CS-GY 6763: Lecture 4
High Dimensional Geometry, the Johnson-Lindenstrauss Lemma

NYU Tandon School of Engineering, Prof. Christopher Musco

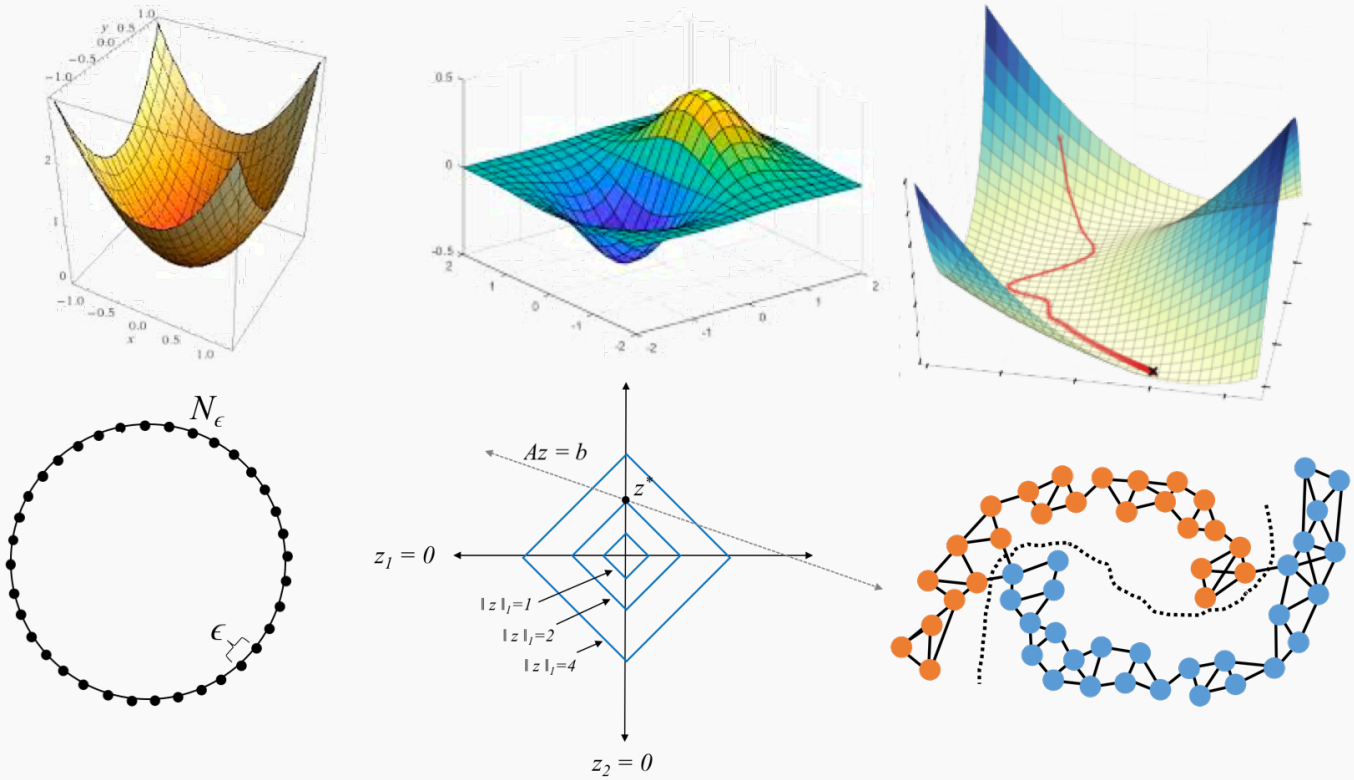
UNIFYING THEME OF THE COURSE

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

- High-dimensional similarity search.
- Iterative methods for optimizing functions in high-dimensions.
- SVD + low-rank approximation to find and visualize low-dimensional structure.
- Convert large graphs to high-dimensional vector data to uncover interesting things.

HIGH DIMENSIONAL IS NOT LIKE LOW DIMENSIONAL

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

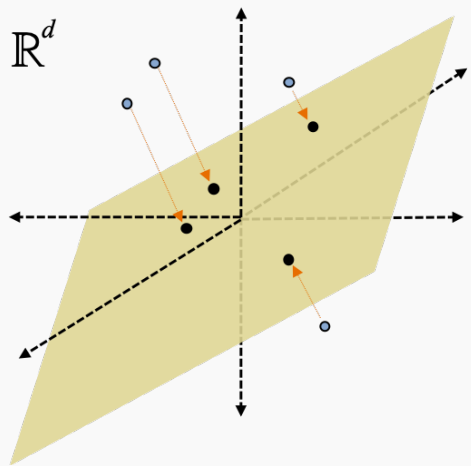


First part of lecture: Prove that high-dimensional space looks very different from low-dimensional space. These images are rarely very informative!

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 (next class) .

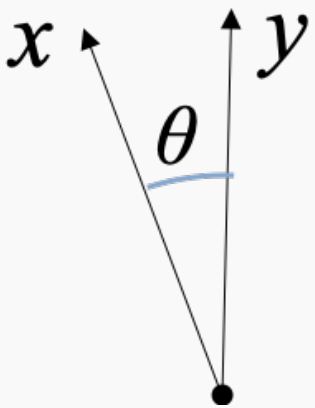


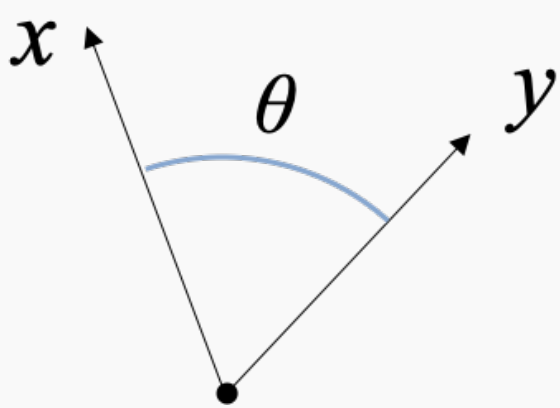
First part of lecture should help you understand the potential and limitations of these methods.

ORTHOGONAL VECTORS

Recall the inner product between two d dimensional vectors:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_{j=1}^{d} x[j] y[j]$$

