Sketching Algorithms for Big Data	Fall 2017

Lecture 10 — Oct. 3, 2017

Prof. Jelani Nelson *Scribe: Zhun Deng*

1 Overview

In the last lecture we mentioned about fast JL transform (FJLT). We said that the time complexity is $O(d \log d + m^3)$, where d is the dimension of the vector x and m is the number of rows of the transform matrix Π. However, in practice, the vector x is often a sparse vector, and we would expect that the time complexity for the transform $x \to \Pi x$ is $O(m\|x\|_0)$, where $\|x\|_0 = |\{i : x_i \neq 0\}|$, and the time complexity of FJLT is terrible if $\|x\|_0$ is small relative to d.

In this lecture, we suggest methods to to speed up JL by making Π sparse.

2 History of various of methods

2.1 The method by [Ach01]

Here, we first introduce a method purposed by [Ach01].

- Make Π sparse, each column of Π has less or equal than s non-zero entries in expectation. The expected time is $O(s\|x\|_0)$ to compute Πx.
- The specific construction is that Π_{ij} 's are independent random variables that

$$\Pi_{ij} = \begin{cases} 0, \text{ w.p. } 1-q \\ \frac{\pm 1}{\sqrt{qm}}, \text{ w.p. } q \end{cases}$$

  where w.p. is the shorthand for with probability.

and [Ach01] proved that to get "$\forall \|x\|_2 = 1, P(\big|\|\Pi x\|_2^2 - 1\big| < \varepsilon) < \delta$", it is sufficient to take

$$m \geq (1 + o(1))\frac{4\ln(\frac{2}{\delta})}{\varepsilon^2},\ q = \frac{1}{3} \text{ (so, } s = \frac{m}{3})$$

2.2 The method by [Mat08]

In [Mat08], it mentions that if the approach is to take i.i.d. sub-gaussian entries, then one must have

$$q = \Omega(1) \text{ for } m = O(\frac{1}{\varepsilon^2}\lg(\frac{1}{\delta})).$$

2.3 The method by [DKS10]

In [DKS10], it mentions that it is possible to achieve "$\forall \|x\|_2 = 1, P(\big| \|\Pi x\|_2^2 - 1\big| < \varepsilon) < \delta$" by

$$m = O(\frac{1}{\varepsilon^2} \lg(\frac{1}{\delta})), \ s = \tilde{O}(\frac{1}{\varepsilon} \lg^3(\frac{1}{\delta}))$$

where $\tilde{O}(f) := f \cdot \text{poly}(\log(f))$.

Specifically, their matrix is constructed in the following way:

$$\Pi = AB$$

where

$$A = \begin{bmatrix} & 0 & \\ & 0 & \\ & \vdots & \\ & \pm 1 & \\ & 0 & \\ & 0 & \\ & \vdots & \end{bmatrix}_{m \times ds}$$

is a matrix with each column has one non-zeros entry with value 1 or -1 and

$$B = \begin{bmatrix} 1 & & & \\ 1 & & & \\ \vdots & & & \\ 1 & & & \\ & 1 & & \\ & 1 & & \\ & \vdots & & \\ & 1 & & \\ & & \ddots & \\ & & \ddots & 1 \\ & & \ddots & 1 \\ & & \ddots & \vdots \\ & & \ddots & 1 \end{bmatrix}_{ds \times d}$$

where each column has s 1's and it can duplicate each element s times for the vector x.

Remark 1. *Also it is worth mentioning that there were other methods improve s to $\tilde{O}(\varepsilon^{-1} \lg^2(1/\delta))$ by [KN10] and [BOR10].*

2.4 The method by [KN14]

Based on the method of [KN14], by noticing the error coming from collision of elements, Prof. Nelson purposed another approach and proved that it is possible to achieve "$\forall \|x\|_2 = 1, P(\big|\|\Pi x\|_2^2 -$

$1\big| < \varepsilon) < \delta$" by

$$m = O(\frac{1}{\varepsilon^2}\lg\frac{1}{\delta}),\ \ s = O(\frac{1}{\varepsilon}\lg\frac{1}{\delta}).$$

The construction of Π can be in the following two forms, both the analysis we will show works.

$$\Pi = \frac{1}{\sqrt{s}}\begin{bmatrix} & \pm 1 & \\ & 0 & \\ & \vdots & \\ & \pm 1 & \\ & \pm 1 & \end{bmatrix}$$

where in each column there are s non-zero entries, being 1 or -1.

