Algorithms, Geometry, and Optimization

Lecture 4: Jan 22, 2024

Last class

- Estimation via sampling: *what fraction of a population can ski?* (Sampling gives approx answers).

- Main parameters: estimate, error of estimate, confidence in estimate

  estimated value.

- Suppose population had N people

  ☆ no N anywhere in bounds.

- Using $\approx \frac{4}{\epsilon^2}$ samples, we can be 90% confident that true fraction is within $\pm\epsilon$ of the estimate

  8

  95%...

Hoeffding's inequality

Theorem. Suppose $X_1, X_2, \ldots, X_m$ are *independent* random variables with the property that $a_i \leq X_i \leq b_i$ always holds. Let $Y = X_1 + X_2 + \cdots + X_n$. Then

$$\Pr[|\overbrace{Y - \mathbb{E}[Y]}^{\text{error}}| > t] \leq 2e^{-\frac{t^2}{\sum_{i=1}^{m}(a_i - b_i)^2}}$$

- Sum of independent random variables is *well concentrated*

- Closely related to the Central Limit Theorem

Dimension reduction

$d, n \sim 10^9$

Suppose we are given points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, where d is large. Can we *embed* these points into a smaller dimensional space, so distances are approximately preserved?

$m \ll d.$

$\underset{\in \mathbb{R}^d}{x_i} \rightsquigarrow \phi(x_i) \in \mathbb{R}^m$ (classical applications of rand. algorithms.).

$$\|\phi(x_i) - \phi(x_j)\| \approx \|x_i - x_j\| \quad \forall i, j$$

Johnson-Lindenstrauss (JL) lemma

x_i

Theorem (1984). There exists an embedding ϕ into $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensional space such that:

for *every* pair x_i, x_j,

sq. distance in embedded space

sq. distance in original space.

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|\phi(x_i) - \phi(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2.$$

- in fact, linear embeddings do it!

$\phi(x) = Ax$

(Euclidean dists are special.)

JL-Lemma

(Larsen Nelson, '2017)

- The bound is tight (cannot be improved)
- As such, stated and proved for the Euclidean norm only
- Weaker versions hold for ℓ_p norms with $p > 1$
- Impossible for $p = 1$ (Brinkman, Charikar '2003).

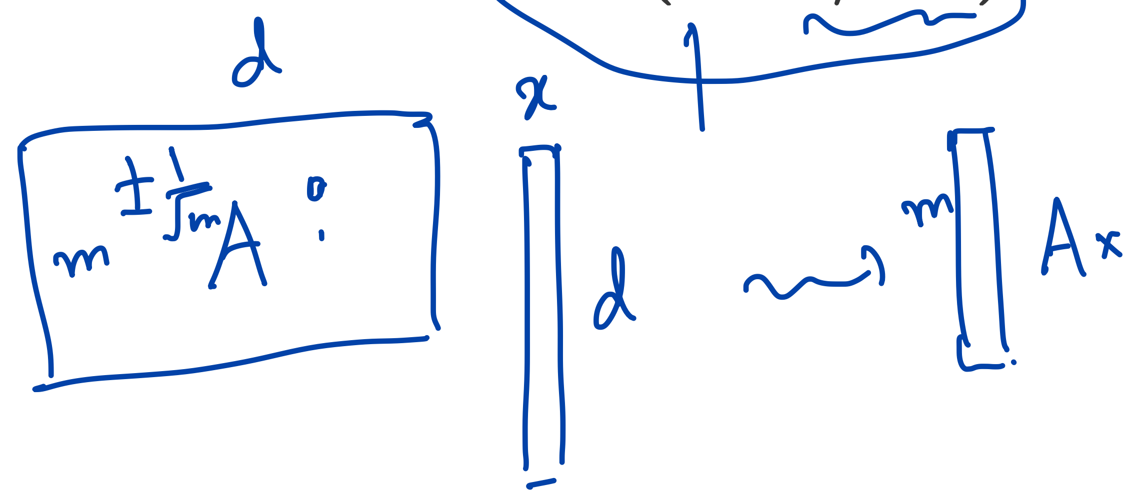
JL-Lemma: Proof Sketch

- Pick A to be an $m \times d$ matrix with each entry A_{ij} uniformly at random!

