CS-GY 6763: Lecture 3
High Dimensional Geometry, the Johnson-Lindenstrauss Lemma, MinHash

NYU Tandon School of Engineering, Prof. Rajesh Jayaram

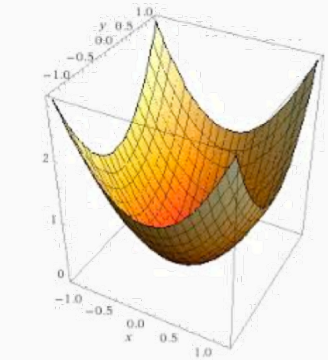
Unifying themes of the course

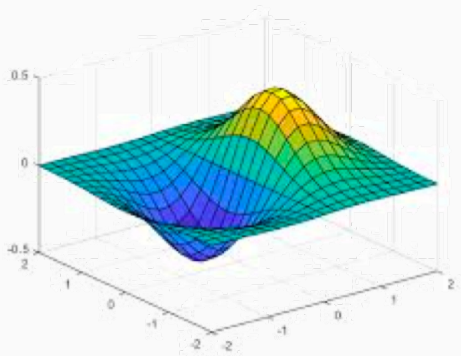
How do we deal with data (vectors) in high dimensions?

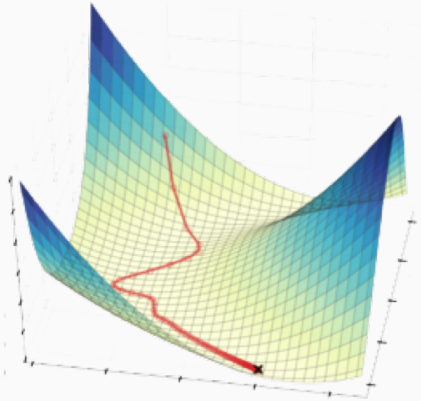
- Distance-preserving dimensionality reduction (JL Lemma)
- Locality sensitive hashing (LSH) for nearest neighbor search.
- Iterative methods for optimizing functions that depend on many variables.
- SVD + low-rank approximation to find and visualize low-dimensional structure.

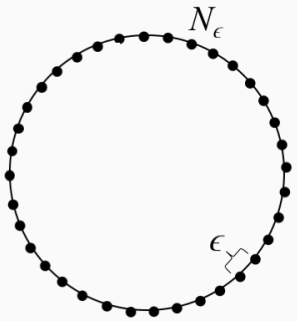
High-dimensional space is not like low-dimensional space

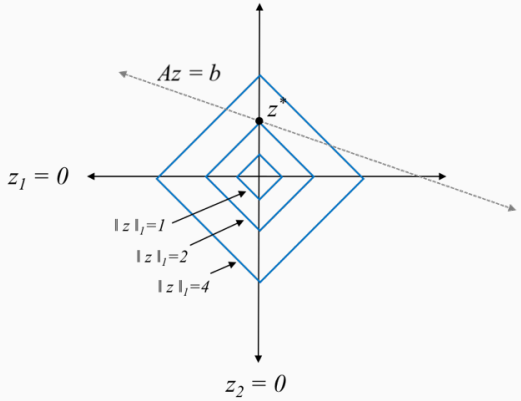
Often visualize data and algorithms in 1,2, or 3 dimensions.

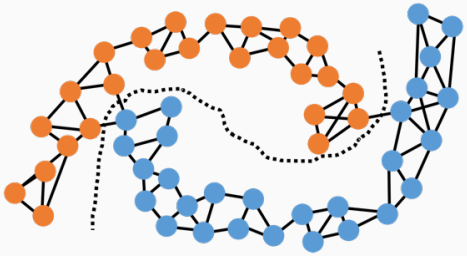












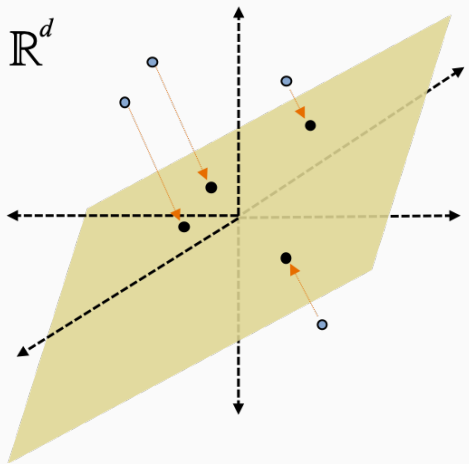
First part of lecture: Prove that high-dimensional space looks **very different** from low-dimensional space. These images are rarely very informative!

Sketching and dimensionality reduction

Second part of lecture: Ignore our own advice.

Learn about **sketching, aka dimensionality reduction** techniques that seek to approximate high-dimensional vectors with much lower dimensional vectors.

- Johnson-Lindenstrauss lemma for ℓ_2 space.
- MinHash for binary vectors.

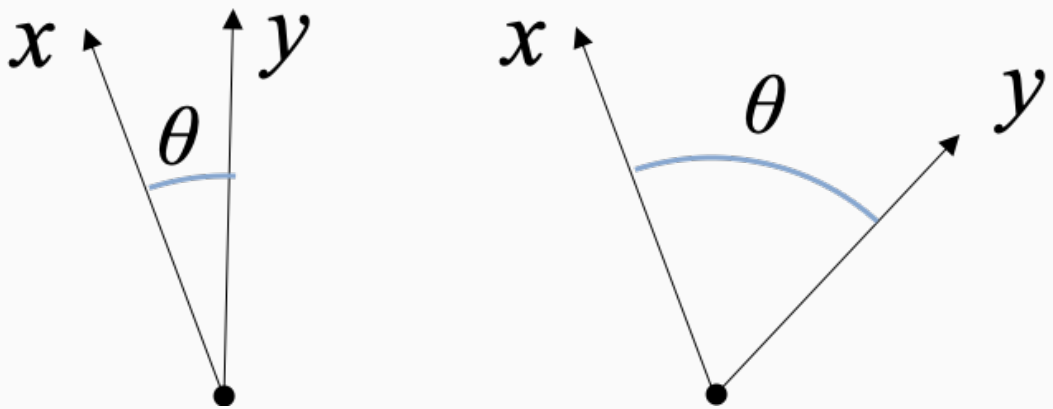


First part of lecture should help you understand the potential and limitations of these methods.

Orthogonal vectors

Recall the inner product between two d dimensional vectors:

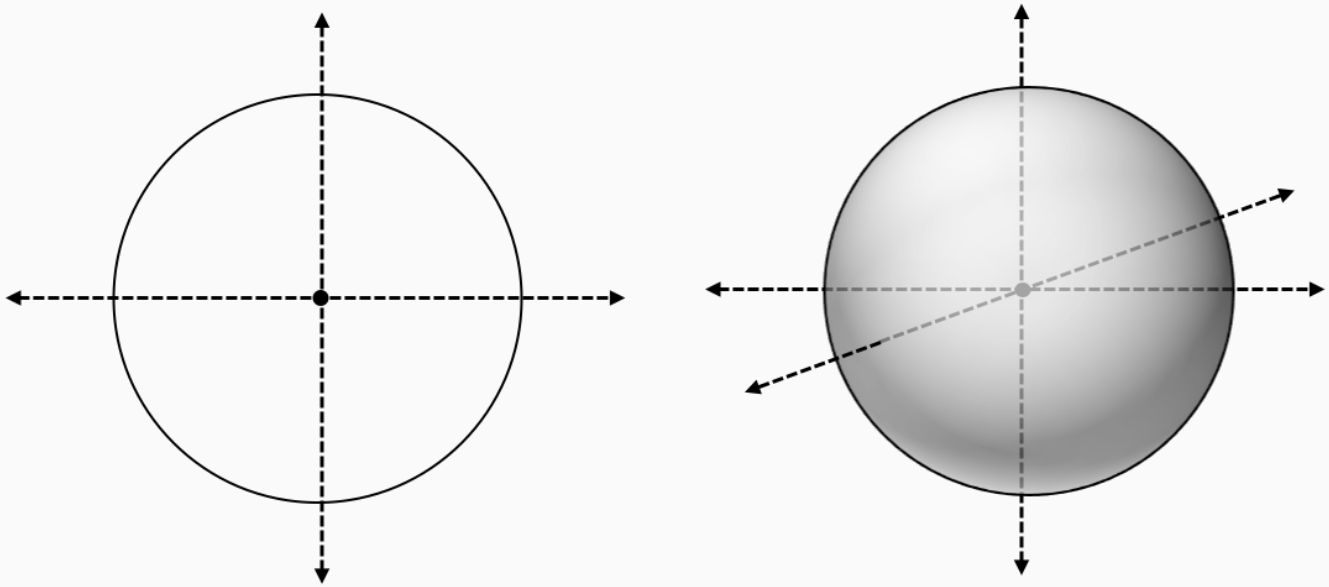
$$\langle x, y \rangle = x^T y = y^T x = \sum_{i=1}^{d} x_i y_i$$