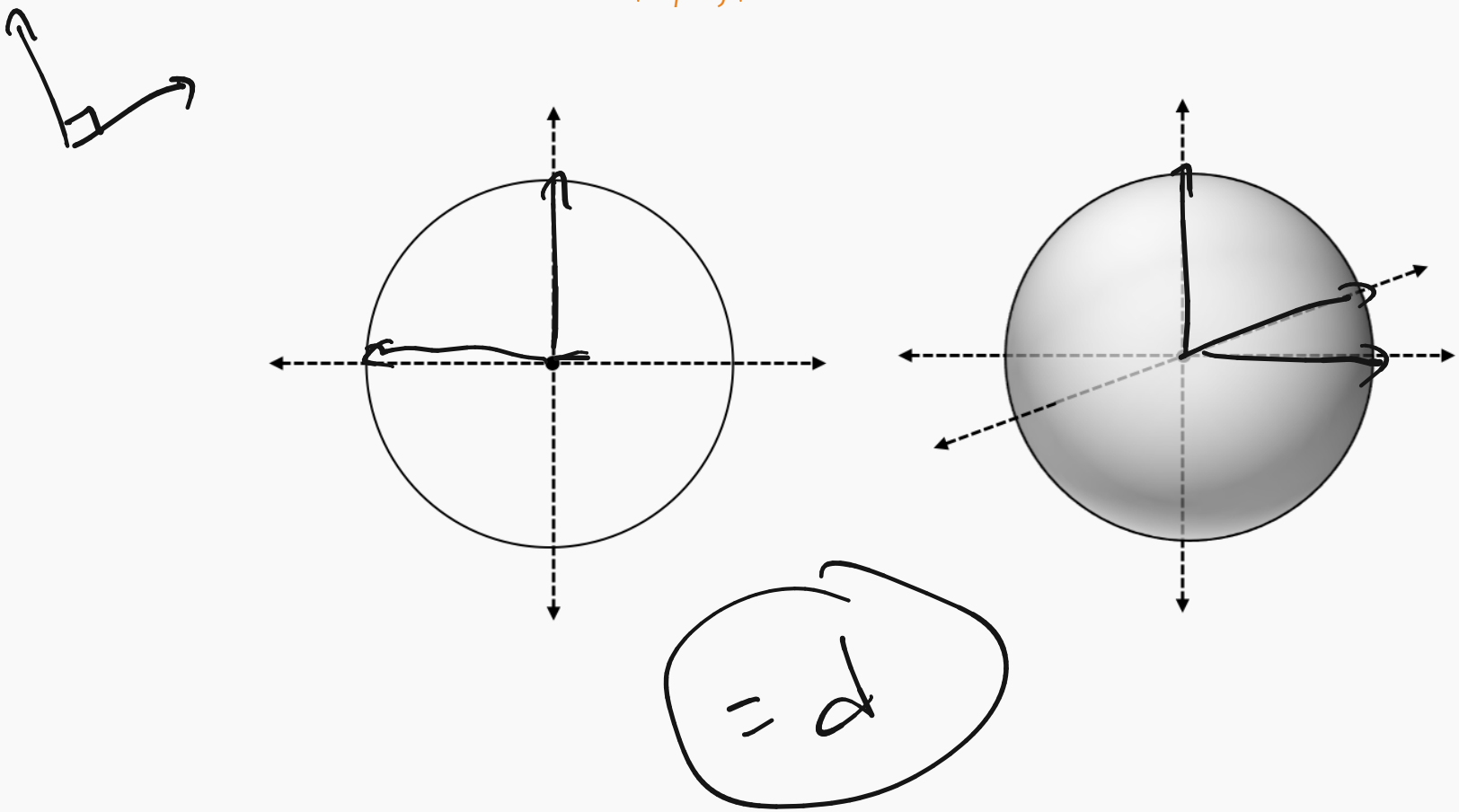




$$\langle \mathbf{x}, \mathbf{y} \rangle = \cos(\theta) \cdot \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2$$

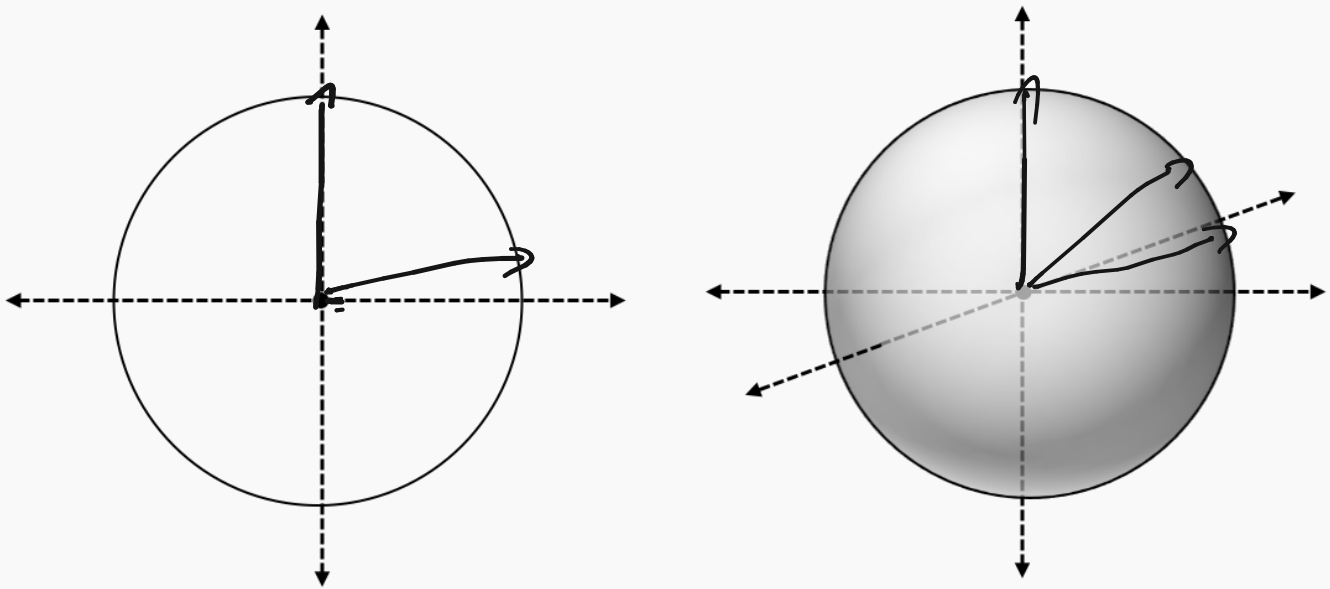
ORTHOGONAL VECTORS

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

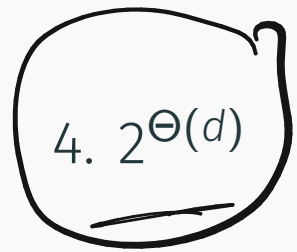
What is the largest set **nearly orthogonal** unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ in d dimensions. I.e., with inner product $|\mathbf{x}_i^T \mathbf{x}_j| \leq \epsilon$ for all i, j. Consider the case when ϵ is a constant. E.g. $\epsilon = 1/10$.



ORTHOGONAL VECTORS

What is the largest set **nearly orthogonal** unit vectors $\mathbf{x}_1, \ldots, \mathbf{x}_t$ in d dimensions. I.e., with inner product $|\mathbf{x}_i^T \mathbf{x}_j| \leq \epsilon$ for all i, j. Consider the case when ϵ is a constant. E.g. $\epsilon = 1/10$.

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



ORTHOGONAL VECTORS

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

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

$2^{\Theta((1/\sqrt{d})^2 \cdot d)} = 2^{O(1)}$

Proof strategy: Use the Probabilistic Method! For $t = 2^{\Theta(\epsilon^2 d)}$, 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.

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

Proof: Let $x_1, \ldots, x_t$ all have independent random entries, each set to $\pm\frac{1}{\sqrt{d}}$ with equal probability.

$x_i = (+1/\sqrt{d} \;\; -1/\sqrt{d} \;\; \cdots \;\; 1/\sqrt{d})$

- $\|x_i\|_2^2 = \sum_{j=1}^{d} x_i[j]^2 = \sum_{j=1}^{d} 1/d = 1.$
- $\mathbb{E}[x_i^T x_j] = \mathbb{E}\left[\sum_{k=1}^{d} x_i[k]\, x_j[k]\right] = \sum_{k=1}^{d} \mathbb{E}[x_i[k]\, x_j[k]] = \sum_{k=1}^{d} 0$

  $= 0$

  want close to 0
- $\mathrm{Var}[x_i^T x_j] = \mathrm{Var}\left(\sum_{k=1}^{d} x_i[k]\, x_j[k]\right) = \sum_{k=1}^{d} \mathrm{Var}[x_i[k]\, x_j[k]]$

$x_i[k]\, x_j[k] = \begin{cases} 1/d & \text{w.p. } 1/2 \\ -1/d & \text{w.p. } 1/2 \end{cases}$

$= \sum_{k=1}^{d} 1/d^2 = \boxed{1/d}$

INFORMAL PROOF

Let $Z = \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{d} C_i$ where each C_i is random $\pm \frac{1}{d}$.

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

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

$\mathbb{E}Z = 0$, $\alpha = \varepsilon$

$\sigma = 1/\sqrt{d}$

$d = \varepsilon \cdot \sqrt{d}$

w.p. $1 - O(e^{-\varepsilon^2 d})$, $|x_i^T x_j| < \varepsilon$.

$x_1, \ldots, x_t$ $\Pr(|x_i^T x_j| \leq \varepsilon$ for all $i,j \in 1,\ldots t] \leq 1 - \binom{t}{2} \cdot e^{-\varepsilon^2 d}$

How large can I set t so that this is ≥ 0, $1 - O(t^2 \cdot e^{-\varepsilon^2 d})$

By a union bound, we can claim that the above holds for all pairs in a set of size: $t^2 \cdot e^{-\varepsilon^2 d} = 1$ $t^2 = e^{\varepsilon^2 d}$ $t = e^{\varepsilon^2 d/2} = e^{\Theta(\varepsilon^2 d)}$

Use an exponential concentration inequality!

Theorem (Chernoff Bound)

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

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

Does not immediately apply because we have random variables that are $\pm 1/d$, not $0, 1$.

Common trick: shift and scale to transform to the binary case.

