LECTURE #2.

1

Algorithms for sparse FT

Let's ~~review~~ define (~~in~~ more formally) the sparse FT problem:

MATH: ~~Can we design a~~ Given any vector $x \in \mathbb{C}^M$ Is there a sampling set $S \subset [N]$ of

size $|S| = m < n$ and an algorithm $\mathcal{A}$ such that the algorithm on input $x(S)$ can find a(n approximate) k term Fourier rep. of x?

CS: Design a sampling set S (or a distribution on sets) and an algorithm $\mathcal{A}$ s.t. on input $x(S)$, the algo. returns $\hat{x}$, a(n approx.) k-term rep. for x.

How fast is algo?

DISCUSS

- How are m = # samples and k = # Fourier terms related?
- How efficient is the algorithm?
- Is there ONE sample set for all vectors? What if we draw a new sample set (from specified distribution) PER signal?
- Can we eff'ly generate the one sample set? How do we get it o.w.?

We'll start with some basic mathematical answers:

① if $\mathcal{F}_S^*$ (= DFT matrix restricted to ~~cols~~ rows in sample set S) satisfies the RIP (= Restricted Isometry Property), then there are several algos. that return $\hat{x}$ with for all x

$$\|x - \hat{x}\|_2 \le \frac{C}{\sqrt{k}} \|x - x_k\|_1$$

where x_k = best k-term Fourier rep., $|S| = O(k \lg^c(n/k))$ $c =$

② Algos.

- ℓ_1 minimization } convex opt.
- CoSaMP
- IHT
- OMP(?)

(CoSaMP, IHT, OMP: iterative greedy algos.)

→ do these proofs?

③ How to produce an RIP Fourier matrix? → do this proof

RIP = Restricted Isometry Property   RIP$(\underbrace{k, 2, \delta_k})$

Def^n: A matrix $\Phi \in \mathbb{C}^{m\times n}$ ~~has~~ satisfies RIP$(k,2,\delta_k)$ if δ_k is the smallest $\delta \geq 0$ s.t. for all k-sparse $x \in \mathbb{C}^n$

$$(1-\delta)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta)\|x\|_2^2 .$$

[3 params
k = sparsity
2 = norm
δ_k = distortion]

[the distinction between $\|\cdot\|_2^2$ and $\|\cdot\|_2$ is not so critical, except we use $\|\cdot\|_2$ later.]

Can express in terms of singular vals/operator norms:

$$\delta_k = \max_{\substack{S \subset [n] \\ |S| \leq k}} \left\| \Phi_S^* \Phi_S - I \right\|_{2\to 2}$$

[See Rauhut & Foucart to make this argument terribly precise.]

The idea is to have δ_k fairly small for k fairly large; i.e., Φ almost preserves the norm of all k-sparse signals, the distortion δ_k is small.

Props:

- $\delta_1 \leq \delta_2 \leq \delta_3 \leq \cdots \leq \delta_n$
- ideally, $\delta_k < 1$ so that Φ is injective on all k-sparse vects.
- We will actually be interested in RIP$(2k, 2, \delta_{2k})$ for a matrix Φ so that we can guarantee for x, x' distinct k-sparse vects,
  $$\|\Phi(x-x')\|_2^2 > 0.$$
- There are other settings where it's critical to change norm from 2 to something else.

[You should think of Φ as a row submatrix of DFT matrix]

Let's assume that we have a matrix Φ that satisfies RIP$(2k, 2, \delta_{2k})$ and we take linear measurements $y = \Phi x$, then there are several algos. that return $\hat{x}$, good approx^n to x. [$\delta_{2k} < 1$]

We'll prove this directly and get a pretty sophisticated result. There are more "machinery-based" proofs with more abstract, modular pieces.

