CS 229r: Algorithms for Big Data	Fall 2013

Lecture 17 — October 29, 2013

Thomas Steinke — *Scribe: Ludwig Schmidt*

1 Overview

In the last lecture we introduced compressive sensing and saw the Restricted Isometry Property (RIP), one of the main tools in compressive sensing. In this lecture, we study a connection between the RIP and the Johnson-Lindenstrauss property (JL).

2 JL implies RIP

We have already established one connection between RIP and JL matrices by proving the following theorem:

Theorem 1 (Pset 5 problem 1 or [2]). *Let D be a distribution on $m \times n$-matrices satisfying the (ε, δ)-distributional JL property with $\delta \ll \frac{1}{\binom{n}{k} C^k}$ for some universal constant $C > 0$, i.e. for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$,*

$$\mathbb{P}\left[\left| \|\Pi x\|_2^2 - 1 \right| > \varepsilon \right] < \delta \, .$$

Then w.h.p. $\Pi \sim D$ is a $(O(\varepsilon), k)$-RIP matrix, i.e. for all $x \in \mathbb{R}^n$ with $\|x\|_0 \le k$,

$$\|\Pi x\|_2^2 \lesseqgtr (1 \pm O(\varepsilon)) \|x\|_2 \, .$$

This theorem shows that distributional JL implies the RIP for suitably chosen parameters. In fact $m = O(\log(1/\delta)/\varepsilon^2) = O(k \log(n/k\varepsilon)/\varepsilon^2)$ suffices.

In this lecture, we show that the converse also holds: given a matrix satisfying the RIP, we can construct a distribution on matrices satisfying the JL property. Note that the RIP without randomization cannot give you the distributional JL property with good paramters, because it will always fail for a vector in the kernel of the RIP matrix.

3 RIP implies JL

The following theorem by Krahmer and Ward (2011) [5] is the main result of today's class.

Theorem 2 ([5]). *Let $\Pi \in \mathbb{R}^{m \times n}$ be a matrix satisfying the $(\varepsilon, 2k)$-RIP. Let $\sigma \in \{+1, -1\}^n$ be uniformly random and D_σ the diagonal matrix with σ on the diagonal. Then ΠD_σ satisfies the $(O(\varepsilon), 2^{-\Omega(k)})$-distributional JL property.*

Combined with theorem 1, this shows that JL and RIP are equivalent up to randomization. We can construct JL matrices from RIP matrices and vice versa.

The work [5] followed immediately after [1], which showed that techniques used to show that a sampled Fourier matrix satisfies RIP [6, 3] could be extended to obtain JL when flipping column signs randomly. Krahmer and Ward then showed that the RIP itself, black box, implies that flipping column signs gives JL (the theorem above).

3.1 Application of the theorem

Before we prove the theorem, we briefly describe an application. Recall the Fast JL Transform (FJLT) construction

$$\Pi = SHD$$

where

- S is a scaled $m \times n$ sampling matrix,
- H is a $n \times n$ Hadamard or Fourier matrix,
- D is a diagonal matrix with random signs (± 1).

In lecture 12 we analyzed this construction by splitting Π into two parts: the "preconditioner" HD and the sampling matrix D. We showed that for $y = HDx$, $\|y\|_\infty$ is small w.h.p. and then used this property to show that $\|Sy\|_2 \approx \|y\|_2$.

For an alternative proof using theorem 2, consider the following result:

Theorem 3 ([6, 3])**.** *If $m \gtrsim \frac{k}{\varepsilon^2} \log(n) \log^3(k)$, then w.h.p. $\Pi = SH$ has the (ε, k)-RIP.*

So this gives an alternative analysis of the FJLT by analyzing the part SH and its interaction with the random sign matrix D.

