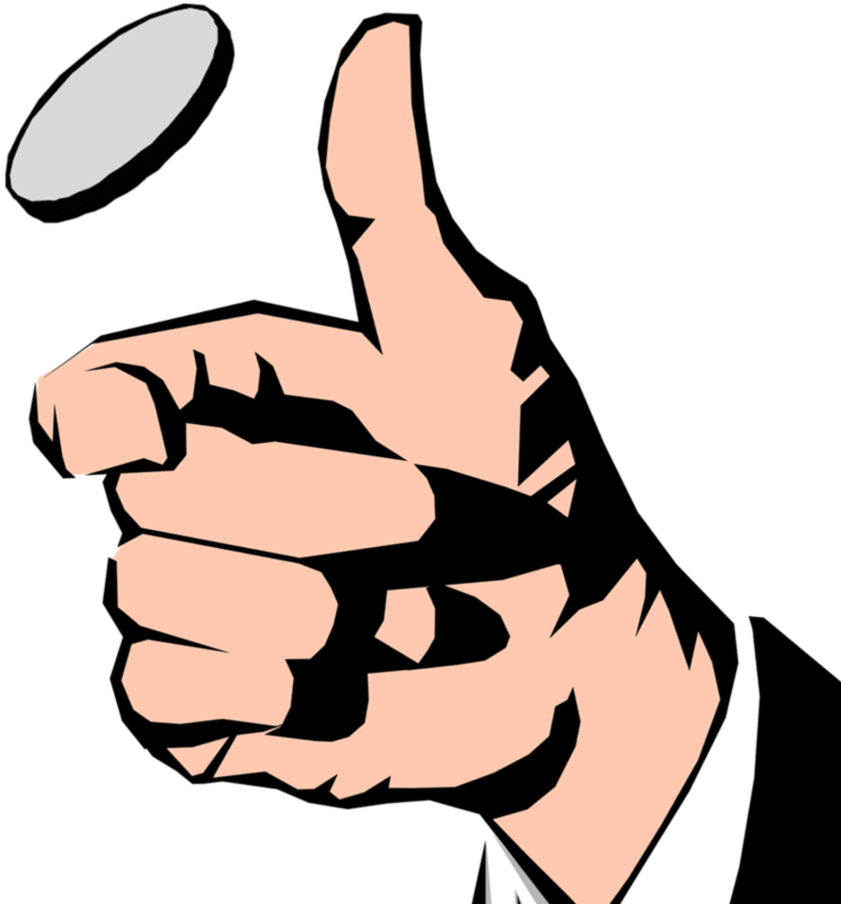
Randomized Algorithms

Ioannis Caragiannis (this time) and Kasper Green Larsen



Multi-dimensional data

- **Documents** as bag of words: # of occurences of word w in a document
- **Network traffic**: number of packets sent by node i to node j
- **User ratings**: rating of user i for service/product/business/etc j

How can we compare documents?

- **Similarity** between two documents is given by the **distance** of their "vectors"
- Claim: **projecting** the document vector in a **smaller space** preserves the similarity between documents
- How? E.g., using the **Johnson-Lindenstrauss** transform

The Johnson-Lindenstrauss transform

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \dots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds
$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \dots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \leq i < j \leq m$, it holds
$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \leq \left\|Lx_i - Lx_j\right\|_2^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$
- The linear transformation L is simply **multiplication** by a matrix whose entries are **sampled independently from a standard Gaussian**, **scaled appropriately**

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds
$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$
- The linear transformation L is simply **multiplication** by a matrix whose entries are **sampled independently from a standard Gaussian**, **scaled appropriately**
- Let A be random $k \times d$ matrix with $A_{i,j} \sim \mathcal{N}(0,1)$, independently from the other entries
- Set $L = \frac{1}{\sqrt{k}} A$

The Johnson-Lindenstrauss lemma

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \dots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds
$$(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$$
- The linear transformation L is simply **multiplication** by a matrix whose entries are **random +1/-1**, **scaled appropriately**
- Let A be a random $k \times d$ matrix with $A_{i,j}$ selected equiprobably from $\{+1, -1\}$, independently from the other entries
- Set $L = \frac{1}{\sqrt{k}} A$