$$A_{ij} = \begin{cases} \frac{1}{\sqrt{m}} & \text{w.p. } \frac{1}{2} \\ -\frac{1}{\sqrt{m}} & \text{w.p. } \frac{1}{2} \end{cases}$$

- Two choices of distribution: $\mathcal{N}(0, 1/m)$ or $\pm \frac{1}{\sqrt{m}}$

- Set $\phi(x) = Ax$



$x \rightsquigarrow Ax$.

$\|Ax\| \approx \|x\|$

JL-Lemma: Proof Sketch

Lemma. Let x be any vector. $\in \mathbb{R}^d$ Then with probability $\geq 1 - 2e^{-m\epsilon^2/4}$, we have

$$(1-\epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1+\epsilon)\|x\|^2.$$

$m = \frac{8 \log n}{\epsilon^2}$

$1 - \frac{1}{4n^2}$

- This implies the JL lemma. Why?

Actual JL lemma requires the above for $x = x_i - x_j \ \forall\ i, j$

$\mathcal{E}_{ij}$: event that norm preservation holds for pair $(i,j) \to 1 - \frac{1}{2\binom{n}{2}}$

* Example of Union Bound Argument :

- Can use union bound!

$$\Pr(A_1 \cup A_2 \cup A_3) \le \Pr(A_1) + \Pr(A_2) + \Pr(A_3)$$

A_{ij} : event that norm preservation <u>fails</u> to hold for pair (i,j)

$\Pr(A_{ij})$ for any i,j is $\le \frac{1}{4n^2}$

$$\Pr\left(\bigcup_{i,j} A_{ij}\right) \le \binom{n}{2} \cdot \frac{1}{4n^2} \le \frac{1}{8}$$

$\therefore \Pr(\text{all norms are preserved}) \ge \frac{7}{8}$ (80%).

$i \quad j$

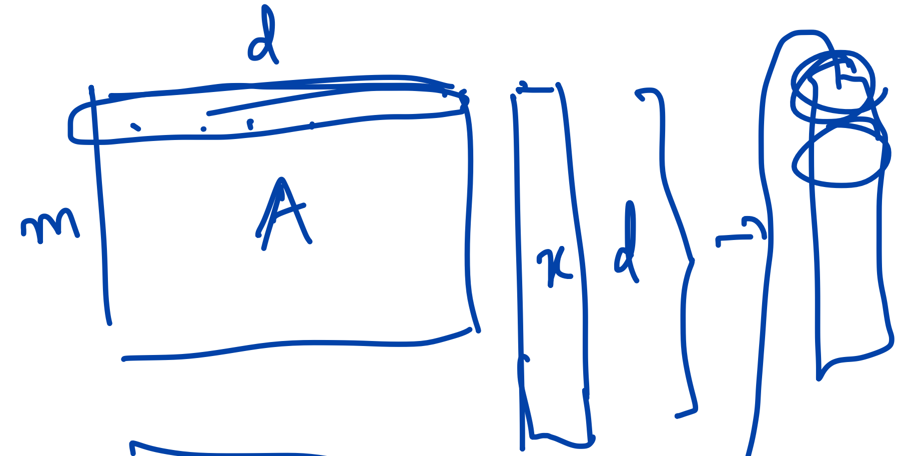
$\phi(i) \quad \phi(j)$.

Expected value

(Lemma wanted to say $\|Ax\|^2 \approx \|x\|^2$)

- What is $\mathbb{E}[\|Ax\|^2]$? fixed.