FORMAL PROOF

$$\mathbf{x}_i^T \mathbf{x}_j = Z = \sum_{i=1}^{d} C_i = \frac{2}{d} \sum_{i=1}^{d} \frac{d}{2} \cdot C_i$$

$$= \frac{2}{d} \cdot \left(\sum_{i=1}^{d} B_i - 1/2 \right)$$

$$= \frac{2}{d} \cdot \left(-\frac{d}{2} + \sum_{i=1}^{d} B_i \right)$$

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

FORMAL PROOF

$$x_i^T x_j = Z = \frac{2}{d} \cdot \left(-\frac{d}{2} + \sum_{i=1}^{d} B_i \right)$$

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

$$\begin{aligned}
\Pr[|Z| > \epsilon] &= \Pr\left[\sum_{i=1}^{d} B_i \geq \frac{d}{2} + \frac{\epsilon d}{2}\right] + \Pr\left[\sum_{i=1}^{d} B_i \leq \frac{d}{2} - \frac{\epsilon d}{2}\right] \\
&= \Pr\left[\sum_{i=1}^{d} B_i \geq (1+\epsilon)\mathbb{E}\left[\sum_{i=1}^{d} B_i\right]\right] \\
&\quad + \Pr\left[\sum_{i=1}^{d} B_i \leq (1-\epsilon)\mathbb{E}\left[\sum_{i=1}^{d} B_i\right]\right]
\end{aligned}$$

Theorem (Chernoff Bound)

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

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

$S = \sum B_i$

Apply with $X_1, \ldots, X_d = B_1, \ldots, B_d$:

$p_1, \ldots, p_d = 1/2$

$$\Pr[|S - \mathbb{E}[S]| \geq \epsilon \mathbb{E}[S]] \leq 2e^{-\epsilon^2 (d/2)/3}$$

$$= 2e^{-\epsilon^2 d/6}$$

$|Z| \leq \epsilon$ w.p. $2e^{-\epsilon^2 d/6}$

Conclusion from Chernoff bound:

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

By a union bound:

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

Can set $t = 2^{O(\epsilon^2 d)}$ and this will be positive.

$\Rightarrow \exists$ set of $t = 2^{O(\epsilon^2 d)}$ unit vectors that ϵ-orthogonal.

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 (and roughly equidistant) in high-dimensions.

$\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2\langle x_j, x_i\rangle$

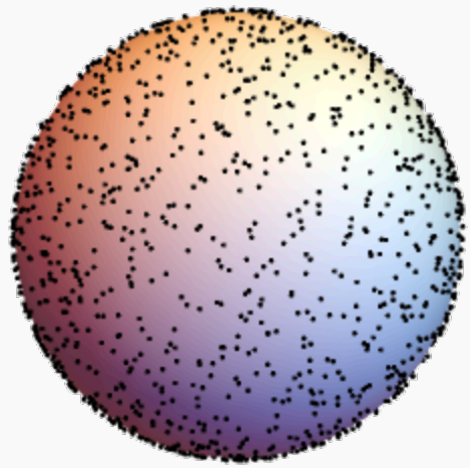
$1 \qquad 1$

$= 2 - 2\langle x_j - x_i \rangle.$

$\downarrow$

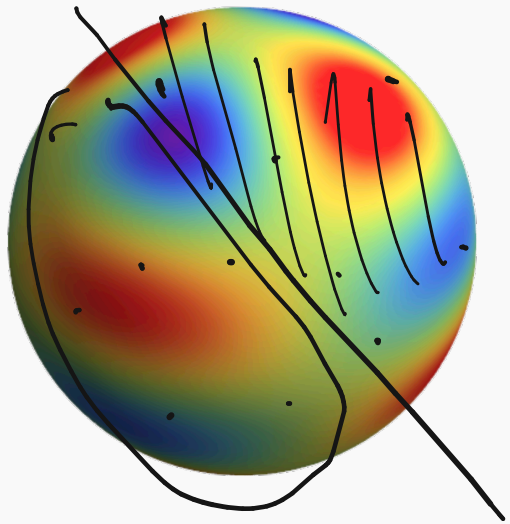
$\leq \epsilon$ for all x_i, x_j

$2^{\Theta(d)}$



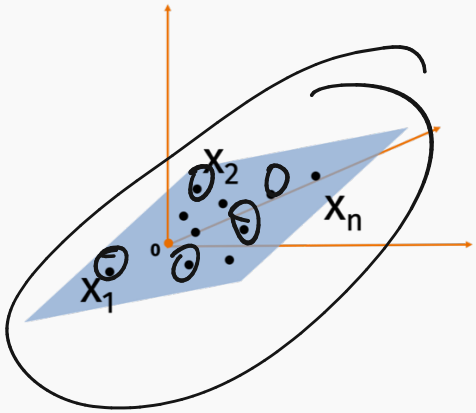
CURSE OF DIMENSIONALITY

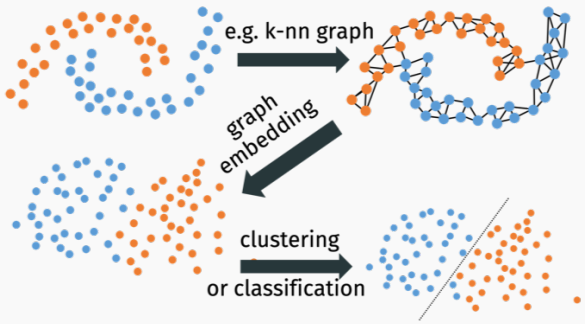
Curse of dimensionality: Suppose we want to use e.g. k-nearest neighbors to learn a function or classify points in $\mathbb{R}^d$. If our data distribution is truly random, we typically need an exponential amount of data.

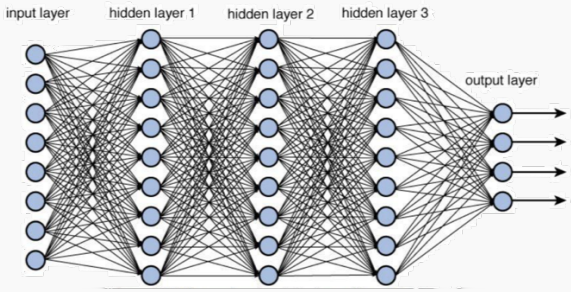


The existence of lower dimensional structure is our data is often the only reason we can hope to learn.

Low-dimensional structure.







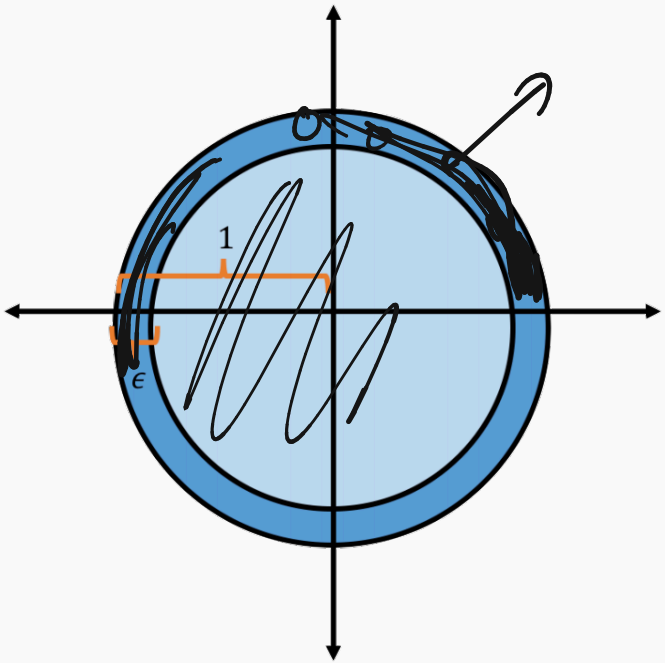
For example, data lies on low-dimensional subspace, or does so after transformation. Or function can be represented by a restricted class of functions, like neural net with specific architecture.

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?



$$\frac{c \cdot 1^d - c \cdot (1-\epsilon)^d}{c \cdot 1^d} = 1 - (1-\epsilon)^d \simeq 1/e$$

$$= 1 - \left((1-\epsilon)^{1/\epsilon}\right)^{\epsilon d} = 1 - \left(\frac{1}{e}\right)^{\epsilon d}$$

$$= 1 - 2^{o(-\epsilon d)}$$

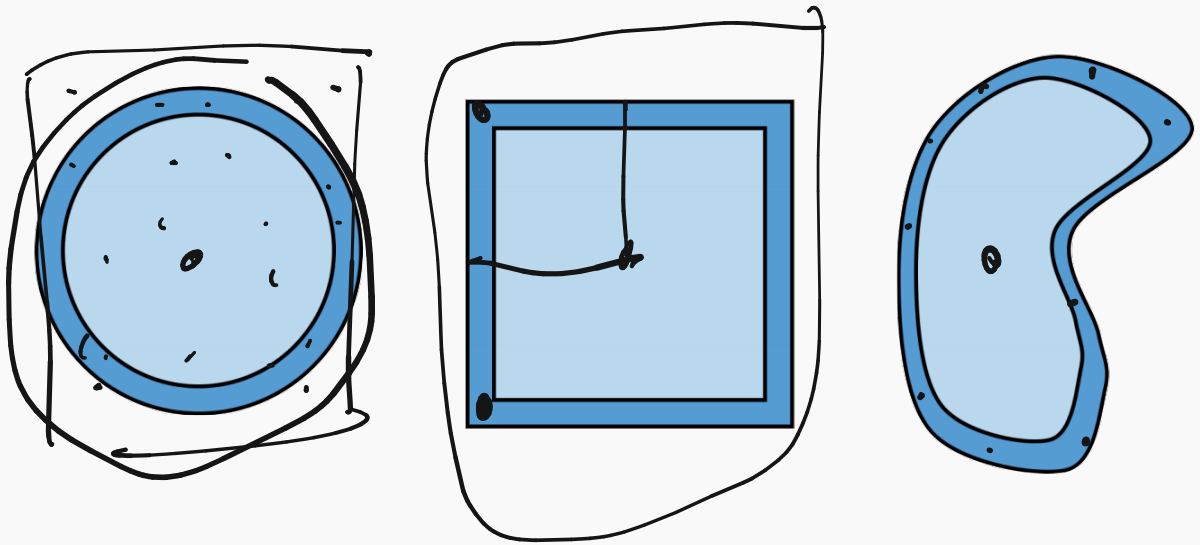
Volume of radius R ball is $\frac{\pi^{d/2}}{(d/2)!} \cdot R^d$.