Lemma: For any unit vector v, $||Av||_2^2$ is sharply concentrated around its expectation

- Let $v \in \mathbb{R}^d$ be a **unit vector**
- Let A be a **random $k \times d$ matrix** with $A_{i,j}$ selected **equiprobably from $\{+1, -1\}$**, independently of the other entries
- Then, the **squared norm $Q = ||Av||_2^2$** has $\mathbb{E}[Q] = k$, and for $\eta \in [0,1/2]$, $\Pr[|Q - k| \geq \eta k] \leq 2\exp(-\eta^2 k/8)$

The easy part of the proof: $\mathbb{E}[Q] = k$

- By **linearity of expectation**:

$$\mathbb{E}[Q] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} A_{i,j}v_j\right)^2\right]$$

$$= \sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d} A_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} A_{i,j}A_{i,j'}v_j v_{j'}\right]$$

$$= \sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d} v_j^2\right] + 2\sum_{i=1}^{k}\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}\mathbb{E}[A_{i,j}]\mathbb{E}[A_{i,j'}]v_j v_{j'} = k$$

Summarizing up to now

- $\Pr\left[\left|\|Av\|_2^2 - k\right| \geq \eta k\right] \leq 2\exp(-\eta^2 k/8)$
- Since $L = \frac{1}{\sqrt{k}} A$, this is equivalent to
$$\Pr\left[\left|\|Lv\|_2^2 - 1\right| \geq \eta\right] \leq 2\exp(-\eta^2 k/8)$$
- I.e., $\boldsymbol{L}$ **does not distort the squared norm of the unit vector** $\boldsymbol{v}$ by much

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|_2}$
- Assume that matrix L is such that $1 - \varepsilon \leq \|Lv_{ij}\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|_2}$
- Assume that matrix L is such that $1 - \varepsilon \leq \|Lv_{ij}\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1 - \varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1 + \varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

- Proof: Notice that

$$\|Lx_i - Lx_j\|_2^2 = \|L(x_i - x_j)\|_2^2 = \left\| \|x_i - x_j\|_2 L \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_2^2 = \|x_i - x_j\|_2^2 \cdot \|Lv_{ij}\|_2^2$$

Lemma: It suffices to focus on unit vectors

- For $1 \leq i < j \leq m$, denote by v_{ij} the unit vector $v_{ij} = \frac{x_i - x_j}{\|x_i - x_j\|_2}$
- Assume that matrix L is such that $1 - \varepsilon \leq \|Lv_{ij}\|_2^2 \leq 1 + \varepsilon$, for $1 \leq i < j \leq m$
- Then, $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

- Proof: Notice that

$$\|Lx_i - Lx_j\|_2^2 = \|L(x_i - x_j)\|_2^2 = \left\| \|x_i - x_j\|_2 L \frac{x_i - x_j}{\|x_i - x_j\|_2} \right\|_2^2 = \|x_i - x_j\|_2^2 \cdot \|Lv_{ij}\|_2^2$$

- Hence $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \le \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \le \|Lx_i - Lx_j\|_2^2 \le (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \le i < j \le m$
- We have shown that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \ge \varepsilon\right] \le 2\exp(-\varepsilon^2 k/8)$

Final push: Proof of JL lemma

- So, we know that if $\left| \|Lv_{ij}\|_2^2 - 1 \right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left| \|Lv_{ij}\|_2^2 - 1 \right| \geq \varepsilon \right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left| \|Lv_{ij}\|_2^2 - 1 \right| \geq \varepsilon \right] \leq \frac{2}{m^3}$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i, j\colon \left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{1}{m}$

Final push: Proof of JL lemma

- So, we know that if $\left|\|Lv_{ij}\|_2^2 - 1\right| \leq \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\|x_i - x_j\|_2^2 \leq \|Lx_i - Lx_j\|_2^2 \leq (1+\varepsilon)\|x_i - x_j\|_2^2$, for $1 \leq i < j \leq m$
- We have shown that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2} \ln m$, we have that $\Pr\left[\left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i, j: \left|\|Lv_{ij}\|_2^2 - 1\right| \geq \varepsilon\right] \leq \frac{1}{m}$
- Equivalently, $\Pr\left[\forall i, j: \left|\|Lv_{ij}\|_2^2 - 1\right| < \varepsilon\right] \geq 1 - \frac{1}{m}$

