Sketching Algorithms for Big Data Fall 2017

Lecture 9 — Sept 29, 2017

Prof. Jelani Nelson *Scribe: Mitali Bafna*

1 Fast JL transform

Typically we have some high-dimensional computational geometry problem, and we use JL to speed up our algorithm in two steps: (1) apply a JL map Π to reduce the problem to low dimension m, then (2) solve the lower-dimensional problem. As m is made smaller, typically (2) becomes faster. However, ideally we would also like step (1) to be as fast as possible. In this section, we investigate two approaches to speed up the computation of Πx.

One of the analyses will make use of the following Chernoff bound.

Theorem 1 (Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be independent random variables in $[0, \tau]$, and write $\mu := \mathbb{E} \sum_i X_i$. Then*

$$\forall \varepsilon > 0,\ \mathbb{P}(|\sum_i X_i - \mu| > \varepsilon\mu) < 2\left(\frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mu/\tau}$$

The approach we cover here was investigated by Ailon and Chazelle [AC09]. This approach gives a running time to compute Πx of roughly $O(d \log d)$. They called their transformation the *Fast Johnson-Lindenstrauss Transform (FJLT)*. A construction similar to theirs, which we will analyze here, is the $m \times n$ matrix Π defined as

$$\Pi = \sqrt{\frac{d}{m}} SHD \tag{1}$$

where S is an $m \times d$ sampling matrix with replacement (each row has a 1 in a uniformly random location and zeroes elsewhere, and the rows are independent), H is a *bounded orthonormal system*, and $D = diag(\alpha)$ for a vector α of d independent Rademachers. A bounded orthonormal system is a matrix $H \in \mathbb{C}^{d \times d}$ such that $H^*H = I$ and $\max_{i,j} |H_{i,j}| \leq 1/\sqrt{d}$. For example, H can be the Fourier matrix or Hadamard matrix.

The motivation for the construction (1) is speed: D can be applied in $O(d)$ time, H in $O(d \log d)$ time (e.g. using the Fast Fourier Transform or divide and conquer in the case of the Hadamard matrix), and S in $O(m)$ time. Thus, overall, applying Π to any fixed vector x takes $O(d \log d)$ time. Compare this with using a dense matrix of Rademachers, which takes $O(md)$ time to apply.

We will now give some intuition behind why such a Π works. Consider the sampling matrix S which samples a random coordinate of x. If the norm of x is spread out among its coordinates then in expectation the norm of Sx is the norm of x. But what do we do in the case where x has mass only on a few coordinates. It is known that a Fourier matrix spreads out the mass of vectors with highly concentrated mass and vice versa. So we multiply S with H and to handle the case where H concentrates the mass of vectors with their mass spread out we finally multiply x in the beginning by D_α.

1.1 Analysis of [AC09]

We will show that for $m \gtrsim \varepsilon^{-2}\log(1/\delta)\log(d/\delta)$, the random Π described in (1) provides DJL. We will consider the case of H as the normalized Hadamard matrix, so that every entry of H is in $\{-1/\sqrt{d}, 1/\sqrt{d}\}$.

Theorem 2. *Let $x \in \mathbb{R}^n$ be an arbitrary unit norm vector, and suppose $0 < \varepsilon, \delta < 1/2$. Also let $\Pi = \sqrt{\frac{d}{m}}SHD$ as described above with a number of rows equal to $m \gtrsim \varepsilon^{-2}\log(1/\delta)\log(n/\delta)$. Then*

$$\mathop{\mathbb{P}}_{\Pi}(|\|\Pi x\|_2^2 - 1| > \varepsilon) < \delta.$$

Proof. Define $y = HDx$. The goal is to first show that $\|HDx\|_\infty = O(\sqrt{\log(d/\delta)/n})$ with probability $1 - \delta/2$, then conditioned on this event, that $(1-\varepsilon) \le \|\frac{d}{m}Sy\|_2^2 \le (1+\varepsilon)$ with probability $1 - \delta/2$.

For the first event, note

$$y_i = (HDx)_i = \sum_{j=1}^{n} \sigma_j \cdot (\frac{1}{\sqrt{d}}\gamma_{i,j}x_j) = \langle \sigma, z^i \rangle,$$

where $|\gamma_{i,j}| = 1$ and z^i is the vector with $(z^i)_j = \frac{1}{\sqrt{d}}\gamma_{i,j}x_j$. Thus by Khintchine's inequality

$$\forall i,\ \mathbb{P}(|y_i| > \sqrt{\frac{2\log(4d/\delta)}{n}}) < 2e^{-\log(d/\delta)} = \frac{\delta}{2d}.$$

Thus by a union bound,

$$\mathbb{P}(\|y\|_\infty > \sqrt{\frac{2\log(4d/\delta)}{n}}) = \mathbb{P}(\exists i :\ |y|_i > \sqrt{\frac{2\log(4d/\delta)}{n}}) < \frac{\delta}{2}.$$