Another construction which is easier to implement is:

$$\Pi = \frac{1}{\sqrt{s}}\left[\begin{array}{c|c|c} \cdots & B_1 & \\ \hline \cdots & B_2 & \\ \hline \cdots & \vdots & \\ \hline \cdots & B_s & \end{array}\right]$$

where each block B_i is a m/s column vector with only one entry non-zero, being 1 or -1.

The corresponding countsketch:

$$h : [d] \times [s] \to [\frac{m}{s}]$$

$$\sigma : [d] \times [s] \to \{-1, 1\}$$

3 Analysis

Now, we analysis the method by section 2.4. Our goal is to prove that for any $\varepsilon > 0$

$$P_\Pi(\big|\|\Pi_x\|_2^2 - 1\big| > \varepsilon) < \delta$$

Before that, we make clear of some notations.

$$\Pi_{r,i} := \frac{\eta_{r,i}\sigma_{r,i}}{\sqrt{s}},\ \ \sigma_{r,i} \in \{-1, 1\},\ \ \eta_{r,i} \in \{0, 1\}.$$

Also we should notice that

$$E\eta_{r,i} = \frac{s}{m}.$$

Now, we begin the analysis.

First, notice that

$$(\Pi x)_r = \sum_{r=1}^{d} \Pi_{r,i} x_i = \frac{1}{\sqrt{s}} \sum_{i=1}^{d} \eta_{r,i}\sigma_{r,i} x_i,$$

then, we can obtain that

$$\|\Pi x\|_2^2 = \sum_{r=1}^{m} (\Pi x)_r^2 = \frac{1}{s} \sum_{r=1}^{m} \sum_{i,j=1}^{d} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j.$$

The last term can be conquered in two parts,

$$\frac{1}{s} \sum_{r=1}^{m} \sum_{i,j=1}^{d} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j = \frac{1}{s} \sum_{r=1}^{m} \big[\sum_{i=1}^{d} x_i^2 \eta_{r,i} + \sum_{i \neq j} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j \big]$$

Notice the first part $\frac{1}{s}\sum_{r=1}^{m}\sum_{i=1}^{d} x_i^2 \eta_{r,i}$ is exactly $\|x\|_2^2$ since $\sum_r \eta_{r,i} = s$, then we only need to analyze the second part.

We denote

$$Z = \frac{1}{s} \sum_{r=1}^{m} \sum_{i \neq j}^{d} \eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j$$

In order to analyze Z, we need some inequalities.

3.1 Some Inequalities We Need

Throughout, for a random variable X, $\|X\|_p$ denotes $(\mathbb{E}\,|X|^p)^{1/p}$. It is known that $\|\cdot\|_p$ is a norm for any $p \geq 1$ (Minkowski's inequality). It is also known $\|X\|_p \leq \|X\|_q$ whenever $p \leq q$. Henceforth, whenever we discuss $\|\cdot\|_p$, we will assume $p \geq 1$.