Final push: Proof of JL lemma

- So, we know that if $\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \le \varepsilon$ for the $m(m-1)/2$ unit vectors v_{ij}, then $(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \le \left\|Lx_i - Lx_j\right\|_2^2 \le (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$, for $1 \le i < j \le m$
- We have shown that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \ge \varepsilon\right] \le 2\exp(-\varepsilon^2 k/8)$
- Selecting $k = 24\varepsilon^{-2}\ln m$, we have that $\Pr\left[\left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \ge \varepsilon\right] \le \frac{2}{m^3}$
- Using the union bound, we have $\Pr\left[\exists i, j: \left|\left\|Lv_{ij}\right\|_2^2 - 1\right| \ge \varepsilon\right] \le \frac{1}{m}$
- Equivalently, $\Pr\left[\forall i, j: \left|\left\|Lv_{ij}\right\|_2^2 - 1\right| < \varepsilon\right] \ge 1 - \frac{1}{m}$
- Hence, with probability at least $1 - 1/m$, we get that, for $1 \le i < j \le m$,

$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \le \left\|Lx_i - Lx_j\right\|_2^2 \le (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$

QED

Sparse embeddings

Literature on JL transforms

- Johnson & Lindenstrauss (1984): statement and **first proof** of the JL lemma
- Larsen & Nelson (2017): **tight lower bound** on the number of dimensions of the host space
- Kane & Nelson (2014): **sparse** JL transform
- Nelson & Nguyen (2013): **lower bound** for sparse embeddings
- Weinberger, Dasgupta, Langford, Smola, & Attenberg (2009): **feature hashing**
- Freksen, Kamma, & Larsen (2018): when does feature hashing work?
- Ailon & Chazelle (2009): **fast** JL transform

Sparse embeddings

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2} \ln m\right)$ and any points $x_1, x_2, \dots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds

$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \le \left\|Lx_i - Lx_j\right\|_2^2 \le (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$

- Embedding time: $O(kd)$ per point
- Usually, vectors have few non-zero entries $\|x\|_0$ (e.g., bag of words data)
- Embedding time: $O(k\|x\|_0)$
- Can be improved further if the matrix L has few non-zero entries

Sparse embeddings

- For any $\varepsilon \in (0,1/2)$ and any integer m, then for integer $k = O\left(\frac{1}{\varepsilon^2}\ln m\right)$ and any points $x_1, x_2, \dots, x_m \in \mathbb{R}^d$, there exists a **linear map** (matrix) $L: \mathbb{R}^d \to \mathbb{R}^k$ such that for any $1 \le i < j \le m$, it holds
$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \le \left\|Lx_i - Lx_j\right\|_2^2 \le (1+\varepsilon)\left\|x_i - x_j\right\|_2^2$$

- For each column of the matrix $A \in \mathbb{R}^{k\times d}$, pick a uniform random set of $t = O(\varepsilon^{-1}\log m)$ rows and assign -1 or $+1$ to the corresponding entries equiprobably and independently
- Set $L = \frac{1}{\sqrt{t}}A$
- Embedding time improved to $O(\|x\|_0 \varepsilon^{-1}\log m)$

Lemma: For any vector v, $||Lv||_2^2$ has expectation $||v||_2^2$

- Let $v \in \mathbb{R}^d$ be a **vector**
- Let A be a $\boldsymbol{k \times d}$ **matrix** defined as follows: For each column of the matrix $A \in \mathbb{R}^{k \times d}$, pick a **uniform random set of $\boldsymbol{t}$ rows** and assign $\boldsymbol{-1}$ **or** $\boldsymbol{+1}$ to the corresponding entries equiprobably and independently
- Then, the **squared norm** $\boldsymbol{||Lv||_2^2}$ satisfies $\mathbb{E}[||Lv||_2^2] = ||v||_2^2$

