CS 395T: Sublinear Algorithms, Fall 2020 Oct. 27, 2020

Lecture 18: Fourier RIP

Prof. Eric Price *Scribe: Anish Acharya, Rudrajit Das*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Background and Motivation

1.1 Compressed Sensing and the RIP

There are two main types of compressed sensing algorithms that we have seen so far:

- *Structure-aware*: Here the recovery algorithm is tied to the matrix structure, e.g. Count-Sketch. These methods are faster and allow for sparse matrices, but are less robust.
- *Structure-oblivious*: As the name suggests, here the recovery algorithm does not depend on the sensing matrix structure. The recovery is formulated as a least squares problem with ℓ_1-norm regularization to account for the signal sparsity. These methods are slower, require dense matrices, but are more robust.

We are going to focus on the latter category in this lecture.
Mathematically, the goal is to recover $\mathbf{x} \in \mathbb{R}^n$ – which is k-sparse – given linear measurements of the form $\mathbf{y} = \Phi\mathbf{x}$, where $\Phi \in \mathbb{R}^{m \times n}$ and $m \ll n$. We have seen earlier that we can solve a convex relaxation of this NP-hard problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} = \Phi\mathbf{x}. \tag{1}$$

The main bottleneck of this approach is the multiplication by Φ or Φ^T. But we know that if the Φ satisfies the Restricted Isometry Property (RIP), then this convex relaxation gives us the exact solution (in the noiseless case). Recall the RIP:

$$(1-\epsilon)\|\mathbf{x}\|^2 \leq \|\Phi\mathbf{x}\|^2 \leq (1+\epsilon)\|\mathbf{x}\|^2 \text{ for all } k\text{-sparse } \mathbf{x} \in \mathbb{R}^n. \tag{2}$$

Ideally, we would like m to be small and the cost of multiplication by Φ or Φ^T to be small. For a random Gaussian matrix, $m = \mathcal{O}(k \log n)$, but the multiplication time is $\mathcal{O}(kn \log n)$. We would like to decrease the multiplication time to $\mathcal{O}(n \log n)$ while having a small m – this is where the Fourier matrix comes into the picture!

1.2 Johnson Lindenstrauss (JL) Transform

Recall that the JL transform can be used to project high dimensional data points to a lower dimension such that their distances are approximately preserved in this lower dimensional space.

Specifically, suppose $S \subset \mathbb{R}^n$. Then, the JL lemma says that there exists a low dimensional sketch of S, say $\Phi(S) \in \mathbb{R}^m$, such that

$$(1-\epsilon)\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq (1+\epsilon)\|\mathbf{x}\|_2 \;\forall\; \mathbf{x} \in S, \text{ and}$$

$$\langle \Phi\mathbf{x}, \Phi\mathbf{y}\rangle \in \langle \mathbf{x}, \mathbf{y}\rangle \pm \epsilon\|\mathbf{x}\|_2\|\mathbf{y}\|_2.$$

Theorem 1. *Let $\mathbf{x} \in \mathbb{R}^n$. A random Gaussian matrix $\Phi \in \mathbb{R}^{m\times n}$ will have*

$$(1-\epsilon)\|\mathbf{x}\|_2 \leq \|\Phi\mathbf{x}\|_2 \leq (1+\epsilon)\|\mathbf{x}\|_2$$

with probability $1-\delta$, so long as

$$m \gtrsim \frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right).$$

Setting $\delta = 1/2^k$, we can embed 2^k points in $\mathcal{O}(k/\epsilon^2)$ dimensions.
Ideally, we would like to have a small m (i.e. target dimension) as well as fast multiplication to use in practical problems such as regression, low-rank approximation, clustering, etc. Random Gaussian allows us to have small $m \sim \mathcal{O}(k/\epsilon^2)$, but the multiplication time is $\mathcal{O}(nk/\epsilon^2)$ which is expensive. It was furthrer shown by Krahmer and Ward (2011) that given a matrix obeying the RIP, we can transform it (by multiplying it with a random diagonal matrix) such that the transformed matrix satisfies the JL lemma with high probability – i.e. RIP $\implies$ JL. Now since the random Fourier matrix obeys the RIP (which we are going to prove in this lecture) and it allows fast multiplication ($\mathcal{O}(n \log n)$ time), we can use it for *fast JL*!