~~Theorem: Suppose Φ satisfies RIP$(2k, 2, \delta_{2k} < \frac{1}{3})$, then every k-sparse vector $x \in \mathbb{C}^n$ is the unique sol^n of~~

$$\cancel{\hat{x} = \operatorname*{argmin}_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{s.t.} \quad \Phi z = \Phi x}$$

~~i.e., $\hat{x} = x$ uniquely.~~

LECTURE #2

2

W

Theorem 3. *Suppose that Φ is RIP$(2, 6k, \delta_{6k})$ with $\delta_{8k} < 1$. Then there is an algorithm $\mathcal{A}$ such that for every $x \in \Sigma_k$, x can be reconstructed exactly from measurements Φx. The algorithm $\mathcal{A}$ is the convex relaxation of the following (non-convex) optimization problem*

$$\hat{x} = \arg\min \|z\|_0 \quad s.t. \quad \Phi z = \Phi x;$$

in particular, the algorithm is a linear program

We'll use 4k and require $\delta_{4k} < \frac{1}{4}$

$$\hat{x} = \arg\min \|z\|_1 \quad s.t. \quad \Phi z = \Phi x. \tag{1.1.1}$$

Signals or vectors in Σ_k are rather special; they have *exactly* k non-zero components. We typically encounter signals that are compressible rather than exactly sparse. They have a few, say k, predominant coordinates and a large number of considerably smaller ones. It makes more sense to strive to recover this type of signal and to do so with an error that is commensurate with the best k-term compression of the original signal. The idea being that we are going to compress the signal anyway, why not reconstruct it up to the error that we would achieve from a compressed version?

Theorem 4. *For every $x \in \mathbb{R}^N$, the optimizer $\hat{x}$ of* (1.1.1) *satisfies*

then $C \geq 6$

$$\|x - \hat{x}\|_2 \leq \frac{C}{\sqrt{k}} \|x - x_k\|_1$$

where x_k is the best k-term approximation to x and C is a constant.

Proof. Let us define $y = \hat{x} - x$ and observe that $\Phi y = 0$. Our goal is to show that

$$\|y\|_2 \leq \frac{C}{\sqrt{k}} \|x - x_k\|_1.$$

We assume, without loss of generality, that the entries of x are sorted in decreasing order. Let us also assume that the "tail" elements of y are also sorted in decreasing order $|y_{k+1}| \geq |y_{k+2}| \geq \ldots \geq |y_N|$. We divide the indices of y into blocks of size k with $S_0 = S$ being the first block, S_1 the next block of k indices, and so forth. Informally, we refer to y_S as the "head" of the vector y, y_{S^c} as the "tail" (and similarly for x). See Figure 1.1 for an illustration. To reiterate, our goal is to bound the ℓ_2 norm of the head and the tail of y with respect to the ℓ_1 norm of the tail of x.

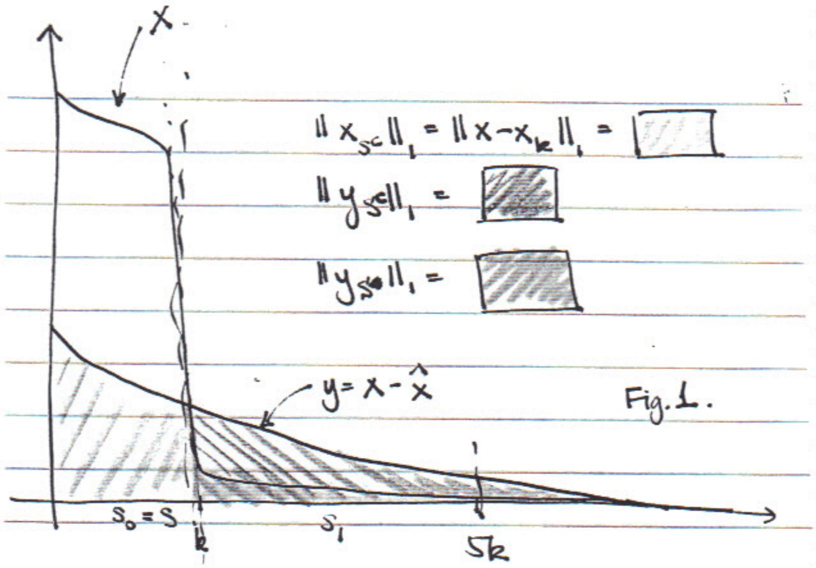


Fig. 1.

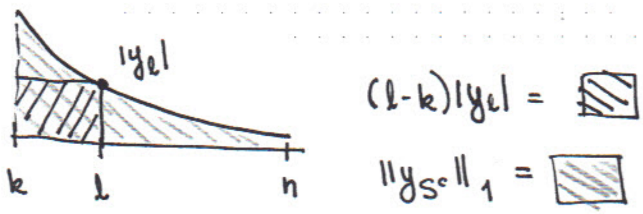
We will prove our main result in three steps, only the last one uses the RIP property of Φ. The first two lemmas control the ℓ_1 and ℓ_2 norms of portions of y. First, we control the ℓ_1 norm of the tail of y in terms of the ℓ_1 norms of its head and the tail of x. Then we pass from the ℓ_1 norm to the ℓ_2 norm. The last lemma uses the RIP property of Φ to control the ℓ_2 norm of the head of y. The final step of the proof is to put together all of the bounds on the pieces of y.

