# COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2020.

Lecture 11

# LOGISTICS

- Problem Set 2 was due yesterday.
- Quiz 5 is due **today at 8pm.**
- The exam will be held next Thursday-Friday. Let me know ASAP if you need accommodations (e.g., extended time).
- My office hours this week and next will focus on exam review and going through practice questions.

**Last Class: The Johnson-Lindenstrauss Lemma**

- Low-distortion embeddings for any set of points via random projection.
- Started on proof of the JL Lemma via the Distributional JL Lemma.

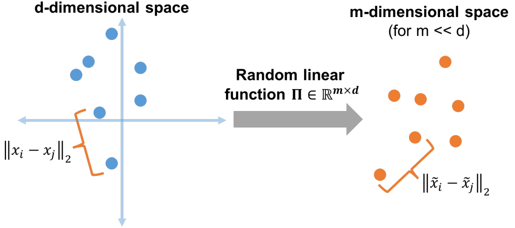
**This Class:**

- Finish Up proof of the JL lemma.
- Example applications to classification and clustering.
- Discuss connections to high dimensional geometry.

**Johnson-Lindenstrauss Lemma:** For any set of points $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ and $\epsilon > 0$ there exists a linear map $\mathbf{\Pi} : \mathbb{R}^d \rightarrow R^m$ such that $m = O\left(\frac{\log n}{\epsilon^2}\right)$ and letting $\tilde{x}_i = \mathbf{\Pi}\vec{x}_i$:

$$\text{For all } i,j:\ (1-\epsilon)\|\vec{x}_i - \vec{x}_j\|_2 \leq \|\tilde{x}_i - \tilde{x}_j\|_2 \leq (1+\epsilon)\|\vec{x}_i - \vec{x}_j\|_2.$$

Further, if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ has each entry chosen i.i.d. from $\mathcal{N}(0, 1/m)$ and $m = O\left(\frac{\log n/\delta}{\epsilon^2}\right)$, $\mathbf{\Pi}$ satisfies the guarantee with probability $\geq 1 - \delta$.

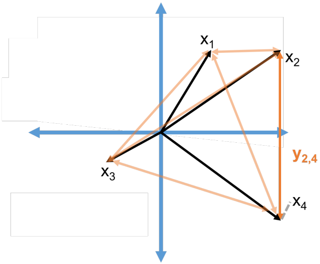


We showed that the Johnson-Lindenstrauss Lemma follows from:

**Distributional JL Lemma:** Let $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $\vec{y} \in \mathbb{R}^d$, with probability $\geq 1 - \delta$

$$(1-\epsilon)\|\vec{y}\|_2 \leq \|\mathbf{\Pi}\vec{y}\|_2 \leq (1+\epsilon)\|\vec{y}\|_2.$$

**Main Idea:** Union bound over $\binom{n}{2}$ difference vectors $\vec{y}_{ij} = \vec{x}_i - \vec{x}_j$.

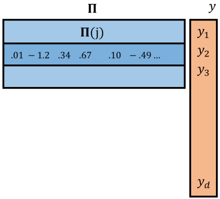


# DISTRIBUTIONAL JL PROOF

**Distributional JL Lemma:** Let $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$. If we set $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then for any $\vec{y} \in \mathbb{R}^d$, with probability $\geq 1 - \delta$

$$(1-\epsilon)\|\vec{y}\|_2 \leq \|\mathbf{\Pi}\vec{y}\|_2 \leq (1+\epsilon)\|\vec{y}\|_2.$$

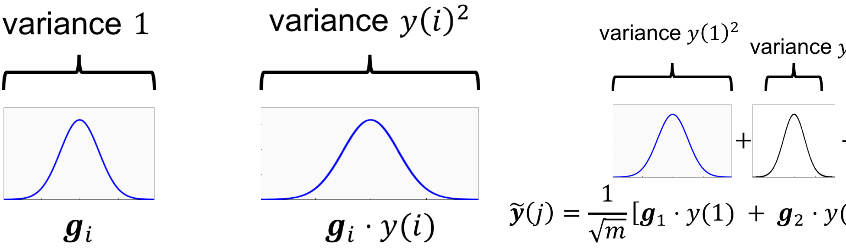
- Let $\tilde{\mathbf{y}}$ denote $\mathbf{\Pi}\vec{y}$ and let $\mathbf{\Pi}(j)$ denote the $j^{th}$ row of $\mathbf{\Pi}$.
- For any $j$, $\tilde{\mathbf{y}}(j) = \langle \mathbf{\Pi}(j), \vec{y} \rangle = \frac{1}{\sqrt{m}} \sum_{i=1}^{d} \mathbf{g}_i \cdot \vec{y}(i)$ where $\mathbf{g}_i \sim \mathcal{N}(0, 1/m1)$.