2 Preliminaries

2.1 The Sub-Sampled Fourier Matrix

Recall the Fourier matrix $\mathbf{F}$ whose entries are given by $\mathbf{F}_{i,j} = \omega^{ij}$ where $\omega = \sqrt{-1}/n$. The Fourier matrix obeys the property that $\mathbf{F}\mathbf{F}^* = n\mathbf{I}_n$. Suppose A contains random rows from the Fourier matrix. Note that multiplication by A can be done in $\mathcal{O}(n \log n)$ time by using the Fast Fourier Transform (FFT).

But an open question is – *how many rows does A need to have to ensure it satisfies the RIP*? Several works have established the bound of $\mathcal{O}(k \log n \log^2 k)$. Recently this was improved to $\mathcal{O}(k \log n \log k)$ for the subsampled Hadamard matrix. Ideally, we would like to reduce this to just $m = \mathcal{O}(k \log n)$.

In this lecture, we are going to discuss the proof technique of Rudelson and Vershynin (2008) to obtain the bound of $m = \mathcal{O}(k \log^4 n)$. Before doing that, we shall discuss several helpful probability tools.

2.2 Concentration of Measure: A Toolbox

Suppose Σ_k is the set of unit-norm k-sparse vectors. We wish to show for our distribution Φ on matrices that:

$$\mathbb{E}\sup_{\mathbf{x}\in\Sigma_k}\left|\|\Phi\mathbf{x}\|^2 - \|\mathbf{x}\|^2\right| < \epsilon.$$

This can be thought of as the expected deviation of $\Phi^T\Phi$ from $\mathbf{I}_n$, in a non-standard norm. Probabilists have lots of tools to analyze this – see Figure 1. The key idea is to somehow convert such problems to Gaussians and then use results for Gaussian concentration. We shall prove Symmetrization and Dudley's entropy integral.

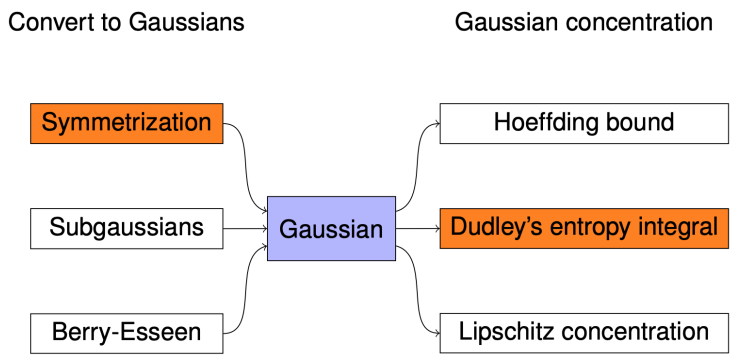

Figure 1: A Probabilist's Toolbox. We are going to be focusing on the highlighted blocks.

2.2.1 Symmetrization

Lemma 1. *Suppose $X_1, \ldots, X_t$ are i.i.d with mean μ. For any norm $\|.\|$*

$$\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i X_i - \mu\Big\|\Big] \leq 2\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i s_i X_i\Big\|\Big] \leq 3\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i g_i X_i\Big\|\Big].$$

where $s_i \in \{\pm 1\}$ independently and $g_i \in \mathcal{N}(0,1)$ independently.

Symmetrization is a measure to quantify how well does X concentrate about its mean.

For example, we can use this to bound the RIP constant of the sub-sampled Fourier matrix $\mathbf{A}$, for some norm $\|\cdot\|$.

Proof. Draw $X_1', \ldots, X_t'$ independently from the same distribution. Then:

$$\begin{aligned}
\mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i X_i - \mu\Big\|\Big] &= \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i X_i - \mathbb{E}\Big[\frac{1}{t}\sum_i X_i'\Big]\Big\|\Big] \\
&\leq \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i (X_i - X_i')\Big\|\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{t}\sum_i s_i(X_i - X_i')\Big\|\Big].
\end{aligned}$$

Applying the triangle inequality, we get the first inequality in the lemma.

