and $0 < t < \sqrt{2}k$. Then

$$\sqrt{\ln(2\mathcal{N}(U, \|\cdot\|_X, t))} \leq \tilde{C} k \sqrt{\ln(9m)\ln(8N)}\; t^{-1}$$ for an absolute constant $\tilde{C}$

~~Proof~~: Fix $x \in U$. The idea is to approximate x by very sparse vectors. This technique is called <u>Maurey's empirical method</u>

<u>Use:</u> $B^*_{\|\cdot\|_1}$ is the convex hull of $V := \{\pm e_j, \pm i e_j : j \in [N]\}$

$$\Rightarrow x = \sum_{v \in V} \lambda_v v \quad \text{with } \lambda_v \geq 0 \text{ and } \sum_{v \in V} \lambda_v = 1. \quad (*)$$

• Define the random vector Z by $\mathbb{P}(Z = v) = \lambda_v$ for $v \in V$

$(*)$ ensures that this defines a probability distribution

Note that $\mathbb{E}\, Z = \sum_{v \in V} \lambda_v v = x$

<u>Goal:</u> Approximate x with $z = \frac{1}{M} \sum_{k=1}^{M} Z_k$

where $Z_1, \dots, Z_M$ are independent copies of Z.

<u>Expected distance of z to x in $\|\cdot\|_X$:</u>

$$\mathbb{E}\|z - x\|_X = \mathbb{E}\left\|\frac{1}{M}\sum_{k=1}^{M} (Z_k - \mathbb{E} Z_k)\right\|_X \underset{\nearrow}{\leq} \frac{2}{M}\mathbb{E}\left\|\sum_{k=1}^{M} \varepsilon_k Z_k\right\|_X$$

•
$$= \frac{2}{M}\mathbb{E} \max_{\ell \in [m]} \left|\sum_{k=1}^{M} \varepsilon_k \langle x_\ell, Z_k\rangle\right| \overset{\text{Lemma 5.1}}{=} \frac{2}{M} \mathbb{E}_{(Z_1,\dots,Z_M)} \mathbb{E}_\varepsilon |\cdots|$$

where ε_ℓ is a Rademacher sequence independent of $(Z_1, \dots, Z_M)$.

~~Hence~~ We will ~~condition~~ consider the <u>inner expectation</u>, that is we will keep $(Z_1, \dots, Z_M)$ fixed for the moment.
Each Z_k has exactly ~~one~~ non-zero component, hence $\|Z_k\|_1 = 1$
So
$$|\langle x_\ell, Z_k\rangle| \leq \|x_\ell\|_\infty \, \|Z_k\|_1 \leq K .$$
$$\Rightarrow \|(\langle x_\ell, Z_k\rangle)_{k=1}^M\|_2 \leq \sqrt{M}\, K \quad , \; \ell \in [m]$$
Thus by Hoeffding's inequality, ~~Theorem~~ Corollary 5.4,
$$\mathbb{P}_\varepsilon\left(\left|\sum_{k=1}^M \varepsilon_k \langle x_\ell, Z_k\rangle\right| \geq \sqrt{M}\, K\, t\right) \leq 2e^{-C_3 t^2}$$
~~and~~ by ~~a~~ union bound
$$\mathbb{P}_\varepsilon\left(\max_{\ell \in [m]} \left|\sum_{k=1}^M \varepsilon_k \langle x_\ell, Z_k\rangle\right| \geq \sqrt{M}\, K\, t\right) \leq 2m\, e^{-C_3 t^2},$$
and by Proposition 3.8
$$\mathbb{E}_\varepsilon \max_{\ell \in [m]} \left|\sum_{k=1}^M \varepsilon_k \langle x_\ell, Z_k\rangle\right| \leq C_4 \sqrt{M}\, K \sqrt{\ln(8m)}$$
where C_3 and C_4 are absolute constants.
Thus
$$\mathbb{E}\, \|z - x\|_X \leq \frac{2}{M} \mathbb{E}_Z \mathbb{E}_\varepsilon \max_{\ell \in [m]} \left|\sum_{k=1}^M \varepsilon_k \langle x_\ell, Z_k\rangle\right| \leq \frac{C_4 K}{\sqrt{M}} \sqrt{\ln(8m)}$$

