27 Apr 2026

"Gaussian Hashing" and Applications

Announcements

1. Problem Set 4
2. TCS Club final meeting

Weds Apr. 29 5–6:30 CS 450

pizza provided

Gaussian Singular Value Lemma.

If A is a matrix with

- m rows
- $n \geq m$ columns
- $N(0, \frac{1}{n})$ i.i.d. entries

mnemonic: total variance per row normalized to 1

then with probability $> 1 - 2e^{-\frac{1}{2}\varepsilon^2 n}$

its singular values $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$

satisfy

$$1 - \varepsilon - \sqrt{\frac{m}{n}} \leq \sigma_m \leq \sigma_1 \leq 1 + \varepsilon + \sqrt{\frac{m}{n}}.$$

Analogy between hash functions

$$[n] \longrightarrow [m] \qquad (n >> m)$$

keys ↗ ↖ buckets

and random matrices viewed as linear functions

$$\mathbb{R}^n \longrightarrow \mathbb{R}^m.$$

Hash functions when $n >> m$:

(1) potential collisions are inevitable

$$\Pr\left(\exists x \neq x' \quad h(x) = h(x') \right) = 1$$

(2) ..... but hopefully rare when hash fune is evaluated on a specific small set of keys

$$\forall x \neq x' \quad \Pr\left(h(x) = h(x') \right) \ll 1.$$

With $m \times n$ matrices when $n > m$

(1) positive dimensional nullspace is inevitable

$$\Pr\left(\exists x \neq 0 \quad Ax = 0 \right) = 1.$$

(2) ... but hopefully every specific vector is mapped far from zero w.h.p.

$$\forall x \neq 0 \quad \Pr\left(\|Ax\| > (1-\epsilon)\|x\| \right) \ll 1.$$

Lemma. ("Gaussian Hashing Lemma") CS 4850/5850 terminology

Suppose $A \in \mathbb{R}^{m \times n}$ with i.i.d. $N(0, \frac{1}{m})$ entries.

mnemonic: columns have total variance 1.

Suppose $W \subseteq \mathbb{R}^n$ is any (non-random) d-dimensional linear subspace, $d \leq \varepsilon^2 m$.

Then with probability $> 1 - 2e^{-\frac{1}{2}\varepsilon m^2}$ the matrix A satisfies

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

Proof. W is d-dimensional so it has an orthonormal basis $\{w_1, \ldots, w_d\} \subset \mathbb{R}^n$. Let $B \in \mathbb{R}^{n \times d}$ the matrix whose columns are $w_1, \ldots, w_d$.

B ($r \times d$)

$$\begin{bmatrix} | & & | \\ w_1 & \cdots & w_r \\ | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

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

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

$$\Updownarrow$$

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

$$\| \quad \quad \quad \|$$

$$(1-2\varepsilon)\|x\| \quad \quad (1+2\varepsilon)\|x\|$$

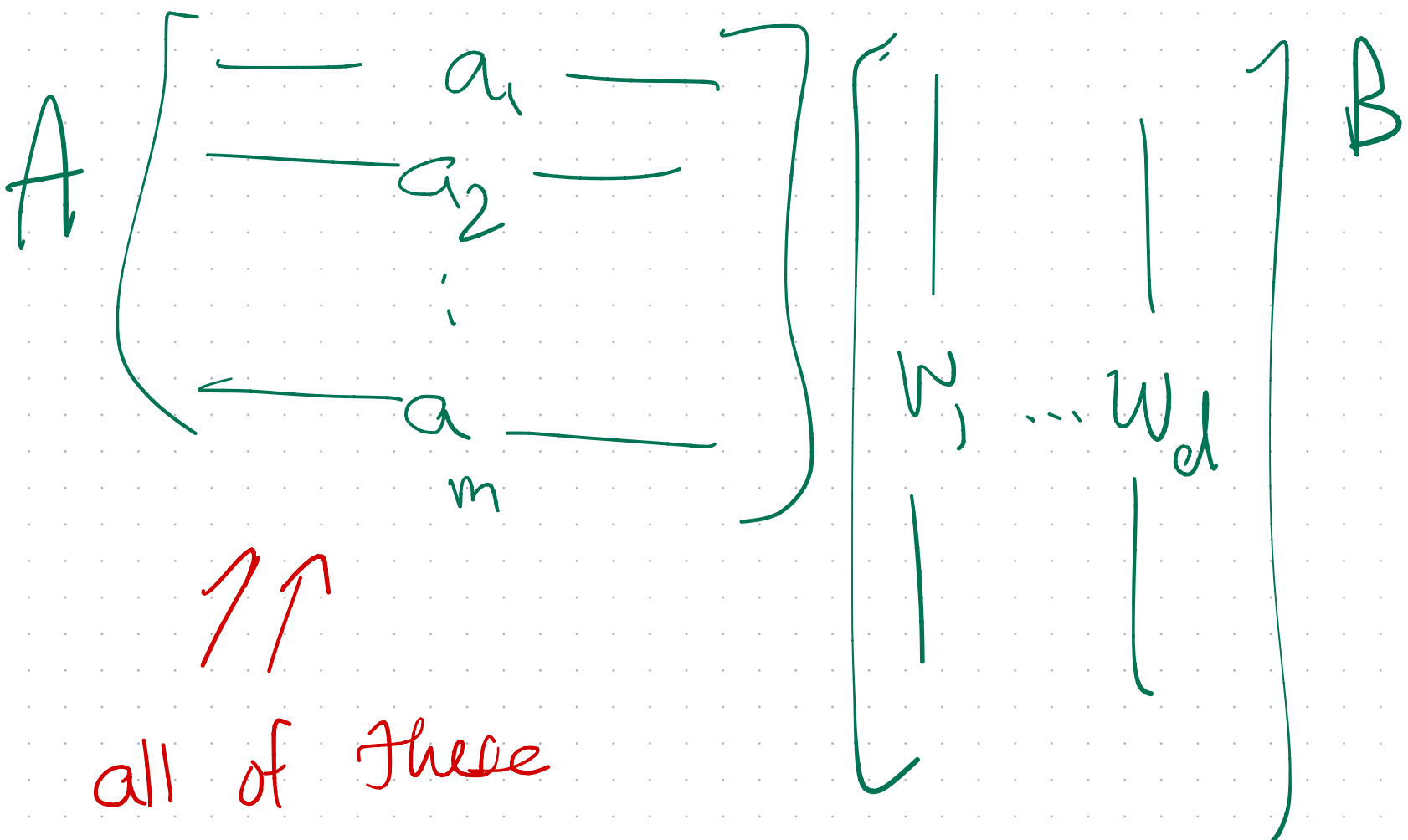
$$\|Bx\|^2 = \Big\| \sum x_i \overrightarrow{w_i} \Big\|^2$$