Now, let us condition on this event that $\|y\|_\infty^2 \le 2\log(4d/\delta)/n := \tau/d$. For $i \in [m]$, define $X_i = d \cdot y_i^2$. By the Chernoff bound above,

$$\mathbb{P}(|\sum_{i=1}^{m} X_i - m| > \varepsilon m) < 2\left(\frac{e^\varepsilon}{(1+\varepsilon)^{1+\varepsilon}}\right)^{m/\tau},$$

which is at most $\delta/2$ for $m \gtrsim \varepsilon^{-2}\log(1/\delta)\log(d/\delta)$. □

Remark 1. Note that the FJLT as analyzed above provides suboptimal m. If one desired optimal m, one can instead use the embedding matrix $\Pi'\Pi$,where Π is the FJLT and Π' is, say, a dense matrix with Rademacher entries having the optimal $m' = O(\varepsilon^{-2}\log(1/\delta))$ rows. The downside is that the runtime to apply our embedding worsens by an additive $m \cdot m'$. [AC09] slightly improved this additive term (by an ε^2 multiplicative factor) by replacing the matrix S with a random sparse matrix P.

Can a better analysis be given? Unfortunately not by much: the quadratic dependence $\log^2(1/\delta)$ needs to be there by an example of Eric Price. The bad case is x has $1/d^{1/4}$ on the first $\sqrt{d}$ coordinates, and imagine $\delta \ll 2^{-\sqrt{d}}$.

1.2 Analysis based on RIP

Here we give a different analysis, based on combining the main results of [KW11] and [RV08] which use the method of chaining, as seen in the last lecture.

First we have to give a definition.

Definition 1. *We say a matrix $\Pi \in \mathbb{R}^{m \times n}$ satisfies the* (ε, k)-restricted isometry property *(or* RIP *for short) if for all k-sparse vectors x of unit Euclidean norm,*

$$1 - \varepsilon \leq \|\Pi x\|_2^2 \leq 1 + \varepsilon.$$

Using the fact that the operator norm of a matrix $\|M\|$ is equal to $\sup_x \|x^T M x\|$, it follows that being (ε, k)-RIP is equivalent to

$$\sup_{T \subset [n] | T| = k} \|I_k - (\Pi^{(T)})^* \Pi^{(T)}\| < \varepsilon,$$

where $\Pi^{(T)}$ is the $m \times |T|$ matrix obtained by restricting Π to the columns in T.

As we will see later in the course, this notion of RIP is useful for *compressed sensing*, which is closely related to the heavy hitters problem. For now, we will just use it to obtain fast JL by combining it with the following theorem of [KW11].

Theorem 3. *There exists a universal constant $C > 0$ such that the following holds. Suppose A satisfies $(\varepsilon/C, k)$-RIP for $k \geq C \log(1/\delta)$, and let $\alpha \in \{-1, 1\}^n$ be chosen uniformly at random. Then for any $x \in \mathbb{R}^n$ of unit norm*

$$\mathbb{P}_{\alpha}(|\|AD_\alpha x\|_2^2 - 1| > \varepsilon) < \delta.$$

In other words, the probability distribution $\Pi = AD_\alpha$ over matrices, induced by α, satisfies the distributional JL property.

We will not prove Theorem 3 here, but we will show that the matrix $\sqrt{\frac{d}{m}} SH$ satisfies RIP with positive probability for fairly small m. That is, there does *some* choice of few rows of a bounded orthonormal system that gives RIP (though unfortunately we do not know which explicit set, though see [BDF+11]).

A number of bounds on the best m to achieve RIP for sampling Fourier/Hadamard rows were given, starting with the work of Candés and Tao [CT06]. Then subsequent works gave better bounds [RV08, Bou14, HR16]. An analysis was also given for a related construction in [NPW14]. We will give the analysis of [RV08] since it is most similar to what we saw in the last lecture.

Recall for $T \subset \mathbb{R}^n$,

$$r(T) := \mathop{\mathbb{E}}_{\sigma} \sup_{x \in T} |\langle \sigma, x \rangle|.$$

Last lecture we did not include the absolute values, but it does not make much of a difference (the Khintchine tail bound only differs by a factor of two). Also recall that we showed

$$r(T) \lesssim \Delta(T, \|\cdot\|_2),$$

where for T a set of vectors of at most unit $\|\cdot\|$ norm,

$$\Delta(T, \|\cdot\|) \simeq \sum_{k=1}^{\infty} \frac{1}{2^k} \cdot \lg^{1/2} \mathcal{N}(T, \|\cdot\|, \frac{1}{2^k}) \simeq \int_0^{\infty} \lg^{1/2} \mathcal{N}(T, \|\cdot\|, u) du \simeq \inf_{\{T_r\}} \sum_{r=1}^{\infty} 2^{r/2} \cdot \sup_{x \in T} \|x - T_r\|.$$

This was the Dudley bound. Let us now show that for RIP, $m = \Omega(\varepsilon^{-2} k \log^4 n)$ suffices.