3.2 Proof

We now prove the main theorem. In the remainder of this section, let $x \in \mathbb{R}^n$ be a fixed unit vector. We analyze $\|\Pi D_\sigma x\|_2^2$:

$$\begin{aligned}
\|\Pi D_\sigma x\|_2^2 &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \Pi_{i,j} \sigma_j x_j \right)^2 \\
&= \|\Pi D_x \sigma\|_2^2 \\
&= \sigma^T (D_x^T \Pi^T \Pi D_x) \sigma \\
&= \sigma^T X \sigma
\end{aligned}$$

We would like to use Hanson-Wright on X. But for this we need to bound $\|X\|$ and $\|X\|_F$, which does not work.

Instead, we partition X into $X = A + B + B^T + C$ as follows and analyse A, B, and C separately.

Assume w.l.o.g. that $|x_i| \geq |x_{i+1}|$ for all i. We partition the indices $\{1, 2, \cdots, n\}$ into n/k blocks $(1) = \{1, \cdots, k\}, (2) = \{k+1, \cdots, k+k\}, \cdots, (i) = \{(i-1)k+1, \cdots, ik\}, \cdots, (n/k) = \{n-k, \cdots, n\}$. Thus $x^T = (x_{(1)}^T x_{(2)}^T \cdots x_{(n/k)}^T)$ and $\Pi = (\Pi_{(1)} \Pi_{(2)} \cdots \Pi_{(n/k)})$, where $\Pi_{(i)}$ is a $m \times k$ matrix corresponding to the i^{th} block of columns of Π.

Graphically we partition X into $(n/k) \times (n/k)$ blocks and divide them into A, B, and C as follows.

$$X = D_x^T \Pi^T \Pi D_x = \begin{pmatrix} A & B & B & B & \cdots & B \\ B^T & A & C & C & \cdots & C \\ B^T & C & A & C & \cdots & C \\ B^T & C & C & A & \cdots & C \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ B^T & C & C & C & \cdots & A \end{pmatrix}$$

Formally,

$$\begin{aligned} \forall i \qquad & A_{(i),(i)} = D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(i)} D_{x_{(i)}}, \\ \forall j > 1 \qquad & B_{(1),(j)} = D_{x_{(1)}}^T \Pi_{(1)}^T \Pi_{(j)} D_{x_{(j)}}, \\ \forall i, j > 1, i \neq j \qquad & C_{(i),(j)} = D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(j)} D_{x_{(j)}}, \end{aligned}$$

and all other blocks of A, B, and C are zero.

Our plan is now the following:

1. Analyze A directly.
2. Analyze B with the Hoeffing bound.
3. Apply Hanson-Wright to C.

3.2.1 Analysis of A

We have

$$\begin{aligned} \sigma^T A \sigma &= \sum_{i=1}^{n/k} \sigma_{(i)}^T D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(i)} D_{x_{(i)}} \sigma_{(i)} \\ &= \sum_{i=1}^{n/k} \|\Pi_{(i)} D_{x_{(i)}} \sigma_{(i)}\|_2^2 \\ &= \sum_{i=1}^{n/k} \|\Pi_{(i)} D_{\sigma_{(i)}} x_{(i)}\|_2^2 \\ &\lesseqgtr \sum_{i=1}^{n/k} (1 \pm \varepsilon) \|D_{\sigma_{(i)}} x_{(i)}\|_2^2 \\ &= (1 \pm \varepsilon) \|x\|_2^2 \\ &= (1 \pm \varepsilon) \;, \end{aligned}$$

where line four follows from line three by using the RIP of Π.

3.2.2 Analysis of B

Let (-1) to be all the blocks except the first. Let B' be the nonzero part of B, that is $B' = D_{x_{(1)}} \Pi_{x(1)}^T \Pi_{(-1)} D_{x_{(-1)}}$.

First, we prove the following lemma:

Lemma 4. *Let* $\Pi = \left[\Pi_{(1)}, \ldots, \Pi_{(n/k)}\right]$ *be a* $(\varepsilon, 2k)$*-RIP matrix. Then* $\|\Pi_{(i)}^T \Pi_{(j)}\| \leq 2\varepsilon$ *for* $i \neq j$.