- The proof that $||Lv||_2^2$ is sharply concentrated around its expectation is considerably more difficult

Proof

$$\mathbb{E}[\|Lv\|_2^2] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}Av\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{k}\left(\sum_{j=1}^{d} A_{i,j}v_j\right)^2\right]$$

Proof

$$\mathbb{E}[\|Lv\|_2^2] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}Av\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{k}\left(\sum_{j=1}^{d}A_{i,j}v_j\right)^2\right]$$

linearity of expectation

$$= \frac{1}{t}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d}A_{i,j}^2v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}A_{i,j}A_{i,j'}v_jv_{j'}\right]$$

Proof

$$\mathbb{E}[\|Lv\|_2^2] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}Av\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{k}\left(\sum_{j=1}^{d} A_{i,j}v_j\right)^2\right]$$

linearity of expectation

$$= \frac{1}{t}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d} A_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} A_{i,j}A_{i,j'}v_j v_{j'}\right]$$

independence

$$= \frac{1}{t}\sum_{i=1}^{k}\sum_{j=1}^{d}\mathbb{E}[A_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}\mathbb{E}[A_{i,j}]\mathbb{E}[A_{i,j'}]v_j v_{j'}$$

Proof

$$\mathbb{E}[\|Lv\|_2^2] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}Av\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{k}\left(\sum_{j=1}^{d}A_{i,j}v_j\right)^2\right]$$

linearity of expectation

$$= \frac{1}{t}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d}A_{i,j}^2v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}A_{i,j}A_{i,j'}v_jv_{j'}\right]$$

independence

$$= \frac{1}{t}\sum_{i=1}^{k}\sum_{j=1}^{d}\mathbb{E}[A_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}\mathbb{E}[A_{i,j}]\mathbb{E}[A_{i,j'}]v_jv_{j'}$$

$\mathbb{E}[A_{i,j}^2] = t/k$
$A_{i,j}^2 = 1$ w.p. t/k,
$A_{i,j}^2 = 0$ otherwise

$\mathbb{E}[A_{i,j}] = 0$

$$= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d}v_j^2 = \sum_{j=1}^{d}v_j^2 = \|v\|_2^2$$

Proof

$$\mathbb{E}[\|Lv\|_2^2] = \mathbb{E}\left[\left\|\frac{1}{\sqrt{t}}Av\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{k}\left(\sum_{j=1}^{d}A_{i,j}v_j\right)^2\right]$$

linearity of expectation

$$= \frac{1}{t}\sum_{i=1}^{k}\mathbb{E}\left[\sum_{j=1}^{d}A_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} A_{i,j}A_{i,j'}v_j v_{j'}\right]$$

independence

$$= \frac{1}{t}\sum_{i=1}^{k}\sum_{j=1}^{d}\mathbb{E}[A_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d}\mathbb{E}[A_{i,j}]\mathbb{E}[A_{i,j'}]v_j v_{j'}$$

$\mathbb{E}[A_{i,j}^2] = t/k$

$A_{i,j}^2 = 1$ w.p. t/k,

$A_{i,j}^2 = 0$ otherwise

$\mathbb{E}[A_{i,j}] = 0$

$$= \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{d}v_j^2 = \sum_{j=1}^{d}v_j^2 = \|v\|_2^2$$

QED

Feature hashing

- Feature hashing: selects **just one row per column** that is $+1/-1$
- Important question: Assuming an optimal number of dimension, what is a sufficient/necessary condition so that fearure hashing preserves distances within $1 \pm \varepsilon$?
- Quantification in terms of $\|\boldsymbol{x}\|_{\infty} / \|\boldsymbol{x}\|_{\mathbf{2}}$ (hopefully small)
- Recall the **bag of word** example
- By ignoring very frequent words (e.g., “the” has frequency 5% in english texts), the document vectors do not have any large coordinate

Fast Johnson-Lindenstrauss transform

High-level idea

- First multiply all vectors with a matrix that **ensures that coordinates are small** (like in feature hashing)
- Then, use a **sparse embedding**