c

ISOPERIMETRIC INEQUALITY

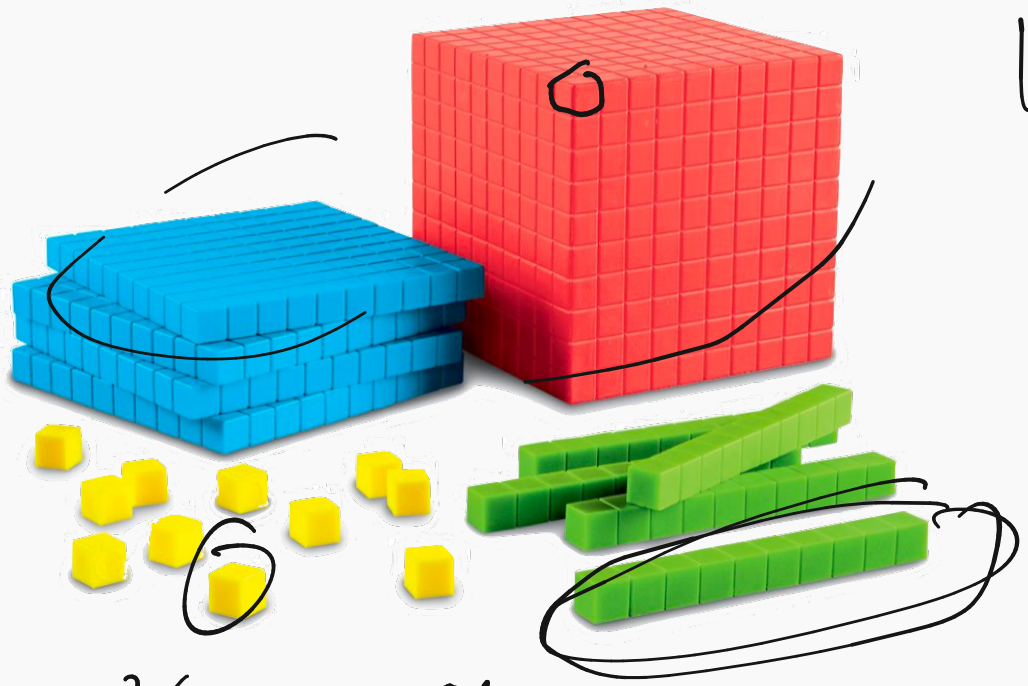
All but a vanishing small $2^{-\Theta(\epsilon d)}$ fraction of a unit ball's volume is within ϵ of its surface.

Isoperimetric Inequality: the ball has the minimum 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.'

INTUITION



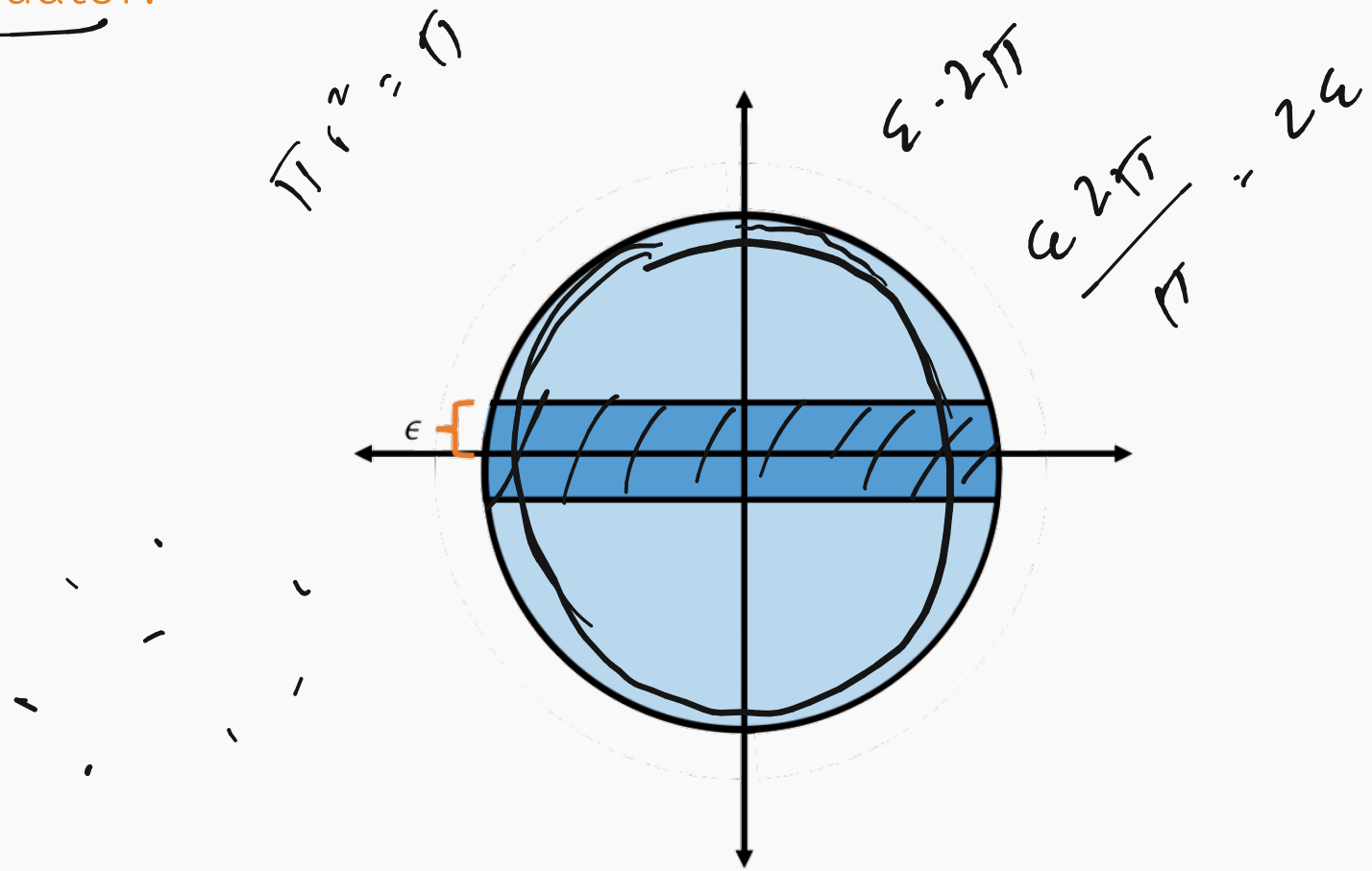
1D: $\frac{\text{surface cubes}}{\text{total cubes}} = \frac{2}{10} = 20\%$

2D: $\frac{\text{surface cubes}}{\text{total cubes}} = \frac{10^2 - 8^2}{10^2} = 36\%$

3D: $\frac{\text{surface cubes}}{\text{total cubes}} = \frac{1000 - 8^3}{1000} = \frac{1000 - 512}{1000} \approx 48.8\%$