$$\langle x, y \rangle = \cos(\theta) \cdot \|x\|_2 \cdot \|y\|_2$$

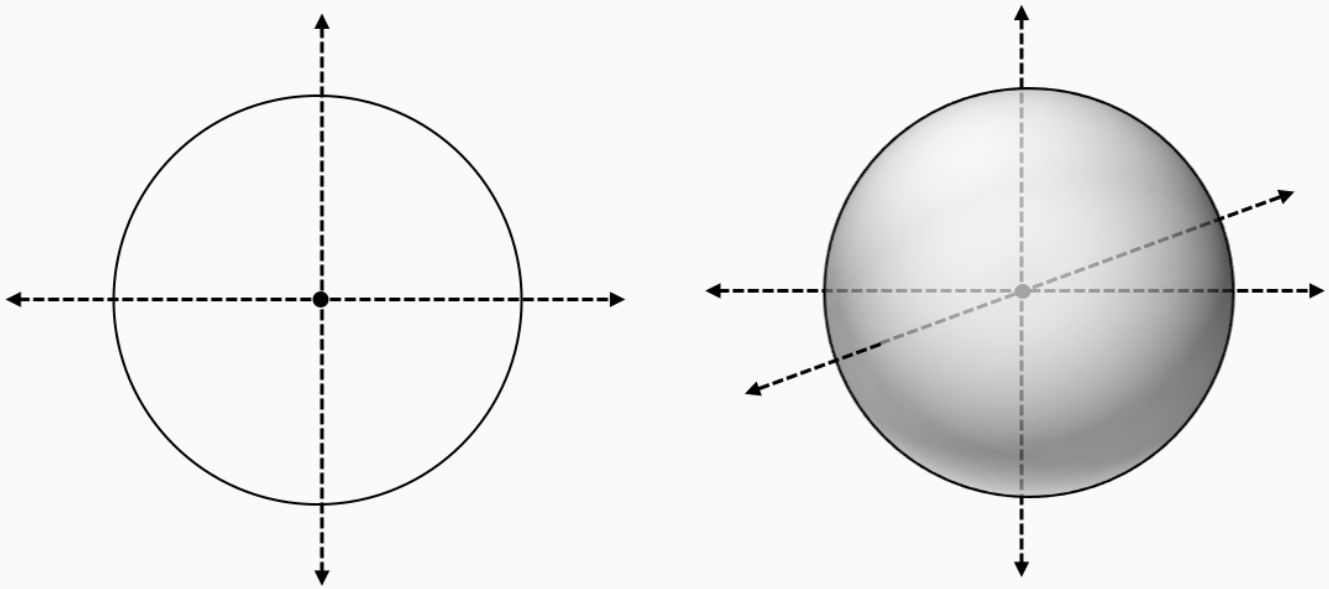
Orthogonal vectors

What is the largest set of **mutually orthogonal** unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ in d-dimensional space? I.e. with inner product $|\mathbf{x}_i^T \mathbf{x}_j| = 0$ for all i, j.



Orthogonal vectors

What is the largest set **nearly orthogonal** unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ in d-dimensional space. I.e., with inner product $|\mathbf{x}_i^T \mathbf{x}_j| \leq \epsilon$ for all i, j.



Orthogonal vectors

What is the largest set **nearly orthogonal** unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ in d-dimensional space. I.e., with inner product $|\mathbf{x}_i^T \mathbf{x}_j| \leq \epsilon$ for all i, j.

1. d
2. $\Theta(d)$
3. $\Theta(d^2)$
4. $2^{\Theta(d)}$

Orthogonal vectors

Claim: There is an exponential number (i.e., $\sim 2^d$) of nearly orthogonal unit vectors in d dimensional space.

Proof strategy: Use the Probabilistic Method! For $t = O(2^d)$, define a random process which generates random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ that are unlikely to have large inner product.

1. Claim that, with non-zero probability, $|\mathbf{x}_i^T \mathbf{x}_j| \leq \epsilon$ for all i, j.
2. Conclude that there must exists some set of t unit vectors with all pairwise inner-products bounded by ϵ.

Probabilistic method

Claim: There is an exponential number (i.e., $\sim 2^d$) of nearly orthogonal unit vectors in d dimensional space.

Proof: Let $\mathbf{x}_1, \ldots, \mathbf{x}_t$ all have independent random entries, each set to $\pm\frac{1}{\sqrt{d}}$ with equal probability.

- $\|\mathbf{x}_i\|_2 =$
- $\mathbb{E}[\mathbf{x}_i^T \mathbf{x}_j] =$
- $\mathrm{Var}[\mathbf{x}_i^T \mathbf{x}_j] =$

Probabilistic method

Let $Z = \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{d} C_i$ where each C_i is $+\frac{1}{d}$ or $-\frac{1}{d}$ with equal probability.

Z is a sum of many i.i.d. random variables, so looks approximately Gaussian. Roughly, we expect that:

$$\Pr[|Z - \mathbb{E}Z| \geq \alpha \cdot \sigma] \leq O(e^{-\alpha^2})$$

Note that we can transform to binary random variable:

$$\begin{aligned} Z = \sum_{i=1}^{d} C_i &= \frac{2}{d} \sum_{i=1}^{d} \frac{d}{2} \cdot C_i \\ &= \frac{2}{d} \cdot \left(-\frac{d}{2} + \sum_{i=1}^{d} B_i \right) \end{aligned}$$

where each B_i is uniform in $\{0, 1\}$.

Chernoff bound

Theorem (Chernoff Bound)

Let $X_1, X_2, \ldots, X_k$ be independent $\{0,1\}$-valued random variables and let $S = \sum_{i=1}^{k} X_i$. We have for any $\epsilon < 1$:

$$\Pr[|S - \mathbb{E}[S]| \geq \epsilon \mathbb{E}[S]] \leq 2e^{\frac{-\epsilon^2 \mathbb{E}[S]}{3}}.$$

$$\Pr[|B - \mathbb{E}[B]| \geq \qquad\qquad] \leq$$

Probabilistic method

Formally, using a Chernoff bound:

$$\Pr[|Z - \mathbb{E}Z| \geq \epsilon] \leq 2e^{-\epsilon^2 d/6}$$

$$\text{For } \underline{\text{any}} \ i, j \text{ pair, } \Pr[|\mathbf{x}_i^T \mathbf{x}_j| < \epsilon] \geq 1 - 2e^{-\epsilon^2 d/6}.$$

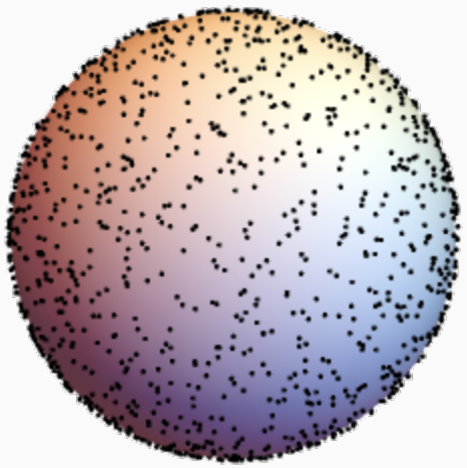
By a union bound:

$$\text{For } \underline{\text{all}} \ i, j \text{ pairs simultaneously, } \Pr[|\mathbf{x}_i^T \mathbf{x}_j| < \epsilon] \geq 1 - \binom{t}{2} \cdot 2e^{-\epsilon^2 d/6}.$$

orthogonal vectors

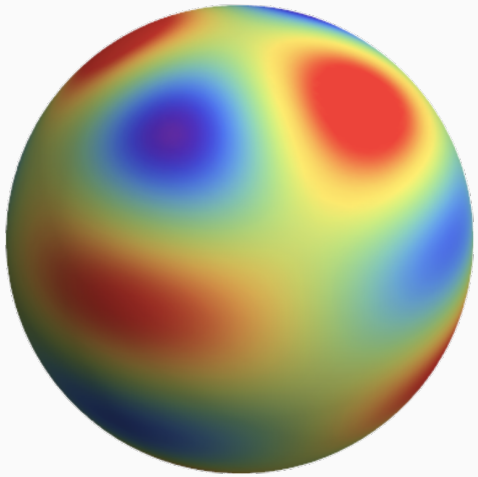
Final result: In d-dimensional space, there are $2^{\theta(\epsilon^2 d)}$ unit vectors with all pairwise inner products $\leq \epsilon$.

Corollary of proof: Random vectors tend to be far apart in high-dimensions.



curse of dimensionality

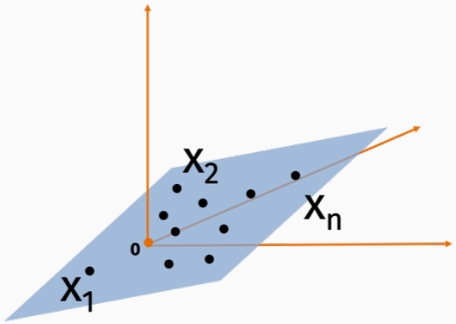
Curse of dimensionality: Suppose we want to use e.g. k-nearest neighbors to learn a function or classify points in $\mathbb{R}^d$. If our data distribution is truly random, we typically need an exponential amount of data before seeing close points!

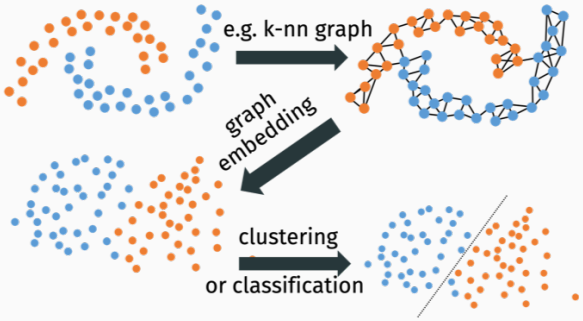


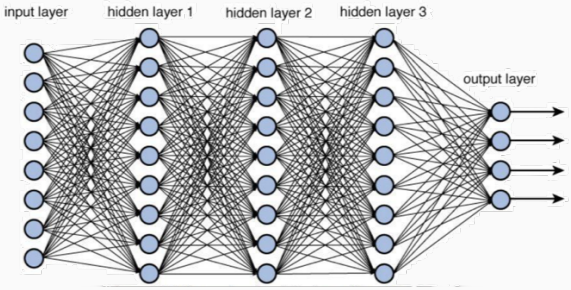
The existence of lower dimensional structure is our data is often the only reason we can hope to learn.

curse of dimensionality

Low-dimensional structure.