Before we begin, there are two important facts to recall. The first is the relationship between the ℓ_2 and ℓ_1 norms of any k-dimensional vector v,

$$\|v\|_1 = \sum_{\ell=1}^{k} |v_\ell| \leq \Big(\sum_{\ell=1}^{k} |v_\ell|^2\Big)^{1/2}\Big(\sum_{\ell=1}^{k} 1\Big)^{1/2} = \sqrt{k}\|v\|_2. \tag{1.1.2}$$

The second fact uses the sorted-ness of the entries in the tail of y. Observe that for any $\ell \in S^c$, $(\ell - k)|y_\ell|$ is the area of the shaded region in Figure . Because the entries of y_{S^c} are sorted in decreasing order $(\ell - k)|y_\ell| \leq \|y_{S^c}\|_1$ and hence,

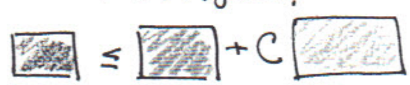
$$|y_\ell| \leq \frac{1}{\ell - k}\|y_{S^c}\|_1. \tag{1.1.3}$$



Lemma 5. *(ℓ_1 concentration)*

$$\|y_{S^c}\|_1 \leq \|y_S\|_1 + O(\|x - x_k\|_1).$$

In Figure 1, we have



Proof. By optimality of $\hat{x}$, we have $\|x\|_1 \geq \|\hat{x}\|_1$. Additionally, $\|x\|_1 = \|x_S\|_1 + \|x_{S^c}\|_1$. Therefore,

$$\begin{aligned}
\|x\|_1 &\geq \|\hat{x}\|_1 \\
&= \|x + y\|_1 \\
&= \|x_S + y_S\|_1 + \|x_{S^c} + y_{S^c}\|_1 \\
&\geq \|x_S\|_1 - \|y_S\|_1 - \|x_{S^c}\|_1 + \|y_{S^c}\|_1,
\end{aligned}$$

where we use the triangle inequality in the last step. Thus,

$$\begin{aligned}
\|y_{S^c}\|_1 &\leq \|y_S\|_1 + \|x\|_1 - \|x_S\|_1 + \|x_{S^c}\|_1 \\
&= \|y_S\|_1 + 2\|x_{S^c}\|_1.
\end{aligned}$$

LECTURE #2 (3)

(12)

Let $S_{01} = S_0 \bigcup S_1$ denote the first two blocks of indices, and set $M = 4k$. This lemma bounds the "far tail of y in terms of its entire tail which we express in terms of the head of y and the tail of x:

Lemma 6. *(Tail Bound)*

$$\|y_{S_{01}^c}\|_2 \le \frac{\|y_{S^c}\|_1}{\sqrt{M}} \le O(\|y_S\|_2 + \frac{1}{\sqrt{k}}\|x - x_k\|_1).$$

Proof. The quantity we are trying to bound is

$$\|y_{S_{01}^c}\|_2^2 = \sum_{\ell \in S_{01}^c} |y_\ell|^2.$$

We will use Equation 1.1.2 for each term in the sum and obtain

$$\begin{aligned}
\|y_{S_{01}^c}\|_2^2 &= \sum_{\ell \in S_{01}^c} |y_\ell|^2 \\
&\le \sum_{\ell=M+k+1}^{N} \frac{1}{(\ell-k)^2}\|y_{S^c}\|_1^2 \quad (*) \\
&= \|y_{S^c}\|_1^2 \sum_{l=M+k+1}^{N} \frac{1}{(\ell-k)^2} \\
&\le \frac{\|y_{S^c}\|_1^2}{M}.
\end{aligned}$$

(*) using Eqn (1.1.3) in its squared form.

- can also argue (**) and get a cleaner structure.

← in order to have M here, need to start at M+k+1 but just need M = ck.