SLICES OF THE UNIT BALL

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

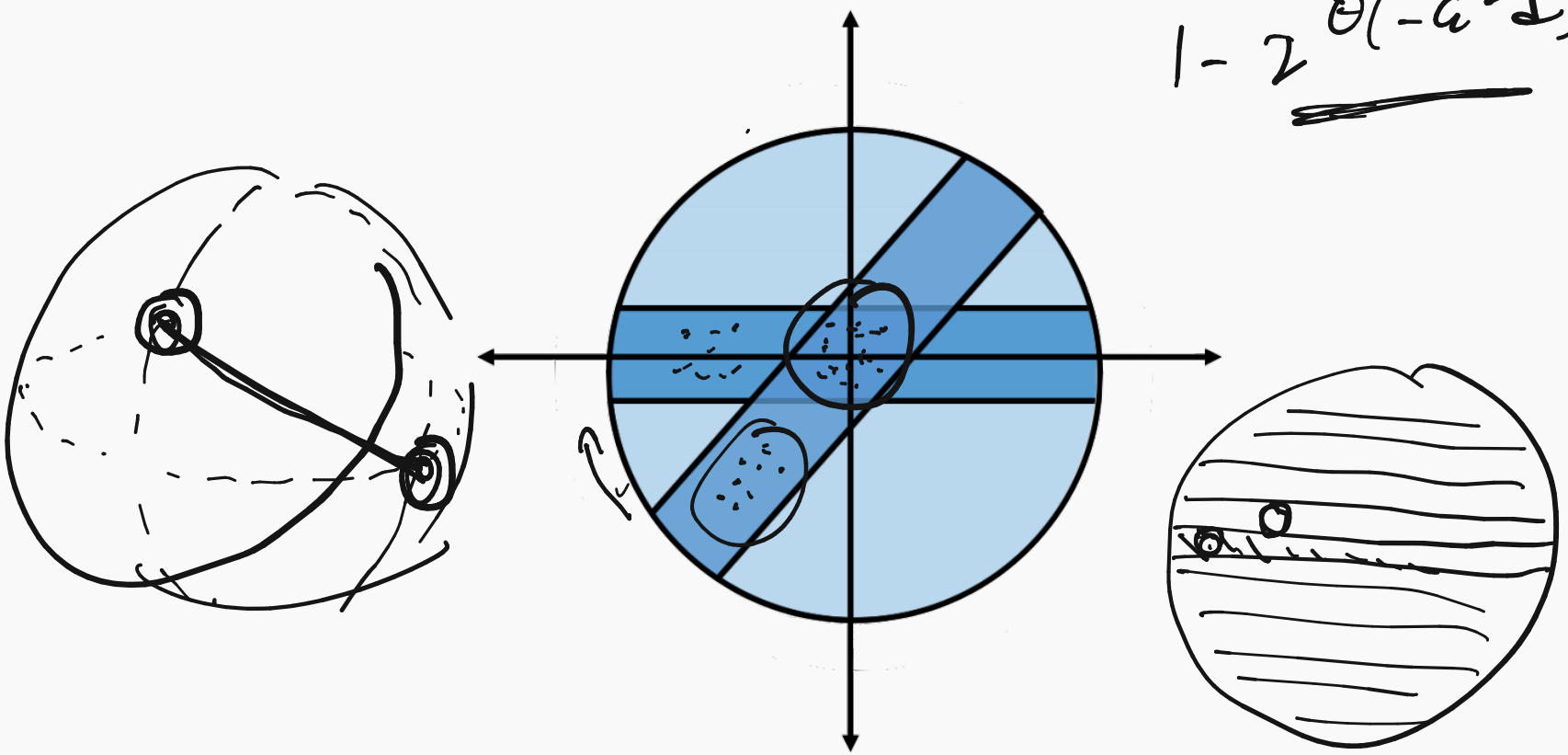


$$S = \{\mathbf{x} \in \mathcal{B}_d : |x_1| \leq \epsilon\}$$

SLICES OF THE UNIT BALL

What percentage of the volume of $\mathcal{B}_d$ falls within ϵ of its equator? **Answer**: all but a $2^{-\Theta(\epsilon^2 d)}$ fraction.

$1 - 2^{\Theta(-\epsilon^2 d)}$

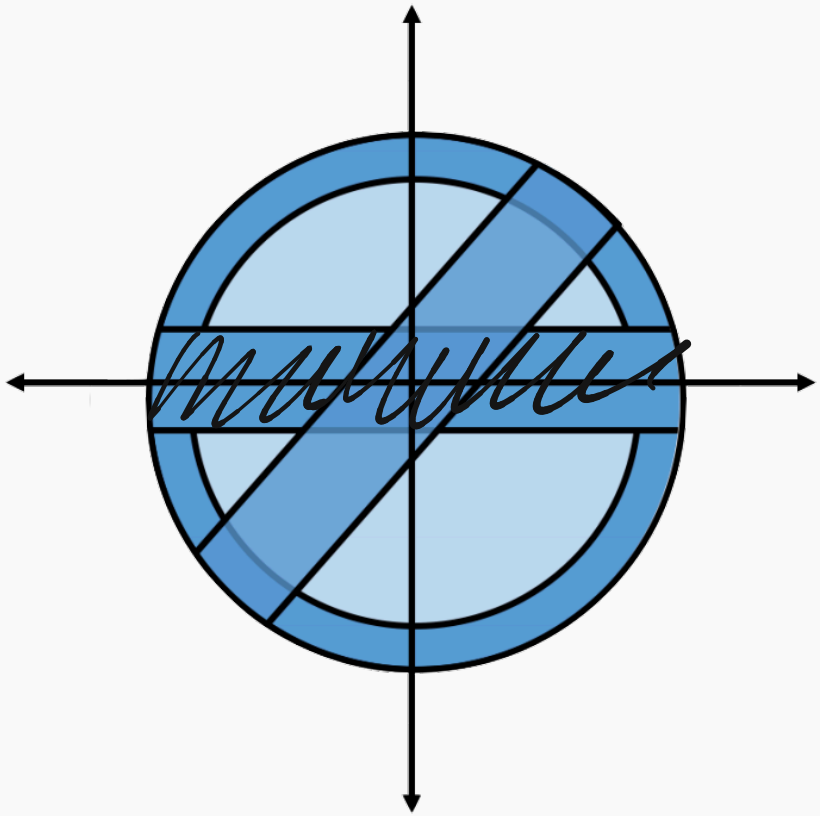


By symmetry, this is true for any equator:

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

BIZARRE SHAPE OF UNIT BALL

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



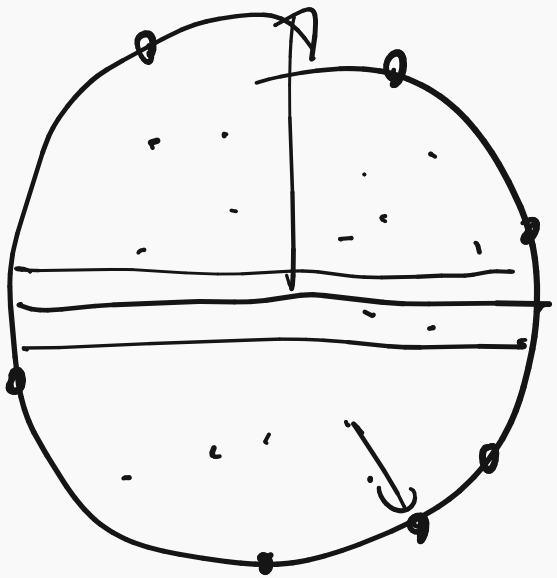
High-dimensional ball looks nothing like 2D ball!

CONCENTRATION AT EQUATOR

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

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

Back on 3:55 pm.

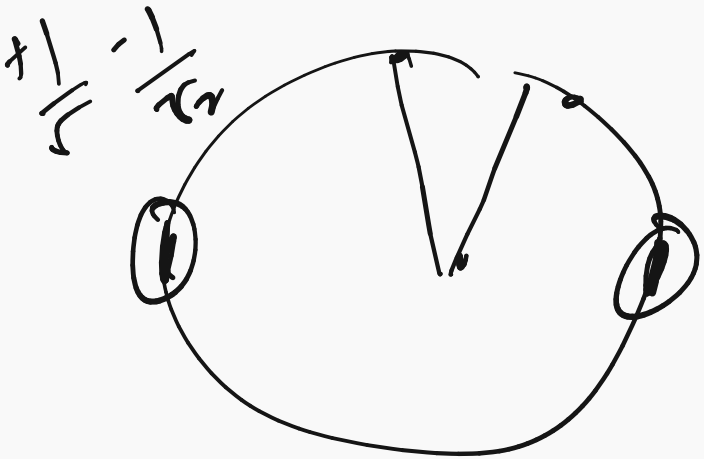


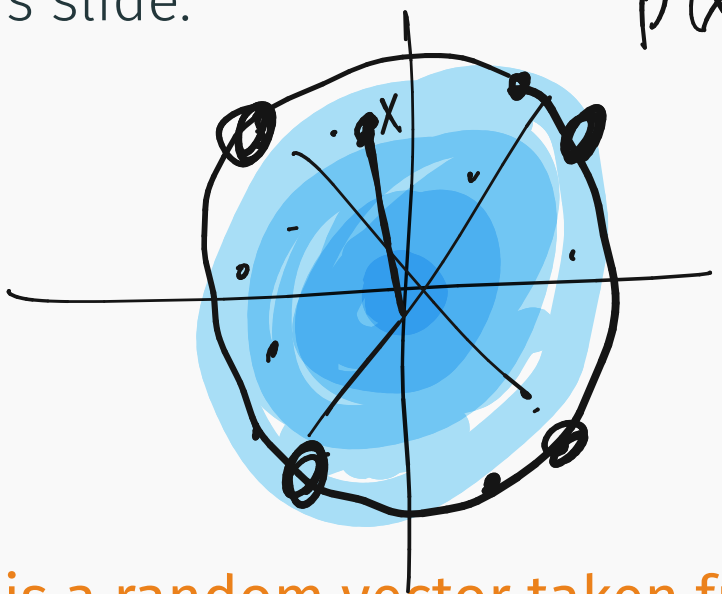
CONCENTRATION AT EQUATOR

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 $\geq 1 - 2^{-\Theta(\epsilon^2 d)}$. This then proves our statement from the previous slide.