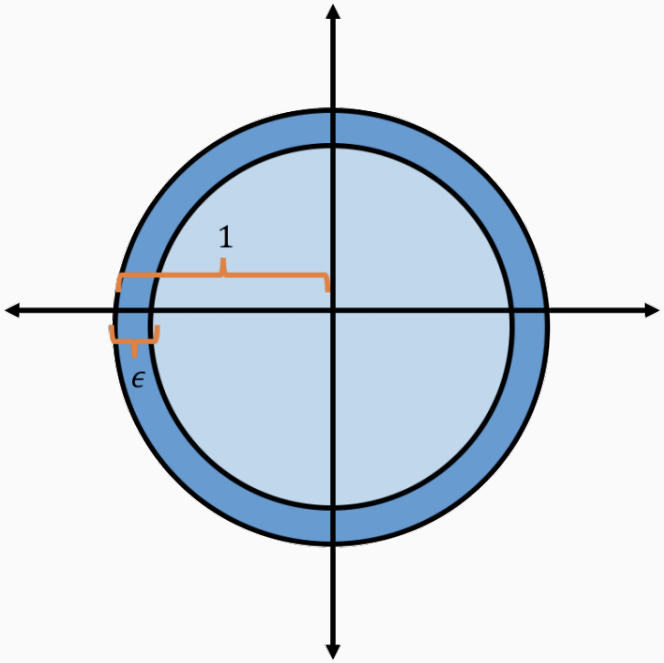
For example, data lies on low-dimensional subspace, or does so after transformation. Or function can be represented by a restricted class of functions, like neural net with specific structure.

unit ball in high dimensions

Let $\mathcal{B}_d$ be the unit ball in d dimensions:

$$\mathcal{B}_d = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}.$$

What percentage of volume of $\mathcal{B}_d$ falls with ϵ of its surface?

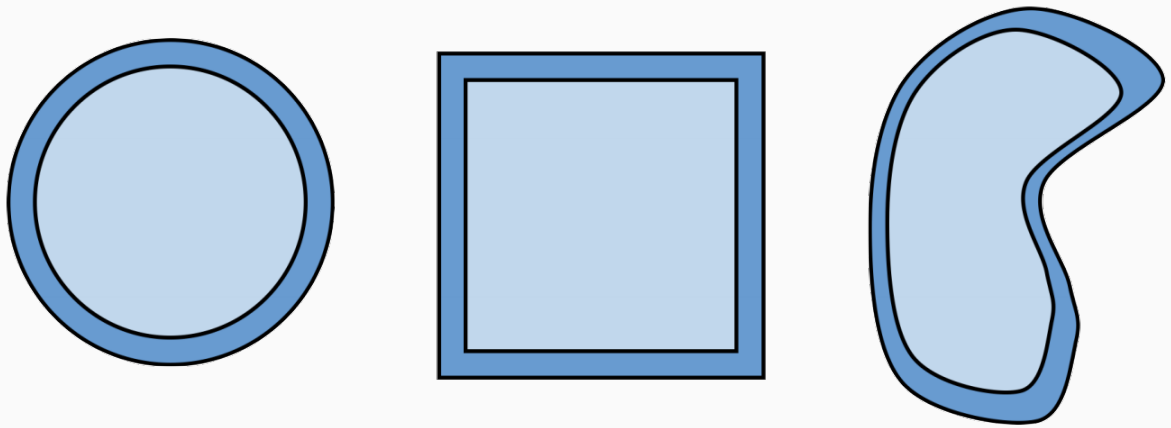


Volume of radius R ball is $\frac{\pi^{d/2}}{(d/2)!} \cdot R^d$.

isoperimetric inequality

All but an $\frac{1}{2}^{\Theta(\epsilon d)}$ fraction of a unit ball's volume is within ϵ of its surface.

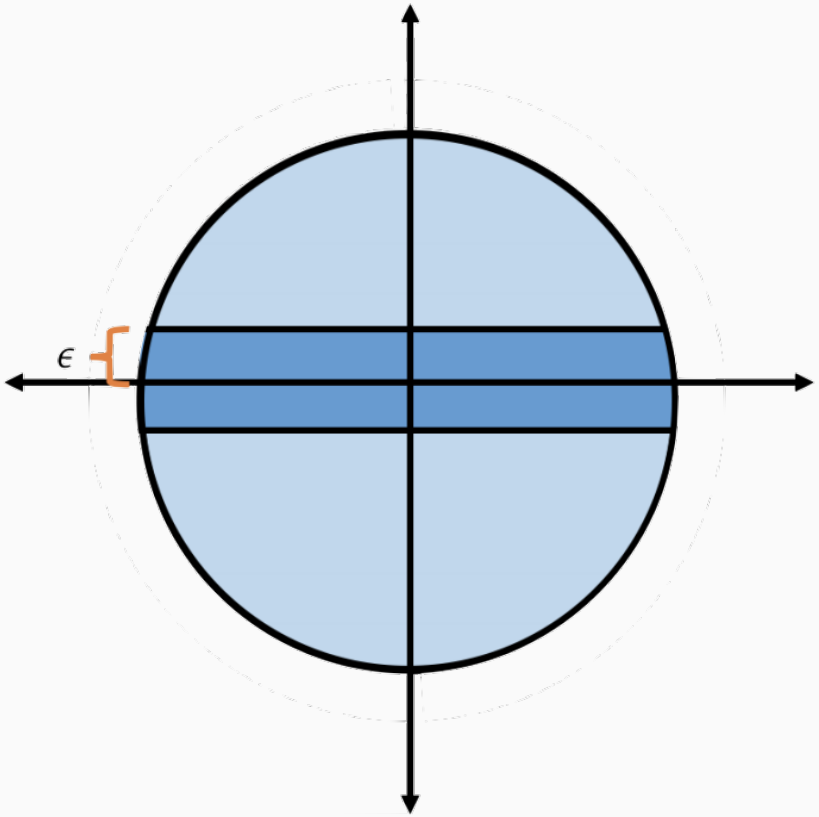
Isoperimetric Inequality: the ball has the maximum surface area/volume ratio of any shape.



- If we randomly sample points from any high-dimensional shape, nearly all will fall near its surface.
- 'All points are outliers.'

slices of the unit ball

What percentage of the volume of $\mathcal{B}_d$ falls within ϵ of its equator?

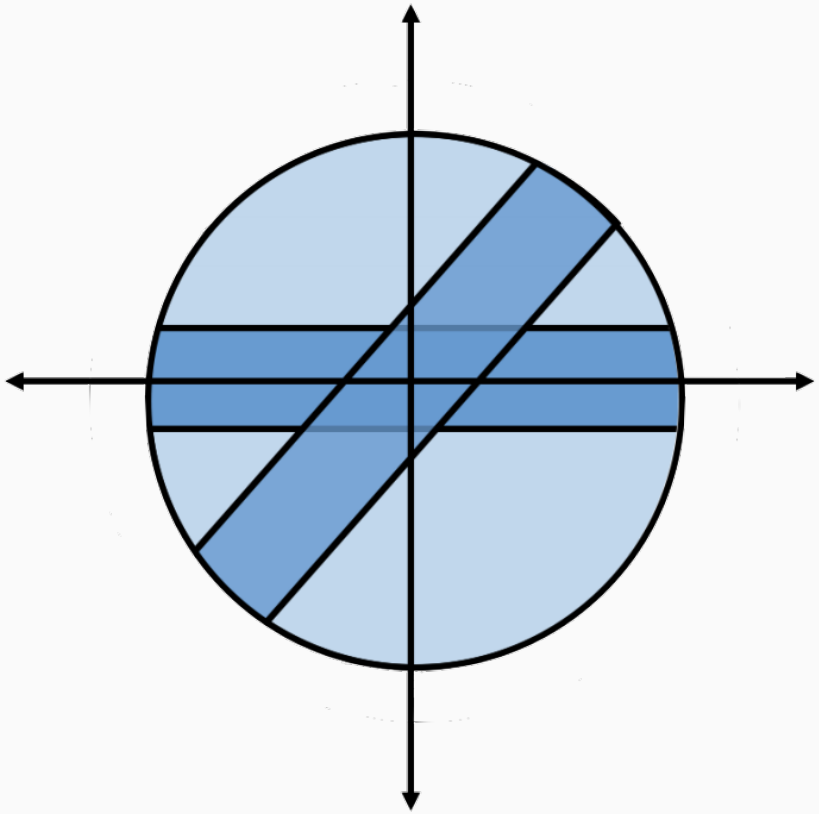


$$S = \{\mathbf{x} \in \mathcal{B}_d : |x_1| \leq \epsilon\}$$

slices of the unit ball

What percentage of the volume of $\mathcal{B}_d$ falls within ϵ of its equator?

Answer: all but a $\frac{1}{2}^{\Theta(\epsilon^2 d)}$ fraction.