If this bound holds in expectation, it must be true at least for some realization $z = \frac{1}{M}\sum_{k=1}^M z_k$ where $z_k \in V$, i.e.,
$$\|z - x\|_X \leq \frac{C_4 K}{\sqrt{M}} \sqrt{\ln(8m)} .$$
~~Then~~ Choosing $M = \left\lfloor \frac{4 C_4^2 K^2}{t^2} \ln(9m) \right\rfloor$
we have
$$M \geq \frac{4C_4^2 K^2}{t^2} \log(9m) - 1 \geq \frac{4C_4^2 K^2}{t^2} \log(9/8) + \frac{4C_4^2 K^2}{t^2} \log(8m) - 1$$
$$\overset{t \leq 2K}{\geq} \frac{4C_4^2 K^2}{t^2} \log(8m) + \underbrace{\frac{4C_4^2 \ln(9/8)}{2} - 1}_{\substack{>0 \\ (\text{for } C_4 \text{ large})}} \geq \frac{4C_4^2 K^2}{t^2} \log(8m)$$
$$\Rightarrow \|z - x\|_X \leq \frac{t}{2}$$

So we found an M- sparse z such that $\|x-z\|_X \le \frac{t}{2}$

Do they form a covering? No, they are not in U necessarily.

To get a t-covering of U:

For each z selected in this way, select $z' \in U$ s.t $\|z-z'\|_X \le \frac{t}{2}$

There must be such z', otherwise z would not have been Selected. Now

$$\forall x \in U: \exists z: \|z-x\|_X \le \tfrac{t}{2}, \exists z': \|z-z'\|_X \le \tfrac{t}{2} \Rightarrow \|x-z'\| \le t.$$

So the z' form a t-covering of U.

Furthermore, there are at most as many z' as there are z.

The number of z can be ~~est~~ bounded by

$$(4N)^{M}$$ as there are $4N$ choices for each z_ℓ.

So we found a t-covering with $(4N)^M$ elements

$$\Rightarrow \sqrt{\ln(2\mathcal{N}(U, \|\cdot\|_X, t))} \le \sqrt{\ln(2(4N)^M)} \le \sqrt{\left[\frac{36 c_4 k^2}{t^2}\ln 9m\right]\ln(8N)}$$

$$\le 4 c_4^2 k \sqrt{\ln(9m)\ln(8N)}\, t^{-1} \quad \square$$

$$\Rightarrow \sqrt{\ln(2\mathcal{N}(D_{S,N}, \|\cdot\|_X, t))} \le \tilde{C} K\sqrt{s}\sqrt{\ln(9m)\ln(8N)}\, t^{-1}, \quad 0 < t \le 2K\sqrt{s}$$

- Covering number estimates for small t

Note $x \in D_{S,N} \Rightarrow \frac{x}{\|x\|_2} \in D_{S,N} \Rightarrow \left\|\frac{x}{\|x\|_2}\right\|_X \le \underbrace{\Delta(D_{S,N}, \|\cdot\|_X)}_{\le K\sqrt{s}}$

$$\Rightarrow \|x\|_X \le K\sqrt{s}\,\|x\|_2$$

We seek to estimate covering numbers via Prop. 3.16

(~~possibly~~ ~~totaling~~ as in last homework).

Problem: Prop 3.16 is for $\mathbb{R}^N$

Our approach: Treat Real and imaginary part separately, consider $\mathbb{C}^N$ as $\mathbb{R}^{2N}$

Then calculate

$$\mathcal{N}(D_{S,N}, \|\cdot\|_X, t) \overset{\text{Lemma 3.14 (b)}}{\le} \sum_{S\subset[N],\,|S|=s} \mathcal{N}(B_S, K\sqrt{s}\|\cdot\|_2, t)$$

$$\overset{\text{Lemma 3.14 (4)}}{\le} \sum_{S\subset[N]} \mathcal{N}\left(B_S, \|\cdot\|_2, \frac{t}{K\sqrt{s}}\right) \overset{\text{Prop 3.16}}{\le} \binom{N}{s}\left(1+\frac{2K\sqrt{s}}{t}\right)^{2s}$$

Bound on Binomial coeff. $\searrow$

$$\le \left(\frac{eN}{s}\right)^s \left(1+\frac{2K\sqrt{s}}{t}\right)^{2s}$$

$$\Rightarrow \sqrt{\ln(2\mathcal{N}(D_{s,N}, \|\cdot\|_X, t)} \leq \sqrt{2s}\sqrt{\ln(2eN/s) + \ln(1+2K\sqrt{s}/t)}$$

$$\leq \sqrt{2s}\left(\sqrt{\ln 2eN/s} + \sqrt{\ln(1+2K\sqrt{s}/t}\right), \quad t>0$$

(0) <u>Dudley integral</u>

We will ~~use~~ this estimate for t up to some value $\kappa \in (0, \frac{\Delta(D_{s,N})}{2}$ and the above estimate for $t \geq \kappa$