How can we generate 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 $\mathbf{g}$ be a random Gaussian vector, with each entry drawn from $\mathcal{N}(0,1)$. Then $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2$ is distributed uniformly on the unit sphere.

Why? Consider the probability density function of a high dimensional Gaussian:

$$p(\mathbf{g}) = p(g[1]) \cdots p(g[d]) = \prod_{i=1}^{d} c e^{-\mathbf{g}[i]^2/2}$$

$$= c^d e^{\sum_{i=1}^{d} -g[i]^2/2}$$

$$= c^d e^{-\|\mathbf{g}\|_2^2/2}$$

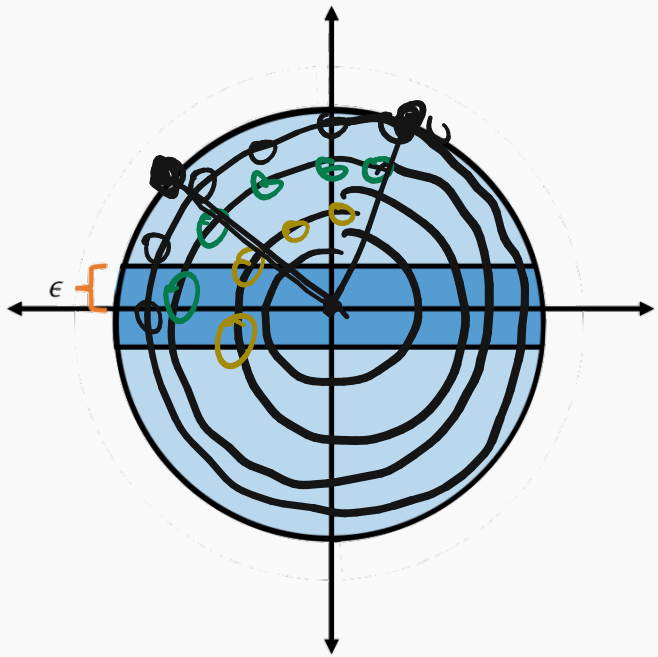
$e^{-\frac{1}{2}\sum_{i=1}^{d} g[i]^2}$

$= e^{-1/2\,\|g\|_2^2}$

$\mathcal{N}(0,1) \cdots \mathcal{N}(0,1)$

$p(x) = c \cdot e^{-x^2/2}$

Draw $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Show that first entry of $\mathbf{w} = \mathbf{g}/\|\mathbf{g}\|_2 \leq \epsilon$ with very high probability.



1. Prove that with high probability, the first entry of $\mathbf{g}/\sqrt{d}$ is small.
2. Prove that $\mathbf{g}/\sqrt{d}$ is very very close to $\mathbf{g}/\|\mathbf{g}\|_2^2$, so this vector also has small first entry.

CONCENTRATION AT EQUATOR

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

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

Excersize for home: Prove that $\Pr\left[\|\mathbf{g}\|_2^2 \leq \frac{1}{2}\mathbb{E}[\|\mathbf{g}\|_2^2]\right] \leq 2^{-\Theta(d)}$.

This should intuitively make sense. Can you tell me why?

$|w_1| \leq a$

$\Pr[\|g\|^2 \leq \frac{1}{2}d] \leq 2^{-\Theta(d)}$

CONCENTRATION AT EQUATOR

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

Recall that $\mathbf{w} = \frac{\mathbf{g}}{\|\mathbf{g}\|_2}$. Given this event:

$$\Pr[|w_1| \leq \epsilon] = \Pr[|g_1/\|\mathbf{g}\|_2| \leq \epsilon]$$

$$\geq \Pr\left[|g_1|/\sqrt{d/2} \leq \epsilon\right]$$

$$= \Pr\left[|g_1| \leq \epsilon \cdot \sqrt{d/2}\right]$$

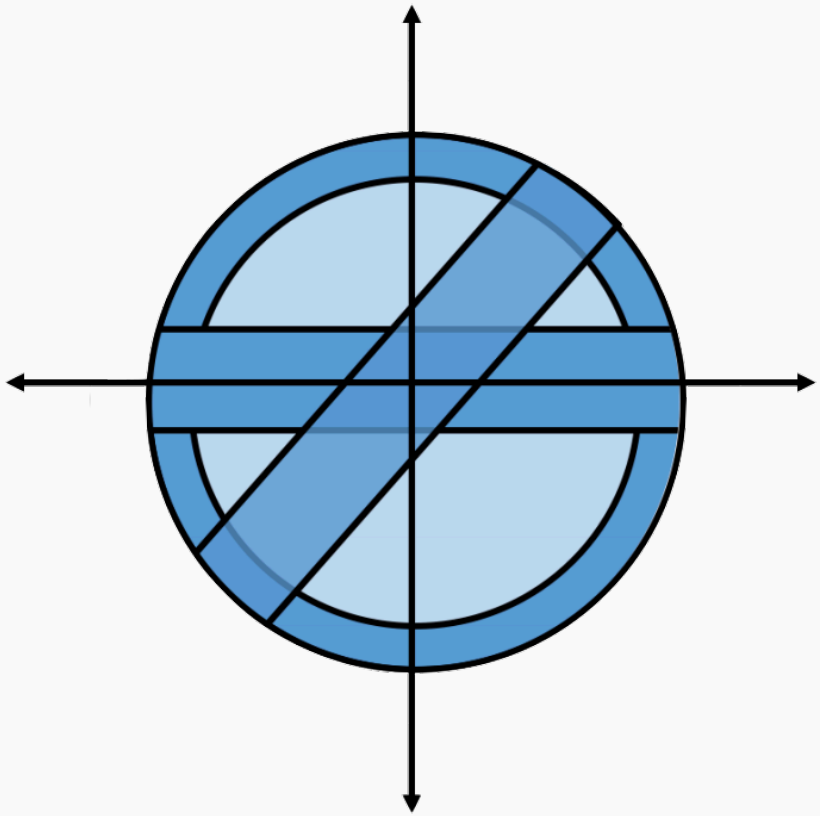
$$\geq 1 - 2^{-\Theta\left((\epsilon \cdot \sqrt{d/2})^2\right)} = 1 - 2^{-\Theta(\epsilon^2 d)}$$

By union bound, overall we have:

$$\Pr[|w_1| \leq \epsilon] \geq 1 - 2^{-\Theta(\epsilon^2 d)} - 2^{-\Theta(d)}$$

BIZARRE SHAPE OF UNIT BALL

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

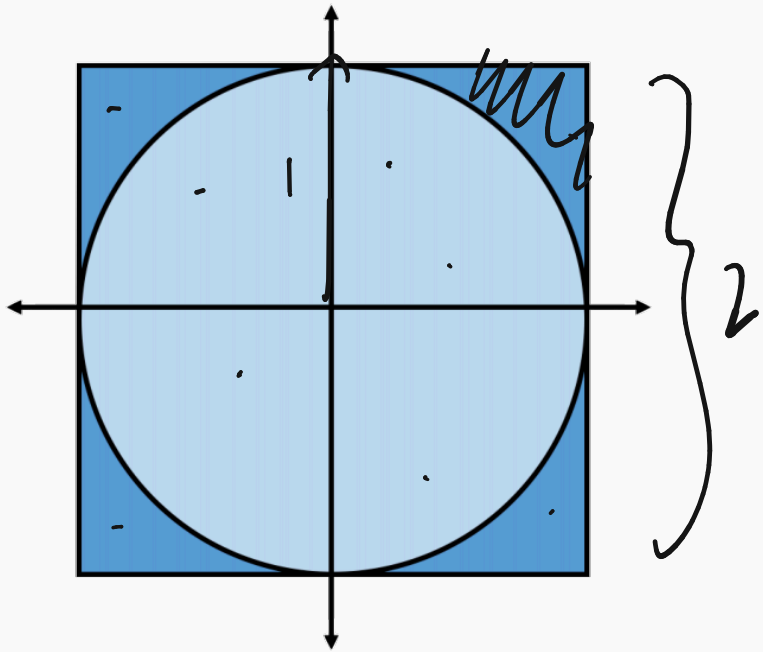


High-dimensional ball looks nothing like 2D ball!