For the second inequality in the lemma, note that $\mathbb{E}[|g_i|] \approx 0.8 > 2/3$. Hence:

$$2\mathbb{E}[\|\sum_i s_i X_i\|] \leq 3\mathbb{E}[\|\sum_i s_i \mathbb{E}[|g_i|] X_i\|]$$

$$\leq 3\mathbb{E}[\|\sum_i s_i|g_i|X_i\|]$$
$$\leq 3\mathbb{E}[\|\sum_i g_iX_i\|].$$

This concludes the proof. ■

2.2.2 Gaussian Processes

Very briefly, we can think of a Gaussian process as follows – for every point $\mathbf{x} \in T$ which is our space, we have a Gaussian $G_{\mathbf{x}}$. We are interested in estimating the following:

$$\mathbb{E} \sup_{\mathbf{x}\in T} G_{\mathbf{x}}.$$

Obviously, this quantity depends on the geometry of T. Also note that for any two points $\mathbf{x}, \mathbf{y} \in T$, we can define a notion of distance between them, i.e. $\|\mathbf{x}-\mathbf{y}\|$, which is just the standard deviation of $G_{\mathbf{x}} - G_{\mathbf{y}}$. In other words, $G_{\mathbf{x}} - G_{\mathbf{y}} \sim \mathcal{N}(0, \|\mathbf{x}-\mathbf{y}\|^2)$.

For example, suppose $\mathbf{A}$ is a random $m \times n$ Gaussian matrix. For $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^n$, define

$$G_{\mathbf{u},\mathbf{v}} := \mathbf{u}^T\mathbf{A}\mathbf{v} := \langle \mathbf{u}\mathbf{v}^T, \mathbf{A}\rangle. \tag{3}$$

Then $G_{\mathbf{u},\mathbf{v}} \sim \mathcal{N}(0, \|\mathbf{u}\mathbf{v}^T\|_F^2)$. Also note that:

$$\mathbb{E}\|\mathbf{A}\|_2 = \mathbb{E} \sup_{\mathbf{u},\mathbf{v}\in\mathbb{R}^{m-1}\times\mathbb{R}^{n-1}} \mathbf{u}^T\mathbf{A}\mathbf{v} = \mathbb{E} \sup_{\mathbf{u},\mathbf{v}\in\mathbb{R}^{m-1}\times\mathbb{R}^{n-1}} G_{\mathbf{u},\mathbf{v}}.$$

Note that for this example, $\|(\mathbf{u},\mathbf{v}) - (\mathbf{u}',\mathbf{v}')\| = \|\mathbf{u}\mathbf{v}^T - \mathbf{u}'\mathbf{v}'^T\|_F$. Coming back to our problem of estimating $\mathbb{E}\sup_{\mathbf{x}\in T} G_{\mathbf{x}}$, we next introduce the chaining technique for computing this.

2.2.3 Chaining and Dudley's Entropy Integral

Let us first consider a simple case where T is finite with $|T| = n$. Suppose $0 \in T$, and $G_0 = 0$. Also let $\sigma_{\max} = \max_{\mathbf{x}\in T} \|\mathbf{x}-0\|$. See Figure 2 for reference. Recalling that the maximum of n Gaussians varies as $\mathcal{O}(\sqrt{\log n})$, we get:

$$\mathbb{E} \sup_{\mathbf{x}\in T} G_{\mathbf{x}} \lesssim \sigma_{\max}\sqrt{\log n}. \tag{4}$$

Note that this technique only works if $|T|$ is finite.

Let us now consider an ϵ-cover of T under the same $\|.\|$ norm. Denote this cover by $\mathcal{C}(T, \|.\|, \epsilon)$ and the covering number $= |\mathcal{C}(T, \|.\|, \epsilon)| = \mathcal{N}(T, \|.\|, \epsilon)$. Then for every $\mathbf{x} \in T$, we have some $c(\mathbf{x}) \in \mathcal{C}(T, \|.\|, \epsilon)$, such that $\|\mathbf{x} - c(\mathbf{x})\| \leq \epsilon$. See Figure 3 for reference. In that case:

$$\begin{aligned}
\mathbb{E} \sup_{\mathbf{x}\in T} G_{\mathbf{x}} &= \mathbb{E} \sup_{\mathbf{x}\in T} [G_{\mathbf{x}} - G_{c(\mathbf{x})} + G_{c(\mathbf{x})}] \\
&\leq \mathbb{E} \sup_{\mathbf{x}\in T} [G_{\mathbf{x}} - G_{c(\mathbf{x})}] + \mathbb{E} \sup_{\mathbf{x}\in T} G_{c(\mathbf{x})} \\
&\lesssim \underbrace{\epsilon\sqrt{\log n}}_{\text{(I)}} + \underbrace{\sigma_{\max}\sqrt{\log \mathcal{N}(T, \|.\|, \epsilon)}}_{\text{(II)}}.
\end{aligned} \tag{5}$$

