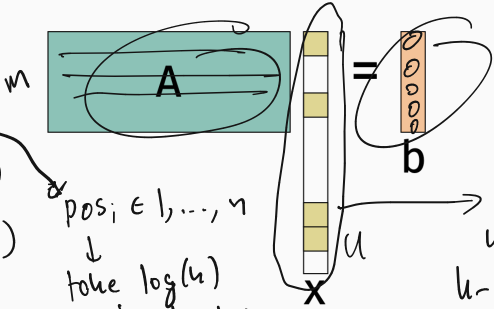
CS-GY 9223 I: Lecture 14
High Dimensional Geometry (+ finish up compressed sensing)

NYU Tandon School of Engineering, Prof. Christopher Musco

SPARS~~ITY~~ RECOVERY/COMPRESSED SENSING

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m < n$, $\mathbf{b} \in \mathbb{R}^m$ with $\mathbf{b} = \mathbf{A}\mathbf{x}$ Recover $\mathbf{x}$, under the assumption that it is k-sparse.



$(pos_1, value_1)$

$(pos_2, value_2)$

$\vdots$

$(pos_k, value_k)$

$pos_i \in 1, \ldots, n$

↓

take $\log(n)$ bits to store.

How much space needed to store k-sparse vector x.

↓

$O(k \log_2 n)$ bits

What properties of $\mathbf{A}$ let us solve this problem efficiently.

- **Measurement efficiency:** Small m.
- **Computational efficiency:** **polynomial in** n, k.

BASIC RESULT FROM LAST CLASS

sparsity → accuracy

If **A** is matrix satisfying the $(O(k), O(1))$- Restricted Isometry Property then **x** can be uniquely recovered from **b** = **Ax** in polynomial time using the basis pursuit linear program:

$O(n^{3.5})$ time to solve linear program.

$$\mathbf{x} = \arg\min_{\mathbf{z}} \|\mathbf{z}\|_1 \qquad \text{subject to} \qquad \mathbf{Az} = \mathbf{b}.$$

→ is always equal to $\arg\min_{z} \|z\|_0$ subject to $Az = b$.

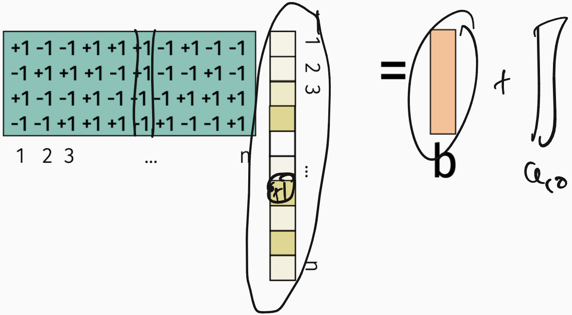
Measurement complexity $m \approx O(k \log n)$ for:

- Random matrices (JL matrices)
- Subsampled Fourier matrices

In other words, this is how many rows these matrices need to satisfy $(O(k), O(1))$-RIP.

$b = [\;]$ } $< k$ entries → $x = [\;]$ → involves k numbers

Random matrices used in algorithms. E.g. to solve the heavy-hitters problem:



$$\begin{array}{cccccccccc}
+1 & -1 & -1 & +1 & +1 & +1 & -1 & +1 & -1 & -1 \\
-1 & +1 & +1 & +1 & -1 & +1 & -1 & -1 & +1 & -1 \\
+1 & -1 & -1 & +1 & -1 & +1 & -1 & +1 & +1 & +1 \\
-1 & -1 & +1 & +1 & +1 & -1 & +1 & -1 & -1 & +1 \\
1 & 2 & 3 & & & \dots & & & & n
\end{array}$$

Subsampled Fourier matrices: $\mathbf{b} = \mathbf{A}\mathbf{x}$ is the evaluation of the Fourier transform $\mathbf{F}\mathbf{x}$ at a set of random frequencies $f_1, \ldots, f_m \in \{0, \ldots, n-1\}$. If we had entirety of $\mathbf{F}\mathbf{x}$ could recover $\mathbf{x}$ using inverse Fourier transform.

BOYER–MOORE MAJORITY VOTE ALGORITHM

$O(1)$

e

Input: Stream of number: 1, 2, 1, 5, 1, 1, 6, ...