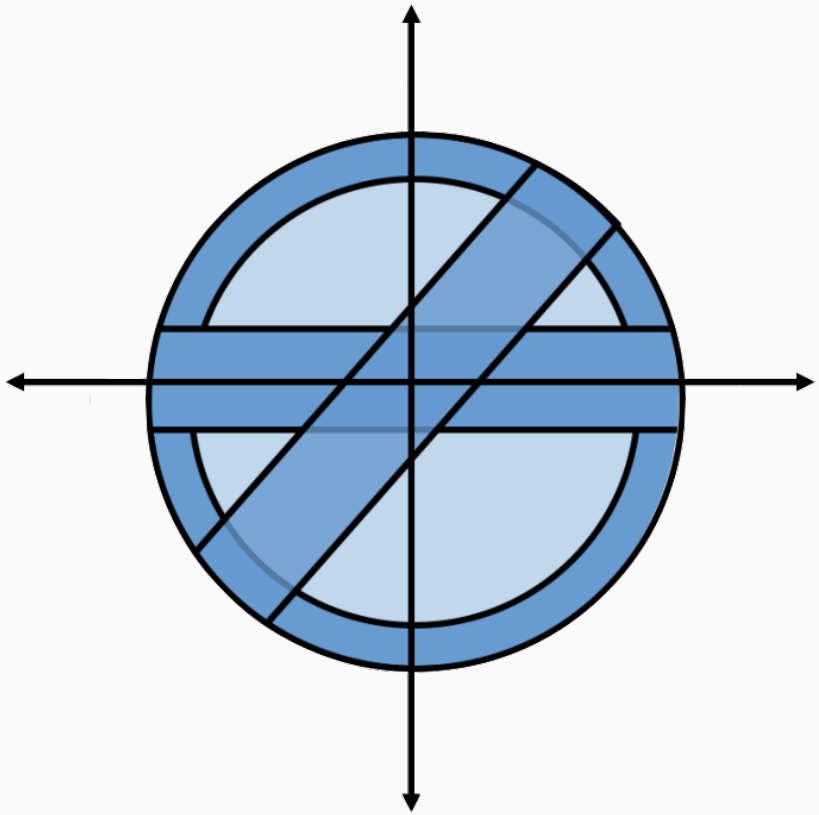


By symmetry, this is true for any equator:

$$S_{\mathbf{t}} = \{\mathbf{x} \in \mathcal{B}_d : \mathbf{x}^T \mathbf{t} \leq \epsilon\}.$$

bizarre shape of unit ball

1. $(1 - \frac{1}{2}^{\Theta(\epsilon d)})$ fraction of volume lies ϵ close to surface.
2. $(1 - \frac{1}{2}^{\Theta(\epsilon^2 d)})$ fraction of volume lies ϵ close to any equator.



High-dimensional ball looks nothing like 2D ball!

concentration at equator

Claim: All but a $\frac{1}{2}^{\Theta(\epsilon^2 d)}$ fraction of the volume of the ball falls within ϵ of its equator.

Equivalent: If we draw a point $\mathbf{x}$ randomly from the unit ball, $|x_1| \leq \epsilon$ with probability $\geq 1 - \frac{1}{2}^{\Theta(\epsilon^2 d)}$.

concentration at equator

Let $\mathbf{w} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. Because $\|\mathbf{x}\|_2 \leq 1$,

$$\Pr\left[|x_1| \leq \epsilon\right] \geq \Pr\left[|w_1| \leq \epsilon\right].$$

Claim: $|w_1| \leq \epsilon$ with probability $\geq 1 - \frac{1}{2}^{\Theta(\epsilon^2 d)}$, which then proves our statement from the previous slide.

How can we generate $\mathbf{w}$, which is a random vector taken from the unit sphere (the surface of the ball)?

important fact in high dimensional geometry

Rotational Invariance of Gaussian distribution: Let $\mathbf{g} \in \mathbb{R}^n$ be a random Gaussian vector $\boldsymbol{g} \sim \mathcal{N}(0, I_n)$, i.e each entry drawn i.i.d. from $\mathcal{N}(0, 1)$. Then $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2$ is distributed uniformly on the unit sphere.

important fact in high dimensional geometry

Rotational Invariance of Gaussian distribution: Let $\mathbf{g} \in \mathbb{R}^n$ be a random Gaussian vector $\boldsymbol{g} \sim \mathcal{N}(0, I_n)$, i.e each entry drawn i.i.d. from $\mathcal{N}(0,1)$. Then $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2$ is distributed uniformly on the unit sphere.

Proof: We can compute the PDF. For a single Gaussian: $p(x) = ce^{-\frac{1}{2}x^2}$. For *independent random variables* X, Y with PDF's $p_X(x), p_Y(y)$, the joint pdf is given by the product $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. Thus:

$$p(\vec{\boldsymbol{g}}) = p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} ce^{-\frac{1}{2}x_i^2} ce^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = ce^{-\frac{1}{2}\|x\|_2^2}$$

PDF $p(\vec{\boldsymbol{g}})$ only depends on norm $\|g\|_2^2$.

concentration at equator

Let $\mathbf{g}$ be a random Gaussian vector and $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2$.

- $\mathbb{E}[\|\mathbf{g}\|_2^2] = \mathbb{E}[\sum_{i=1}^d g_i^2] = \sum_{i=1}^d \mathrm{Var}[g_i] = d$

- $\Pr\left[\|\|\mathbf{g}\|_2^2 \leq \frac{1}{10}\mathbb{E}[\|\mathbf{g}\|_2^2]\right] \leq \frac{1}{2}^{\theta(d)}$

concentration at equator

Let $\mathbf{g}$ be a random Gaussian vector and $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2$.

- $\mathbb{E}[\|\mathbf{g}\|_2^2] = \mathbb{E}[\sum_{i=1}^d g_i^2] = \sum_{i=1}^d \mathrm{Var}[g_i] = d$
- $\Pr\left[\|\|\mathbf{g}\|_2^2 \leq \frac{1}{10}\mathbb{E}[\|\mathbf{g}\|_2^2]\right] \leq \frac{1}{2}^{\theta(d)}$

Theorem (Chernoff Bound)

Let $X_1, X_2, \ldots, X_k$ be independent $\{0,1\}$-valued r.v.s. Set $S = \sum_{i=1}^d X_i$, $\mu = \mathbb{E}[S]$. Then for $0 < \epsilon < 1$:

$$\Pr[S \leq (1-\epsilon)\mu] \leq e^{\frac{-\epsilon^2\mu}{2}}.$$

Proof: $X_i = 1 \iff g_i^2 \geq 1$. Then $\Pr[X_i = 1] > .3$, so $\mathbb{E}[\sum_{i=1}^d X_i] > .3d$. Can set $\epsilon = 1/2$ above.

concentration at equator

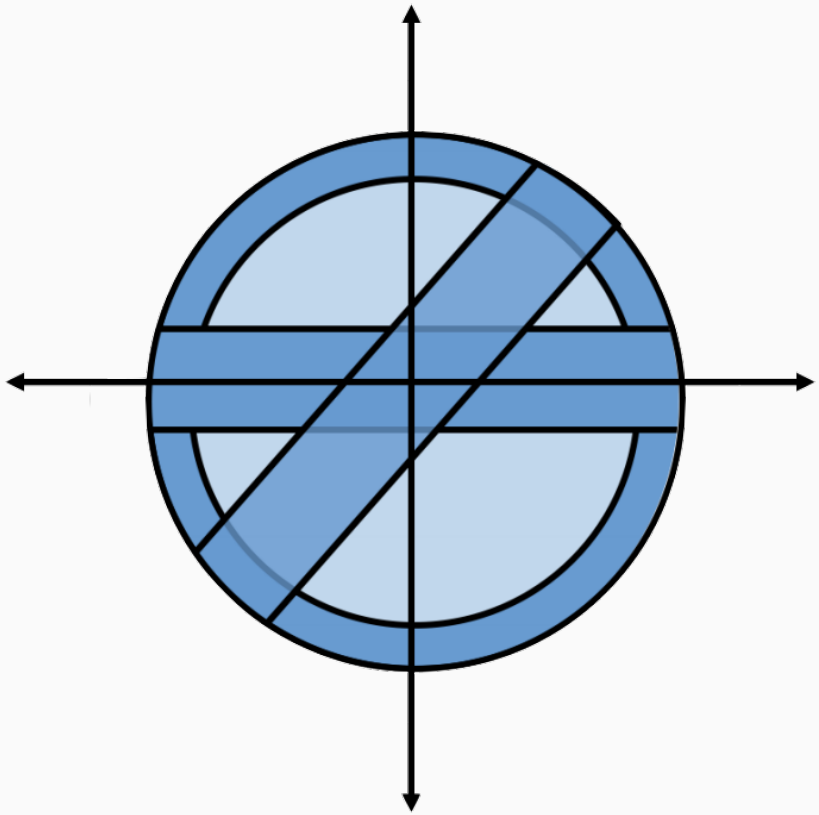
For $1 - \frac{1}{2}^{\theta(d)}$ fraction of vectors $\mathbf{g}$, $\|\mathbf{g}\|_2 \geq \sqrt{d/10}$. Condition on the event that we get a random vector in this set.

$$\begin{aligned}
\Pr\left[|w_1| \leq \epsilon\right] &= \Pr\left[|w_1| \cdot \sqrt{d/10} \leq \epsilon\sqrt{d/10}\right] \\
&\geq \Pr\left[|g_1| \leq \epsilon\sqrt{d/10}\right] \\
&\geq 1 - \frac{1}{2}^{\theta\left((\epsilon\cdot\sqrt{d/10})^2\right)}
\end{aligned}$$

Recall: $\mathbf{w}_1 = \frac{\mathbf{g}_1}{\|\mathbf{g}\|_2}$. So after conditioning, we have $\mathbf{w}_1 \leq \frac{\mathbf{g}_1}{\sqrt{d/10}}$.

bizarre shape of unit ball

1. $(1 - \frac{1}{2}^{\Theta(\epsilon d)})$ fraction of volume lies ϵ close to surface.
2. $(1 - \frac{1}{2}^{\Theta(\epsilon^2 d)})$ fraction of volume lies ϵ close to any equator.



High-dimensional ball looks nothing like 2D ball!

dimensionality reduction

Despite the fact that low-dimensional space behaves nothing like high-dimensional space, next we will demonstrate how to **compress high dimensional vectors to low dimensions,** without distorting distances between them too much.

In particular, a celebrated and simple method known as Johnson-Lindenstrauss Random Projection achieves an *optimal* compression of high-dimensional vectors into low-dimensional space.