Thus, using Lemma 5, we have

$$\|y_{S_{01}^c}\|_2 \le \frac{\|y_{S^c}\|_1}{\sqrt{M}} \le \frac{1}{2\sqrt{k}}(\|y_S\|_1 + 2\|x - x_k\|_1).$$

With Equation 1.1.3, we can change to the ℓ_2 norm of y_S and complete the proof

$$\|y_{S_{01}^c}\|_2 \le O(\|y_S\|_2 + \frac{1}{\sqrt{k}}\|x - x_k\|_1).$$

□

The last lemma is the only piece of this proof that uses the RIP property of Φ.

Lemma 7. *(Bound on the head)*

$$\|y_{S_{01}}\|_2 \le O(\frac{1}{\sqrt{k}}\|x - x_k\|_1).$$

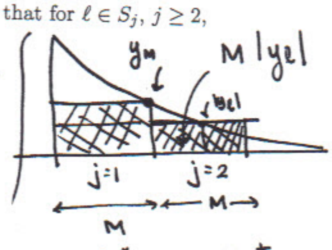
Proof. Using the RIP property of Φ and the definition of y, we find

$$0 = \|\Phi y\|_2 \ge \|\Phi y_{S_{01}}\|_2 - \sum_{j \ge 2} \|\Phi y_{S_j}\|_2 \tag{1.1.4}$$

$$\ge (1-\delta)\|y_{S_{01}}\|_2 - (1+\delta)\sum_{j \ge 2} \|y_{S_j}\|_2. \tag{1.1.5}$$

Let us look at $\|y_{S_j}\|_2^2 = \sum_{\ell \in S_j} |y_\ell|^2$. Because the entries in blocks $S_1, S_2, \ldots$ are sorted in decreasing order, we argue in a similar fashion to Equation 1.1.3, that for $\ell \in S_j$, $j \ge 2$,

$$|y_l| \le \frac{1}{M}\|y_{S_{j-1}}\|_1, \longrightarrow$$



and hence,

$$(**) \quad \|y_{S_j}\|_2 \le \frac{\|y_{S_{j-1}}\|_1}{\sqrt{M}},$$

$M|y_\ell| \overset{*}{\le} M|y_M| \overset{+}{\le} \|y_{S_1}\|_1$

13

(**)

$$\text{for } j \geq 2 \quad \|y_{S_j}\|_2^2 = \sum_{\ell \in S_j} |y_\ell|^2 \leq \frac{1}{M^2} \underbrace{\sum_{\ell \in S_j} \|y_{S_{j-1}}\|_1^2}_{M \text{ copies}} = \frac{1}{M} \|y_{S_{j-1}}\|_1^2$$

$$\Rightarrow \quad \|y_{S_j}\|_2 \leq \frac{1}{\sqrt{M}} \|y_{S_{j-1}}\|_1 .$$

Therefore, the second term in Equation 1.1.6 is bounded above by

$$\begin{aligned}
\sum_{j\geq 2} \|y_{S_j}\|_2 &\leq \frac{1}{\sqrt{M}} \sum_{j \geq 1} \|y_{S_j}\|_1 \\
&= \frac{1}{\sqrt{M}} \|y_{S^c}\|_1 \\
&\leq \frac{1}{\sqrt{M}} \Big(\|y_S\|_1 + O(\|x - x_k\|_1) \Big) \quad \text{(by Lemma 5)} \\
&\leq \sqrt{\frac{k}{M}} \|y_S\|_2 + \frac{1}{\sqrt{M}} O(\|x - x_k\|_1) \quad \text{(by Equation 1.1.2)} \\
&\leq \frac{1}{2} \|y_S\|_2 + \frac{c}{\sqrt{M}} \|x - x_k\|_1
\end{aligned}$$

Let's rewrite Eqn (1.1.15)

$$\|y_{S_{01}}\|_2 \leq \left(\frac{1+\delta}{1-\delta}\right) \sum_{j \geq 2} \|y_{S_j}\|_2$$ and plug in $\delta = \frac{1}{4}$ to sort out constant

$$\leq \frac{5}{3} \sum_{j\geq 2} \|y_{S_j}\|_2$$ and then plug in, ignoring constant on x tail term

$$\leq \frac{5}{3} \cdot \frac{1}{2} \|y_S\|_2 + \frac{c}{\sqrt{M}} \|x - x_k\|_1$$

$$\leq \frac{5}{6} \|y_{S_{01}}\|_2 + \frac{c}{\sqrt{M}} \|x - x_k\|_1$$

$$\Rightarrow \quad \|y_{S_{01}}\|_2 \leq \frac{6 \cdot c}{\sqrt{M}} \|x - x_k\|_1 = \frac{c}{\sqrt{k}} \|x - x_k\|_1 .$$