Fast JL transform

- Assume t is power of 2
- $\overline{H}_t$ is a **Walsch-Hadamard matrix**, defined as
- $\overline{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
- $\overline{H}_{2t} = \begin{bmatrix} \overline{H}_t & \overline{H}_t \\ \overline{H}_t & -\overline{H}_t \end{bmatrix}$

- Property: all rows of $\overline{H}_t$ are **orthogonal** (differ in exactly half of the entries)

Fast JL transform

- Assume d is power of 2
- $\overline{H}_t$ is a **Walsch-Hadamard matrix**, defined as
- $\overline{H}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
- $\overline{H}_{2t} = \begin{bmatrix} \overline{H}_t & \overline{H}_t \\ \overline{H}_t & -\overline{H}_t \end{bmatrix}$

- Property: all rows of $\overline{H}_t$ are **orthogonal** (differ in exactly half of the entries)
- Property: the matrix-vector product $\overline{H}_t v$ can be computed **efficiently** (faster than the naïve $O(t^2)$ time)

Computing the matrix-vector product $\overline{H}_t v$

- Write v as $\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$

Computing the matrix-vector product $\overline{H}_t v$

- Write v as $\binom{v_1}{v_2}$
- Then, $\overline{H}_t v = \begin{bmatrix} \overline{H}_{t/2} & \overline{H}_{t/2} \\ \overline{H}_{t/2} & -\overline{H}_{t/2} \end{bmatrix} \binom{v_1}{v_2} = \begin{pmatrix} \overline{H}_{t/2} v_1 + \overline{H}_{t/2} v_2 \\ \overline{H}_{t/2} v_1 - \overline{H}_{t/2} v_2 \end{pmatrix}$

Computing the matrix-vector product $\overline{H}_t v$

- Write v as $\binom{v_1}{v_2}$
- Then, $\overline{H}_t v = \begin{bmatrix} \overline{H}_{t/2} & \overline{H}_{t/2} \\ \overline{H}_{t/2} & -\overline{H}_{t/2} \end{bmatrix} \binom{v_1}{v_2} = \begin{pmatrix} \overline{H}_{t/2} v_1 + \overline{H}_{t/2} v_2 \\ \overline{H}_{t/2} v_1 - \overline{H}_{t/2} v_2 \end{pmatrix}$
- So, to compute $\overline{H}_t v$, it suffices to compute the vectors $\overline{H}_{t/2} v_1$ and $\overline{H}_{t/2} v_2$ and their addition and difference

Computing the matrix-vector product $\overline{H}_t v$

- Write v as $\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$
- Then, $\overline{H}_t v = \begin{bmatrix} \overline{H}_{t/2} & \overline{H}_{t/2} \\ \overline{H}_{t/2} & -\overline{H}_{t/2} \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \overline{H}_{t/2} v_1 + \overline{H}_{t/2} v_2 \\ \overline{H}_{t/2} v_1 - \overline{H}_{t/2} v_2 \end{pmatrix}$
- So, to compute $\overline{H}_t v$, it suffices to compute the vectors $\overline{H}_{t/2} v_1$ and $\overline{H}_{t/2} v_2$ and their addition and difference
- I.e., time $T(t) = 2T(t/2) + O(t)$, or $T(t) = O(t \log t)$

Fast JL transform

- Let H be a **normalized Walsch-Hadamard matrix**, i.e., $H = d^{-1/2}\overline{H}_d$
- Draw a **random diagonal matrix** $D \in \mathbb{R}^{d\times d}$ where each diagonal entry is equiprobably and independently -1 and $+1$
- **Draw a matrix $\boldsymbol{P}$** $\in \mathbb{R}^{k\times d}$ with $k = O\left(\frac{1}{\varepsilon^2}\ln m\right)$
- Each entry of P is 0 with probability $1 - q$ and drawn from $\mathcal{N}(0, (kq)^{-1})$ otherwise
- **Embedding of point $\boldsymbol{v}$ at $\boldsymbol{PHDv}$**: takes time $O(d)$ to compute Dv, $O(d\log d)$ to compute HDv, and $O(qkd)$ to compute $PHDv$
- Achieves the JL guarantee by setting $q = O(\log^2 m / d)$
- Embedding time: $O(d\log d + \varepsilon^{-2}\log^3 m)$

Properties of HDv

- Multiplication Dv leaves the norm of v unaffected
- H is an orthogonal matrix with all rows having norm 1
- Hence, $||HDv||_2^2 = ||v||_2^2$ for any vector $v \in \mathbb{R}^d$

Proof of the JL guarantee

- Step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p
- Step 2: Assuming $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$, multiplication with matrix P preserves the norms with $1 \pm \varepsilon$

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.