Then

$$I := \int_0^{\Delta(D_{s,N})/2} \sqrt{\log(2\mathcal{N}(D_{s,N}, \|\cdot\|_X, t)}\,dt$$

$$\leq \sqrt{2s}\int_0^{\kappa}\left(\sqrt{\log(2eN/s)} + \sqrt{\log(1+2K\sqrt{s}/t)}\right)dt$$

$$+ 6K\sqrt{2s\log(9m)\ln 8N}\int_\kappa^{K\sqrt{s}/2} t^{-1}\,dt$$

Estimating integral using integration by parts (notes)

$$\leq \sqrt{2s}\left(\kappa\sqrt{\log(2eN/s)} + \kappa\sqrt{\log(e(1+2K\sqrt{s}/\kappa))}\right) + \tilde{C}K\sqrt{\log(9m)\log(8N)}\cdot\log\left(\frac{K\sqrt{s}}{2\kappa}\right)$$

Fix $\kappa = \frac{K}{3}$. Then

$$I \leq \sqrt{2s}\,K\left(\frac{1}{3}\sqrt{\log(2eN/s)} + \frac{1}{3}\sqrt{\log(e(1+6\sqrt{s}))} + \tilde{C}\sqrt{\log(9m)\log(8N)}\cdot\log(\sqrt{4s})\right)$$

$$\leq \sqrt{2s}\,C_5\sqrt{\log(9m)\log(8N)}\cdot\log(4s)$$

assuming that N, m not too small (≥ 4 or the like)

<u>Together</u> with $\boxed{0}$, this proves Lemma 6.3 $\square$

(Back to Theorem 6.2)

We had
$$E := \mathbb{E}\,\delta_s \leq \frac{2}{m}\mathbb{E}_X\mathbb{E}_\varepsilon \left|\left|\left| \sum_{\ell=1}^m \varepsilon_\ell (X_\ell X_\ell^*) \right|\right|\right|_s$$

$$\leq \frac{2C_1 K\sqrt{s}\log(4s)\sqrt{\log(8N)\log(9m)}}{\sqrt{m}}\,\mathbb{E}_X\sqrt{\left|\left|\left| m^{-1}\sum_{\ell=1}^m X_\ell X_\ell^* \right|\right|\right|_s}$$

Now by the triangle inequality

$$\cancel{E} \; \mathbb{E} ||| m^{-1} \sum_{e=1}^{m} X_e X_e^* |||_s$$

$$\leq \mathbb{E}\, m^{-1} ||| \sum_{e=1}^{m} (X_e X_e^* - \mathrm{Id}) |||_s + ||| \mathrm{Id} |||_s$$

$$= \mathbb{E}\; m^{-1} ||| \sum_{e=1}^{m} (X_e X_e^* - \mathrm{Id}) |||_s + 1$$

$$= \quad E \qquad + 1$$

$$\Rightarrow E \leq \underbrace{\frac{2 C_1 K \sqrt{s}\, \log(4s) \sqrt{\log(8N) \log(9m)}}{\sqrt{m}}}_{=: D} \sqrt{E+1}$$

$$\Rightarrow E^2 \leq D^2 (E+1)$$

$$\Rightarrow \left(E - \frac{D^2}{2}\right)^2 \leq D^2 + D^4/4$$

$$\Rightarrow E \leq \sqrt{D^2 + D^4/4} + D^2/2 \leq D + D^2.$$

Sufficient condition for $E \leq \delta$: $D < \frac{\delta}{2}$ (as $\delta < 1$)

$$D < \frac{\delta}{2} \iff \frac{m}{\log 9m} \geq \underbrace{8 C_1}_{=: C} K^2 s \log^2(4s) \log(8N)$$

$\square$

<u>Corollary 6.5</u>: Under the same assumptions as in Theorem 6.2 but with $\frac{m}{\log(9m)} \geq C \eta^{-2} \delta^{-2} K^2 s \log^2(4s) \log(8N)$, ($\Diamond$)
one has $\mathbb{P}(\delta_s \geq \delta) \leq \eta$

<u>Proof:</u> By Theorem 6.2, ($\Diamond$) implies that $\mathbb{E}\, \delta_s \leq \eta \delta$
Then by Markow's inequality, Theorem 3.4.,

$$\mathbb{P}(\delta_s \geq \delta) \leq \frac{\mathbb{E}\, \delta_s}{\delta} \leq \frac{\eta\delta}{\delta} = \eta$$

$\square$

<u>Remark:</u> Corollary 6.5 is suboptimal by a large margin.
Much better requirements for the embedding dimension are

$$\frac{m}{\log 9m} \gtrsim \bar{C} \delta^{-2} K^2 s \log^2(4s) \log(8N)$$

<u>and</u> $m \geq \bar{C} \delta^{-2} K^2 s \log(\eta^{-1})$ ← much weaker dependence on η