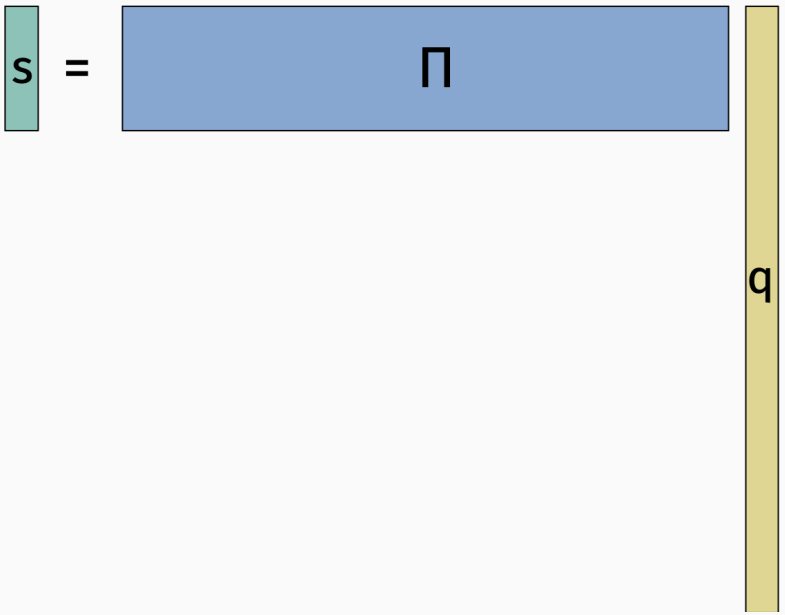
break

euclidean dimensionality reduction

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$ there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

$$(1-\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2 \leq \|\mathbf{\Pi}\mathbf{q}_i - \mathbf{\Pi}\mathbf{q}_j\|_2 \leq (1+\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2.$$



euclidean dimensionality reduction

Please remember: This is equivalent to:

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$ there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

$$(1-\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2^2 \leq \|\mathbf{\Pi q}_i - \mathbf{\Pi q}_j\|_2^2 \leq (1+\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2^2.$$

because for small ϵ, $(1+\epsilon)^2 = 1 + O(\epsilon)$ and $(1-\epsilon)^2 = 1 - O(\epsilon)$.

euclidean dimensionality reduction

And this is equivalent to:

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$ there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

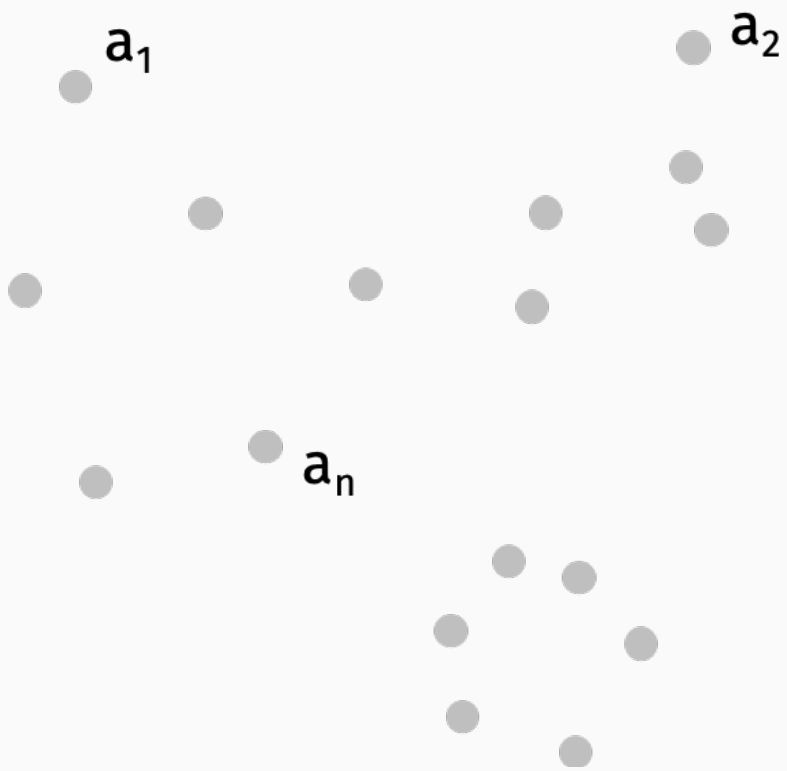
$$(1-\epsilon)\|\mathbf{\Pi q}_i - \mathbf{\Pi q}_j\|_2^2 \leq \|\mathbf{q}_i - \mathbf{q}_j\|_2^2 \leq (1+\epsilon)\|\mathbf{\Pi q}_i - \mathbf{\Pi q}_j\|_2^2.$$

because for small ϵ, $\frac{1}{1+\epsilon} = 1 - O(\epsilon)$ and $\frac{1}{1-\epsilon} = 1 + O(\epsilon)$.

sample application

k-means clustering: Give data points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ to minimize:

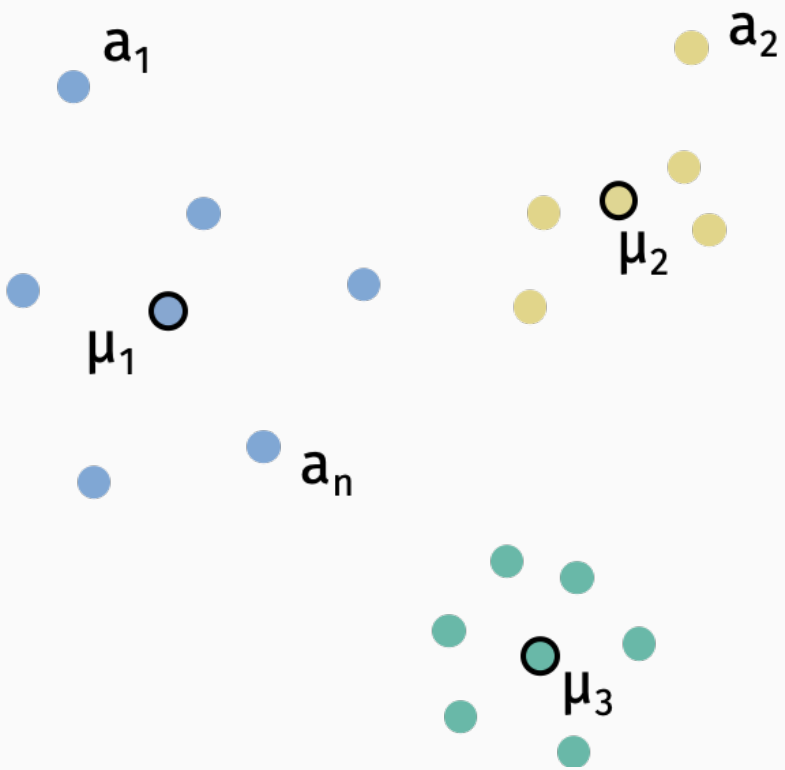
$$Cost(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\boldsymbol{\mu}_j - \mathbf{a}_i\|_2^2$$



sample application

k-means clustering: Give data points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ to minimize:

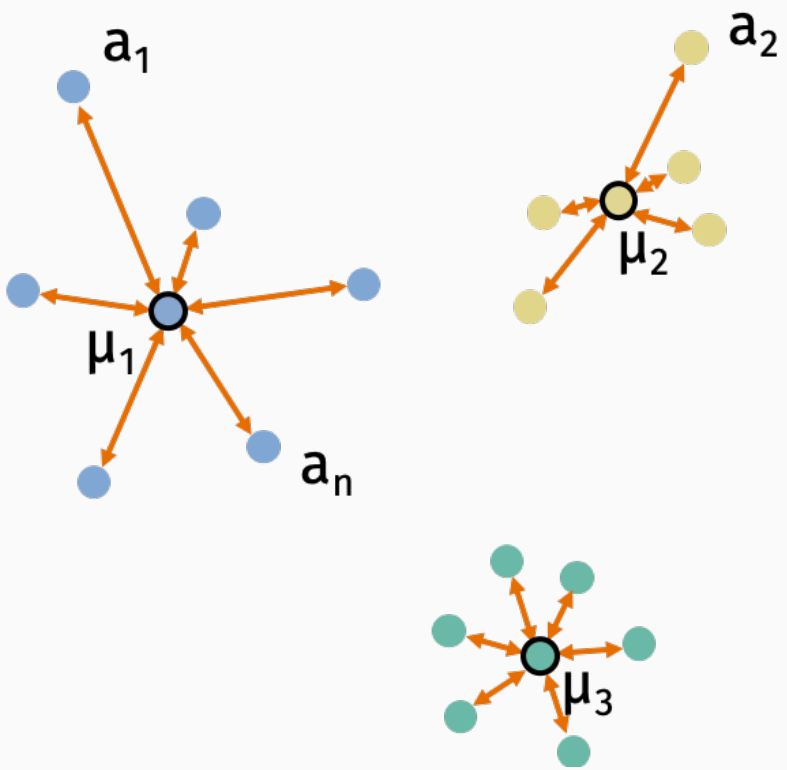
$$Cost(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\boldsymbol{\mu}_j - \mathbf{a}_i\|_2^2$$



sample application

k-means clustering: Give data points $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$, find centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d$ to minimize:

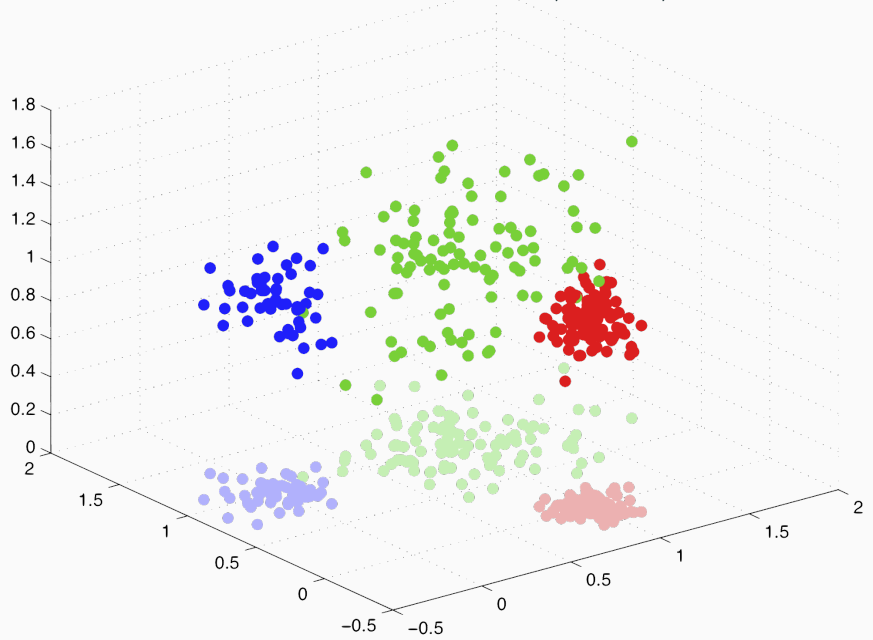
$$Cost(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\boldsymbol{\mu}_j - \mathbf{a}_i\|_2^2$$



k-means clustering

NP hard to solve exactly, but there are many good approximation algorithms. All depend at least linearly on the dimension d.