- First observe that the entries in each row of the $d \times d$ matrix HD are either $d^{-1/2}$ or $-d^{-1/2}$ and independent (why?)
- Hence, $(HDv)_i$ is a **sum of independent random variables** $X_1, X_2, \dots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ (equiprobably)
- Hence, $\mathbb{E}[(HDv)_i] = 0$
- How large can $(HDv)_i$ be?

Hoeffding inequality

- Let $X_1, X_2, \dots, X_d$ be independent random variables where X_j takes values in $\left[a_j, b_j\right]$. Let $X = \sum_{j=1}^{d} X_j$. Then

$$\Pr[|X - \mathbb{E}[X]| > t] < 2\exp\left(-\frac{2t^2}{\sum_{j=1}^{d}(b_j - a_j)^2}\right)$$

Proof (of Hoeffding inequality)

- Let $\lambda > 0$. Using **Markov inequality**,

$$\Pr[X - \mathbb{E}[X] \geq t] = \Pr[\exp\big(\lambda(X - \mathbb{E}[X])\big) \geq \exp(\lambda t)] \leq \frac{\mathbb{E}\big[\exp\big(\lambda(X - \mathbb{E}[X])\big)\big]}{\exp(\lambda t)}$$

- The numerator becomes

$$\mathbb{E}\big[\exp\big(\lambda(X - \mathbb{E}[X])\big)\big] = \mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^{d} (X_j - \mathbb{E}[X_j])\right)\right]$$

$$= \mathbb{E}\left[\prod_{j=1}^{d} \exp(\lambda(X_j - \mathbb{E}[X_j]))\right] = \prod_{j=1}^{d} \mathbb{E}[\exp(\lambda(X_j - \mathbb{E}[X_j]))]$$

Hoeffding’s lemma

- Let Y be a random variable taking values in $[a, b]$
- Then, for every $\lambda \in \mathbb{R}$, it is

$$\mathbb{E}[\exp(\lambda Y)] \leq \exp\left(\lambda \mathbb{E}[Y] + \lambda^2 \frac{(b-a)^2}{8}\right)$$

Proof (of Hoeffding inequality)

- Random variable $X_j - \mathbb{E}[X_j] \in \left[a_j - \mathbb{E}[X_j], b_j - \mathbb{E}[X_j]\right]$ has expectation 0
- By Hoeffding's lemma, we get that the numerator is upper-bounded by $\prod_{j=1}^{d} \exp\left(\frac{\lambda^2 (b_j - a_j)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8} \sum_{j=1}^{d} (b_j - a_j)^2\right)$

Proof (of Hoeffding inequality)

- Random variable $X_j - \mathbb{E}[X_j] \in \left[a_j - \mathbb{E}[X_j], b_j - \mathbb{E}[X_j]\right]$ has expectation 0
- By Hoeffding's lemma, we get that the numerator is upper-bounded by $\prod_{j=1}^{d} \exp\left(\frac{\lambda^2 (b_j - a_j)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8} \sum_{j=1}^{d} (b_j - a_j)^2\right)$
- Putting everything together

$$\Pr[X - \mathbb{E}[X] \geq t] \leq \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{j=1}^{d} (b_j - a_j)^2\right)$$

Proof (of Hoeffding inequality)

- Random variable $X_j - \mathbb{E}[X_j] \in \left[a_j - \mathbb{E}[X_j], b_j - \mathbb{E}[X_j]\right]$ has expectation 0
- By Hoeffding's lemma, we get that the numerator is upper-bounded by $\prod_{j=1}^{d} \exp\left(\frac{\lambda^2 (b_j - a_j)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8} \sum_{j=1}^{d} (b_j - a_j)^2\right)$
- Putting everything together

$$\Pr[X - \mathbb{E}[X] \geq t] \leq \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{j=1}^{d} (b_j - a_j)^2\right)$$