HIGH DIMENSIONAL CUBE

Let $\mathcal{C}_d$ be the d-dimensional cube:

$$\mathcal{C}_d = \{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}(i)| \leq 1 \ \forall i\}.$$



2d

$\sqrt{\pi}^d$

$(d/2)!$

$O(d)^{O(d)}$

In two dimensions, the cube is pretty similar to the ball.

But volume of $\mathcal{C}_d$ is 2^d while volume of unit ball is $\frac{\sqrt{\pi}^d}{(d/2)!}$.

This is a huge gap! Cube has $O(d)^{O(d)}$ more volume.

HIGH DIMENSIONAL CUBE

Some other ways to see these shapes are very different:

- $\max_{x \in \mathcal{B}_d} \|x\|_2^2 =$ 1
- $\max_{x \in \mathcal{C}_d} \|x\|_2^2 =$ d

HIGH DIMENSIONAL CUBE

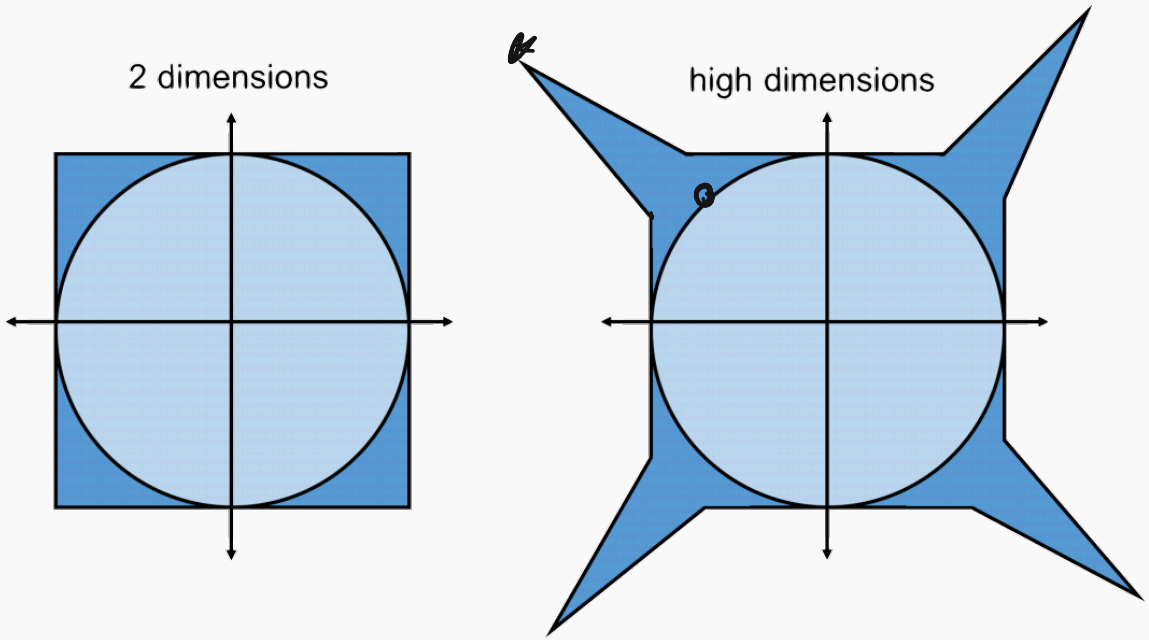
Some other ways to see these shapes are very different:

- $\mathbb{E}_{\mathbf{x}\sim\mathcal{B}_d}\|\mathbf{x}\|_2^2$ ≈ 1
- $\mathbb{E}_{\mathbf{x}\sim\mathcal{C}_d}\|\mathbf{x}\|_2^2 = \mathbb{E}\left(\sum_{i=1}^{d} x_i^2\right) = \sum_{i=1}^{d} \mathbb{E}[x_i^2]$ ($\mathbb{E}[x_i^2] = 1/3$)

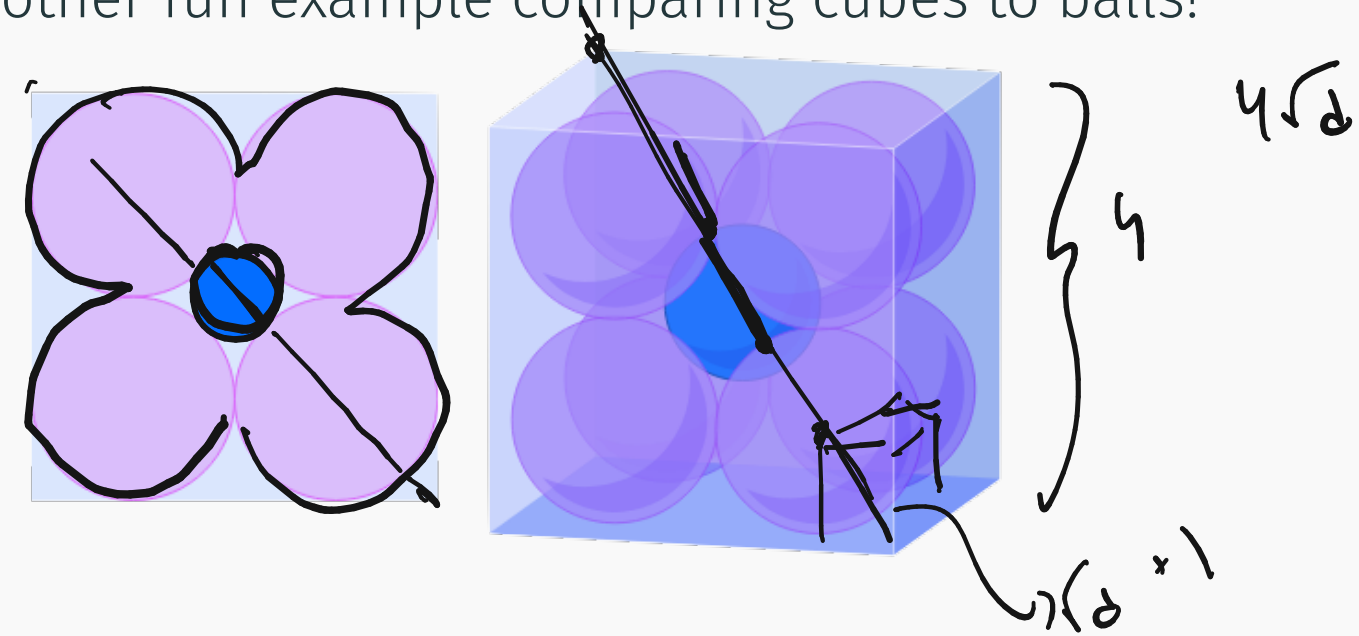
$= d/3$

HIGH DIMENSIONAL CUBE

Almost all of the volume of the unit cube falls in its corners, and these corners lie far outside the unit ball.



See **The Journey to Define Dimension** from Quanta Magazine for another fun example comparing cubes to balls!



Place 2^d unit balls in box with side length 4. Look at sphere they enclose. It has radius $\sqrt{d} - 1$.

So for $d > 9$, it sticks out of the box...

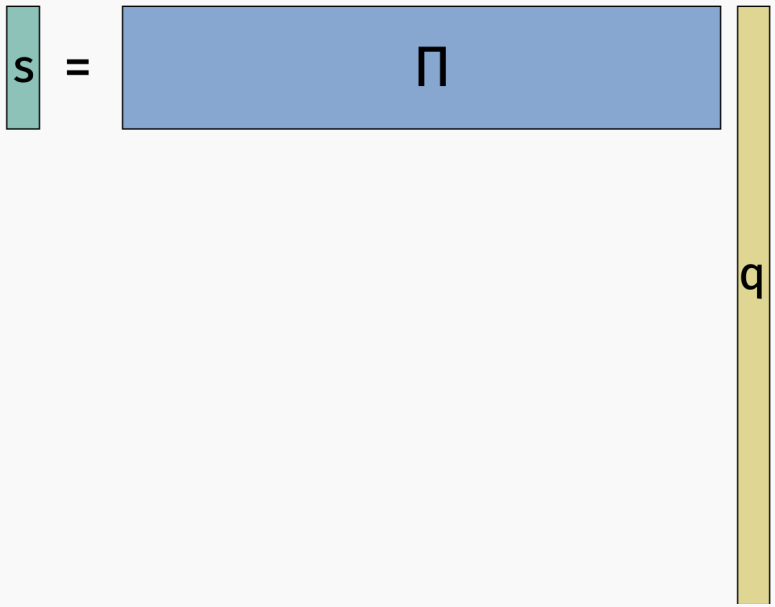
Despite **all this** warning that low-dimensional space looks nothing like high-dimensional space, next we are going to learn about how to **compress high dimensional vectors to low dimensions.**

We will be very careful not to compress things too far. An extremely simple method known as Johnson-Lindenstrauss Random Projection pushes right up to the edge of how much compression is possible.