$\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$: random projection. $d$: original dim. $m$: compressed dim, $\epsilon$: error, $\delta$: failure prob.

# DISTRIBUTIONAL JL PROOF

- Let $\tilde{\mathbf{y}}$ denote $\mathbf{\Pi}\vec{y}$ and let $\mathbf{\Pi}(j)$ denote the $j^{th}$ row of $\mathbf{\Pi}$.
- For any $j$, $\tilde{\mathbf{y}}(j) = \langle \mathbf{\Pi}(j), \vec{y} \rangle = \frac{1}{\sqrt{m}} \sum_{i=1}^{d} \mathbf{g}_i \cdot \vec{y}(i)$ where $\mathbf{g}_i \sim \mathcal{N}(0,1)$.
- $\mathbf{g}_i \cdot \vec{y}(i) \sim \mathcal{N}(0, \vec{y}(i)^2)$: a normal distribution with variance $\vec{y}(i)^2$.




What is the distribution of $\tilde{\mathbf{y}}(j)$? Also Gaussian!

> $\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$: random projection mapping $\vec{y} \to \tilde{\mathbf{y}}$. $\mathbf{\Pi}(j)$: $j^{th}$ row of $\mathbf{\Pi}$, $d$: original dimension. $m$: compressed dimension, $\mathbf{g}_i$: normally distributed random variable.

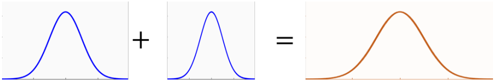
# DISTRIBUTIONAL JL PROOF

Letting $\tilde{\mathbf{y}} = \mathbf{\Pi}\vec{y}$, we have $\tilde{\mathbf{y}}(j) = \langle \mathbf{\Pi}(j), \vec{y} \rangle$ and:

$$\tilde{\mathbf{y}}(j) = \frac{1}{\sqrt{m}} \sum_{i=1}^{d} \mathbf{g}_i \cdot \vec{y}(i) \text{ where } \mathbf{g}_i \cdot \vec{y}(i) \sim \mathcal{N}(0, \vec{y}(i)^2).$$

**Stability of Gaussian Random Variables.** For independent $a \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $b \sim \mathcal{N}(\mu_2, \sigma_2^2)$ we have:

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




Thus, $\tilde{\mathbf{y}}(j) \sim \frac{1}{\sqrt{m}} \mathcal{N}(0, \vec{y}(1)^2 + \vec{y}(2)^2 + \ldots + \vec{y}(d)^2 \|\vec{y}\|_2^2) \mathcal{N}(0, \|\vec{y}\|_2^2/m)$. I.e., $\tilde{\mathbf{y}}$ itself is a random Gaussian vector. Rotational invariance of the Gaussian distribution.

Stability is another explanation for the central limit theorem.

**So far:** Letting $\mathbf{\Pi} \in \mathbb{R}^{d \times m}$ have each entry chosen i.i.d. as $\mathcal{N}(0, 1/m)$, for any $\vec{y} \in \mathbb{R}^d$, letting $\tilde{\mathbf{y}} = \mathbf{\Pi}\vec{y}$:

$$\tilde{\mathbf{y}}(j) \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m).$$

What is $\mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2]$?

$$\mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \mathbb{E}\left[\sum_{j=1}^{m} \tilde{\mathbf{y}}(j)^2\right] = \sum_{j=1}^{m} \mathbb{E}[\tilde{\mathbf{y}}(j)^2]$$

$$= \sum_{j=1}^{m} \frac{\|\vec{y}\|_2^2}{m} = \|\vec{y}\|_2^2$$

So $\tilde{\mathbf{y}}$ has the right norm in expectation.

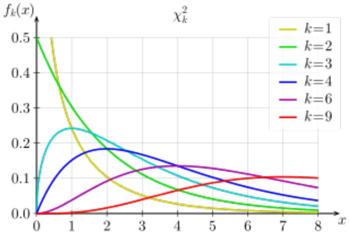
How is $\|\tilde{\mathbf{y}}\|_2^2$ distributed? Does it concentrate?

