CIS 700: “algorithms for Big Data”

Lecture 5: Dimension Reduction

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
UNIVERSITY of PENNSYLVANIA

L_p-norm Estimation

- Stream: $\boldsymbol{m}$ 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: $||f||_p = (\sum_i |f|^p)^{\frac{1}{p}}$

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 sets S
  - $||f||_2^2$: $\sum_i \sigma_i f_i$ for random signs σ_i

AMS as dimensionality reduction

- Maintain a "linear sketch" vector

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

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

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

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

- "Dimensionality reduction": $x \to Rx$, "heavy" tail

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

Normal Distribution

- Normal distribution $N(0,1)$
  - Range: $(-\infty, +\infty)$
  - Density: $\mu(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$
  - 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:
  - $\mathbb{E}[\sigma_i]\mathbb{E}[\sigma_j] = 0;\ \ \mathbb{E}[\sigma_i^2] =$ "variance of σ_i " $= 1$
- Define $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ and define:

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

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

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

Proof of JL Lemma

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

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

- Assume $\left||f|\right|_2^2 = 1$.
- We have $\boldsymbol{Z}_i = \sum_j \sigma_{ij} f_i$ and $\boldsymbol{Z} = (\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_k})$

$$\mathbb{E}\left[\left||\boldsymbol{Z}|\right|_2^2\right] = k\left||f|\right|_2^2 = k$$

- Alternative form of JL Lemma:

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

Proof of JL Lemma

- Alternative form of JL Lemma:

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

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

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

- By Markov and independence of $\boldsymbol{Z}_i's$:

$$\Pr[e^{\boldsymbol{s}\boldsymbol{Y}} > e^{\boldsymbol{s}\alpha}] \le \frac{\mathbb{E}[e^{\boldsymbol{s}\boldsymbol{Y}}]}{e^{\boldsymbol{s}\alpha}} = e^{-\boldsymbol{s}\alpha}\mathbb{E}\left[e^{\boldsymbol{s}\sum_i \boldsymbol{Z}_i^2}\right] = e^{-\boldsymbol{s}\alpha}\prod_{i=1}^{k}\mathbb{E}\left[e^{\boldsymbol{s}\boldsymbol{Z}_i^2}\right]$$

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

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

Proof of JL Lemma

- Alternative form of JL Lemma:

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

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

$$\Pr[\boldsymbol{Y} > \alpha] \le e^{-\boldsymbol{s}\alpha} \prod_{i=1}^{k} \mathbb{E}\left[e^{\boldsymbol{s}\boldsymbol{Z}_i^2}\right] = e^{-\boldsymbol{s}\alpha}(1 - 2\boldsymbol{s})^{-\frac{k}{2}}$$

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

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

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

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

Random Variables and Norms

- For a random variable $\boldsymbol{X}$ and $\mathrm{p} \geq 1$ let:

$$\left||\boldsymbol{X}|\right|_p = \mathbb{E}[X^p]^{1/p}$$

Facts:

- For any c: $\left||c\boldsymbol{X}|\right|_p = c\left||\boldsymbol{X}|\right|_p$
- $||\cdot||_p$ is a norm (Minkowski's inequality)
- $||\cdot||_p \leq ||\cdot||_q$ for $p \leq q$ (Monotonicity of norms)
- Jensen's inequality (used a lot for $F = |x|^p$):

If F is convex then $F(\mathbb{E}[\boldsymbol{X}]) \leq \mathbb{E}[F(\boldsymbol{X})]$

Khintchine Inequality

- [Khintchine]For $p \geq 1, x \in \mathbb{R}^n$ and σ_i i.i.d. Rademachers:
$$\left|\left|\sum_i \sigma_i x_i\right|\right|_p \lesssim \sqrt{p}\, \left|\left|x\right|\right|_2$$
- For r_i (either σ_i or $g_i \sim N(0,1)$) expand $\mathbb{E}[(\sum_i r_i x_i)^p]$
- All odd powers of r_i are zero
- All even moments for σ_i are 1, and for g_i are ≥ 1
- $\left|\left|\sum_i \sigma_i x_i\right|\right|_p \leq \left|\left|\sum_i g_i x_i\right|\right|_p$
- $\sum_i g_i x_i \sim N\left(0, \left|\left|x\right|\right|_2^2\right) \Rightarrow \left|\left|\sum_i g_i x_i\right|\right|_p \lesssim \sqrt{p}\left|\left|x\right|\right|_2$