Proof. Let $x_{(i)}, x_{(j)} \in \mathbb{R}^k$ be unit vectors and $x = \left[0, \ldots, 0, x_{(i)}^T, 0, \ldots, 0, x_{(j)}^T, 0, \ldots, 0\right]^T$. Then we have

$$\begin{aligned}
(1 \pm \varepsilon)\|x\|_2^2 &\gtreqless \|\Pi x\|_2^2 \\
&= \|\Pi_{(i)} x_{(i)}\|_2^2 + \|\Pi_{(j)} x_{(j)}\|_2^2 + 2x_{(i)}^T \Pi_{(i)}^T \Pi_{(j)} x_{(j)} \\
&\gtreqless (1 \mp \varepsilon)\|x_{(i)}\|_2^2 + (1 \mp \varepsilon)\|x_{(j)}\|_2^2 + 2x_{(i)}^T \Pi_{(i)}^T \Pi_{(j)} x_{(j)} \ .
\end{aligned}$$

Hence

$$\begin{aligned}
2x_{(i)}^T \Pi_{(i)}^T \Pi_{(j)} x_{(j)} &\lesseqgtr (1 \pm \varepsilon)2 - (1 \mp \varepsilon) - (1 \mp \varepsilon) \\
&= \pm 4\varepsilon \ .
\end{aligned}$$

And so $x_{(i)}^T \Pi_{(i)}^T \Pi_{(j)} x_{(j)} \lesseqgtr \pm 2\varepsilon$, which imples the statement of the lemma. ∎

Let $u \in \mathbb{R}^k$ satisfy $\|u\|_\infty = 1$ and $v = (v_{(2)} v_{(3)} \cdots v_{(n/k)}) \in \mathbb{R}^{n-k}$ satisfy $\|v\|_2 = 1$. Then

$$\begin{aligned}
u^T B' v &= \sum_{j>1} u^T D_{x_{(1)}}^T \Pi_{(1)}^T \Pi_{(j)} D_{x_{(j)}} v_{(j)} \\
&= \sum_{j>1} x_{(1)}^T D_u^T \Pi_{(1)}^T \Pi_{(j)} D_{x_{(j)}} v_{(j)} \\
&\leq \sum_{j>1} \left\|x_{(1)}\right\|_2 \|D_u\| \left\|\Pi_{(1)}^T \Pi_{(j)}\right\| \left\|D_{x_{(j)}}\right\| \left\|v_{(j)}\right\|_2 \\
&= \sum_{j>1} \left\|x_{(1)}\right\|_2 \|u\|_\infty \left\|\Pi_{(1)}^T \Pi_{(j)}\right\| \left\|x_{(j)}\right\|_\infty \left\|v_{(j)}\right\|_2 \\
&\leq \frac{2\varepsilon}{\sqrt{k}} \sum_{j>1} \left\|x_{(j-1)}\right\|_2 \left\|v_{(j)}\right\|_2 .
\end{aligned}$$

This follows from our ordering of x:

$$\left\|x_{(j)}\right\|_\infty = \max_{i \in (j)} |x_i| \leq \min_{i \in (j-1)} |x_i| \leq \underset{i \in (j-1)}{\text{average}} |x_i| = \frac{1}{k} \left\|x_{(j-1)}\right\|_1 \leq \frac{1}{\sqrt{k}} \left\|x_{(j)}\right\|_2 .$$

Now we apply the arithmetic-mean-geometric-mean inequality: For all $a, b \geq 0$,

$$0 \leq (a-b)^2 = a^2 - 2ab + b^2 \implies ab \leq \frac{1}{2}(a^2 + b^2).$$

So

$$u^T B' v \leq \frac{\varepsilon}{\sqrt{k}} \sum_{j>1} \left\| x_{(j-1)} \right\|_2^2 + \left\| v_{(j)} \right\|_2^2 \leq \frac{2\varepsilon}{\sqrt{k}}.$$

We will analyse $\sigma^T B \sigma = \sigma_{(1)}^T B' \sigma_{(-1)}$ using the following concentration bound.