Putting the three lemmas together, we have

$$\begin{aligned}
\|y\|_2 &\le \|y_{S_{01}}\|_2 + \|y_{S_{01}^c}\|_2 \\
&\le \cancel{O(}\frac{C}{\sqrt{M}}\|x - x_k\|_1) + \cancel{O}(\frac{C}{\sqrt{M}}\|x - x_k\|_1) + \cancel{O}C(\|y_S\|_2) \\
&\le \cancel{O}(\frac{C}{\sqrt{M}}\|x - x_k\|_1) + \cancel{O}C(\|y_{S_{01}}\|_2) \\
&\le \cancel{O(}\frac{C}{\sqrt{M}}\|x - x_k\|_1);
\end{aligned}$$

and the proof of the theorem is complete. □

We proved that

$$\|\hat{x} - x\|_2 \le \frac{C}{\sqrt{k}}\|x - x_k\|_1.$$

Observe that $\hat{x}$ is a vector of length N, it need not be k-sparse. In practice, it is frequently not important that $\hat{x}$ be compressed so reconstructing a vector of length N is perfectly acceptable (e.g., image processing applications in which one needs to render an image). If, however, we want to reconstruct a *compressed* version of our original vector, then a simple (repeated) application of the triangle inequality shows that by setting $\hat{x}_k$ equal to the k largest (in magnitude) entries in $\hat{x}$, $\hat{x}_k$ is a good k-sparse approximation to the original vector x. In which case, we have

Corollary 8. *The k-sparse approximation $\hat{x}_k$ satisfies the reconstruction guarantee*

$$\|\hat{x}_k - x\|_2 \le \|x - x_k\|_2 + \frac{C}{\sqrt{k}}\|x - x_k\|_1.$$

Proof.

$$\begin{aligned}
\|\hat{x}_k - x\|_2 &\le \|x - \hat{x}\|_2 + \|\hat{x} - \hat{x}_k\|_2 \quad (*) \\
&\le \|x - \hat{x}\|_2 + \|\hat{x} - x_k\|_2 \\
&\le \|x - \hat{x}\|_2 + \|x - \hat{x}\|_2 + \|x - x_k\|_2 \\
&\le 2\|x - \hat{x}\|_2 + \|x - x_k\|_2 \\
&\le \|x - x_k\|_2 + \frac{C}{\sqrt{k}}\|x - x_k\|_1.
\end{aligned}$$

In the second inequality, we (*) used the fact that $\hat{x}_k$ is the optimal k-term approximation for $\hat{x}$ while x_k is a sub-optimal one so $\|\hat{x} - \hat{x}_k\|_2 \le \|\hat{x} - x_k\|_2$. □

Please note the ℓ_1 optimization is not an implementation nor an algorithm for solving the opt. problem. It's just a mathematical formulation! There are many different ways to carry out the opt. We won't cover these. They are all iterative and require the multiplication of Φ^* with a vector. To speed up these multiplications, the full DFT matrix is frequently used. ~~This~~ so that the algos. may take advantage of the FFT. The running times of these algos. is typically $O(\sqrt{n}\,T)$ where T is the time to compute the matrix-vector product $\Phi^* v$, possibly via the FFT.

There are other greedy, iterative algorithms that are not based on ℓ_1 optimization but still enjoy the same theoretical performance, both in approximation quality of the output and (faster) running time.

CoSaMP

y = measurements
k = sparsity parameter

Input: Φ, y, k

Initialize: $x^{(0)} = 0, t=0, \Lambda^{(0)} = \emptyset$

While not done {

$$\Lambda^{(t+1)} = \text{supp}(x^{(t)}) \cup \underbrace{\text{Supp}_{2k}\left(\Phi^*(y - \Phi x^{(t)})\right)}_{\text{supp of } 2k \text{ largest (abs. val.) entries of the residual signal}}$$

$$\Phi^*(y - \Phi x^{(t)}) = \Phi^* y - \Phi^* \Phi x^{(t)}$$

$$w^{(t+1)} = \underset{\substack{z \in \mathbb{C}^n \\ \text{supp}(z) \subset \Lambda^{(t+1)}}}{\text{argmin}} \|y - \Phi z\|_2$$

proxy = best ℓ_2 fit to meas. using the $2k$ vects. in supp $\Lambda^{(t+1)}$ | solve via least squares! It's OVERDET'd.

$$x^{(t+1)} = H_k(w^{(t+1)})$$

threshold the proxy to the top k (abs. val) entries.

t ++

}

return: $x^{(t)}$ — we return a k-sparse vector.

