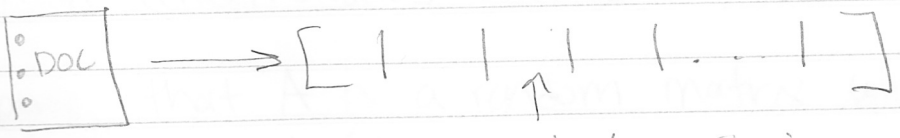
6.854 / 18.415

Last Time: distinct elements and heavy hitters

ie. approximating statistics of a stream w/o storing it

Today: dimensionality reduction

Many types of data are high-dimensional, e.g. documents as bag-of-words




Let $\|u-v\|_2^2 = \sum_{i=1}^{d} (u_i - v_i)^2$ called $\ell_2^2$-distance

All-Pairs Distances: Given n documents, compute all pairs of distances (e.g. for hierarchical clustering) — Kruskal

Time: $O(n^2 d)$

Can we find representations in lower dimensions that preserve the geometry? — approximately

why does n show up and not d?

Theorem: [Johnson, Lindenstrauss] For any $\varepsilon \log_2 \leq n$ set $S$ of $n$ points in $d$-dimensions there is a matrix $A \in \mathbb{R}^{k \times d}$ with

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|Au-Av\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2$$

JL-property

for all $u, v \in S$, with $k = \frac{20 \log n}{\varepsilon^2}$

Time via JL: $O\left(\underbrace{\frac{d n \log n}{\varepsilon^2}}_{(dk)\cdot n} + \frac{n^2 \log n}{\varepsilon^2}\right)$

matrix-vec multiply

Moreover we can choose $A$ randomly with independent Gaussian entries (not only does it exist, but almost every choice works)

Johnson-Lindenstrauss reduces to preserving norms

Suppose that $A$ is a random matrix which for any fixed $x$ satisfies

$$(*) \quad \Pr\left[(1-\varepsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\varepsilon)\|x\|_2^2\right] \ge 1 - \frac{1}{n^3}$$

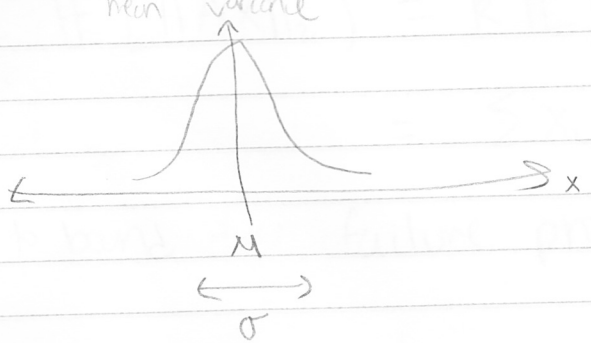
Then by union bounding over $X = u - v$, for all $u, v \in S$ we get the theorem (since there's $\binom{n}{2}$ of them)

---

The Gaussian distribution

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

($\mu$: mean, $\sigma^2$: variance)




Ubiquitous in probability and statistics

normalized appropriately

Central limit theorem: Sums of independent random variables converge to a Gaussian

The Chernoff bound shows such sums concentrate, but another explation is they converge to Gaussian

Fact: If $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

then $Z_1 + Z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

(think of $Z_1, Z_2$ as representing sums of indep. r.vs)

---

Now returning to JL, Choose entries of $A$ to be $\mathcal{N}(0, \frac{1}{k})$

Then

$$(Ax)_1^2 = \left(\underbrace{\sum_{i=1}^{d} A_{1i} x_i}_{\mathcal{N}(0, \frac{\sum x_i^2}{k})}\right)^2$$

using fact

Hence $\mathbb{E}\left[\|Ax\|_2^2\right] = k\,\mathbb{E}\left[((Ax)_1)^2\right]$

$= \sum x_i^2 = \|x\|_2^2$

Now to bound the failure probability:

one side:

$$\Pr\left[\|Ax\|_2^2 > (1+\varepsilon)\|x\|_2^2\right]$$

$$= \Pr\left[\sum_{i=1}^{k} z_i^2 > (1+\varepsilon)\|x\|_2^2\right]$$

where $z_i \sim N\left(0, \frac{\|x\|_2^2}{k}\right)$, now dividing thru by $\frac{\|x\|_2^2}{k}$, we get

$$= \Pr\left[\sum_{i=1}^{k} Y_i^2 > (1+\varepsilon)k\right]$$

where $Y_i \sim N(0,1)$, and similarly for other side

Lemma:

$$\Pr\left[\sum_{i=1}^{k} Y_i^2 > (1+\varepsilon)k\right] \leq e^{-\frac{k}{4}(\varepsilon^2-\varepsilon^3)}$$

$$\Pr\left[\sum_{i=1}^{k} Y_i^2 < (1-\varepsilon)k\right] \leq e^{-\frac{k}{4}(\varepsilon^2-\varepsilon^3)}$$

The proof is similar to usual proof of Chernoff via mgf (see link)

Now if we plug in $k = \frac{20\log n}{\varepsilon^2}$, we can invoke the lemma to show that whp $\|Ax\|_2^2 \sim \|x\|_2^2$

then union bound over $x = u - v$

## Extensions:

What really happened here is for i.i.d mean-zero, variance one $z_1, z_2, \ldots z_d$ the quantity

$$\left(\sum_{i=1}^{d} z_i x_i\right)^2$$

is an unbiased estimator for $\|x\|_2^2$, and we boosted it by repeating

**Theorem** [Achlioptas] 2003 If the entries of A are chosen as

$$R_{ij} = \sqrt{3} \begin{cases} +1 & \text{with probability } 1/6 \\ 0 & \text{" } \quad 2/3 \\ -1 & \text{" } \quad 1/6 \end{cases}$$

for $1 \le i \le k$, $1 \le j \le d$

and $A = \frac{1}{\sqrt{k}} R$, then A has the JL-property

multiplying by A is easier in many settings

This idea of finding sparse dimension reduction can be taken much further

**Theorem:** [Kane, Nelson] 2014 non i.i.d there are distributions on $A \in \mathbb{R}^{k \times d}$ with $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ and $s = O\left(\frac{\log n}{\varepsilon}\right)$ non-zeros per column that have the JL-property

Multiplying sparse vectors by sparse matrices is fast, enables many applications

## Fast Johnson-Lindenstrauss:

An alternative approach due to Ailon, Chazelle (2006) is

$$A = PHD$$

P: sparse, H: hadamard, D: diagonal

where $H_1 = [1]$ $1 \times 1$ hadamard

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$ $2 \times 2$ hadamard

$$H_{2^r} = \begin{bmatrix} H_{2^{r-1}} & H_{2^{r-1}} \\ H_{2^{r-1}} & -H_{2^{r-1}} \end{bmatrix}$$ $2^r \times 2^r$ hadamard

and $D = \begin{bmatrix} \pm 1 & & & \\ & \pm 1 & & \\ & & \ddots & \\ & & & \pm 1 \end{bmatrix}$

Idea: Multiplying a sparse vector by a sparse matrix can result in a poor embedding, so make the vectors dense

FJLT relies on fact that multiplying by H can be done by divide-and-conquer, and uncertainty principles

distribution on

**Theorem** [Ailon-Chazelle] There is a matrices $A \in \mathbb{R}^{k \times d}$ with $k = O(\frac{\log n}{\varepsilon^2})$ [1] that has the JL-property, [2] and matrix-vector multiplications can be done in time

$$O(d \log d + \frac{d \log n}{\varepsilon^2})$$

provided $n < 2^{d^{1/3}}$

Applications to nearest neighbor (we'll cover this problem later)