University of Wrocław: Algorithms for Big Data (Spring'22) 30/03/2022

Lecture 5: Dimensionality reduction

Lecturer: *Przemysław Uznański*

1 Dimensionality reduction recap

Two versions of the JL lemma.

Theorem 1 (Distributional JL)**.** *For any integer n, and $0 < \varepsilon, \delta < \frac{1}{2}$, and $m = \mathcal{O}(\log(1/\delta)/\varepsilon^2)$ there is distribution $\mathcal{D}$ over matrices $\mathbb{R}^{m \times n}$ such that for every $x \in \mathbb{R}^n$ where $\|x\|_2 = 1$:*

$$\Pr_{\Pi \sim \mathcal{D}}(|\Pi \cdot x|_2^2 - 1 > \varepsilon) < \delta$$

The second version:

Theorem 2 (Metric JL)**.** *For X, a set of n points in dimension n, there exists linear $f : X \to \mathbb{R}^m$ for $m = \mathcal{O}(\log(n)/\varepsilon^2)$ that preserves distances approximately, that is*

$$\forall_{i,j} |f(x_i) - f(x_j)|_2 \approx (1 \pm \varepsilon)|x_i - x_j|_2.$$

Theorem 2 follows from Theorem 1 by setting $\delta < 1/n^2$ and taking the union bound.
Those theorems are existential, but we know that setting each matrix field to:

- $\mathcal{N}(0,1) \cdot 1/\sqrt{m}$ iid coefficients works
- scaled rademacher $\{+1/-1\} \cdot 1/\sqrt{m}$ iid coefficients works

Issue: applying a single projection takes $\mathcal{O}(nm)$ time.

2 Fast JL [AC09]

Main idea is to structurize the matrix, so the time to apply matrix is $\mathcal{O}(m + n \log n)$.
We set $\Pi = \frac{1}{\sqrt{m}} \cdot S \cdot H \cdot D$ where

- S is $m \times n$ sampling matrix (each row has single 1 in a random position and 0's everywhere else, rows are independent)
- H is a $n \times n$ Fourier matrix or Hadamard matrix (we need $HH = I$ and $\max_{i,j} |H_{i,j}| \leq 1/\sqrt{n}$)
- D is a $n \times n$ diagonal matrix $\mathrm{diag}(\sigma)$ where σ is a vector of independent Rademachers

D is applied in time $\mathcal{O}(n)$, H is applied in time $\mathcal{O}(n \log n)$ and S is applied in time $\mathcal{O}(m)$.

Theorem 3. *For $m = \mathcal{O}(\log(1/\delta) \cdot \log(n/\delta) \cdot \varepsilon^{-2})$ and $\|x\|_2 = 1$ we have*

$$\Pr_{\Pi}(|\Pi \cdot x|_2^2 - 1 > \varepsilon) < \delta$$

We need:

Theorem 4 (Khintchine inequality)**.** *For any $p \geq 1$, $x \in \mathbb{R}^n$ and (σ_i) independent Rademacher,*

$$\left(\mathbb{E} \left| \sum_i x_i \sigma_i \right|^p \right)^{1/p} \leq \mathcal{O}(\sqrt{p}) \|x\|_2$$

Theorem 5 (Chernoff bound)**.** *$X_1, \ldots, X_n$ are independent random variables and $X_i \in [0, \tau]$. Let $\mu = \mathbb{E} \sum_i X_i$. Then*

$$\Pr[|\sum_i X_i - \mu| > \varepsilon\mu] < 2 \exp(-\frac{\varepsilon^2 \mu}{2\tau})$$

Proof of the main result. Denote $y = \frac{1}{\sqrt{n}} HDx$ and $z = \sqrt{\frac{n}{m}} \cdot S \cdot y$.

First our goal is to show bound on $\|y\|_\infty$.

$$y_i = (\frac{1}{\sqrt{n}} HDx)_i = \sum_{j=1}^{n} \sigma_j \frac{1}{\sqrt{n}} \gamma_{i,j} x_j = \sigma \odot w^{(i)}$$

where $w^{(i)}$ is a vector $w^{(i)}_j = \frac{1}{\sqrt{n}} \gamma_{i,j} x_j$.