Lemma 1 (Khintchine Inequality). *For any $p \geq 1$, $x \in \mathbb{R}^n$, and (σ_i) independent Rademachers,*

$$\|\sum_i \sigma_i x_i\|_p \lesssim \sqrt{p} \cdot \|x\|_2$$

Lemma 2 (Jensen Inequality). *For F convex, $F(\mathbb{E}\,X) \leq \mathbb{E}\,F(X)$.*

Lemma 3 (Markov Inequality).

$$\mathbb{P}(Z > \lambda) \leq \lambda^{-p} \cdot \mathbb{E}\,|Z|^p.$$

Lemma 4 (Decoupling [DlPG12]). *Let $x_1, \ldots, x_n$ be independent and mean zero, and $x'_1, \ldots, x'_n$ identically distributed as the x_i and independent of them. Then for any $(a_{i,j})$ and for all $p \geq 1$*

$$\|\sum_{i \neq j} a_{i,j} x_i x_j\|_p \leq 4 \|\sum_{i,j} a_{i,j} x_i x'_j\|_p$$

Theorem 5 (Hanson-Wright inequality). *For $\sigma_1, \ldots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n \times n}$ real and symmetric, for all $p \geq 1$*

$$\|\sigma^T A \sigma - \mathbb{E}\,\sigma^T A \sigma\|_p \lesssim \sqrt{p} \cdot \|A\|_F + p \cdot \|A\|.$$

Proof. Without loss of generality we assume in this proof that $p \geq 2$ (so that $p/2 \geq 1$). Then

$$\begin{aligned}
\|\sigma^T A\sigma - \mathbb{E}\,\sigma^T A\sigma\|_p &\lesssim \|\sigma^T A\sigma'\|_p \text{ (by decoupling)} && (1)\\
&\lesssim \sqrt{p}\cdot \|\|Ax\|_2\|_p \text{ (Khintchine)} && (2)\\
&= \sqrt{p}\cdot \|\|Ax\|_2^2\|_{p/2}^{1/2} && (3)\\
&\leq \sqrt{p}\cdot \|\|Ax\|_2^2\|_p^{1/2}\\
&\leq \sqrt{p}\cdot (\|A\|_F^2 + \|\|Ax\|_2^2 - \mathbb{E}\,\|Ax\|_2^2\|_p)^{1/2} \text{ (triangle inequality)}\\
&\leq \sqrt{p}\cdot \|A\|_F + \sqrt{p}\cdot \|\|Ax\|_2^2 - \mathbb{E}\,\|Ax\|_2^2\|_p^{1/2}\\
&\lesssim \sqrt{p}\cdot \|A\|_F + \sqrt{p}\cdot \|x^T A^T Ax'\|_p^{1/2} \text{ (by decoupling)}\\
&\lesssim \sqrt{p}\cdot \|A\|_F + p^{3/4}\cdot \|\|A^T Ax\|_2\|_p^{1/2} \text{ (Khintchine)}\\
&\lesssim \sqrt{p}\cdot \|A\|_F + p^{3/4}\cdot \|A\|^{1/2}\cdot \|\|Ax\|_2\|_p^{1/2} && (4)
\end{aligned}$$

Writing $E = \|\|Ax\|_2\|_p^{1/2}$ and comparing 2 and 4, we see that for some constant $C > 0$,

$$E^2 - Cp^{1/4}\|A\|^{1/2}E - C\|A\|_F \leq 0.$$

Thus E must be smaller than the larger root of the above quadratic equation, implying our desired upper bound on E^2. □

Theorem 6 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables that are each at most K almost surely, and where*

$$\sum_{i=1}^{n} \mathbb{E}(X_i - \mathbb{E}\,X_i)^2 = \sigma^2.$$

Then for all $p \geq 1$

$$\|\sum_{i=1}^{n} X_i - \mathbb{E}\sum_i X_i\|_p \lesssim \sigma\sqrt{p} + Kp.$$

3.2 Analysis of Z

Theorem 7. *As long as $m \simeq \varepsilon^{-2}\log(1/\delta)$ and $s \simeq \varepsilon m$,*

$$\forall x : \|x\|_2 = 1,\ \mathbb{P}_{\Pi}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta. \tag{5}$$

Proof. Abusing notation and treating σ as an mn-dimensional vector,

$$Z = \|\Pi x\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\eta_{r,i}\eta_{r,j}\sigma_{r,i}\sigma_{r,j}x_i x_j := \sigma^T A_{x,\eta}\sigma,$$