Symmetrization

- [Symmetrization]: If $Z_1, \ldots, Z_n$ are independent and σ_i are i.i.d. Rademachers:

$$\left\|\sum_i Z_i - \mathbb{E}\sum_i Z_i\right\|_p \leq 2\left\|\sum_i \sigma_i Z_i\right\|_p$$

- Let $Y_1 \ldots Y_n$ be independent with the same distribution as Z_i
- $\left||\sum_i Z_i - \mathbb{E}\sum_i Z_i|\right|_p = \left||\sum_i Z_i - \mathbb{E}_Y \sum_i Y_i|\right|_p$

$\leq \left||\sum_i (Z_i - Y_i)|\right|_p$ (Jensen)

$= \left||\sum_i \sigma_i (Z_i - Y_i)|\right|_p$ ($\mathrm{Z_i - Y_i}$ are independent and symmetric)

$\leq 2\left||\sum_i \sigma_i Z_i|\right|_p$(triangle inequality)

Decoupling

- Let $x_1, \ldots x_n$ be independent with mean 0 and $x_1', \ldots x_n'$ identically distributed as x_i and independent of them. For any a_{ij} and $p \geq 1$:

$$\left\| \sum_{i \neq j} a_{ij} x_i x_j \right\|_p \leq 4 \left\| \sum_{i,j} a_{ij} x_i x_j' \right\|_p$$

- Let $\eta_1, \ldots, \eta_n$ be i.i.d. Bernoullis (0/1 w.p. 1/2):

$$\left\| \sum_{i \neq j} a_{ij} x_i x_j \right\|_p = 4 \left\| \mathbb{E}_\eta \sum_{i \neq j} a_{ij} x_i x_j \eta_i (1 - \eta_j) \right\|_p$$

$$\leq 4 \left\| \sum_{i \neq j} a_{ij} x_i x_j \eta_i (1 - \eta_j) \right\|_p \text{ (Jensen)}$$

- There exists $\eta' \in \{0,1\}^n$ such that:

$$\left\| \sum_{i \neq j} a_{ij} x_i x_j \eta_i (1 - \eta_j) \right\|_p \leq \left\| \sum_{i \in S} \sum_{j \in \bar{S}} a_{ij} x_i x_j \right\|_p$$

where $S = \{i : \eta' = 1\}$.

Decoupling (continued)

Let x_S be an S-dimensional vector of x_i for $i \in S$.

- $\left|\left|\sum_{i\in S}\sum_{j\in\bar{S}} a_{ij}x_ix_j\right|\right|_p = \left|\left|\sum_{i\in S}\sum_{j\in\bar{S}} a_{ij}x_ix'_j\right|\right|_p$

$= \left|\left|\mathbb{E}_{x_{\bar{S}}}\mathbb{E}_{x'_S}\sum_{i,j} a_{ij}x_ix'_j\right|\right|_p$ $(\mathbb{E}[\mathrm{x_i}] = \mathbb{E}[\mathrm{x'_i}] = 0)$

$\leq ||\sum_{i,j} a_{ij}\, x_i\, x'_j\, ||_p$ (Jensen)

- Overall:

$$\left\|\sum_{i\neq j} a_{ij}x_ix_j\right\|_p \leq 4\left\|\sum_{i,j} a_{ij}x_ix'_j\right\|_p$$

Hanson-Wright Inequality

- For $\sigma_1, \dots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n\times n}$ real and symmetric for all $p \geq 1$:
$$\left|\left|\sigma^T A\sigma - \mathbb{E}[\sigma^T A\sigma]\right|\right|_p \lesssim \sqrt{p}\,\left|\left|A\right|\right|_F + p\,\left|\left|A\right|\right|$$
- Recall:
  - $\left|\left|A\right|\right|_F = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{Tr(A^T A)}$
  - $\left|\left|A\right|\right| = \sup_{\{v\neq 0\}} \frac{\left|\left|Av\right|\right|_2}{\left|\left|v\right|\right|_2}$