Output: Majority element (if one exists)

```
• $item = 0$; $count = 0$;
• For each $e$ in our stream:
    • If $count == 0$, $item = e$; $count = 1$;
    • Else if $item == e$; $count = count + 1$;
    • Else $count = count - 1$;
• Return $item$
```

If m is majority item, c increases more than it decreases. So its > 0 at end of stream

If item m arrives in the stream

$c \leftarrow c+1$.

always

If another item arrives

$c \leftarrow c \pm 1$

C = count if item = m where m is the true majority item

otherwise $c = -$count

Misra-Gries algorithm

Slick generalization of this algorithm for finding any element e which appears more than $1/k$ of the time. Uses $O(k)$ space.

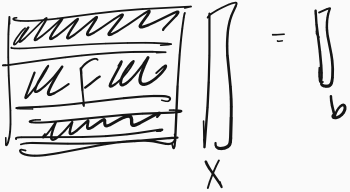
Gives very similar bounds to what you get from applying compressed sensing methods.

FASTER METHODS

A lot of interest in developing even faster algorithms that avoid using the “heavy hammer” of linear programming and run in even faster than $O(n^{3.5})$ time.

- **Iterative Hard Thresholding**: Looks a lot like projected gradient descent. Solve $\min_{\mathbf{z}} \|\mathbf{A}\mathbf{z} - \mathbf{b}\|$ while continually projecting z back to the set of k-sparse vectors. Runs in time $\sim O(nk \log n)$ for Gaussian measurement matrices and $O(n \log n)$ for subsampled Fourer matrices.
- Other “first order” type methods: Orthogonal Matching Pursuit, CoSaMP, Subspace Pursuit, etc.

FASTER METHODS



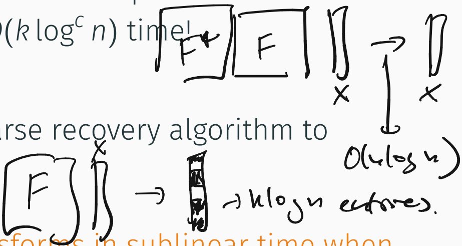
When A is a subsampled Fourier matrix, there are now methods that run in $O(k \log^c n)$ time [Hassanieh, Indyk, Kapralov, Katabi, Price, Shi, etc. 2012+].

$kn \to \underline{k \log^c n}$ $< O(n \log n)$

SPARSE FOURIER TRANSFORM

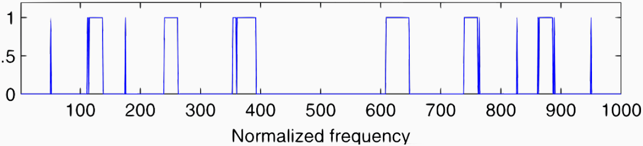
Corollary: When $\mathbf{x}$ is k-sparse, we can compute the inverse Fourier transform $\mathbf{F}^*\mathbf{F}\mathbf{x}$ of $\mathbf{F}\mathbf{x}$ in $O(k \log^c n)$ time!

- Randomly subsample $\mathbf{F}\mathbf{x}$.
- Feed that input into our sparse recovery algorithm to extract $\mathbf{x}$.



Fourier and inverse Fourier transforms in <u>sublinear time</u> when the output is sparse.

O(k log n) instead O(n log n)



Applications in: Wireless communications, GPS, protein imaging, radio astronomy, etc. etc.

HIGH DIMENSIONAL GEOMETRY

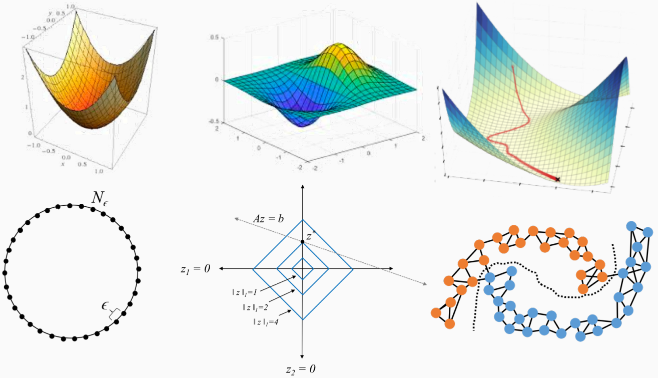
UNIFYING THEME OF THE COURSE

How do we deal with data in high dimensions?