Iterative Hard Thresholding

Input: Φ, y, k

Initialize: $x^{(0)} = 0, t=0$

While not done {

$$x^{(t+1)} = H_k\Big(\underbrace{x^{(t)}}_{\text{current approx.}} + \underbrace{\Phi^*(y - \Phi x^{(t)})}_{\text{residual signal}}\Big)$$

threshold the updated approx. to the top k (abs. val.) entries.

t ++

}

Return: $x^{(t)}$ — we return a k-sparse vector

- common features in these iterative, greedy algorithms

let's look at basic op'ns + quantities of interest to really understand algos. (and how/why speedups work)

① they're greedy: retain top k entries at each iteration rather than iteratively updating n entries in vector and deciding at the end which k entries to return (or returning all n) ℓ_1 opt. delays such decisions.

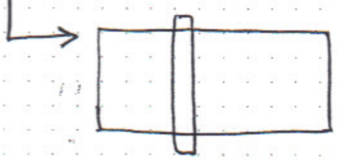
② algos. include same basic operations:

"remeasuring" current approx^n : $\Phi x^{(t)}$

applying adjoint Φ^* to measurements : $\Phi^* y$ or $\Phi^* \Phi x^{(t)}$

"computing" measurements of resid. signal : $\Phi(y - \Phi x^{(t)})$

$= \Phi(x - x^{(t)})$



↳ you can also think of this procedure as computing dot prods of cols of Φ with the measurements (typically followed by thresholding...).

These procedures Φ, Φ^* are EXPENSIVE! They take time $\Omega(n)$ all for information that's $O(k)$ roughly.

② LET'S DO SOMETHING FAR MORE EFFICIENT!

① HOW DO WE GET an RIP- FOURIER MATRIX ?!

Bounded orthonormal systems

(BOS)

Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix. U is a <u>bounded orthonormal system</u> with bound K if

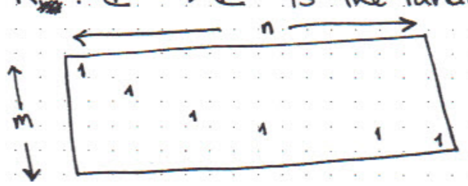
$$\sqrt{n} \max_{k,j\in[n]} |U_{k,j}| = K.$$

For $U = \text{DFT matrix} = \mathcal{F}$, $K = 1$.

Choose the m sampling points $\ell \in [n]$ indep'ly and unif'ly at random $\ell_1, \ell_2, \ldots, \ell_m$ and generate the (structured) random matrix Φ

$$\Phi = \sqrt{n}\, R\, \mathcal{F}$$

where $R : \mathbb{C}^n \to \mathbb{C}^m$ is the random subsampling operator



$$R_{t,j} = \begin{cases} 1, & \text{if } j = \ell_t \\ 0, & \text{otherwise} \end{cases}$$

Then $\sqrt{n} R \mathcal{F} x$ = selected random Fourier coeffs of x. (scaled by $\sqrt{n}$...).

$y = \Phi x$

The computation $R\mathcal{F}x$ can be done quickly via FFT and then discarding the extraneous FCoeffs.

Theorem: Let Φ be the random sampling matrix assoc'ed with a BOS with constant $K \geq 1$. If, for $\delta \in (0,1)$,

$$m \geq C K^2 \delta^{-2} k \lg^4(n),$$

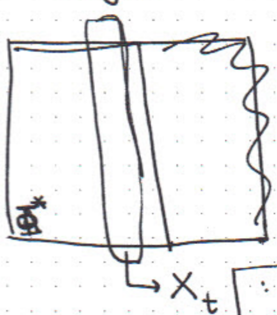
then with probability $\geq 1 - n^{-\lg^3(n)}$, the RIP (2) constant δ_k of $\frac{1}{\sqrt{m}}\Phi$ satisfies $\delta_k < \delta$.