$\vec{y} \in \mathbb{R}^d$: arbitrary vector, $\tilde{\mathbf{y}} \in \mathbb{R}^m$: compressed vector, $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$: random projection mapping $\vec{y} \to \tilde{\mathbf{y}}$. $\mathbf{\Pi}(j)$: $j^{th}$ row of $\mathbf{\Pi}$, $d$: original dimension. $m$: compressed dimension, $\mathbf{g}_i$: normally distributed random variable

**So Far:** Each entry of our compressed vector $\tilde{\mathbf{y}}$ is Gaussian with :

$$\tilde{\mathbf{y}}(j) \sim \mathcal{N}(0, \|\vec{y}\|_2^2/m) \text{ and } \mathbb{E}[\|\tilde{\mathbf{y}}\|_2^2] = \|\vec{y}\|_2^2$$

$\|\tilde{\mathbf{y}}\|_2^2 = \sum_{i=1}^m \tilde{\mathbf{y}}(j)^2$ a Chi-Squared random variable with $m$ degrees of freedom (a sum of $m$ squared independent Gaussians)

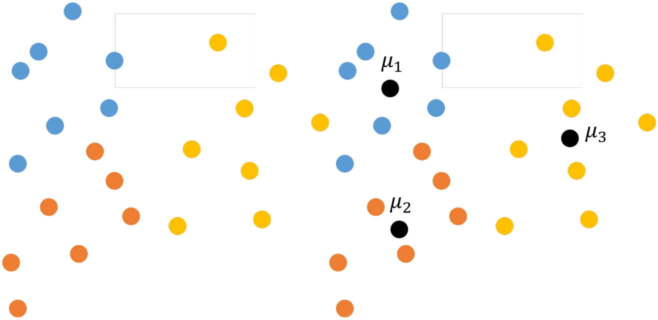



**Lemma:** (Chi-Squared Concentration) Letting $\mathbf{Z}$ be a Chi-Squared random variable with $m$ degrees of freedom,

$$\Pr\left[|\mathbf{Z} - \mathbb{E}\mathbf{Z}| \geq \epsilon \mathbb{E}\mathbf{Z}\right] \leq 2e^{-m\epsilon^2/8}.$$

# EXAMPLE APPLICATION: $k$-MEANS CLUSTERING

**Goal:** Separate $n$ points in $d$ dimensional space into $k$ groups.




**k-means Objective:** $Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x} \in \mathcal{C}_k} \|\vec{x} - \mu_j\|_2^2.$

**Write in terms of distances:**

$$Cost(\mathcal{C}_1, \ldots, \mathcal{C}_k) = \min_{\mathcal{C}_1, \ldots \mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1, \vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$$

**k-means Objective:** $Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\vec{x}_1,\vec{x}_2 \in \mathcal{C}_k} \|\vec{x}_1 - \vec{x}_2\|_2^2$ If we randomly project to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions, for all pairs $\vec{x}_1, \vec{x}_2$,

$$(1-\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2 \le \|\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\|_2^2 \le (1+\epsilon)\|\vec{x}_1 - \vec{x}_2\|_2^2 \implies$$

Letting $\overline{Cost}(\mathcal{C}_1,\ldots,\mathcal{C}_k) = \min_{\mathcal{C}_1,\ldots\mathcal{C}_k} \sum_{j=1}^{k} \sum_{\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2 \in \mathcal{C}_k} \|\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2\|_2^2$

$$(1-\epsilon)Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k) \le \overline{Cost}(\mathcal{C}_1,\ldots,\mathcal{C}_k) \le (1+\epsilon)Cost(\mathcal{C}_1,\ldots,\mathcal{C}_k).$$

**Upshot:** Can cluster in $m$ dimensional space (much more efficiently) and minimize $\overline{Cost}(\mathcal{C}_1,\ldots,\mathcal{C}_k)$. The optimal set of clusters will have true cost within $1 + c\epsilon$ times the true optimal. Good exercise to prove this.

# The Johnson-Lindenstrauss Lemma and High Dimensional Geometry

- High-dimensional Euclidean space looks *very different* from low-dimensional space. So how can JL work?
- Is Euclidean distance in high-dimensional meaningless, making JL useless? (The curse of dimensionality)

What is the largest set of mutually orthogonal unit vectors in $d$-dimensional space?

a) 1 b) $\log d$ c) $\sqrt{d}$ d) $d$

What is the largest set of unit vectors in $d$-dimensional space that have all pairwise dot products $|\langle \vec{x}, \vec{y} \rangle| \leq \epsilon$? (think $\epsilon = .01$)

a) $d$ b) $\Theta(d)$ c) $\Theta(d^2)$ d) $2^{\Theta(d)}$

In fact, an exponentially large set of random vectors will be nearly pairwise orthogonal with high probability!