- Randomized sketching + dimensionality reduction.
- Locality sensitive hashing for similarity search.
- Iterative methods for optimizing functions that depend on many variables.
- SVD + low-rank approximation to find and visualize low-dimensional structure.
- Convert large graphs to high dimensional vector data.

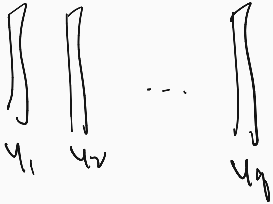
VISUALIZING HIGH DIMENSIONAL DATA

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

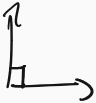


This is not always a good thing to do: high-dimensional space looks very different from low-dimensional space.

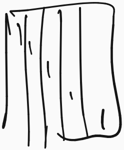
What is the largest set of mutually orthogonal unit vectors in d-dimensional space?



want $u_i^T \underline{u_j} = 0$ for all i,j.

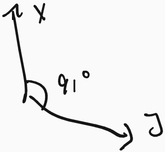


How large can q be?



ORTHOGONAL VECTORS

What is the largest set of unit vectors in d-dimensional space with inner product $|\mathbf{x}^T\mathbf{y}| \leq \epsilon$? → = .01



3d → 3

~~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.

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.

- $\|\mathbf{x}_i\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\sum 1/d} = 1.$

  $x_i^2 = 1/d$ for all i

- $\mathbb{E}[\mathbf{x}_i^T\mathbf{x}_j] = \mathbb{E}\sum_{\ell=1}^{d} x_i(\ell)\,x_j(\ell) = \sum_{\ell=1}^{d} \mathbb{E}\, x_i(\ell)\,x_j(\ell)$

  for $i \neq j$

  $= 0$

  $1/d$ w/ prob. $1/2$ $\quad -1/d$ w/ prob. $1/2$

  $= 0$

ORTHOGONAL VECTORS

$\rightarrow \mathbb{E}[x_i^T x_j] = 0$

For any i, j pair, $\Pr[|\mathbf{x}_i^T\mathbf{x}_j| < \epsilon] \geq 1 - 2e^{-\epsilon^2 d/3}$.

By a union bound:

For all i, j pairs simultaneously, $\Pr[|\mathbf{x}_i^T\mathbf{x}_j| < \epsilon] \geq 1 - t^2 \cdot 2e^{-\epsilon^2 d/3}$.

$i, j \in 1, \ldots, t$ $\quad O(t^2)$ pairs i, j $\rightarrow \geq 1/2$

$$\boxed{t = \frac{1}{2} e^{\epsilon^2 d/6}} \qquad t^2 = \frac{1}{4} e^{\epsilon^2 d/3}$$

$$t^2 \cdot 2e^{-\epsilon^2 d/3} = \frac{1}{4} \cdot 2 = 1/2 \qquad t = O(e^{\epsilon^2 d})$$

2

$2.3\ldots$

ORTHOGONAL VECTORS

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

$\log(d)$

$$2^{\varepsilon^2 d} = e^{1.\dots} \qquad e^{(1.1)\varepsilon^2 d} \qquad 2^{O(\varepsilon^2 d)} = e^{O(\varepsilon^2 d)}$$

Corollary: Random vectors are all approximately the same distance from each other.

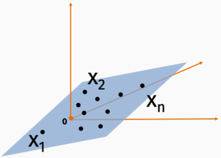
$$\|x_i - x_i\|_2^2 = \underbrace{\|x_i\|_2^2}_{1} + \underbrace{\|x_j\|_2^2}_{1} - \underbrace{2x_i^T x_j}_{\in[-\varepsilon,\varepsilon]}$$

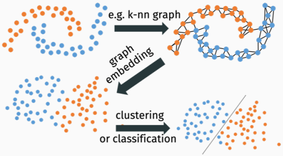
$$2-2\varepsilon \leq \|x_i - x_j\|_2^2 \leq 2+2\varepsilon$$

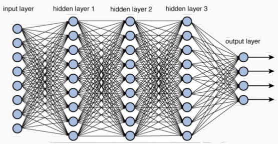
CURSE OF DIMENSIONALITY

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.