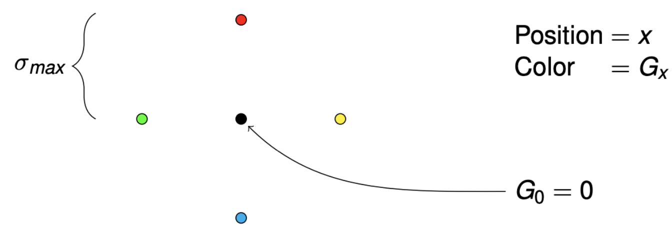

Figure 2: A simple case of T with $|T| = n = 4$.

In eq. (5), (I) is obtained as follows: the maximum variance of each $G_{\mathbf{x}} - G_{c(\mathbf{x})}$ is $\leq \epsilon$ as $\mathcal{C}(T, \|.\|, \epsilon)$ is an ϵ-cover of T, and $|T| = n$. Then we can directly apply the result of eq. (4).
For obtaining (II), note that $\mathbb{E} \sup_{\mathbf{x} \in T} G_{c(\mathbf{x})} = \mathbb{E} \sup_{\mathbf{y} \in \mathcal{C}(T, \|.\|, \epsilon)} G_{\mathbf{y}}$. Recalling that $|\mathcal{C}(T, \|.\|, \epsilon)| = \mathcal{N}(T, \|.\|, \epsilon)$, and that $\|\mathbf{y} - 0\| \leq \sigma_{\max}$ for all $\mathbf{y} \in \mathcal{C}(T, \|.\|, \epsilon) \subset T$, we again apply the result of eq. (4).

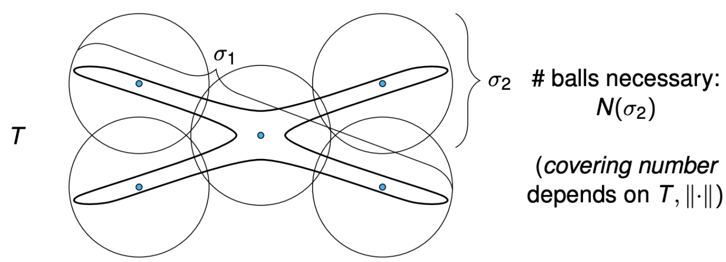

Figure 3: The chaining example with two levels. Note that $\sigma_1 = \sigma_{\max}$ and $\sigma_2 = \epsilon$ here.

Henceforth, let $\sigma_{\max} = \sigma_1$ and $\epsilon = \sigma_2$. Then, Equation (5) becomes:

$$\mathbb{E} \sup_{\mathbf{x} \in T} G_{\mathbf{x}} \lesssim \sigma_1 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_2)} + \sigma_2 \sqrt{\log n}. \tag{6}$$

In the above example, we stopped at two levels – but nothing is stopping us from repeating the above idea. Specifically, we can keep finding covers of T with progressively smaller cover sizes (i.e. ϵ) and keep extending (6). As an example, now if we find a σ_3-cover of T of size $\mathcal{N}(T, \|.\|, \sigma_3)$, where $\sigma_3 < \sigma_2$, then just by extending the idea of (6), we get:

$$\mathbb{E} \sup_{\mathbf{x} \in T} G_{\mathbf{x}} \lesssim \sigma_1 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_2)} + \sigma_2 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_3)} + \sigma_3 \sqrt{\log n}. \tag{7}$$

Doing this infinitely many times, we get:

$$\mathbb{E} \sup_{\mathbf{x} \in T} G_{\mathbf{x}} \lesssim \sigma_1 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_2)} + \sigma_2 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_3)} + \sigma_3 \sqrt{\log \mathcal{N}(T, \|.\|, \sigma_4)} + \ldots \tag{8}$$

Let us choose $\sigma_r = \sigma_1 / 2^{r-1}$. Then, we get:

$$\mathbb{E} \sup_{\mathbf{x} \in T} G_{\mathbf{x}} \lesssim \sum_{r=0}^{\infty} \frac{\sigma_1}{2^r} \sqrt{\log \mathcal{N}\Big(T, \|.\|, \frac{\sigma_1}{2^{r+1}}\Big)} \lesssim \int_0^{\infty} \sqrt{\log \mathcal{N}(T, \|.\|, \sigma)} d\sigma. \tag{9}$$