Approximation scheme: Find clusters $\tilde{C}_1, \ldots, \tilde{C}_k$ for the $k = O\left(\frac{\log n}{\epsilon^2}\right)$ dimension data set $\mathbf{\Pi a}_1, \ldots, \mathbf{\Pi a}_n$.

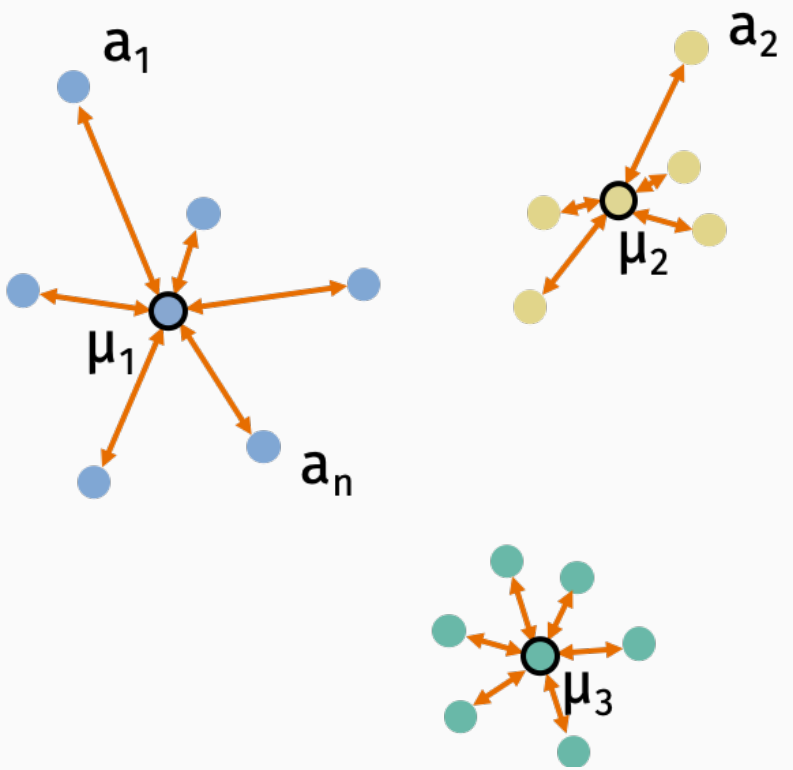


Argue these clusters are near optimal for $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

k-means clustering

Equivalent formulation: Find clusters $C_1, \ldots, C_k \subseteq \{1, \ldots, n\}$ to minimize:

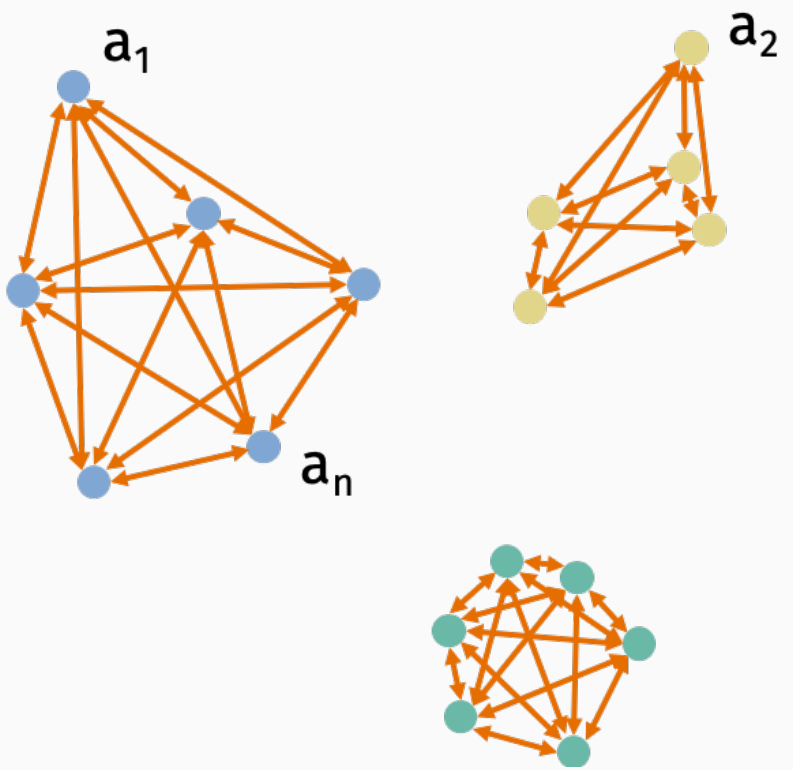
$$Cost(C_1, \ldots, C_k) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|\mathbf{a}_u - \mathbf{a}_v\|_2^2.$$



k-means clustering

Equivalent formulation: Find clusters $C_1, \ldots, C_k \subseteq \{1, \ldots, n\}$ to minimize:

$$Cost(C_1, \ldots, C_k) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|\mathbf{a}_u - \mathbf{a}_v\|_2^2.$$



k-means clustering

$$Cost(C_1, \ldots, C_k) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|\mathbf{a}_u - \mathbf{a}_v\|_2^2$$

$$\widetilde{Cost}(C_1, \ldots, C_k) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|\Pi\mathbf{a}_u - \Pi\mathbf{a}_v\|_2^2$$

Claim: For any clusters $C_1, \ldots, C_k$:

$$(1-\epsilon)Cost(C_1, \ldots, C_k) \leq \widetilde{Cost}(C_1, \ldots, C_k) \leq (1+\epsilon)Cost(C_1, \ldots, C_k)$$

k-means clustering

Suppose we use an approximation algorithm to find clusters $B_1, \ldots, B_k$ such that:

$$\widetilde{Cost}(B_1, \ldots, B_k) \leq (1 + \alpha)\widetilde{Cost}^*$$

Then:

$$\begin{aligned} Cost(B_1, \ldots, B_k) &\leq \frac{1}{1-\epsilon}\widetilde{Cost}(B_1, \ldots, B_k) \\ &\leq (1+\alpha)(1+O(\epsilon))\widetilde{Cost}^* \\ &\leq (1+\alpha)(1+O(\epsilon))(1+\epsilon)Cost^* \\ &= 1 + O(\alpha+\epsilon)Cost^* \end{aligned}$$

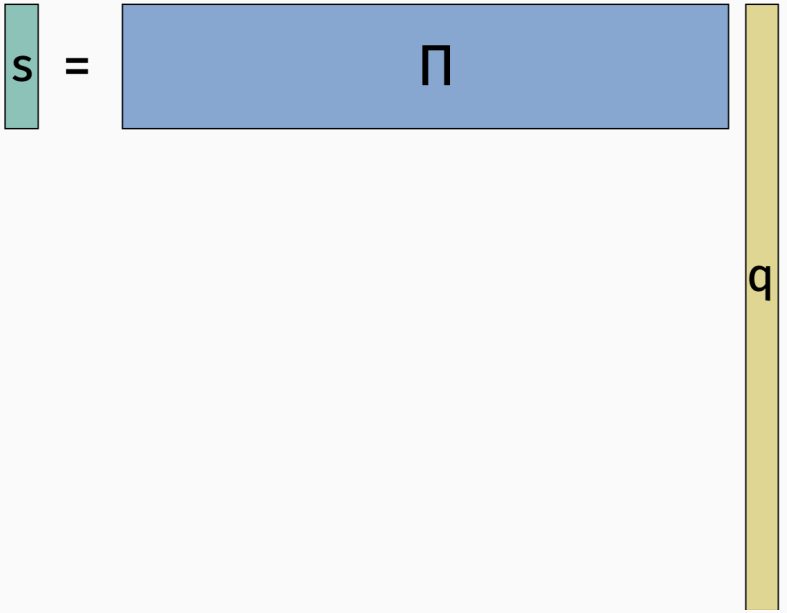
$Cost^* = \min_{C_1,\ldots,C_k} Cost(C_1, \ldots, C_k)$ and
$\widetilde{Cost}^* = \min_{C_1,\ldots,C_k} \widetilde{Cost}(C_1, \ldots, C_k)$

euclidean dimensionality reduction

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$ there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

$$(1-\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2 \leq \|\mathbf{\Pi}\mathbf{q}_i - \mathbf{\Pi}\mathbf{q}_j\|_2 \leq (1+\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2.$$



euclidean dimensionality reduction

Remarkably, Π can be chosen completely at random!

One possible construction: Random Gaussian.

$$\mathbf{\Pi}_{i,j} = \frac{1}{\sqrt{k}}\mathcal{N}(0,1)$$

The map $\mathbf{\Pi}$ is **oblivious to the data set**. This stands in contrast to e.g. PCA, amoung other differences.