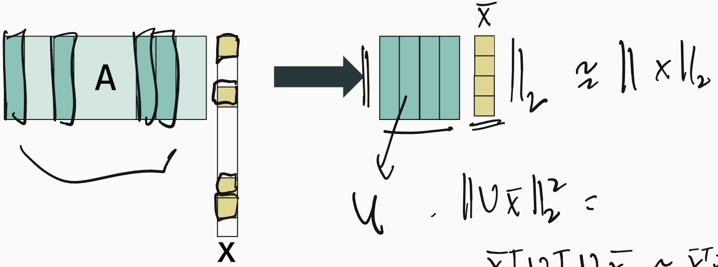


VALUE OF MANY ORTHOGONAL VECTORS

Definition ((q, ϵ)-Restricted Isometry Property)

A matrix $\mathbf{A}$ satisfies (q, ϵ)-RIP if, for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq q$,

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



$\| U \bar{x} \|_2 \approx \| x \|_2$

$\| U \bar{x} \|_2^2 =$

$\bar{x}^T \underbrace{U^T U}_{\approx I} \bar{x} \approx \bar{x}^T \bar{x}$

Every subset of k columns $\mathbf{U} \in \mathbb{R}^{m \times k}$ is approximate isometry.

$$\underline{\mathbf{U}^T\mathbf{U} \approx \mathbf{I}.}$$

VALUE OF MANY ORTHOGONAL VECTORS

$2^{.1^2 O(\log d)} = 2^{O(\log d)}$ $u_i^T u_j \leq .1$

If $\mathbf{U}^T\mathbf{U} \approx \mathbf{I}$, it better be that any two columns $\mathbf{u}_i, \mathbf{u}_j$ are approximately orthogonal.

Deduce: All pairs of columns in **A** are approximately orthogonal.

Think of k as a constant: $k = O(1)$. We have d nearly orthogonal vectors living in $O(k \log d) = O(\log d)$ dimensional space.

$O(\log d)$ dimensions

d vectors

$2^{O(\log d)} \geq d$ vectors in $O(\log d)$ dimensions that are approximately orthogonal

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?

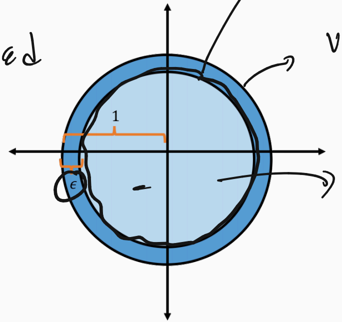
$\rightarrow 1 - e^{-\varepsilon d}$

fraction of volume

$$(1-\varepsilon)^d = \left(\underbrace{(1-\varepsilon)^{1/\varepsilon}}_{\approx 1/e}\right)^{\varepsilon d}$$

$$\downarrow$$

$$\approx \boxed{(1/e)^{\varepsilon d}}$$



volume $\frac{\pi^{d/2}}{(d/2)!}$

volume $\boxed{\frac{\pi^{d/2}}{(d/2)!}(1-\varepsilon)^d}$

$\overline{\pi^{d/2}/(d/2)!}$

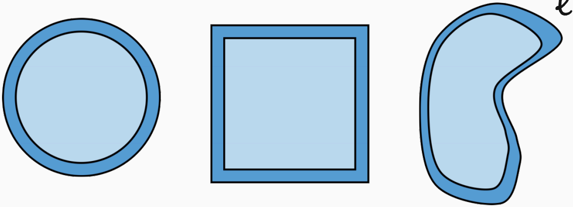
$= (1-\varepsilon)^d$

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

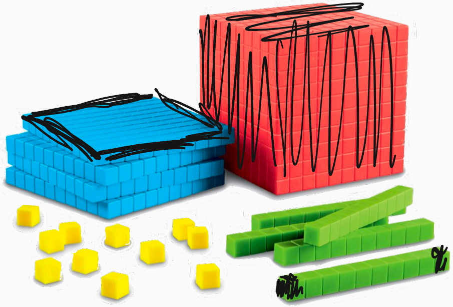
ISOPERIMETRIC INEQUALITY

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

Isoperimetric Inequality: For *any* shape, the ratio of $\frac{\text{surface area}}{\text{volume}}$ is *minimized* by the ℓ_2 ball.