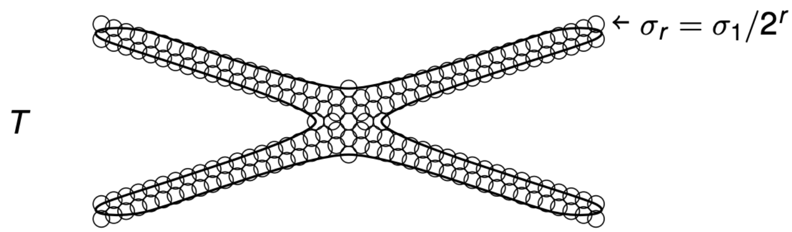

Figure 4: Repeated chaining.

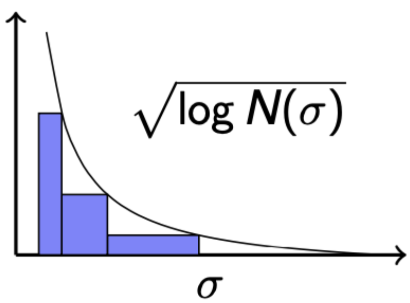

Figure 5: Reimann sum approximation.

The second inequality follows by Reimann sum approximation (see Figure 5).

The second inequality in (9) is what is known as Dudley's Entropy Integral. We state this as a theorem now:

Theorem 2. ***(Dudley's Entropy Integral)*** *Define the norm $\|.\|$ of a Gaussian process G by*

$$\|\mathbf{x} - \mathbf{y}\| = \textit{standard deviation of } G_{\mathbf{x}} - G_{\mathbf{y}}.$$

Then

$$\gamma_2(T, \|.\|) = \mathbb{E} \sup_{\mathbf{x} \in T} G_{\mathbf{x}} \lesssim \int_0^\infty \sqrt{\log \mathcal{N}(T, \|.\|, u)} du.$$

This is useful for bounding a random variable over a space using its geometry.

3 RIP constant of Sub-sampled Fourier Matrix

Theorem 3. *Suppose the sensing matrix $\mathbf{F}$ contains random rows sampled from a Fourier matrix. Let $\Omega \subseteq [n]$, be a set of $m = \mathcal{O}(\frac{k}{\epsilon^2} \log^4 n)$ i.i.d uniform indices. Then $\frac{1}{\sqrt{m}} \mathbf{F}_\Omega$ satisfies (k, ϵ)-RIP in expectation. More formally if $\mathbf{A} = \frac{1}{\sqrt{m}} \mathbf{F}_\Omega$ and $\sum_k$ denote unit-norm k-sparse vectors, then:*

$$\mathbb{E}_\Omega \sup_{\mathbf{x} \in \sum_k} |\|\mathbf{A}\mathbf{x}\|^2 - \|\mathbf{x}\|^2| < \epsilon. \tag{10}$$

Proof. We will follow the proof described in Rudelson and Vershynin (2008).