We will analyze a slightly different construction, just for ease of notation. Instead of sampling m rows from H, we will simply keep each row with probability m/d, independently. Let η_i be an indicator for whether we keep row i. Also, let us define x^i to equal the ith row of $\sqrt{d} \cdot H$, so $x^i \in \{-1, 1\}^n$.

We let $\beta = \mathbb{E}_\mu \sup_{|T|=k} \|I_k - \frac{1}{m} \sum_i \mu_i z_i^{(T)} (z_i^{(T)})^T\|$ and we will now get an upper bound for β in terms of $\sqrt{\beta}$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mu} \sup_{|T|=k} \|I_k - \frac{1}{m} \sum_i \mu_i z_i^{(T)} (z_i^{(T)})^T\| \\
=&\mathop{\mathbb{E}}_{\mu} \sup_{|T|=k} \| \mathop{\mathbb{E}}_{\mu'} (\frac{1}{m} \sum_i \mu_i' z_i^{(T)} (z_i^{(T)})^T - \frac{1}{m} \sum \mu_i z_i^{(T)} (z_i^{(T)})^T)\| \\
\leq &\mathop{\mathbb{E}}_{\mu, \mu'} \frac{1}{m} \sup_{|T|=k} \| \sum_i \mu_i' z_i^{(T)} (z_i^{(T)})^T - \sum \mu_i z_i^{(T)} (z_i^{(T)})^T)\| && \text{Jensen's inequality} \\
=&\frac{1}{m} \mathop{\mathbb{E}}_{\mu, \mu', \sigma} \sup_{|T|=k} \| \sum_i \sigma_i (\mu_i' - \mu_i) z_i^{(T)} (z_i^{(T)})^T\| && \text{By symmetrization over } \sigma \\
\leq &\frac{2}{m} \mathop{\mathbb{E}}_{\mu} \mathop{\mathbb{E}}_{\sigma} \sup_{|T|=k} \| \sum_i \sigma_i \mu_i z_i^{(T)} (z_i^{(T)})^T\| && \text{Triangle inequality} \\
=&\frac{2}{m} \mathop{\mathbb{E}}_{\mu} \mathop{\mathbb{E}}_{\sigma} \sup_{|T|=k} \sup_{x \in \mathbb{R}^n} \| \sum_{i \in [d]} \sigma_i \mu_i \langle x, z_i^{(T)} \rangle^2\| && \text{Using the defn of operator norm of a matrix} \\
=&\frac{2}{m} \mathop{\mathbb{E}}_{\mu} \mathop{\mathbb{E}}_{\sigma} \sup_{|T|=k} \sup_{x \in D_2^{d,k}} \| \sum_{i \in [d]} \sigma_i \mu_i \langle x, z_i^{(T)} \rangle^2\| && \text{where } D_2^{d,k} = \text{ set of all k-sparse unit vectors in } \mathbb{R}^d
\end{aligned}$$

We let, $T_\mu = \{\mu_1 \langle x, z_1 \rangle^2, \ldots, \mu_d \langle x, z_d \rangle^2, x \in D_2^{d,k}\}$ and $r(T_\mu) = \mathbb{E} \sup_{z \in T_\mu} |\langle \sigma, z \rangle|$.

Dudley's inequality gives us that $r(T) \leq \Delta(T, l_2)$.

Let $g(x) = (\mu_1 \langle x, z_1 \rangle, \ldots, \mu_d \langle x, z_d \rangle)$ and $g(y)$ is defined similarly.

We have that,

$$\|g(x) - g(y)\|_2 \leq \max_{1 \leq j \leq d} |\langle z_j, x - y \rangle| \cdot 2\sqrt{m} \cdot (\beta + 1)^{1/2}.$$

So we get that,

$$\beta \leq \sqrt{\beta + 1} \frac{\Delta(D_2^{d,k}, \|\cdot\|)}{\sqrt{m}},$$

which implies that $\beta^2 - CR\beta - CR \leq 0$.

References

[AC09] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[BDF+11] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, and Denka Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke Mathematical Journal*, 159(1):145–185, 2011.

[Bou14] Jean Bourgain. An improved estimate in the restricted isometry problem. *Geometric Aspects of Functional Analysis*, 2116:65–70, 2014.

[CT06] Emmanuel J. Candés and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[HR16] Ishay Haviv and Oded Regev. The restricted isometry property of subsampled fourier matrices. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 288–297, 2016.

[KW11] Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

[NPW14] Jelani Nelson, Eric Price, and Mary Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1515–1528, January 2014.

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.