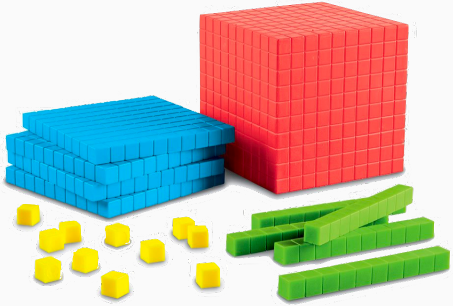


INTUITION



$$\frac{\text{surface cubes}}{\text{total cubes}} =$$

INTUITION



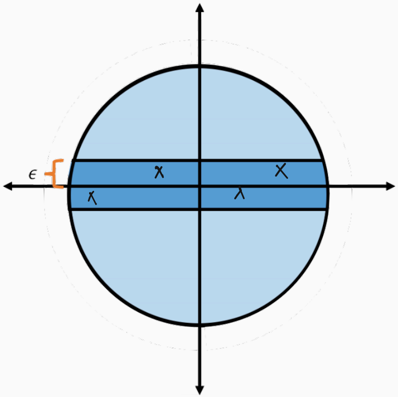
$\frac{\text{surface cubes}}{\text{total cubes}} =$

$2^{d/10}$

- **1 dimension:** 2/10 = .2
- **2 dimension:** 38/100 = .38
- **3 dimension:** 484/1000 = .484

SLICES OF THE UNIT BALL

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



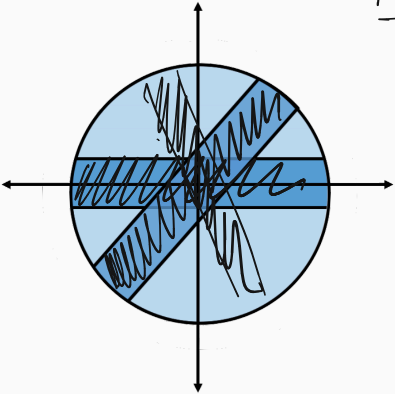
$$S = \{\underline{\mathbf{x} \in \mathcal{B}_d} : \underline{|\mathbf{x}(1)|} \leq \epsilon\}$$

SLICES OF THE UNIT BALL

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

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

volume

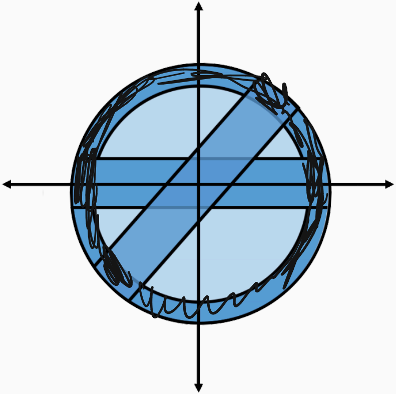


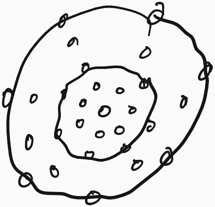
By symmetry, this is true for any equator:

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

BIZARRE SHAPE OF UNIT BALL

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





High-dimensional ball looks nothing like 2D ball!

CONCENTRATION AT EQUATOR

Claim: All but a $e^{\Theta(-\epsilon^2 d)}$ 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, $|\mathbf{x}(1)| \leq \epsilon$ with probability $\geq 1 - e^{\Theta(-\epsilon^2 d)}$.

CONCENTRATION AT EQUATOR

Let $\mathbf{w} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$.

$$\Pr\left[|\mathbf{x}(1)| \leq \epsilon\right] \geq \Pr\left[|\mathbf{w}(1)| \leq \epsilon\right].$$

How can we generate $\mathbf{w}$, which is a random vector taken by scaling a random $\mathbf{x} \in \mathcal{B}_d$?

CONCENTRATION AT EQUATOR

Let $\underline{\mathbf{g}}$ be a random Gaussian vector – each entry is $\underline{\mathcal{N}(0,1)}$. Set $\mathbf{w} = \underline{\mathbf{g}/\|\mathbf{g}\|_2}$.   want to show $|w(1)|$ is small.

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

  $1/2 \cdot d$ $\mathbb{E}(g_i^2)$ = variance = 1

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

With prob. $1 - 2^{-O(d)}$, $\|g\|_2^2 \geq d/2$

CONCENTRATION AT EQUATOR

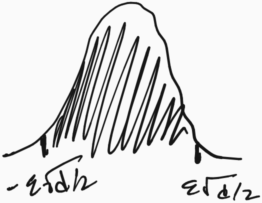
For $1 - 2^{\Theta(d)}$ fraction of vectors $\mathbf{g}$, $\|\mathbf{g}\|_2 \geq \sqrt{d/2}$. Condition on even that we get a random vector in this set.