First, $\|y\|_2 = \|x\|_2 = 1$ since (normalized) Hadamard transform preserves L_2 norms, but one can also prove that $\mathbb{E}|y_i|^2 = \mathbb{E}|\sigma \odot w^{(i)}|^2 = \|w^{(i)}\|_2^2 = \sum_j (\frac{1}{\sqrt{n}} \gamma_{i,j} x_j)^2 = \frac{1}{n} \|x\|_2 = \frac{1}{n}$ where we only used that H has $-1/+1$ coefficients and that σ is Rademacher.

By Khintchine inequality, and using bound on length of x and fact that $\gamma_{i,j} \in \{-1, +1\}$, for some absolute constant C:

$$\mathbb{E}|y_i|^p \leq \left(C\sqrt{p} \|w^{(i)}\|_2 \right)^p = \left(\sqrt{\frac{\mathcal{O}(p)}{n}} \right)^p$$

by Markov's inequality:

$$\Pr \left[|y_i| \geq \sqrt{\frac{\mathcal{O}(p)}{n}} \cdot \left(\frac{2n}{\delta} \right)^{1/p} \right] = \Pr \left[|y_i|^p \geq \left(C \cdot \sqrt{\frac{p}{n}} \right)^p \cdot \frac{2n}{\delta} \right] \leq \frac{\delta}{2n}$$

Optimizing, the term

$$\varphi(p) = \sqrt{\frac{\mathcal{O}(p)}{n}} \cdot \left(\frac{2n}{\delta} \right)^{1/p} \sim \exp \left(\mathcal{O}(1) + \frac{1}{2} \ln p + \frac{1}{p} \ln \left(\frac{2n}{\delta} \right) \right)$$

minimizes when $\frac{1}{2} \frac{1}{p} - \frac{1}{p^2} \ln(2n/\delta) = 0$ (the derivative of the exponent) or equivalently $p = 2 \ln \left(\frac{2n}{\delta} \right)$, so then $\varphi_{\min}(p) = \sqrt{\frac{\mathcal{O}(p)}{n}} \cdot \sqrt{e}$ and so

$$\Pr \left[|y_i| \geq \mathcal{O} \left(\sqrt{\frac{\ln(2n/\delta)}{n}} \right) \right] \leq \frac{\delta}{2n},$$

and taking the union bound

$$\Pr \left[\|y\|_\infty \geq \mathcal{O} \left(\sqrt{\frac{\ln(2n/\delta)}{n}} \right) \right] \leq \frac{\delta}{2}.$$

So in the following we condition on the event that $\|y\|_\infty = \mathcal{O}\left(\sqrt{\frac{\ln(2n/\delta)}{n}}\right)$.

Now consider $X_i = (z_i)^2$ as a random variable. We have the following:

$$\mathbb{E}X_i = \mathbb{E}(z_i)^2 = \frac{n}{m} \cdot \frac{\|y\|_2^2}{n} = \frac{1}{m}.$$

And from bound on $\|y\|_\infty$ there is $X_i \leq \frac{n}{m} \cdot \mathcal{O}(\frac{\ln(2n/\delta)}{n}) = \mathcal{O}(\frac{\ln(2n/\delta)}{m}) = \tau$.

We now apply Chernoff bound, with $\mu = 1$ (keep in mind that $\sum_i X_i = \|z\|_2^2$)

$$\Pr[|\sum_i X_i - 1| > \varepsilon] < 2\exp(-\frac{\varepsilon^2}{2\tau}) = 2\exp\left(-\frac{\varepsilon^2 m}{\mathcal{O}(\ln(2n/\delta))}\right)$$

Now, usually when applying Chernoff bound, $\tau = 1$ and then it is enough to set $m = \mathcal{O}(\varepsilon^{-2}\log(1/\delta))$ to have the probability in the bound be δ. Unfortunately in our case, $\tau \gg 1$ so we have to offset for the log term in the denominator. So we need to set $m = \mathcal{O}(\varepsilon^{-2}\log(1/\delta)\log(n/\delta)$ so we can bound the probability in the Chernoff by $\delta/2$.

