

Sublinear Algorithms for Big Data

Lecture 4

Grigory Yaroslavtsev

<http://grigory.us>



Today

- Dimensionality reduction
 - AMS as dimensionality reduction
 - Johnson-Lindenstrauss transform
- Sublinear time algorithms
 - Definitions: approximation, property testing
 - Basic examples: approximating diameter, testing properties of images
 - Testing sortedness
 - Testing connectedness

L_p -norm Estimation

- Stream: m updates $(x_i, \Delta_i) \in [n] \times \mathbb{R}$ that define vector f where $f_j = \sum_{i:x_i=j} \Delta_i$.
- **Example:** For $n = 4$

$$\langle (1,3), (3, 0.5), (1,2), (2, -2), (2,1), (1, -1), (4,1) \rangle$$
$$f = (4, -1, 0.5, 1)$$

- L_p -norm: $\|f\|_p = (\sum_i |f_i|^p)^{\frac{1}{p}}$

L_p -norm Estimation

- L_p -norm: $\|f\|_p = (\sum_i |f|^p)^{\frac{1}{p}}$
- Two lectures ago:
 - $\|f\|_0 = F_0$ -moment
 - $\|f\|_2^2 = F_2$ -moment (via AMS sketching)
- Space: $O\left(\frac{\log n}{\epsilon^2} \log \frac{1}{\delta}\right)$
- Technique: linear sketches
 - $\|f\|_0$: $\sum_{i \in S} f_i$ for random set S
 - $\|f\|_2^2$: $\sum_i \sigma_i f_i$ for random signs σ_i

AMS as dimensionality reduction

- Maintain a “linear sketch” vector

$$\mathbf{Z} = (Z_1, \dots, Z_k) = Rf$$

$$Z_i = \sum_{j \in [n]} \sigma_j f_j, \text{ where } \sigma_j \in_R \{-1, 1\}$$

- Estimator \mathbf{Y} for $\|f\|_2^2$:

$$\frac{1}{k} \sum_{i=1}^k Z_i^2 = \frac{\|Rf\|_2^2}{k}$$

- “Dimensionality reduction”: $x \rightarrow Rx$, “heavy” tail

$$\Pr \left[\left| \mathbf{Y} - \|f\|_2^2 \right| \geq c \left(\frac{2}{k} \right)^{\frac{1}{2}} \|f\|_2^2 \right] \leq \frac{1}{c^2}$$

Normal Distribution

- Normal distribution $N(0,1)$
 - Range: $(-\infty, +\infty)$
 - Density: $\mu(x) = \left(\sqrt{2\pi}\right)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$
 - Mean = 0, Variance = 1
- Basic facts:
 - If X and Y are independent r.v. with normal distribution then $X + Y$ has normal distribution
 - $Var[cX] = c^2 Var[X]$
 - If X, Y are independent, then $Var[X + Y] = Var[X] + Var[Y]$

Johnson-Lindenstrauss Transform

- Instead of ± 1 let σ_i be i.i.d. random variables from normal distribution $N(0,1)$

$$Z = \sum_i \sigma_i f_i$$

- We still have $\mathbb{E}[Z^2] = \sum_i f_i^2 = \|f\|_2^2$ because:
 - $\mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j] = 0$; $\mathbb{E}[\sigma_i^2] = \text{“variance of } \sigma_i \text{”} = 1$
- Define $\mathbf{Z} = (Z_1, \dots, Z_k)$ and define:

$$\|\mathbf{Z}\|_2^2 = \sum_j Z_j^2 \quad (\mathbb{E}[Y] = k\|f\|_2^2)$$

- **JL Lemma:** There exists $C > 0$ s.t. for small enough $\epsilon > 0$:

$$\Pr \left[\left| \|\mathbf{Z}\|_2^2 - k \|f\|_2^2 \right| > \epsilon k \|f\|_2^2 \right] \leq \exp(-C\epsilon^2 k)$$

Proof of JL Lemma

- **JL Lemma:** $\exists C > 0$ s.t. for small enough $\epsilon > 0$:
$$\Pr \left[\left| \|\mathbf{Z}\|_2^2 - k \|f\|_2^2 \right| > \epsilon k \|f\|_2^2 \right] \leq \exp(-C \epsilon^2 k)$$
- Assume $\|f\|_2^2 = 1$.
- We have $\mathbf{Z}_i = \sum_j \sigma_{ij} f_j$ and $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_k)$
$$\mathbb{E} \left[\|\mathbf{Z}\|_2^2 \right] = k \|f\|_2^2 = k$$
- **Alternative form of JL Lemma:**
$$\Pr \left[\|\mathbf{Z}\|_2^2 > k(1 + \epsilon)^2 \right] \leq \exp(-\epsilon^2 k + O(k \epsilon^3))$$

Proof of JL Lemma

- Alternative form of JL Lemma:

$$\Pr \left[\|\mathbf{Z}\|_2^2 > k(1 + \epsilon)^2 \right] \leq \exp(-\epsilon^2 k + O(k \epsilon^3))$$

- Let $Y = \|\mathbf{Z}\|_2^2$ and $\alpha = k(1 + \epsilon)^2$
- For every $s > 0$ we have:

$$\Pr[Y > \alpha] = \Pr[e^{sY} > e^{s\alpha}]$$

- By Markov and independence of Z'_i s:

$$\Pr[e^{sY} > e^{s\alpha}] \leq \frac{\mathbb{E}[e^{sY}]}{e^{s\alpha}} = e^{-s\alpha} \mathbb{E} \left[e^{s \sum_i Z_i^2} \right] = e^{-s\alpha} \prod_{i=1}^k \mathbb{E} \left[e^{s Z_i^2} \right]$$

- We have $Z_i \in N(0,1)$, hence:

$$\mathbb{E} \left[e^{s Z_i^2} \right] = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{s t^2} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{1 - 2s}}$$

Proof of JL Lemma

- Alternative form of JL Lemma:

$$\Pr \left[\|\mathbf{Z}\|_2^2 > k(1 + \epsilon)^2 \right] \leq \exp(-\epsilon^2 k + O(k \epsilon^3))$$

- For every $\mathbf{s} > 0$ we have:

$$\Pr[\mathbf{Y} > \alpha] \leq e^{-\mathbf{s}\alpha} \prod_{i=1}^k \mathbb{E} \left[e^{\mathbf{s}Z_i^2} \right] = e^{-\mathbf{s}\alpha} (1 - 2\mathbf{s})^{-\frac{k}{2}}$$

- Let $\mathbf{s} = \frac{1}{2} \left(1 - \frac{k}{\alpha} \right)$ and recall that $\alpha = k(1 + \epsilon)^2$
- A calculation finishes the proof:

$$\Pr[\mathbf{Y} > \alpha] \leq \exp(-\epsilon^2 k + O(k \epsilon^3))$$

Johnson-Lindenstrauss Transform

- Single vector: $k = O\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$
 - Tight: $k = \Omega\left(\frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$ [Woodruff'10]
- n vectors simultaneously: $k = O\left(\frac{\log n \log \frac{1}{\delta}}{\epsilon^2}\right)$
 - Tight: $k = \Omega\left(\frac{\log n \log \frac{1}{\delta}}{\epsilon^2}\right)$ [Molinaro, Woodruff, Y. '13]
- Distances between n vectors = $O(n^2)$ vectors:

$$k = O\left(\frac{\log n \log \frac{1}{\delta}}{\epsilon^2}\right)$$