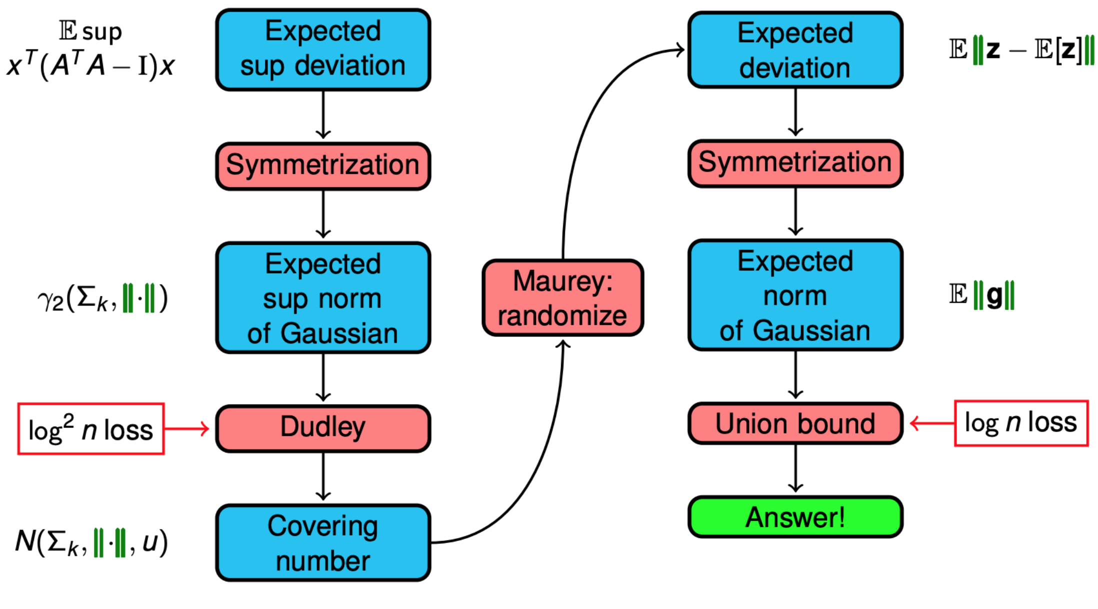

Figure 6: The proof outline.

(Step-0. Setup).
Let $\delta_i = \mathbf{I}_{i \in \Omega}$. Then, we have: $\mathbb{P}[\delta_i] = \frac{m}{n}$. Then we have,

$$\mathbf{A}\mathbf{x} = \frac{1}{m} \sum_{i=1}^{n} \delta_i \mathbf{F}_i \mathbf{x}$$

Then we would like to analyze the RIP constant:

$$\mathbf{R}_\Omega := \sup_{\mathbf{x} \in \sum_k} |\mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x} - 1|$$

Now, for any fixed $\mathbf{x}$:

$$\mathbb{E}_\Omega[\mathbf{x}^T \mathbf{A}^T \mathbf{A}\mathbf{x}] = \frac{1}{n} \mathbf{x}^T \mathbf{F}^T \mathbf{F}\mathbf{x} = \|\mathbf{x}\|_2^2$$

(Step-I. Symmetrization)

$$\begin{aligned}
\mathbb{E}_\Omega[\mathbf{R}_\Omega] &= \mathbb{E}_\Omega \sup_{\mathbf{x}\in\sum_k} |\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2| \\
&= \mathbb{E}_\Omega \sup_{\mathbf{x}\in\sum_k} |\|\mathbf{A}\mathbf{x}\|_2^2 - \mathbb{E}_\Omega[\|\mathbf{A}\mathbf{x}\|_2^2] \\
&= \frac{1}{m}\mathbb{E}_\delta \sup_{\mathbf{x}\in\sum_k} \Big| \sum_{i=1}^{k} \delta_i < \mathbf{F}_i, \mathbf{x} >^2 - \mathbb{E}_\Omega[\sum_{i=1}^{k} \delta_i < \mathbf{F}_i, \mathbf{x} >^2] \Big| \\
&\leq \frac{1}{m} 3\mathbb{E}_{\delta,\mathbf{g}} \sup_{\mathbf{x}\in\sum_k} \Big| \sum_{i=1}^{k} \mathbf{g}_i \delta_i < \mathbf{F}_i, \mathbf{x} >^2 \Big| \quad \text{(Using Lemma 1)} \\
&\leq \frac{1}{m} 3\mathbb{E}_\delta \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{x}\in\sum_k} \Big| \sum_{i=1}^{k} \mathbf{g}_i \delta_i < \mathbf{F}_i, \mathbf{x} >^2 \Big|
\end{aligned}$$

(Step-II. Bounding the Gaussian process metric)
From symmetrization we have:

$$m\mathbb{E}_\Omega[\mathbf{R}_\Omega] \lesssim \mathbb{E}_\delta \mathbb{E}_{\mathbf{g}} \sup_{\mathbf{x}\in\sum_k} \Big| \sum_{i=1}^{k} \mathbf{g}_i < \mathbf{F}_i, \mathbf{x} >^2 \Big|$$

Now, we fix Ω and define the Gaussian process:

$$\mathbf{G}_x = \sum_{i\in\Omega} \mathbf{g}_i < \mathbf{F}_i, \mathbf{x} >^2$$

$\mathbf{G}_x$ induces the norm:

$$\begin{aligned}
\|\mathbf{x}-\mathbf{y}\|_{\mathbf{G}}^2 =& \mathbb{E}[(\mathbf{G}_x - \mathbf{G}_y)^2] \\
=& \sum_{i\in\Omega} (< \mathbf{F}_i, \mathbf{x} >^2 - < \mathbf{F}_i, \mathbf{y} >^2)^2 \\
=& \sum_{i\in\Omega} (< \mathbf{F}_i, \mathbf{x}+\mathbf{y} > \cdot < \mathbf{F}_i, \mathbf{x}-\mathbf{y} >)^2 \\
\leq& \sum_{i\in\Omega} (< \mathbf{F}_i, \mathbf{x}+\mathbf{y} >^2) \max_{i\in\Omega} < \mathbf{F}_i, \mathbf{x}-\mathbf{y} >^2 \\
\leq& 4 \sup_{\hat{\mathbf{x}}\in\sum_k} (< \mathbf{F}_i, \hat{\mathbf{x}}+\mathbf{y} >^2) \max_{i\in[n]} < \mathbf{F}_i, \mathbf{x}-\mathbf{y} >^2 \\
\leq& 4m(1+\mathbf{R}_\Omega)\|\mathbf{F}(\mathbf{x}-\mathbf{y})\|_\infty^2
\end{aligned}$$

Let us define: $\|\mathbf{x}\|_{\mathbf{F}} = \|\mathbf{F}\mathbf{x}\|_\infty$. Then we have,

$$
\begin{aligned}
m\mathbb{E}[\mathbf{R}] &\lesssim \mathbb{E}_\Omega \mathbb{E} \sup_{\mathbf{x}\in\sum_k} \mathbf{G_x} \\
&= \mathbb{E}_\Omega \gamma_2(\sum_k, \|\cdot\|_{\mathbf{G}}) \\
&\leq \mathbb{E}_\Omega \int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{G}}, u)} du \quad \text{(Using Theorem 2 above)} \\
&\lesssim \mathbb{E}_\Omega \int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, \frac{u}{\sqrt{1+\mathbf{R}_\Omega}})} du \\
&\leq \mathbb{E}_\Omega \sqrt{m(1+\mathbf{R}_\Omega)} \int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, u)} du
\end{aligned}
$$

(Road map for the rest of the proof).
We will show: $\int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, \mathbf{u})} du \lesssim \epsilon\sqrt{m}$. This would imply for $\epsilon < 1$:

$$
\begin{aligned}
\mathbb{E}[\mathbf{R}_\Omega] &\lesssim \epsilon \mathbb{E}[\sqrt{1+\mathbf{R}_\Omega}] \\
\Rightarrow \mathbb{E}[\mathbf{R}_\Omega] &\lesssim \epsilon
\end{aligned}
$$

Thus this would imply the result.

(Step - III. Bounding $\mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, \mathbf{u})$ - Covering Number Bound).
Recall $\|\mathbf{y}\|_{\mathbf{F}} = \|\mathbf{F}\mathbf{y}\|_\infty$. Note that $\sum_k \subseteq \mathbf{B}_1.\sqrt{k}$ where $\mathbf{B}_1 = \{\mathbf{x} | \|\mathbf{x}\|_1 \leq 1\}$. Therefore,

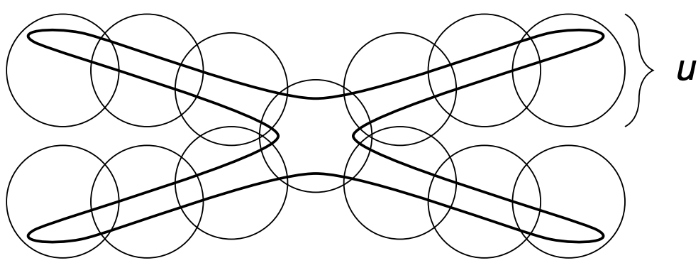

Figure 7: How many l_2 balls of radius u required to cover $\mathbf{B}_1$?

$$
\int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, u)} du \leq \sqrt{k} \int_0^\infty \sqrt{\log \mathbf{N}(\mathbf{B}_1, \|\cdot\|_{\mathbf{F}}, u)} du
$$

We can bound $\mathbf{N}(\mathbf{B}_1, \|\cdot\|_{\mathbf{F}}, u)$ by an easy volume argument as seen previously in earlier lectures on Covering numbers -

$$
\begin{aligned}
\mathbf{N}(\mathbf{B}_1, \|\cdot\|_{\mathbf{F}}, u) &\leq \mathbf{N}(\mathbf{B}_1, \|\cdot\|_1, u) \\
&\leq \left(1 + \frac{2}{u}\right)^n
\end{aligned}
$$

Plugging this into the integral we obtain

$$\begin{aligned}
\int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, u)}du &\leq \sqrt{k}\int_0^\infty \sqrt{\log \mathbf{N}(\mathbf{B}_1, \|\cdot\|_{\mathbf{F}}, u)}du \\
&\leq \int_0^1 \sqrt{n}\sqrt{(1+\frac{2}{u})}du \\
&\lesssim \sqrt{n}
\end{aligned}$$

which is not a very good bound. (Remember, we are shooting for a bound that is $\mathcal{O}(\sqrt{k})$.