Theorem 5 (Hoeffding [4]). *Let $x \in \mathbb{R}^n$ be fixed and $\sigma \in \{-1, +1\}^n$ uniformly random. Then*

$$\mathbb{P}\left[|x^T \sigma| > \lambda \right] \leq 2e^{-\lambda^2 / \|x\|_2^2}.$$

Consider a fixed value of $\sigma(1)$ and let $x = \sigma_{(1)}^T B'$. Then, by the above calculation, $\|x\|_2 \leq 2\varepsilon/\sqrt{k}$. So

$$\mathbb{P}\left[|\sigma^T B \sigma| \geq \varepsilon \right] = \mathbb{P}\left[|x^T \sigma_{(-1)}| \geq \varepsilon \right] \leq 2e^{\varepsilon^2 / \|x\|_2^2} = 2e^{-\varepsilon^2/(4\varepsilon^2/k)} = 2e^{-k/4}.$$

Likewise $\mathbb{P}\left[|\sigma^T B^T \sigma| \geq \varepsilon \right] \leq 2e^{-k/4}$.

3.2.3 Analysis of C

We now use Hanson-Wright on C, which contains all remaining blocks, i.e., all blocks except those on the digonal (A) and in the first row and column of blocks (B and B^T). For this, we first bound the Frobenius norm of C. We use the fact that $\|AB\|_F \leq \|A\| \|B\|_F$, which we leave as an exercise for the reader.

$$\begin{aligned}
\|C\|_F^2 &= \sum_{\substack{i,j>1 \\ i \neq j}} \|D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(j)} D_{x_{(j)}}\|_F^2 \\
&= \sum_{\substack{i,j>1 \\ i \neq j}} \left(\|D_{x_{(i)}}\| \|\Pi_{(i)}^T \Pi_{(j)}\| \|D_{x_{(j)}}\|_F \right)^2 \\
&\leq \sum_{\substack{i,j>1 \\ i \neq j}} \|x_{(i)}\|_\infty^2 (2\varepsilon)^2 \|x_{(j)}\|_2^2 \\
&\leq \frac{4\varepsilon^2}{k} \sum_{\substack{i,j>1 \\ i \neq j}} \|x_{(i-1)}\|_2^2 \|x_(j)\|_2^2 \\
&\leq \frac{4\varepsilon^2}{k}
\end{aligned}$$

In line three we used lemma 4 and in line four shelling.

We now bound the operator norm of C. Since C is symmetric, there is a unit vector $y \in \mathbb{R}^n$ such that $\|C\| = y^T C y$. Hence

$$
\begin{aligned}
\|C\| &= y^T C y \\
&= \sum_{\substack{i,j>1 \\ i \neq j}} y_{(i)}^T D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(j)} D_{x_{(j)}} y_{(j)} \\
&\leq \sum_{\substack{i,j>1 \\ i \neq j}} \|y_{(i)}\|_2 \|D_{x_{(i)}}\| \|\Pi_{(i)}^T \Pi_{(j)}\| \|D_{x_{(j)}}\| \|y_{(j)}\|_2 \\
&\leq 2\varepsilon \sum_{i,j>1} \|y_{(i)}\|_2 \|x_{(i)}\|_\infty \|x_{(j)}\|_\infty \|y_{(j)}\|_2 \\
&\leq \frac{2\varepsilon}{k} \sum_{i,j>1} \|y_{(i)}\|_2 \|x_{(i-1)}\|_2 \|x_{(j-1)}\|_2 \|y_{(j)}\|_2 \\
&= \frac{2\varepsilon}{k} \left(\sum_{i>1} \|y_{(i)}\|_2 \|x_{(i-1)}\|_2 \right)^2 \\
&\leq \frac{2\varepsilon}{k} \left(\frac{1}{2} \sum_{i>1} \|y_{(i)}\|_2^2 + \|x_{(i-1)}\|_2^2 \right)^2 \\
&\leq \frac{2\varepsilon}{k} .
\end{aligned}
$$