A_{ij} are random $\pm \frac{1}{\sqrt{m}}$



$\langle A_1, x \rangle$ row #1 of A

$$Ax = \begin{bmatrix} \langle A_1, x\rangle \\ \langle A_2, x\rangle \\ \vdots \\ \langle A_m, x\rangle \end{bmatrix}$$

$$\mathbb{E}\left[\|Ax\|^2\right] = \mathbb{E}\left[\langle A_1, x\rangle^2 + \langle A_2, x\rangle^2 + \dots + \langle A_m, x\rangle^2\right]$$

$$= \mathbb{E}\left[\langle A_1, x\rangle^2\right] + \mathbb{E}\left[\langle A_2, x\rangle^2\right] + \dots + \mathbb{E}\left[\langle A_m, x\rangle^2\right]$$

$$= m\,\mathbb{E}\left[\langle A_1, x\rangle^2\right] = \|x\|^2$$

Want: $\mathbb{E}\left[\langle A_1, x\rangle^2\right]$

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1d} \end{bmatrix} \qquad \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} = x$$

$$= \mathbb{E}\left[\left(A_{11}u_1 + A_{12}u_2 + \cdots + A_{1d}u_d\right)^2\right]$$

$$(x+y+z)^2 = x^2+y^2+z^2+2xy+2yz+2zx$$

$$= \mathbb{E}\left[(A_{11}u_1)^2 + \cdots + (A_{1d}u_d)^2 + 2\sum_{i<j}(A_{1i}u_i)(A_{1j}u_j)\right]$$

$$\mathbb{E}\left[A_{1i}A_{1j}u_iu_j\right] = 0$$

$$A_{1i} \to \pm\frac{1}{\sqrt{m}} \qquad A_{1j} \to \pm\frac{1}{\sqrt{m}}$$

$$= \mathbb{E}\left[\frac{1}{m}\cdot u_1^2 + \frac{1}{m}u_2^2 + \cdots + \frac{1}{m}\cdot u_d^2\right]$$

$$= \frac{1}{m}\left(u_1^2+u_2^2+\cdots+u_d^2\right) = \frac{\|x\|^2}{m}$$

Concentration bound

$0 \le X_i \le 1$

(Hoeffding...)

- Define $\boxed{Y_i = \langle A_i, x \rangle^2}$
- What is its range?

$a_i \le Y_i \le b_i$

$0 \le Y_i \le \cancel{d \frac{\|x\|^2}{m}}$ (too bad..)

pretend $\rightarrow \frac{2\|x\|^2}{m} \rightarrow$ Turns out to imply the lemma.

Is this concentrated?

$\mathbb{E}\left[\langle A_1, x\rangle^2 + \langle A_2, x\rangle^2 + \dots + \langle A_m, x\rangle^2\right] = \|x\|^2$

$\frac{\|x\|^2}{m}$ in expectation.

Example: linearity of expectation

- Max 3-SAT problem

boolean variables $x_1, \ldots, x_n$

$$\underbrace{(x_1 \vee x_2 \vee \bar{x}_3)}_{C_1}, \quad \underbrace{(x_3 \vee \bar{x}_5 \vee x_7)}_{C_2}, \ \ldots$$

Surprising fact: if you set each x_i to true/false uniformly at random, then $\frac{7}{8}$ fraction of the clauses get satisfied in expectation.

Streaming algorithms: basic model

- Data: values (phone numbers, IP addresses, etc.) arrives one after another

$v_1, v_2, \ldots \ldots, v_T$

- Full data set is too hard to store

- Router must compute aggregate statistics

low resource PC.

Distinct elements problem

- How many distinct values does the stream have?

$\{1, \dots, n\}$

Stream $v_1, v_2, \dots, v_T$

$n << T$

numbers get repeated.

Suppose answer is k.

processing time

- Hash set or equiv: $O(k)$ space, $\log k$ lookup / element.

→ Bad if k is very large... $\log k$ space!!