$|g(1)|/\sqrt{d/2}$

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

$$|w(1)| = \frac{|g(1)|}{\|g\|_2} \leq \frac{|g(1)|}{\sqrt{d/2}}.$$

$\|g\|_2 \geq \sqrt{d/2}$

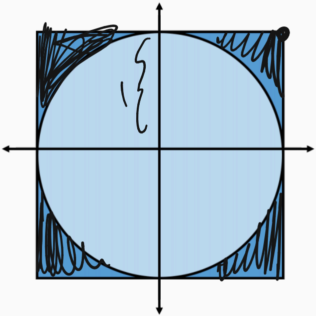


$$\boxed{1 - 2^{\theta(-\epsilon^2 d)}}$$

HIGH DIMENSIONAL CUBE

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

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



$$\left(\frac{\pi}{(d/2)}\right)\left(\frac{\pi}{(d/2-1)}\right) \cdots \frac{\pi}{\cdots} \approx O\left(\left(1/d\right)^{O(d)}\right)$$

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

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

This is a huge gap!.

HIGH DIMENSIONAL CUBE

Some other ways to see these shapes are very different:

- $\max_{\mathbf{x}\in\mathcal{B}_d} \|\mathbf{x}\|_2^2 =$ 1
- $\max_{\mathbf{x}\in\mathcal{C}_d} \|\mathbf{x}\|_2^2 =$ $\| [1, 1, 1, 1, 1, 1, 1, 3] \|_2^2 = d$

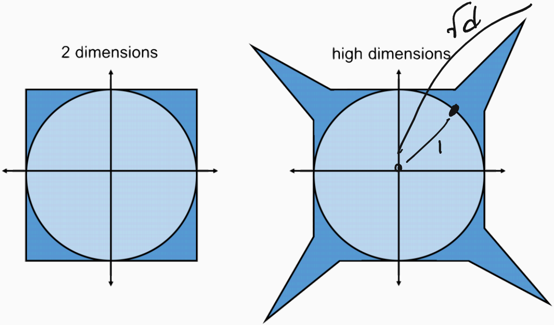
HIGH DIMENSIONAL CUBE

Some other ways to see these shapes are very different:

- $\mathbb{E}_{\mathbf{x}\sim\mathcal{B}_d}\|\mathbf{x}\|_2^2 \leq 1$
- $\mathbb{E}_{\mathbf{x}\sim\mathcal{C}_d}\|\mathbf{x}\|_2^2 = \sum_{i=1}^{d} \overbrace{\mathbb{E}(\underbrace{x(i)}_{\text{uniform from } [-1,\ 1]})^2}^{1/3} = d/3$

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.



CONNECTION TO 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?

CONNECTION TO DIMENSIONALITY REDUCTION

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

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

From our result earlier, in $\boxed{O(\log n/\epsilon^2)}$ dimensions, there exists $2^{O(\epsilon^2 \cdot \log n/\epsilon^2)} \geq n$ unit vectors that are close to mutually orthogonal.

$O(\log n/\epsilon^2)$ = <u>just enough</u> dimensions.

In d dimension there always exists $2^{O(\epsilon^2 d)}$ unit vectors that are ϵ close to orthogonal.

$2^{\epsilon^2 \log n/(\epsilon^2)} \geq n$

nearly orthogonal unit vectors.

THANK YOUR FOR A GREAT SEMESTER!

END OF SEMESTER

Exam format:

- Can have 1 double sided sheet of notes/equations.
- Will be designed for 1.5 hours, but won't cut off until 2 hours.

END OF SEMESTER

Exam topics:

- Very end of convex optimization (preconditioning, coordinate descent, + gradient descent for non-convex functions).
- Singular value decomposition and low-rank approximation.
- Spectral graph theory (stochastic block model, matrix perturbation, etc).
- Randomized linear algebra (subspace embeddings, approximate regression, ϵ-nets
- Sparse recovery (restricted isometry property, basis pursuit)

END OF SEMESTER

Please write a course review!

- `https://m.albert.nyu.edu/app/student/nyuCrseEval/crseEval/1198/24247/Y_LEC/10`