The proof of this is COMPLICATED! I'll just give a brief sketch of the major moving parts:

(1) use the operator norm char= of RIP constant:

$$\delta_k = \max_{\substack{\Lambda \subset [n] \\ |\Lambda| \leq k}} \left\| \Phi_\Lambda^* \Phi_\Lambda - I \right\|_{2\to 2}$$

Φ_Λ = col. submatrix of Φ restricted to supp set Λ.

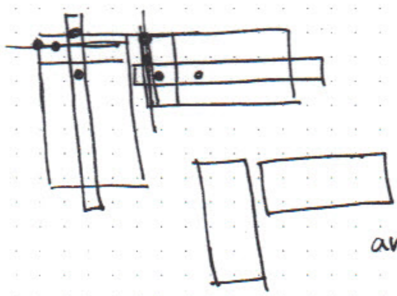


Then $\Phi_\Lambda^* \Phi_\Lambda = \sum_{t=1}^{m} X_t X_t^*$ and

the orthogonality rel^ns tell us $\mathbb{E}(\Phi_\Lambda^* \Phi_\Lambda) = I$.

: Let X_t be a random col. of Φ^* with t corresponding to a sampling pt. l_t; i.e., $X_t = \Phi_t^*$

Check: $\mathbb{E}\left(\frac{1}{m}\Phi_\Lambda^* \Phi_\Lambda\right) = \mathbb{E}\left(\sum_{t=1}^{m} \frac{1}{m} X_t X_t^*\right) = I$



proof: $\frac{1}{m}\left(\Phi_\Lambda^* \Phi_\Lambda\right)_{l,j} = \frac{n}{m}\sum_{t=1}^{m} u_{l,t}^* u_{t,j}$

$= \frac{n}{m}\sum_{t=1}^{m} u_{l,t}^* u_{t,j}$

and $\mathbb{E}\left(\frac{n}{m}\sum_{t=1}^{m} u_{l,t}^* u_{t,j}\right) = \frac{n}{m}\sum_{t=1}^{m} \frac{1}{n} \underbrace{\sum_{s=1}^{n} u_{l,s}^* u_{s,j}}$

$\frac{1}{n}$ = prob. that $t = s$

$= \begin{cases} \frac{n}{m} \cdot \frac{1}{n} \sum_{t=1}^{m} 1 = 1, & \text{if } l = j \end{cases}$

$= \langle u_l, u_j \rangle = \begin{cases} 1, & \text{if } l = j \end{cases}$

then we define a semi-norm on matrices:

- set $D_{k,n} = \{ z \in \mathbb{C}^n \mid \|z\|_2 \le 1,\ \|z\|_0 \le k \} = \bigcup_{\substack{\Lambda \subset [n] \\ |\Lambda| = k}} B^2_\Lambda$ = unit ℓ_2 ball in $\mathbb{C}^k$

- $\|M\|_{\mathcal{K}} = \sup_{z \in D_{k,n}} |\langle Mz, z \rangle|$

so that $\delta_k = \left\| \frac{1}{m} \sum_{t=1}^m X_t X_t^* - I \right\|_{\mathcal{K}} = \frac{1}{m} \left\| \sum_{t=1}^m \left[X_t X_t^* - \mathbb{E}(X_t X_t^*) \right] \right\|_{\mathcal{K}}$

Then ① check $\mathbb{E}(\delta_k)$ and bound

$$\mathbb{E} \left\| \sum_{t=1}^m X_t X_t^* - \mathbb{E}(X_t X_t^*) \right\|_{\mathcal{K}} \le 2\,\mathbb{E} \left\| \sum_{t=1}^m \varepsilon_t X_t X_t^* \right\|_{\mathcal{K}}$$

= Rademacher seq., independent of the choice of t

$\varepsilon_t = \pm 1$

TECHNICAL!

② And bound the probab. that such a process deviates from its mean by a large amount.

TECHNICAL!

Note! DUALITY

Let's summarize:

If we draw indep'ly and unif'ly at random ~~samples of~~ $m = O(k \lg^4(n))$ F coeffs of a vector x of length n, then whp the resulting rand. row submatrix of $\mathcal{F}^* = \mathcal{F}^{-1}\Phi$, will satisfy RIP and ℓ_1 opt, spike