Thus by Hanson-Wright

$$\|Z\|_p \leq \|\sqrt{p}\cdot \|A_{x,\eta}\|_F + p\cdot \|A_{x,\eta}\|\|_p \leq \sqrt{p}\cdot \|\|A_{x,\eta}\|_F\|_p + p\cdot \|\|A_{x,\eta}\|\|_p.$$

$A_{x,\eta}$ is a block diagonal matrix with m blocks, where the rth block is $(1/s)x^{(r)}(x^{(r)})^T$ but with the diagonal zeroed out. Here $x^{(r)}$ is the vector with $(x^{(r)})_i = \eta_{r,i}x_i$. Now we just need to bound $\|\|A_{x,\eta}\|_F\|_p$ and $\|\|A_{x,\eta}\|\|_p$.

Since $A_{x,\eta}$ is block-diagonal, its operator norm is the largest operator norm of any block. The eigenvalue of the rth block is at most $(1/s)\cdot\max\{\|x^{(r)}\|_2^2, \|x^{(r)}\|_\infty^2\} \le 1/s$, and thus $\|A_{x,\eta}\| \le 1/s$ with probability 1.

Next, define $Q_{i,j} = \sum_{r=1}^m \eta_{r,i}\eta_{r,j}$ so that

$$\|A_{x,\eta}\|_F^2 = \frac{1}{s^2}\sum_{i\neq j} x_i^2 x_j^2 \cdot Q_{i,j}.$$

We will show for $p \simeq s^2/m$ that for all i,j, $\|Q_{i,j}\|_p \lesssim p$, where we take the p-norm over η. Therefore for this p,

$$\begin{aligned}
\|\|A_{x,\eta}\|_F\|_p &= \|\|A_{x,\eta}\|_F^2\|_{p/2}^{1/2}\\
&\le \|\frac{1}{s^2}\sum_{i\neq j} x_i^2x_j^2\cdot Q_{i,j}\|_p^{1/2}\\
&\le \frac{1}{s}\left(\sum_{i\neq j} x_i^2x_j^2\cdot\|Q_{i,j}\|_p\right)^{1/2} \quad \text{(triangle inequality)}\\
&\le \frac{1}{\sqrt{m}}
\end{aligned}$$

Then by Markov's inequality and the settings of p, s, m,

$$\mathbb{P}(|\|\Pi x\|_2^2 - 1| > \varepsilon) = \mathbb{P}(|\sigma^T A_{x,\eta}\sigma| > \varepsilon) < \varepsilon^{-p}\cdot C^p(m^{-p/2} + s^{-p}) < \delta.$$

We now show $\|Q_{i,j}\|_p \lesssim p$, for which we use Bernstein's inequality.

Suppose $\eta_{a_1,i},\ldots,\eta_{a_s,i}$ are all 1, where $a_1 < a_2 < \ldots < a_s$. Now, note $Q_{i,j}$ can be written as $\sum_{t=1}^s Y_t$, where Y_t is an indicator random variable for the event that $\eta_{a_t,j} = 1$. The Y_t are not independent, but for any integer $p \ge 1$ their pth moment is upper bounded by the case that the Y_t are independent Bernoulli each of expectation s/m (this can be seen by simply expanding $(\sum_t Y_t)^p$ then comparing with the independent Bernoulli case monomial by monomial in the expansion). Thus Bernstein applies, and as desired we have

$$\|Q_{i,j}\|_p = \|\sum_t Y_t\|_p \lesssim \sqrt{s^2/m}\cdot\sqrt{p} + p \simeq p.$$

□

References

[Ach01] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[BOR10] Vladimir Braverman, Rafail Ostrovsky, and Yuval Rabani. Rademacher chaos, random eulerian graphs and the sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1011.2590*, 2010.

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[DlPG12] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence.* Springer Science & Business Media, 2012.

[KN10] Daniel M Kane and Jelani Nelson. A derandomized sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1006.3585*, 2010.

[KN14] Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

[Mat08] Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.