[Indyk, Motwani 1998] [Arriage, Vempala 1999] [Achlioptas 2001] [Dasgupta, Gupta 2003].

Many other possible choices suffice – you can use random $\{+1, -1\}$ variables, sparse random matrices, pseudorandom Π. Each with different advantages.

randomized jl constructions

Let $\mathbf{\Pi} \in \mathbb{R}^{k \times d}$ be chosen so that each entry equals $\frac{1}{\sqrt{k}}\mathcal{N}(0,1)$.

... or each entry equals $\frac{1}{\sqrt{k}} \pm 1$ with equal probability.

-2.1384	2.9080	-0.3538	0.0229	0.5201	-0.2938	-1.3320	-1.3617	-0.1952
-0.8396	0.8252	-0.8236	-0.2620	-0.0200	-0.8479	-2.3299	0.4550	-0.2176
1.3546	1.3790	-1.5771	-1.7502	-0.0348	-1.1201	-1.4491	-0.8487	-0.3031
-1.0722	-1.0582	0.5080	-0.2857	-0.7982	2.5260	0.3335	-0.3349	0.0230
0.9610	-0.4686	0.2820	-0.8314	1.0187	1.6555	0.3914	0.5528	0.0513
0.1240	-0.2725	0.0335	-0.9792	-0.1332	0.3075	0.4517	1.0391	0.8261
1.4367	1.0984	-1.3337	-1.1564	-0.7145	-1.2571	-0.1303	-1.1176	1.5270
-1.9609	-0.2779	1.1275	-0.5336	1.3514	-0.8655	0.1837	1.2607	0.4669
-0.1977	0.7015	0.3502	-2.0026	-0.2248	-0.1765	-0.4762	0.6601	-0.2097
-1.2078	-2.0518	-0.2991	0.9642	-0.5890	0.7914	0.8620	-0.0679	0.6252

```
>> Pi = randn(m,d);
>> s = (1/sqrt(m))*Pi*q;
```

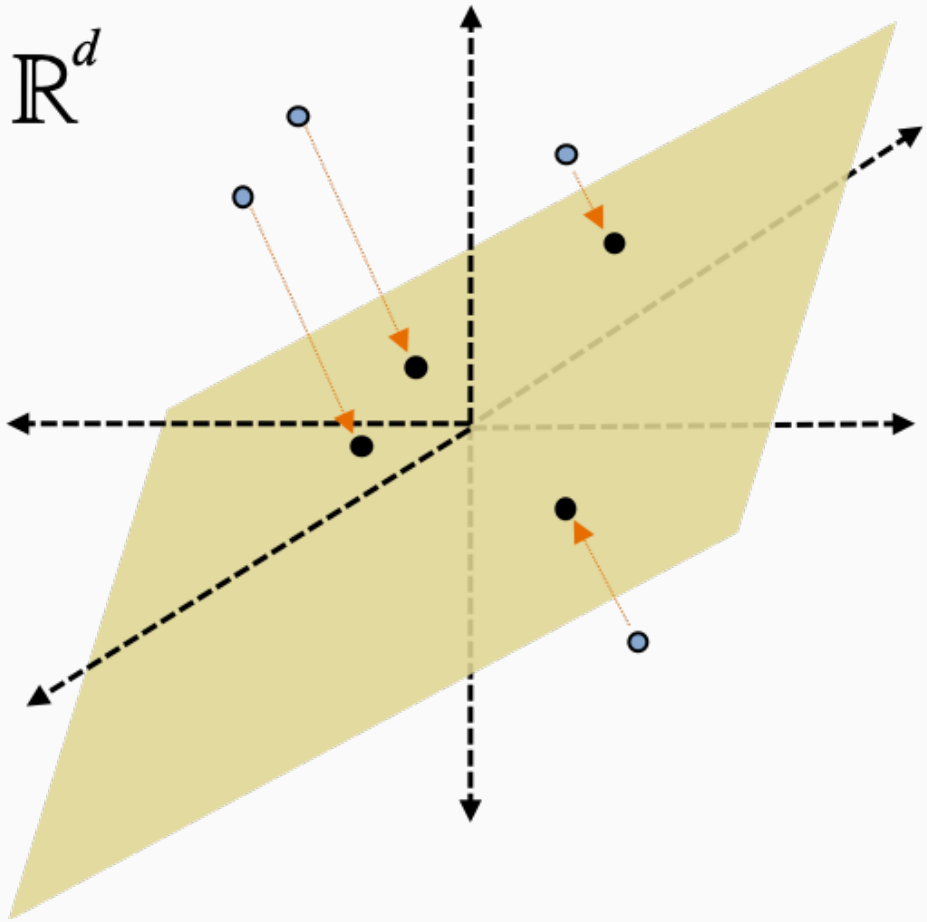
1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	1	1	-1
1	1	1	-1	1	-1	-1	-1	1	1	1	1	-1	1	-1	-1	-1
1	1	-1	-1	-1	1	-1	-1	1	1	-1	1	-1	1	-1	1	-1
-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1
1	-1	-1	1	-1	1	1	-1	-1	-1	1	-1	-1	-1	1	1	1
1	1	-1	1	1	-1	1	-1	1	-1	1	-1	1	1	1	-1	-1
-1	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	1	-1	-1	1	1
-1	-1	1	1	1	1	-1	-1	1	-1	1	1	1	-1	1	-1	1
-1	1	-1	1	-1	1	1	-1	-1	1	-1	1	-1	-1	1	-1	1

```
>> Pi = 2*randi(2,m,d)-3;
>> s = (1/sqrt(m))*Pi*q;
```

A random orthogonal matrix also works. I.e. with $\Pi\Pi^T = \mathbf{I}_{k \times k}$.

For this reason, the JL operation is often called a "random projection", even though it technically isn't a projection when entries are i.i.d.

random projection



Intuitively, close points will remain close after projection, and far points will remain far.

Intermediate result:

Lemma (Distributional JL Lemma)

Let $\mathbf{\Pi} \in \mathbb{R}^{k \times d}$ *be chosen so that each entry equals* $\frac{1}{\sqrt{k}}\mathcal{N}(0,1)$, *where* $\mathcal{N}(0,1)$ *denotes a standard Gaussian random variable.*

If we choose $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, *then for any vector* $\mathbf{x}$, *with probability* $(1-\delta)$*:*

$$(1-\epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Pi x}\|_2^2 \leq (1+\epsilon)\|\mathbf{x}\|_2^2$$

Given this lemma, how do we prove the traditional Johnson-Lindenstrauss lemma?

JL from distributional JL

We have a set of vectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Fix $i, j \in 1, \ldots, n$.

Let $\mathbf{x} = \mathbf{q}_i - \mathbf{q}_j$. By linearity, $\mathbf{\Pi x} = \mathbf{\Pi}(\mathbf{q}_i - \mathbf{q}_j) = \mathbf{\Pi q}_i - \mathbf{\Pi q}_j$.

By the Distributional JL Lemma, with probability $1 - \delta$,

$$(1-\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2 \leq \|\mathbf{\Pi q}_i - \mathbf{\Pi q}_j\|_2 \leq (1+\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2.$$

Finally, set $\delta = \frac{1}{n^2}$. Since there are $< n^2$ total i, j pairs, by a union bound we have that with probability $9/10$, the above will hold for all i, j, as long as we compress to:

$$k = O\left(\frac{\log(1/(1/n^2))}{\epsilon^2}\right) = O\left(\frac{\log n}{\epsilon^2}\right) \text{ dimensions.} \quad \square$$

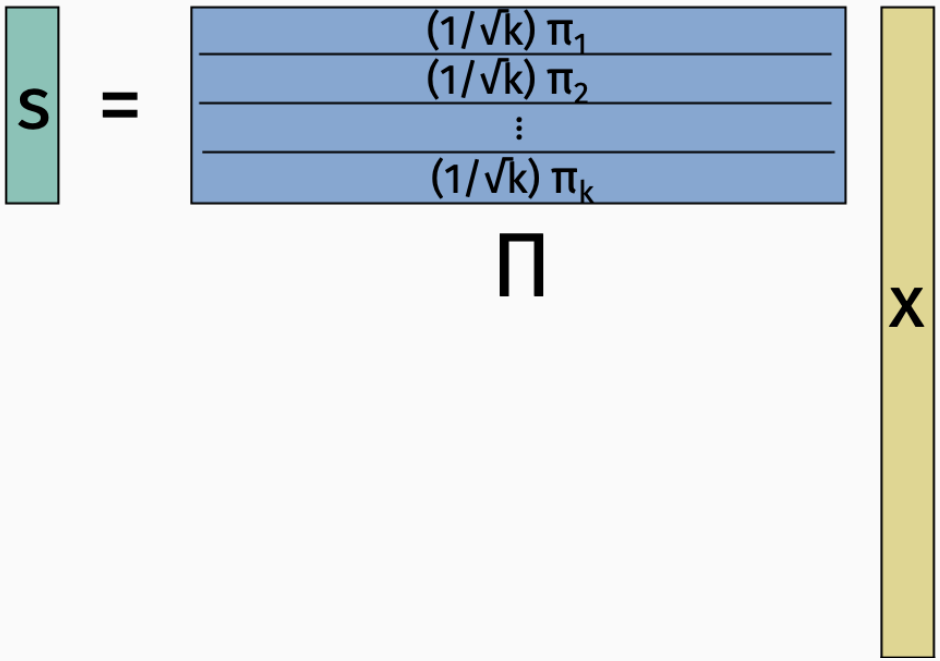
proof of distributional jl

Want to argue that, with probability $(1 - \delta)$,

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq |\mathbf{\Pi x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2$$

Claim: $\mathbb{E}\|\mathbf{\Pi x}\|_2^2 = \|\mathbf{x}\|_2^2$.