We can now apply the Hanson-Wright inequality. Note that

$$
\begin{aligned}
\mathbb{E}\left[\sigma^T C \sigma\right] &= \sum_{\substack{i,j>1 \\ i \neq j}} \mathbb{E}\left[\sigma_{(i)}^T D_{x_{(i)}}^T \Pi_{(i)}^T \Pi_{(j)} D_{x_{(j)}} \sigma_{(j)}\right] \\
&= 0
\end{aligned}
$$

because the $\sigma_{(i)}$ and $\sigma_{(j)}$ are independent for $i \neq j$.

Hence we have

$$
\begin{aligned}
\mathbb{P}\left[|\sigma^T C \sigma| > \varepsilon\right] &\leq 2^{-\Omega(\min\left\{\frac{\varepsilon^2}{\|C\|_F^2}, \frac{\varepsilon}{\|C\|}\right\})} \\
&\leq 2^{-\Omega(\min\left\{\frac{\varepsilon^2 k}{4\varepsilon^2}, \frac{\varepsilon k}{2\varepsilon}\right\})} \\
&\leq 2^{-\Omega(k)}.
\end{aligned}
$$

3.2.4 Putting things together

We now have all the parts for proving that ΠD_σ has the distributional JL property.

$$
\begin{aligned}
\mathbb{P}\left[|\|\Pi D_\sigma x\|_2^2 - 1| > 4\varepsilon\right] &= \mathbb{P}\left[|\sigma^T X \sigma - 1| > 4\varepsilon\right] \\
&\leq \mathbb{P}\left[|\sigma^T A \sigma - 1| > \varepsilon\right] + 2\,\mathbb{P}\left[|\sigma^T B \sigma| > \varepsilon\right] + \mathbb{P}\left[|\sigma^T C \sigma| > \varepsilon\right] \\
&\leq 2^{-\Omega(k)}
\end{aligned}
$$

This completes the proof.

4 Comments

Did we use the full power of $(\varepsilon, 2k)$-RIP to prove the result?

No. We only needed RIP to show that each of the diagonal blocks of $\Pi^T \Pi$ approximately preserved vectors (the analysis of A) and that every off-diagonal block has operator norm at most 2ε (Lemma 4). Thus we only needed Π to preserve vectors supported on one or two blocks, rather than on all $2k$-subsets of coordinates. There are $O((n/k)^2)$ such pairs of blocks, which is much smaller than the $\binom{n}{2k}$ possible choices of coordinates for vectors to be supported on.

To obtain a distribution satisfying the (ε, δ)-JL property using this result, one needs Π satisfying the $(\varepsilon/4, O(\log(1/\delta))$-RIP property. An optimal Π has $m = O(k \log(n/k)/\varepsilon^2) = O(\log(1/\delta)/\varepsilon^2 \cdot \log(n)$, which is optimal up to the $\log(n)$ factor. So this this result is nearly optimal.

References

[1] Nir Ailon, Edo Liberty. An Almost Optimal Unrestricted Fast Johnson-Lindenstrauss Transform. *ACM Transactions on Algorithms* 9(3): 21, 2013.

[2] Richard Baraniuk, Mark Davenport, Ronald DeVore, Michael Wakin. A Simple Proof of the Restricted Isometry Property for Random Matrices *Constructive Approximation* 28(3):253–263, 2008.

[3] Mahdi Cheraghchi, Venkatesan Guruswami , Ameya Velingker. Restricted Isometry of Fourier Matrices and List Decodability of Random Linear Codes. *SODA (ACM-SIAM Symposium on Discrete Algorithms)*, 2013.

[4] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[5] Felix Krahmer, Rachel Ward. New and Improved Johnson-Lindenstrauss Embeddings via the Restricted Isometry Property. *SIAM J. Math. Analysis*, 43(3):1269–1281, 2011.

[6] Mark Rudelson, Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.