$\boxed{\text{30 Apr}}$ Random Projections

Reminder Matrix Chernoff

$$X_1, \ldots, X_N \in \mathbb{R}^{n \times n}$$ indep, symmetric, random matrices

$$0 \preceq X_i \preceq \mathbb{1} \quad \forall i$$

$$X = \sum_{i=1}^{N} X_i, \qquad a \cdot \mathbb{1} \preceq \mathbb{E}[X] \preceq b \cdot \mathbb{1}$$

$$\implies \Pr\left(X \not\succeq (1-\varepsilon) a \mathbb{1}\right) \leq n e^{-\varepsilon^2 a/2}$$

$$\Pr\left(X \not\preceq (1+\varepsilon) b \mathbb{1}\right) \leq n e^{-\varepsilon^2 b/3}$$

Interpretation 1. $n = 1$ Exactly the usual Chernoff bd.

2. $n > 1$ and $X_1, \ldots, X_N$ all diagonal matrices

The matrix Chernoff bound is encoding using ordinary Chernoff on each diagonal entry followed by union bound over all n diagonal entries.

3. We could instead define an entry-wise ordering $A \succeq_{ew} B$ if $a_{ij} \geq b_{ij} \quad \forall i, j$.

Matrix Chernoff w.r.t. $\succeq_{ew}$ follows using Chernoff-plus-union-bound

(with factor n^2 on RHS
because union bound over n^2 entries
rather then n.)

Gaussian Singular Value Inequality

If $A \in \mathbb{R}^{m \times n}$ $(m \leq n)$
and entries a_{ij} are indep
samples from $N(0, \frac{1}{n})$ then $\forall \varepsilon > 0$
with probability at least $1 - 2e^{-\varepsilon^2 n/2}$

$$1 + \varepsilon + \sqrt{\frac{m}{n}} \geq \sigma_1(A) \geq \dots \geq \sigma_m(A) \geq 1 - \varepsilon - \sqrt{\frac{m}{n}}.$$

Specialized to $m = 1$:

says same thing as Problem 2c
on PSet 4, scaled down by $\sqrt{n}$.

Recall $\sigma_1^2, \dots, \sigma_m^2$ are the
eigvals of AA^T.

If a_i denotes i^{th} column of A,

$$AA^T = \sum_{i=1}^{n} \boxed{a_i a_i^T}$$

independent random matrices.

Applications

Random matrices as analogues of hash functions.

Hash $[n] \longrightarrow [m]$ $n \gg m \gg 1$

keys values

Collisions are unavoidable

$$\Pr(\exists x, y \quad h(x) = h(y)) = 1$$

... but unpredictable

$$\forall x, y \quad \Pr(h(x) = h(y)) \ll 1.$$

Random matrix $A \in \mathbb{R}^{m \times n}$ is a linear transformation $\mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(n \gg m)$

Must have positive-dimensional nullspace (hence "collisions") but one hopes that any "small set of test vectors" is mapped into $\mathbb{R}^m$ "probably without distortion!"

CS 4850/5850 term

<u>Lemma</u>: ("Gaussian Hashing Lemma")

Suppose $A \in \mathbb{R}^{m \times n}$ has independent $N(0, \frac{1}{m})$ entries. If $W \subseteq \mathbb{R}^n$ is any (fixed, independent of A) subspace of dimension $d \leq \varepsilon^2 m$ then with probability $\geq 1 - 2e^{-\varepsilon^2 m/2}$

$$\forall w \in W \quad (1-2\varepsilon)\|w\|_2 \leq \|Aw\|_2 \leq (1+2\varepsilon)\|w\|_2$$

<u>Proof</u>. Let B = a matrix, $m \times d$, whose columns are an orthonormal basis of W.

$$W = \{ w = Bx : x \in \mathbb{R}^d \}.$$

$$\forall w \in W \quad (1-2\varepsilon)\|w\|_2 \leq \|Aw\|_2 \leq (1+2\varepsilon)\|w\|_2$$

$$\Updownarrow$$

$$\forall x \in \mathbb{R}^d \quad (1-2\varepsilon)\|x\|_2 \leq \|ABx\|_2 \leq (1+2\varepsilon)\|x\|_2$$

since $\|Bx\|_2 = \|x\|_2$.

The second inequality says

$$\sigma_1(AB) \leq 1+2\varepsilon$$

$$\sigma_d(AB) \geq 1-2\varepsilon.$$

Plan: Show AB has indep Gaussian entries.

Let Q be an orthogonal $n \times n$ matrix whose first d columns are B.

AB = first d columns of AQ.

AQ and A are ident. distrib. so entries of AQ are independent $N(0, \frac{1}{m})$.

AB = $m \times d$ matrix with indep $N(0, \frac{1}{m})$ entries.

Gaussian SV Ineq applied to $(AB)^T$

says with prob $\geq 1-2e^{-\varepsilon^2 m/2}$

$$\sigma_1(AB) \leq 1+\varepsilon+\sqrt{\frac{d}{m}} \leq 1+2\varepsilon$$

$$\sigma_d(AB) \geq 1-\varepsilon-\sqrt{\frac{d}{m}} \geq 1-2\varepsilon.$$

Recall $d \leq \varepsilon^2 m$, $\sqrt{\frac{d}{m}} \leq \varepsilon$.

Dimensionality Reduction of finite point sets (Johnson-Lindenstrauss)

If $x_1, \ldots, x_N \in \mathbb{R}^n$

and $A \in \mathbb{R}^{m \times n}$ random

matrix with $N(0, \frac{1}{m})$ entries.

and $$m \geq \frac{16 \ln(N/\delta)}{\varepsilon^2}$$

then w. prob. $\geq 1-\delta^2$

all pairwise distances

preserved within $1 \pm \varepsilon$ factor.

$$\forall i,j \quad (1-\varepsilon)\|x_i - x_j\| \leq \|Ax_i - Ax_j\|$$

$$\leq (1+\varepsilon)\|x_i - x_j\|$$