Taking union bound over both $\delta/2$ failure probability finishes the proof. □

Using this approach, this dependency is roughly optimal (that is, we are losing one log vs. optimal JL). However, one trick to reduce ε to optimal at the cost of slower runtime is to apply $m' \times m$ "naive" JL at the end of the chain of matrices, for $m' = \mathcal{O}(\log(1/\delta)\varepsilon^{-2})$. This adds $\mathcal{O}(m' \cdot m) = \mathcal{O}(\varepsilon^{-4}\log^2(1/\delta)\log(n/\delta))$ to the running time though.

3 Sparse JL [DKS10]

Motivation: if x is sparse (that is, $\|x\|_0$ is small), we expect time proportional to $\|x\|_0$.

We consider distributions $\mathcal{D}$ of matrices such that:

- Each column has only s non-zero elements, for some s. (Either deterministically or in expectation.)
- They still provide good dimensionality reduction.

The time to compute Πx is then $s \cdot \|x\|_0$, since each column determines "where" each x_i contributes and can be processed in $\mathcal{O}(s)$ time.

3.1 Dasgupta etal. construction

$s = \mathcal{O}(\varepsilon^{-1}\log(1/\delta)\log^2(m/\delta))$,
$h : [sn] \to [m]$ be a random hash function , and let $H \in \{-1, 0, 1\}^{m \times sn}$ be such that $H_{ij} = \delta_{i,h(j)} r_j$. (all r are indep. Rademacher). $P \in \{0, 1\}^{sn \times n}$ be such that

$$P_{i,j} = \begin{cases} 1 \text{ for } (j-1)s + 1 \leq i \leq js \\ 0 \end{cases}$$

Intuition: P creates s copies of each element of input, H hashes each element (after duplication) into $[m]$ together with $+1/-1$ coef. This can be evaluated implicitly without expanding the matrices.

Theorem 6. *$\Pi = \frac{1}{\sqrt{s}} HP$ has JL guarantees. Πx can be evaluated in time $\mathcal{O}(s\|x\|_0)$.*

We skip the proof.

3.2 Kane, Nelson construction

$s = \mathcal{O}(\varepsilon m) = \mathcal{O}(\varepsilon^{-1} \log(1/\delta))$

Construction 1: matrix $m \times n$, where in each column we place s random $-1/+1$ (sample without replacement), normalized with $\frac{1}{\sqrt{s}}$ coef.

Construction 2: group each column into s blocks, each of size m/s. Pick in each block one $-1/+1$, normalize with $\frac{1}{\sqrt{s}}$ coef.

Construction 2 is effectively the same as CountSketch. The proof that it works shows that analysis using more than just 2-independence can show a very good concentration (CountSketch uses median, here we can use average).

4 Missing proofs

Proof of Khintchine Inequality. For any variable X, $\mathbb{E}[|X|^p]^{1/p}$ is increasing with p (see: generalized average inequality), so we can round-up p to even integer. Consider $g_i \sim \mathcal{N}(0,1)$. Expand $\mathbb{E}[(\sum_i \sigma_i x_i)^p]$ into sum of monomials. Any monomial with odd-exponents vanishes. Similarly in $\mathbb{E}[(\sum_i g_i x_i)^p]$. For any even-exponents $\alpha_1, \ldots$, $\mathbb{E}\prod_i \sigma_i^{\alpha_i} = 1$ while $\mathbb{E}\prod_i g_i^{\alpha_i} \geq 1$, so the gaussian case dominates the rademacher case.[1]

But $\sum_i g_i x_i$ is itself normal variable $\mathcal{N}(0, \|x\|_2^2)$, so

$$\mathbb{E}[(\sum_i \sigma_i x_i)^p] \leq \mathbb{E}[(\sum_i g_i x_i)^p] = (p-1)!!\|x\|_2^p$$

Asymptotically, $((p-1)!!)^{1/p} \approx (2^{p/2} \cdot (p/2)!)^{1/p} \approx \sqrt{2} \cdot \left(\frac{(p/2)^{p/2}}{e^{p/2}}\right)^{1/p} = \sqrt{\frac{p}{e}}$, so the hidden constant in Khintchine inequality is $\mathcal{O}(\sqrt{p})$.

□

References

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

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson-lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 341–350, 2010.

[1] Useful fact is that for $g \sim \mathcal{N}(0,1)$ and even p, $\mathbb{E}[g^p] = (p-1)!!$.