$$= \sum_i x_i^2 \underbrace{\|w_i\|^2}_{1} + 2\sum_{i \neq j} x_i x_j \underbrace{(w_i \cdot w_j)}_{0}$$

$$= \sum_i x_i^2 = \|x\|^2$$

WTS: $\sigma_1(AB) \le 1+2\varepsilon$

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



↑↑ all of these are $N(0, \frac{1}{m})$, i.i.d.

FACT: AB is a random $m \times d$ matrix with iid $N(0,1)$ entries.

Why? Let Q be $n \times n$ orthogonal mtx whose first d columns are B.

AQ is a random $m \times n$ matrix.

Its rows are independent.

(the rows of Λ are indep.)

Each row $a_i \cdot Q$ is

a sample from $N(0, \frac{1}{m} \mathbb{1}_n)$,

rotated by Q.

But $N(0, \frac{1}{m} \mathbb{1}_n)$ is rotation

invariant.

Conclusion. AQ has independent

random rows each with

$N(0, \frac{1}{m} \mathbb{1}_n)$ distribution.

$\Longrightarrow$ AQ is composed of

i.i.d random $N(0, \frac{1}{m})$

matrix entries.

AB is first ℓ columns of AQ

$\Rightarrow$ AB is composed of i.i.d random $N(0, \frac{1}{m})$ entries.

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

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

We'll show these properties for $(AB)^T$.

Singular values preserved by transposition because...

$$AB = U \Sigma V^T$$

$$(AB)^T = V \Sigma^T U^T$$

The values appearing on the diagonal of Σ, Σ^T are the same.

$(AB)^T$ is $d \times m$ $\left(d \leq \varepsilon^2 m\right)$

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

Aaahh... the total variance of the rows is now 1.

with probability $> 1 - 2e^{-\frac{1}{2}\varepsilon^2 m}$

$$1 - \varepsilon - \sqrt{\frac{d}{m}} \leq \sigma_d\left((AB)^T\right) \leq \sigma_1\left((AB)^T\right) \leq 1 + \varepsilon + \sqrt{\frac{d}{m}}$$

$1 - 2\varepsilon$ $\qquad$ $1 + 2\varepsilon$.

Johnson-Lindenstrauss Lemma (JL)

If $x_1, \ldots, x_k \in \mathbb{R}^n$ arbitrary vectors,

and $A \in \mathbb{R}^{m \times n}$ has iiid $N(0, \frac{1}{m})$

entries, where $m > \frac{16 \ln(k/\delta)}{\varepsilon^2}$

then with probability $\geq 1-\delta$

$$\forall_{i,j}$$
$$(1-\varepsilon)\|x_i - x_j\| \leq \|Ax_i - Ax_j\| \leq (1+\varepsilon)\|x_i - x_j\|$$

Proof. Each of the $\binom{k}{2}$ vector $x_i - x_j$ spans a 1-dimensional subspace W_{ij}.

Invoke Gaussian Hash Lemma on each W_{ij}.

$$\Pr\left(W_{ij} \text{ is distorted more than } \varepsilon\right) < 2e^{-\frac{1}{8}\varepsilon^2 m}$$

$$\Pr\left(\exists_{i,j}\ W_{ij} \text{ distorted}\right) < 2\binom{k}{2} e^{-\frac{1}{8}\varepsilon^2 m}$$

Choose m large enough that RHS $< \delta$.