**Claim:** $2^{\Theta(\epsilon^2 d)}$ random $d$-dimensional unit vectors will have all pairwise dot products $|\langle \vec{x}, \vec{y} \rangle| \leq \epsilon$ (be nearly orthogonal).

**Proof:** Let $\vec{x}_1, \ldots, \vec{x}_t$ each have independent random entries set to $\pm 1/\sqrt{d}$.

- What is $\|\vec{x}_i\|_2$? Every $\vec{x}_i$ is always a unit vector.
- What is $\mathbb{E}[\langle \vec{x}_i, \vec{x}_j \rangle]$? $\mathbb{E}[\langle \vec{x}_i, \vec{x}_j \rangle] = 0$
- By a Chernoff bound, $\Pr[|\langle \vec{x}_i, \vec{x}_j \rangle| \geq \epsilon] \leq 2e^{-\epsilon^2 d/6}$ (great exercise).
- If we chose $t = \frac{1}{2} e^{\epsilon^2 d/12}$, using a union bound over all $\binom{t}{2} \leq \frac{1}{8} e^{\epsilon^2 d/6}$ possible pairs, with probability $\geq 3/4$ all will be nearly orthogonal.

**Up Shot:** In $d$-dimensional space, a set of $2^{\Theta(\epsilon^2 d)}$ random unit vectors have all pairwise dot products at most $\epsilon$ (think $\epsilon = .01$)

$$\|\vec{x}_i - \vec{x}_j\|_2^2 = \|\vec{x}_i\|_2^2 + \|\vec{x}_j\|_2^2 - 2\vec{x}_i^T\vec{x}_j \geq 1.98.$$

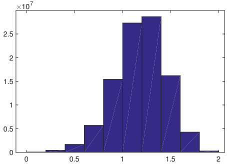
Even with an exponential number of random vector samples, we don't see any nearby vectors.

- Can make methods like nearest neighbor classification or clustering useless.

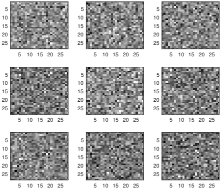
Curse of dimensionality for sampling/learning functions in high-dimensional space – samples are very 'sparse' unless we have a huge amount of data.

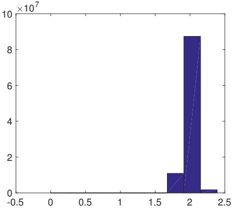
- Only hope is if we lots of structure (which we typically do...)

Distances for MNIST Digits:


Distances for Random Images:




**Another Interpretation:** Tells us that random data can be a very bad model for actual input data.

**Recall:** The Johnson Lindenstrauss lemma states that if $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is a random matrix (linear map) with $m = O\left(\frac{\log n}{\epsilon^2}\right)$, for $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^d$ with high probability, for all $i, j$:

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

**Implies:** If $\vec{x}_1, \ldots, \vec{x}_n$ are nearly orthogonal unit vectors in $d$-dimensions (with pairwise dot products bounded by $\epsilon/8$), then $\frac{\mathbf{\Pi}\vec{x}_1}{\|\mathbf{\Pi}\vec{x}_1\|_2}, \ldots, \frac{\mathbf{\Pi}\vec{x}_n}{\|\mathbf{\Pi}\vec{x}_n\|_2}$ are nearly orthogonal unit vectors in $m$-dimensions (with pairwise dot products bounded by $\epsilon$).

- Algebra is a bit messy but a good exercise to partially work through.

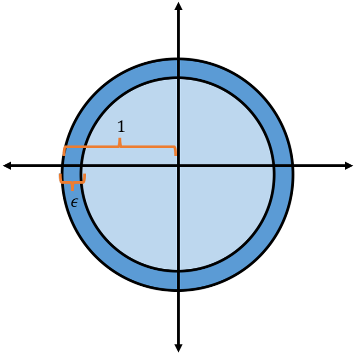
**Claim 1:** $n$ nearly orthogonal unit vectors can be projected to $m = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions and still be nearly orthogonal.

**Claim 2:** In $m$ dimensions, there are at most $2^{O(\epsilon^2 m)}$ nearly orthogonal vectors.

- For both these to hold it might be that $n \leq 2^{O(\epsilon^2 m)}$.
- $2^{O(\epsilon^2 m)} = 2^{O(\log n)} \geq n$. Tells us that the JL lemma is optimal up to constants.
- $m$ is chosen just large enough so that the odd geometry of $d$-dimensional space still holds on the $n$ points in question after projection to a much lower dimensional space.