(Maurey's empirical method).
We can apply Maurey's Empirical Method to get a better bound. Consider any $\mathbf{x} \in \mathbf{B}_1^+$. Let $\hat{\mathbf{z}}_i$

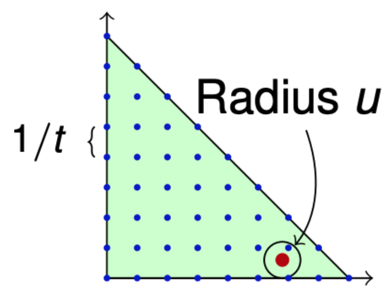

Figure 8: How many balls of radius u required to cover $\mathbf{B}_1^+$

be i.i.d. randomized roundings of $\mathbf{x}$ to simplex. The sample mean, $\mathbf{z} = \frac{1}{t}\sum_{r=1}^t \mathbf{z}_r$ converges to $\mathbf{x}$ as $t \to \infty$. We can let t to be large enough that, regardless of $\mathbf{x}$, we can bound the expected deviation as:

$$\mathbb{E}\left[\left\|\frac{1}{t}\sum_{r=1}^t \mathbf{z}_r - \mathbf{x}\right\|_{\mathbf{F}}\right] \leq u$$

All $\mathbf{x}$ lie within u of at least one possible $\mathbf{z}$. Note that, this implies,

$$N(B_1, \|\cdot\|, u) \leq \|\mathbf{z}\|_0 \leq (n+1)^t$$

Thus we see that for such an t, we would have $N \leq (n+1)^t$ which is a much better bound on as long as t is not too big. We can bound the deviation using Symmetrization as follows,

$$\begin{aligned}
\sigma_R &= \mathbb{E}\left[\left\|\frac{1}{t}\sum \mathbf{z}_i - \mathbf{x}\right\|_{\mathbf{F}}\right] \\
&\lesssim \mathbb{E}\left[\left\|\frac{1}{t}\sum g_i \mathbf{z}_i\right\|_{\mathbf{F}}\right] \\
&= \mathbb{E}\left[\left\|\mathbf{g}\right\|_{\mathbf{F}}\right],
\end{aligned}$$

where $\mathbf{g} \in \mathbb{R}^n$ has

$$\mathbf{g}_j \sim \mathcal{N}(0, \frac{\text{number of } \mathbf{z}_i \text{ at } e_j}{t}),$$

independently in each coordinate.
This implies, for each i, $(\mathbf{Fg})_i \sim \mathcal{N}(0,1)$. Hence $\|\mathbf{g}\|_{\mathbf{F}} = \|\mathbf{Fg}\|_\infty \lesssim \sqrt{\log n}$ with high probability and in expectation. Thus setting $t = \frac{\log n}{u^2}$ suffices and we get:

$$\mathbf{N}(\mathbf{B}_1, \|\cdot\|_{\mathbf{F}}, u) \leq (n+1)^{\mathcal{O}(\frac{\log n}{u^2})} \tag{11}$$

(Step - IV Putting everything together).
Now plugging the bound into our entropy integral we obtain:

$$\int_0^\infty \sqrt{\log \mathbf{N}(\sum_k, \|\cdot\|_{\mathbf{F}}, \mathbf{u})} du \lesssim \sqrt{k \log^4 n} \leq \epsilon \sqrt{m}$$

This gives us the RIP constant $\lesssim \sqrt{\frac{k \log^4 n}{m}}$
This concludes the proof. ■

(Note on $\log^3 n$ factor loss)
As depicted in Figure 6, Dudley's entropy integral results in a loss of a factor of $\log^2 n$ and the union bound finally results in a a loss of a factor of $\log n$. Thus, we are off a by a factor of a loss of a factor of $\log^3 n$ with respect to the optimal bound of $m = \mathcal{O}(k \log n)$.

References

Krahmer, F. and Ward, R. (2011). New and improved johnson–lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281.

Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 61(8):1025–1045.