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 $\mathbf{\Pi} : \mathbb{R}^d \to \mathbb{R}^k$ where $k = O\left(\frac{\log n}{\epsilon^2}\right)$ such that for all i, j,

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



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,

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

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

TONS OF APPLICATIONS

Make pretty much any computation involving vectors faster and more space efficient.

- Faster vector search (used in image search, AI-based web search, Retrieval Augmented Generation (RAG), etc.).
- Faster machine learning (next class we will see an application to speeding up clustering).
- Faster numerical linear algebra.

Only useful if we can explicity construct a JL map Π and apply efficiently to vectors.

Remarkably, Π can be chosen completely at random!

One possible construction: Random Gaussian.

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

The map Π is oblivious to the data set. This stands in contrast to other vector compression methods you might know like PCA.

[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

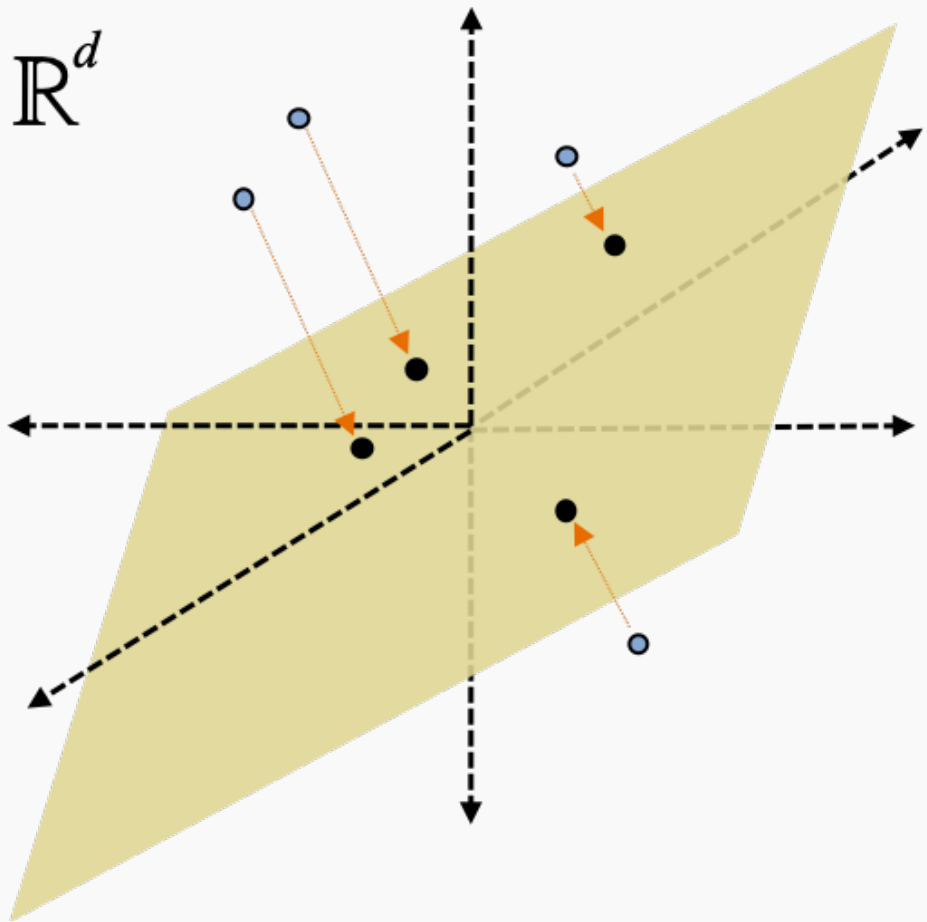
```
>> Pi = 2*randi(2,m,d)-3;
>> s = (1/sqrt(m))*Pi*q;
```

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

For this reason, the JL operation is often called a "random projection", even though it technically is not a projection when $\mathbf{\Pi}$'s entries are i.i.d.

Can anyone see why $\mathbf{\Pi}$ is similar to a projection matrix? I.e., a matrix satisfying $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_{k \times k}$.

RANDOM PROJECTION



Intuition: Multiplying by a random matrix mimics the process of projecting onto a random k dimensional subspace in d dimensions.

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* $(1-\delta)$:

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

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:

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

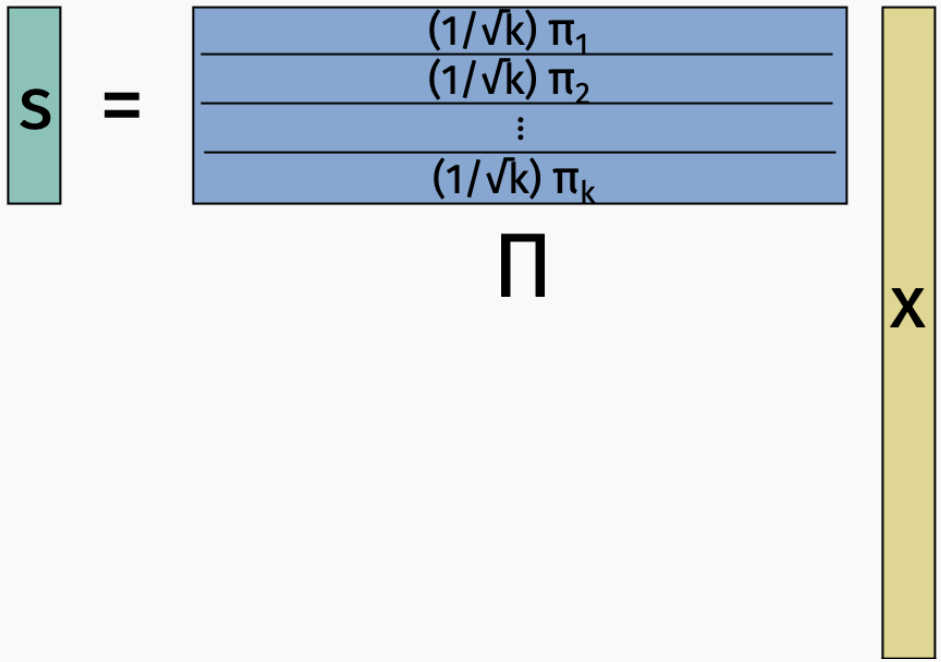
PROOF OF DISTRIBUTIONAL JL

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

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

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

Some notation:



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

Intermediate Claim: Let $\boldsymbol{\pi}$ be a length d vector with $\mathcal{N}(0,1)$ entries.

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

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

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

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

We have that $Z_i \cdot x[i]$ is a normal $\mathcal{N}(0, x[i]^2)$ random variable.

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

What type of random variable is $\langle \boldsymbol{\pi}, \mathsf{x} \rangle$?

Fact (Stability of Gaussian random variables)

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

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

So $\mathbb{E}\|\boldsymbol{\Pi}\mathbf{x}\|_2^2 = \mathbb{E}\left[(\langle \boldsymbol{\pi}, \mathbf{x} \rangle)^2\right] = \mathbb{E}\left[\mathcal{N}(0, \|\mathbf{x}\|_2^2)\right] = \|\mathbf{x}\|_2^2$, as desired.

PROOF OF DISTRIBUTIONAL JL

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

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

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

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

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

CONCENTRATION OF CHI-SQUARED RANDOM VARIABLES

Lemma

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

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

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

CONNECTION TO EARLIER PART OF LECTURE

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?

Hard case: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are all mutually orthogonal unit vectors:

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

When we reduce to k dimensions with JL, we still expect these vectors to be nearly orthogonal. Why?

Hard case: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ are all mutually orthogonal unit vectors:

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

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 that are close to mutually orthogonal. $O(\log n/\epsilon^2)$ = just enough dimensions.

Alternative view: Without additional structure, we expect that learning/inference in d dimensions requires $2^{O(d)}$ data points. If we really had a data set that large, then the JL bound would be vacous, since $\log(n) = O(d)$.