Hanson-Wright Inequality

- For $\sigma_1, \ldots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n\times n}$ real and symmetric for all $p \geq 1$:

$$\left|\left|\sigma^T A\sigma - \mathbb{E}[\sigma^T A\sigma]\right|\right|_p \preccurlyeq \sqrt{p}\,\left|\left|A\right|\right|_F + p\,\left|\left|A\right|\right|$$

$$\left|\left|\sigma^T A\sigma - \mathbb{E}[\sigma^T A\sigma]\right|\right|_p \preccurlyeq \left|\left|\sigma^T A\sigma'\right|\right|_p \text{ (decoupling)}$$

$$\preccurlyeq \sqrt{p}\,\left|\left|\left|\left|A\sigma\right|\right|_2\right|\right|_p \text{ (Khintchine)}$$

$$= \sqrt{p}\,\left|\left|\left|\left|A\sigma\right|\right|_2^2\right|\right|_{p/2}^{\frac{1}{2}}$$

$$\leq \sqrt{p}\,\left|\left|\left|\left|A\sigma\right|\right|_2^2\right|\right|_p^{\frac{1}{2}} \text{ (monotonicity of norms)}$$

Hanson-Wright (continued)

$$\sqrt{p}\ \Big\|\big\||A\sigma|\big|_2\Big\|_p \le \cdots \le \sqrt{p}\ \Big\|\big\||A\sigma|\big|_2^2\Big\|_p^{\frac{1}{2}}$$

$$\le \sqrt{p}\left(\mathbb{E}\left[\big||A\sigma|\big|_2^2\right] + \Big\|\big||A\sigma|\big|_2^2 - \mathbb{E}\left[\big||A\sigma|\big|_2^2\right]\Big\|_p\right)^{\frac{1}{2}} \text{ (triangle ineq.)}$$

$$= \sqrt{p}\left(\big||A|\big|_F^2 + \Big\|\big||A\sigma|\big|_2^2 - \mathbb{E}\left[\big||A\sigma|\big|_2^2\right]\Big\|_p\right)^{\frac{1}{2}}$$

$$\le \sqrt{p}\big||A|\big|_F + \sqrt{p}\Big\|\big||A\sigma|\big|_2^2 - \mathbb{E}\left[\big||A\sigma|\big|_2^2\right]\Big\|_p^{\frac{1}{2}}$$

$$\preccurlyeq \sqrt{p}\big||A|\big|_F + \sqrt{p}\big||\sigma^T A^T A\sigma'|\big|_p^{\frac{1}{2}} \text{ (decoupling)}$$

$$\preccurlyeq \sqrt{p}\big||A|\big|_F + p^{\frac{3}{4}}\ \big|\big|\big||A^T A\sigma|\big|_2\big|\big|_p^{1/2} \text{ (Khintchine)}$$

$$\preccurlyeq \sqrt{p}\big||A|\big|_F + p^{\frac{3}{4}}\big||A|\big|^{\frac{1}{2}}\Big\|\big||Ax|\big|_2\Big\|_p^{\frac{1}{2}}$$

Hanson-Wright (continued)

$$\sqrt{p}\ \left\| \left| \left| A\sigma \right| \right|_2 \right\|_p \lesssim \sqrt{p} \left| \left| A \right| \right|_F + p^{\frac{3}{4}} \left| \left| A \right| \right|^{\frac{1}{2}} \left\| \left| \left| A\sigma \right| \right|_2 \right\|_p^{\frac{1}{2}}$$

Let E = $\left\| \left| \left| Ax \right| \right|_2 \right\|_p^{\frac{1}{2}}$ then $E^2 - Cp^{\frac{1}{4}} \left| \left| A \right| \right|^{\frac{1}{2}} E \ - C \left| \left| A \right| \right|_F \leq 0$

- E $\leq$ larger root of the quadratic equation above
- $E^2 \leq \sqrt{p} \left| \left| A \right| \right|_F + p \left| \left| A \right| \right|$
- (Hanson-Wright) For $\sigma_1, \ldots, \sigma_n$ independent Rademachers and $A \in \mathbb{R}^{n \times n}$ real and symmetric for all $p \geq 1$:

$$\left| \left| \sigma^T A\sigma \ - \ \mathbb{E}[\sigma^T A\sigma] \right| \right|_p \lesssim \sqrt{p} \left| \left| A \right| \right|_F + p \left| \left| A \right| \right|$$