Some notation:



So each π_i contains $\mathcal{N}(0, 1)$ entries.

proof of distributional jl

$$\|\mathbf{\Pi x}\|_2^2 = \sum_i^k \mathbf{s}(i)^2 = \sum_i^k \left(\frac{1}{\sqrt{k}}\langle \boldsymbol{\pi}_i, \mathbf{x}\rangle\right)^2 = \frac{1}{k}\sum_i^k \left(\langle \boldsymbol{\pi}_i, \mathbf{x}\rangle\right)^2$$

$$\mathbb{E}\left[\|\mathbf{\Pi x}\|_2^2\right] = \frac{1}{k}\sum_i^k \mathbb{E}\left[\left(\langle \boldsymbol{\pi}_i, \mathbf{x}\rangle\right)^2\right]$$

$$= \mathbb{E}\left[\left(\langle \boldsymbol{\pi}_i, \mathbf{x}\rangle\right)^2\right]$$

Goal: Prove $\mathbb{E}\|\mathbf{\Pi x}\|_2^2 = \|\mathbf{x}\|_2^2$.

proof of distributional jl

$$\langle \boldsymbol{\pi}_i, \mathbf{x} \rangle = Z_1 \cdot \mathbf{x}(1) + Z_2 \cdot \mathbf{x}(2) + \ldots + Z_d \cdot \mathbf{x}(d)$$

where each $Z_1, \ldots, Z_d$ is a standard normal $\mathcal{N}(0, 1)$ random variable.

This implies that $Z_i \cdot \mathbf{x}(i)$ is a normal $\mathcal{N}(0, \mathbf{x}(i)^2)$ random variable.

Goal: Prove $\mathbb{E}\|\boldsymbol{\Pi}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$. Established: $\mathbb{E}\|\boldsymbol{\Pi}\mathbf{x}\|_2^2 = \mathbb{E}\left[(\langle \boldsymbol{\pi}_i, \mathbf{x} \rangle)^2\right]$

What type of random variable is $\langle \pi_i, \mathbf{x} \rangle$?

Fact (Stability of Gaussian random variables)

$$\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$\begin{aligned}\langle \pi_i, \mathbf{x} \rangle &= \mathcal{N}(0, \mathbf{x}(1)^2) + \mathcal{N}(0, \mathbf{x}(2)^2) + \ldots + \mathcal{N}(0, \mathbf{x}(d)^2) \\ &= \mathcal{N}(0, \|\mathbf{x}\|_2^2).\end{aligned}$$

So $\mathbb{E}\|\mathbf{\Pi x}\|_2^2 = \mathbb{E}\left[(\langle \pi_i, \mathbf{x} \rangle)^2\right] = \|\mathbf{x}\|_2^2$, as desired.

proof of distributional jl

Want to argue that, with probability $(1 - \delta)$,

$$(1 - \epsilon)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Pi x}\|_2^2 \leq (1 + \epsilon)\|\mathbf{x}\|_2^2$$

1. $\mathbb{E}\|\mathbf{\Pi x}\|_2^2 = \|\mathbf{x}\|_2^2$.
2. Need to use a concentration bound.

$$\|\mathbf{\Pi x}\|_2^2 = \frac{1}{k}\sum_{i=1}^{k}(\langle \boldsymbol{\pi}_i, \mathbf{x}\rangle)^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\mathcal{N}(0, \|\mathbf{x}\|_2^2)\right)^2$$

"Chi-squared random variable with k degrees of freedom."

concentration of chi-squared random variables

Lemma

Let Z be a Chi-squared random variable with k degrees of freedom.

$$\Pr[|\mathbb{E}[Z] - Z| \geq \epsilon \mathbb{E}[Z]] \leq 2e^{-k\epsilon^2/8}$$

Goal: Prove $\|\mathbf{\Pi x}\|_2^2$ concentrates within $1 \pm \epsilon$ of its expectation, which equals $\|\mathbf{x}\|_2^2$.

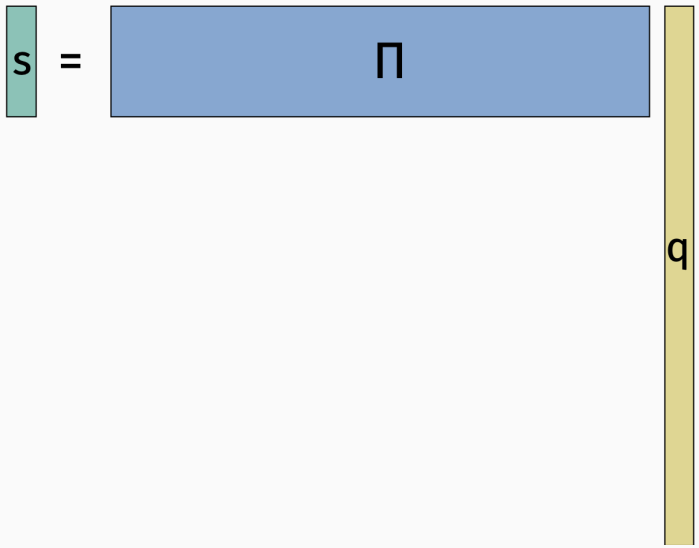
JL Lemma

Putting together the pieces, we have just proven:

Lemma (Johnson-Lindenstrauss, 1984)

For any set of n data points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$ there exists a linear map $\Pi : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

$$(1-\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2 \leq \|\mathbf{\Pi}\mathbf{q}_i - \mathbf{\Pi}\mathbf{q}_j\|_2 \leq (1+\epsilon)\|\mathbf{q}_i - \mathbf{q}_j\|_2.$$



Limits of Dimensionality Reduction

If high dimensional geometry is so different from low-dimensional geometry, why is dimensionality reduction possible? Doesn't Johnson-Lindenstrauss tell us that high-dimensional geometry can be approximated in low dimensions?

Johnson-Lindenstrauss preserves only distances between the n points $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^d$, not all points in $\mathbb{R}^d$!

Limits of Dimensionality Reduction

Hard case: $\vec{0}$ and mutually orthogonal unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ (e.g. $\boldsymbol{x}_i = (0, \ldots, 0, 1, 0, \ldots, 0) = e_i$):

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = 2 \qquad \text{for all } i, j.$$

From our result earlier, in $O(\log n/\epsilon^2)$ dimensions, there exists $2^{O(\epsilon^2 \cdot \log n/\epsilon^2)} \geq n$ unit vectors $\{a_i\}$ that are ϵ-nearly orthogonal, i.e. $\langle a_i, a_j \rangle \leq \epsilon$.

$O(\log n/\epsilon^2)$ is <u>just enough</u> dimensions to fit n ϵ-nearly orthogonal unit vectors.

And now: Linear Sketching

Linear Sketches

Definition

Given any high-dimensional vector $f \in \mathbb{R}^n$, a *linear sketch* is a matrix-vector product $\mathbf{S}f \in \mathbb{R}^k$, $k \ll n$, where $\mathbf{S} \in \mathbb{R}^{k \times n}$ is called a *sketching matrix*.

- Usually, $\mathbf{S}$ is a *random* matrix, who entries can be easily stored (e.g., $\mathbf{S}$ is generated by 2-wise independent hash functions).
- **Goal:** from knowledge only of $\mathbf{S}$ and $\mathbf{S}f$, approximation some function of f (i.e. $\|f\|_2^2$, find heavy hitters $f_i > \epsilon\|f\|_2$, ect.)
- **Benefits:** $\mathbf{S}$ much smaller to store than f, can be maintain in a stream!

We have seen these before: Count-Min, Count-Sketch, Johnson-Lindenstrauss...

Main Benefit of linear sketches:

Linear sketches are Linear!

$$\mathbf{S}x + \mathbf{S}y = \mathbf{S}(x + y)$$

How is this useful for streaming? Suppose we have $\mathbf{S}f$ stored, and f gets an update $(i, \Delta) \in [n] \times \mathbb{Z}$.

$$\mathbf{S}f \leftarrow \mathbf{S}f +$$

Main Benefit of linear sketches:

Linear Sketches are linear!

$$\mathbf{S}x + \mathbf{S}y = \mathbf{S}(x + y)$$

How is this useful for streaming? Suppose we have $\mathbf{S}f$ stored, and f gets an update $(i, \Delta) \in [n] \times \mathbb{Z}$.

$$\mathbf{S}f \leftarrow \mathbf{S}f + \mathbf{S}_i \cdot \Delta$$

Also very useful for **distributed computation**. Machines $m_1, \ldots, m_k$ have data $x_1, ..., x_k \in \mathbb{R}^n$. Each machine can sketch $\mathbf{S}x_i \in \mathbb{R}^k$, send to aggregator, which computes $\sum_i \mathbf{S}x_i = \mathbf{S}(\sum_i x_i)$.

Johnson Lindenstrauss for Streaming

JL-Lemma says that there exists a linear sketch $\mathbf{S} \in \mathbb{R}^{k \times n}$, where $k = \frac{1}{\epsilon^2} \log(1/\delta)$, such that for any fixed $f \in \mathbb{R}^n$, we have $\|\mathbf{S}f\|_2^2 = (1 \pm \epsilon)\|f\|_2^2$ with prob $1 - \delta$. So:

Theorem

There is a turnstile streaming algorithm (in the random oracle model) which estimates the ℓ_2 norm $\|f\|_2^2$ of the frequency vector to $(1 \pm \epsilon)$-multiplicative error with probability $> 1 - \delta$, using $O(\frac{1}{\epsilon^2} \log(1/\delta))$ bits of space.

Recall: Random oracle model of streaming means we can store random bits for free.

Count-Sketch is a Linear Sketch

Linear Sketches

Linear Sketches

Linear Sketches

Linear Sketches