- Setting $\lambda = 4t \left(\sum_{j=1}^{d} (b_j - a_j)^2\right)^{-1}$ we get $\Pr[X - \mathbb{E}[X] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^{d} (b_j - a_j)^2}\right)$

Proof (of Hoeffding inequality)

- Need also to prove that $\Pr[X - \mathbb{E}[X] \leq -t] \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^{d}(b_j - a_j)^2}\right)$
- Very similar!

QED

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p

- $(HDv)_i$ is a sum of independent random variables $X_1, X_2, \ldots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ and $\mathbb{E}[(HDv)_i] = 0$

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p

- $(HDv)_i$ is a sum of independent random variables $X_1, X_2, \dots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ and $\mathbb{E}[(HDv)_i] = 0$
- By Hoeffding inequality,

$$\Pr[|(HDv)_i| \geq t] \leq 2 \exp\left(-\frac{t^2 d}{2 \sum_{j=1}^{d} v_j^2}\right) = 2 \exp\left(-\frac{t^2 d}{2\|v\|_2^2}\right)$$

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p

- $(HDv)_i$ is a sum of independent random variables $X_1, X_2, \ldots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ and $\mathbb{E}[(HDv)_i] = 0$
- By Hoeffding inequality,

$$\Pr[|(HDv)_i| \geq t] \leq 2\exp\left(-\frac{t^2 d}{2\sum_{j=1}^{d} v_j^2}\right) = 2\exp\left(-\frac{t^2 d}{2\|v\|_2^2}\right)$$

- Setting $t = \sqrt{\frac{2\ln dm^3}{d}}\|HDv\|_2$ (recall that $\|HDv\|_2 = \|v\|_2$), we get that the probability that $(HDv)_i / \|HDv\|_2 \geq \sqrt{\frac{2\ln dm^3}{d}}$ is $\frac{2}{dm^3}$

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p

- $(HDv)_i$ is a sum of independent random variables $X_1, X_2, \dots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ and $\mathbb{E}[(HDv)_i] = 0$
- By Hoeffding inequality,

$$\Pr[|(HDv)_i| \geq t] \leq 2\exp\left(-\frac{t^2 d}{2\sum_{j=1}^{d} v_j^2}\right) = 2\exp\left(-\frac{t^2 d}{2\|v\|_2^2}\right)$$

- Setting $t = \sqrt{\frac{2\ln dm^3}{d}} \|HDv\|_2$ (recall that $\|HDv\|_2 = \|v\|_2$), we get that the probability that $(HDv)_i / \|HDv\|_2 \geq \sqrt{\frac{2\ln dm^3}{d}}$ is $\frac{2}{dm^3}$
- Hence, the probability that $\|HDv\|_\infty / \|HDv\|_2 \leq \sqrt{\frac{2\ln dm^3}{d}}$ is $1 - \frac{2}{m^3}$

Proof of step 1: $\|HDv\|_\infty / \|HDv\|_2 = O\left(\sqrt{\frac{\log m}{d}}\right)$ w.h.p

- $(HDv)_i$ is a sum of independent random variables $X_1, X_2, \dots, X_d$ with X_j takes values either $d^{-1/2} v_j$ or $-d^{-1/2} v_j$ and $\mathbb{E}[(HDv)_i] = 0$
- By Hoeffding inequality,

$$\Pr[|(HDv)_i| \geq t] \leq 2\exp\left(-\frac{t^2 d}{2\sum_{j=1}^{d} v_j^2}\right) = 2\exp\left(-\frac{t^2 d}{2\|v\|_2^2}\right)$$

- Setting $t = \sqrt{\frac{2\ln dm^3}{d}} \|HDv\|_2$ (recall that $\|HDv\|_2 = \|v\|_2$), we get that the probability that $(HDv)_i / \|HDv\|_2 \geq \sqrt{\frac{2\ln dm^3}{d}}$ is $\frac{2}{dm^3}$
- Hence, the probability that $\|HDv\|_\infty / \|HDv\|_2 \leq \sqrt{\frac{2\ln dm^3}{d}}$ is $1 - \frac{2}{m^3}$ QED

Lemma: Multiplication with P preserves norms in expectation

- For every vector $v \in \mathbb{R}^d$, it is $\mathbb{E}[\|Pv\|_2^2] = \|v\|_2^2$

- The proof that $\|PHDv\|_2^2$ is sharply concentrated around its expectation (which would complete the proof of Step 2) is considerably more difficult

Lemma: Multiplication with P preserves norms in expectation

- For every vector $v \in \mathbb{R}^d$, it is $\mathbb{E}[\|Pv\|_2^2] = \|v\|_2^2$
- Proof: Entry $P_{i,j}$ can be written as $b_{i,j}n_{i,j}$, where $b_{i,j}$ takes value 1 with probability q and 0 with probability $1-q$, and $n_{i,j} \sim \mathcal{N}(0, (kq)^{-1})$

Proof

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right]$$

Proof

linearity of expectation

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} P_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} P_{i,j}P_{i,j'}v_j v_{j'}\right]$$

Proof

linearity of expectation

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} P_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} P_{i,j}P_{i,j'}v_j v_{j'}\right]$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{d} \mathbb{E}[b_{i,j}^2]\mathbb{E}[n_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} \mathbb{E}[b_{i,j}]\mathbb{E}[b_{i,j}]\mathbb{E}[n_{i,j}]\mathbb{E}[n_{i,j'}]v_j v_{j'}\right)$$

linearity of expectation
and independence

Proof

linearity of expectation

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} P_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} P_{i,j}P_{i,j'}v_j v_{j'}\right]$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{d} \mathbb{E}[b_{i,j}^2]\mathbb{E}[n_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} \mathbb{E}[b_{i,j}]\mathbb{E}[b_{i,j}]\mathbb{E}[n_{i,j}]\mathbb{E}[n_{i,j'}]v_j v_{j'}\right)$$

linearity of expectation and independence

$$= \sum_{i=1}^{k}\sum_{j=1}^{d} q(kq)^{-1}v_j^2$$

definition of variance of $n_{i,j} \sim \mathcal{N}(0,(kq)^{-1})$

Proof

linearity of expectation

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} P_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} P_{i,j}P_{i,j'}v_j v_{j'}\right]$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{d} \mathbb{E}[b_{i,j}^2]\mathbb{E}[n_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} \mathbb{E}[b_{i,j}]\mathbb{E}[b_{i,j}]\mathbb{E}[n_{i,j}]\mathbb{E}[n_{i,j'}]v_j v_{j'}\right)$$

linearity of expectation and independence

$$= \sum_{i=1}^{k}\sum_{j=1}^{d} q(kq)^{-1}v_j^2 = \sum_{j=1}^{d} v_j^2 = \|v\|_2^2$$

definition of variance of $n_{i,j} \sim \mathcal{N}(0, (kq)^{-1})$

Proof

linearity of expectation

$$\mathbb{E}[\|Pv\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{k}\left(\sum_{j=1}^{d} P_{i,j}\, v_j\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} P_{i,j}^2 v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} P_{i,j}P_{i,j'}v_j v_{j'}\right]$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{d} \mathbb{E}[b_{i,j}^2]\mathbb{E}[n_{i,j}^2]v_j^2 + 2\sum_{j=1}^{d-1}\sum_{j'=j+1}^{d} \mathbb{E}[b_{i,j}]\mathbb{E}[b_{i,j}]\mathbb{E}[n_{i,j}]\mathbb{E}[n_{i,j'}]v_j v_{j'}\right)$$

linearity of expectation
and independence

$$= \sum_{i=1}^{k}\sum_{j=1}^{d} q(kq)^{-1}v_j^2 = \sum_{j=1}^{d} v_j^2 = \|v\|_2^2$$

definition of variance
of $n_{i,j} \sim \mathcal{N}(0,(kq)^{-1})$

QED

Last slide

- Johnson-Lindenstrauss transform (alternative proof using random coins)
- Sparse embeddings
- Fast JL transform