Let $\mathcal{B}_d$ be the unit ball in $d$ dimensions. $\mathcal{B}_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$.

What percentage of the volume of $\mathcal{B}_d$ falls within $\epsilon$ distance of its surface? Answer: all but a $(1-\epsilon)^d \leq e^{-\epsilon d}$ fraction. Exponentially small in the dimension $d$!

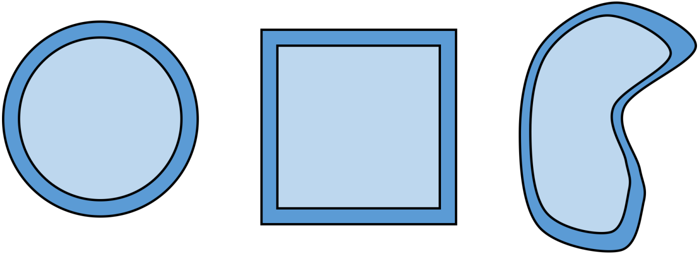



Volume of a radius $R$ ball is $\frac{\pi^{\frac{d}{2}}}{(d/2)!} \cdot R^d$.

# BIZARRE SHAPE OF HIGH-DIMENSIONAL BALLS

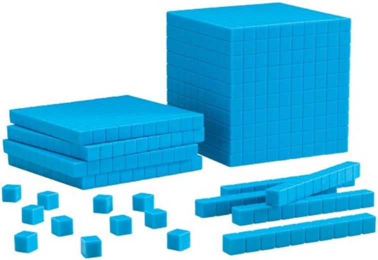
All but an $e^{-\epsilon d}$ fraction of a unit ball's volume is within $\epsilon$ of its surface. If we randomly sample points with $\|x\|_2 \leq 1$, nearly all will have $\|x\|_2 \geq 1 - \epsilon$.

- **Isoperimetric inequality**: the ball has the minimum surface area/volume ratio of any shape.



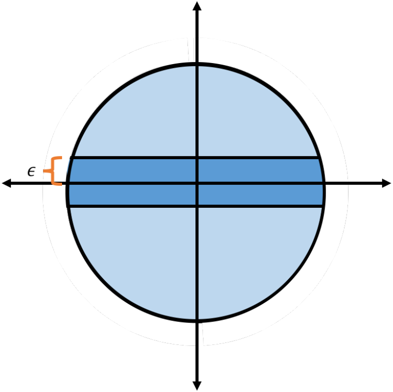
- If we randomly sample points from any high-dimensional shape, nearly all will fall near its surface.
- 'All points are outliers.'

What fraction of the cubes are visible on the surface of the cube?



$$\frac{10^3 - 8^3}{10^3} = \frac{1000 - 512}{1000} = .488.$$

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

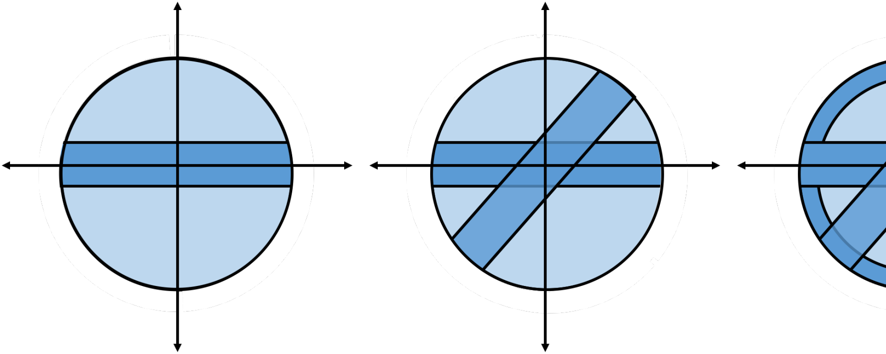



Formally: volume of set $S = \{x \in \mathcal{B}_d : |x(1)| \leq \epsilon\}$.

By symmetry, all but a $2^{\Theta(-\epsilon^2 d)}$ fraction of the volume falls within $\epsilon$ of any equator! $S = \{x \in \mathcal{B}_d : |\langle x, t\rangle| \leq \epsilon\}$

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

**Claim 2:** All but a $2^{\Theta(-\epsilon d)}$ fraction falls within $\epsilon$ of its surface.



How is this possible? High-dimensional space looks nothing like this picture!

Summary: