CS-GY 9223 D: Lecture 3
Surprises in High Dimensional Geometry

NYU Tandon School of Engineering, Prof. Christopher Musco

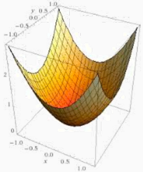
UNIFYING THEME OF THE COURSE

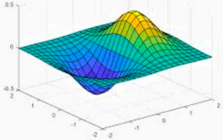
How do we deal with data (vectors) in high dimensions?

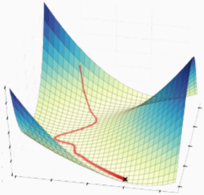
- Randomized sketching + dimensionality reduction.
- Locality sensitive hashing for similarity search.
- Iterative methods for optimizing functions that depend on many variables.
- SVD + low-rank approximation to find and visualize low-dimensional structure.
- Convert large graphs to high dimensional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

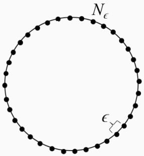
VISUALIZING HIGH DIMENSIONAL DATA

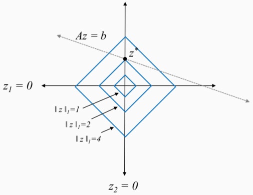
Often visualize data and algorithms in 1,2, or 3 dimensions.

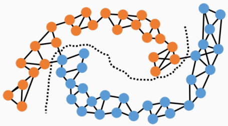












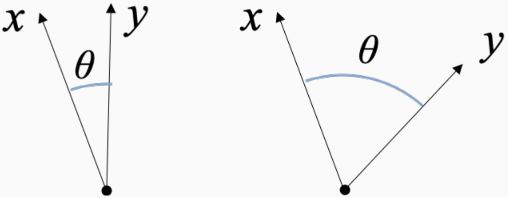
Warning for the rest of the semester: these images are rarely very informative! High-dimensional space looks very different from low-dimensional space.

ORTHOGONAL VECTORS

Recall the inner product between two d dimensional vectors:

$x = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \quad y = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$

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

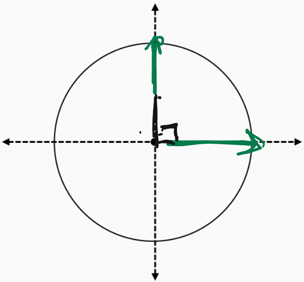


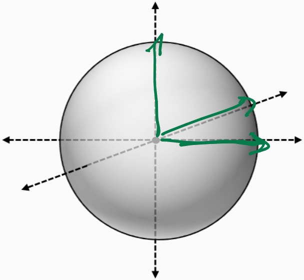
$$\langle x, y\rangle = \cos(\theta) \cdot \underline{\|x\|_2} \cdot \underline{\|y\|\_2}$$

ORTHOGONAL VECTORS

d is dimension

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.



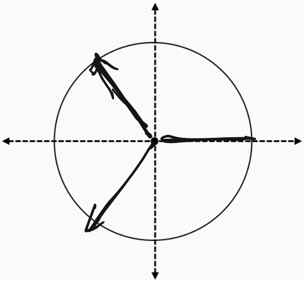


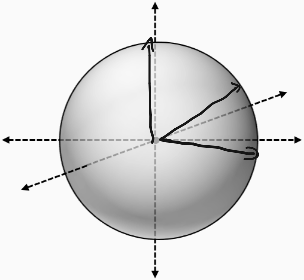
Exactly $\underline{\underline{d}}$.

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.

$\epsilon = .01$





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.

$\varepsilon = .01$

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 nearly orthogonal unit vectors with this property.

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.

- $\underline{\|\mathbf{x}_i\|_2} = \left(\sum_{i=1}^{d} x_i(\ell)^2\right)^{1/2} = \left(\sum_{i=1}^{d} \frac{1}{d}\right)^{1/2} = 1.$
- $\mathbb{E}[\mathbf{x}_i^T\mathbf{x}_j] = \sum_{\ell=1}^{d} \underbrace{x_i(\ell)\, x_j(\ell)}_{=0} = 0$
- $\mathrm{Var}[\mathbf{x}_i^T\mathbf{x}_j] = 1/d$

$$\mathrm{Var}\Big[\sum_{\ell=1}^{d} x_i(\ell)\, x_j(\ell)\Big] = \sum_{\ell=1}^{d} \overset{1/d^2}{\mathrm{Var}\big[x_i(\ell)\, x_j(\ell)\big]}$$

$$= 1/d$$

PROBABILISTIC METHOD

Let $Z = \underline{\mathbf{x}_i^T \mathbf{x}_j} = \sum_{i=1}^d C_i$ 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:

$$\underline{\Pr[|Z - \mathbb{E}Z| \geq \alpha \cdot \underset{1/\sqrt{d}}{\sigma}] \leq O(e^{-\alpha^2})} \rightarrow O(e^{-\varepsilon^2 d})$$

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

$\alpha \cdot \sigma = \varepsilon$

Note that we can transform to binary random variable:

$$Z = \sum_{i=1}^d C_i = \frac{2}{d}\sum_{i=1}^d \frac{d}{2} \cdot C_i$$

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

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

PROBABILISTIC METHOD

Formally, using a Chernoff bound:

$$\Pr[|Z - \mathbb{E}Z| \underline{\geq \epsilon] \leq \boxed{2e^{-\epsilon^2 d/3}}}$$

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

By a union bound:

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

$$\parallel$$

$$O(t)^2$$

$$t^2 \cdot 2e^{-\varepsilon^2 d/3} \leq 1/2$$

$$t = \frac{1}{2}\sqrt{\frac{1}{2e^{-\varepsilon^2 d/3}}} \approx \boxed{2^{O(\varepsilon^2 d)}}$$

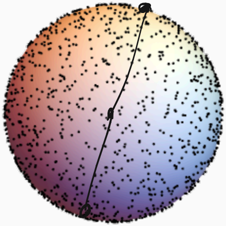
ORTHOGONAL VECTORS

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

Corollary: Most pairs of random vectors are far apart.

$\|x-y\|_2 \le 2$

$\sqrt{\|x\|_2^2 + \|y\|_2^2 - 2\underbrace{\langle x,y\rangle}_{\varepsilon}}$

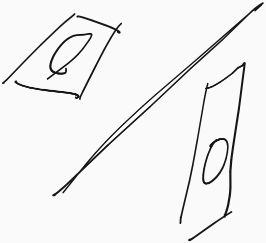


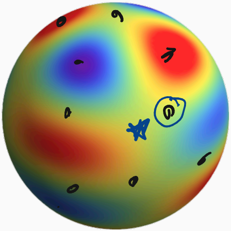
$\sqrt{1+1-2\varepsilon}$

$= \sqrt{2-2a}$

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.

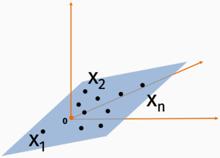


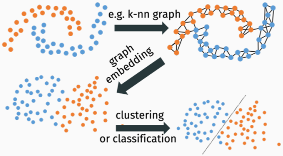


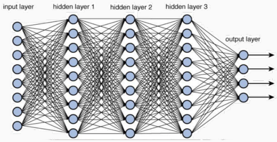
The existence of lower dimensional structure is our data is often the only reason we can hope to learn.

CURSE OF DIMENSIONALITY

Low-dimensional structure.







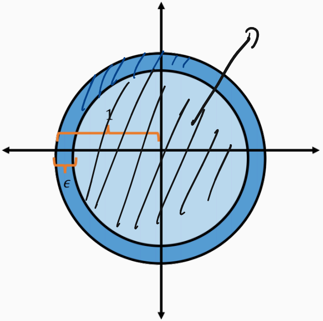
For example, data lies on low-dimensional subspace, or does so after transformation. Or function can be represented by a restricted class of functions, like neural net with specific structure.

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?



$$\frac{c \cdot (1-\varepsilon)^d}{c \cdot \cancel{1^d}}$$

$$\approx 1/e$$

$$\left((1-\varepsilon)^{1/\varepsilon}\right)^{\varepsilon d}$$

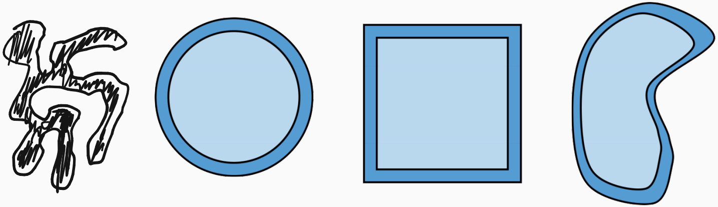
$$= \left(\frac{1}{e}\right)^{\varepsilon d}$$

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

ISOPERIMETRIC INEQUALITY

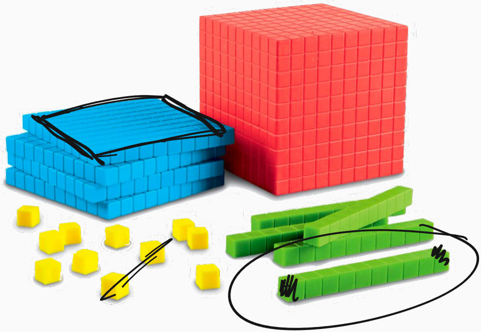
All but an $\underline{e^{\Theta(-\epsilon d)}}$ fraction of a unit ball's volume is within ϵ of its surface.

Isoperimetric Inequality: the ball has the ~~maximum~~ 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: $\frac{\text{surface cubes}}{\text{total cubes}} = .2$

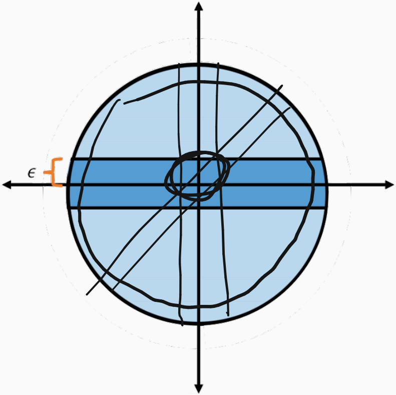
2D: $\frac{\text{surface cubes}}{\text{total cubes}} = .36$

3D: $\frac{\text{surface cubes}}{\text{total cubes}} = \frac{1000-512}{1000} = .49$

SLICES OF THE UNIT BALL

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

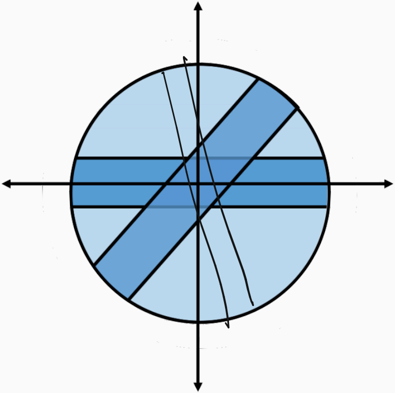
$\chi(1)$ first entry



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

SLICES OF THE UNIT BALL

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

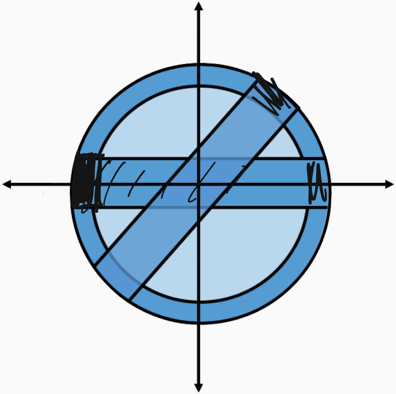


By symmetry, this is true for any equator:

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

BIZARRE SHAPE OF UNIT BALL

1. $(1 - e^{\Theta(-\epsilon d)})$ 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 x randomly from the unit ball, $|\mathbf{x}(1)| \leq \epsilon$ with probability $\geq 1 - e^{\Theta(-\epsilon^2 d)}$.

CONCENTRATION AT EQUATOR

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

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

How can we generate $\mathbf{w}$, which is a random vector taken by scaling a random $\mathbf{x} \in \mathcal{B}_d$? I.e., a random vector on the surface of the ball?

$$X = \left[\underbrace{\quad}_{\text{unif}(0,1)} \; \underbrace{\quad}_{\text{unif}(0,1)} \; \cdots \cdots \right]$$

CONCENTRATION AT EQUATOR

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

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

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

  $\frac{1}{2}d$

Rotationally Invariant

CONCENTRATION AT EQUATOR

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

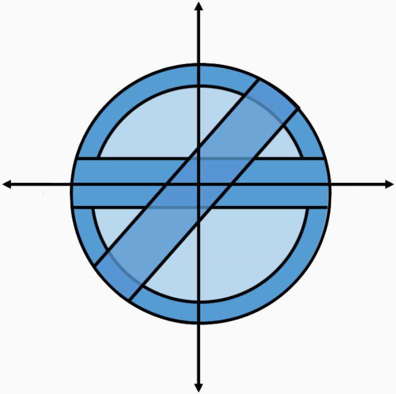
$$
\begin{aligned}
\underline{\Pr\left[|\mathbf{w}(1)| \leq \epsilon\right]} &= \Pr\left[|\mathbf{w}(1)| \cdot \sqrt{d/2} \leq \epsilon \cdot \sqrt{d/2}\right] \\
&\geq \boxed{\Pr\left[|\mathbf{g}(1)| \leq \boxed{\epsilon \cdot \sqrt{d/2}}\right]} \\
&\geq 1 - 2^{\theta\left(-(\epsilon \cdot \sqrt{d/2})^2\right)}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

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

Recall: $\underline{\mathbf{w} = \frac{\mathbf{g}}{\|\mathbf{g}\|_2}}$. So after conditioning, we have $\underline{\mathbf{w}_i} \leq \frac{\mathbf{g}_i}{\sqrt{d/2}}$.

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.

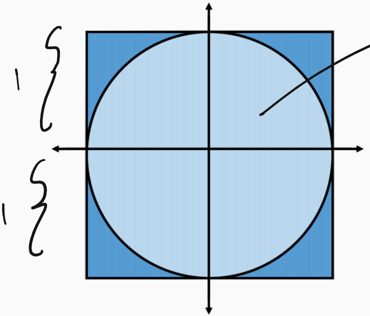


High-dimensional ball looks nothing like 2D ball!

HIGH DIMENSIONAL CUBE

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

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



$\frac{\sqrt{\pi}^d}{(d/2)!}$, 1

2^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 $\boxed{O(d)^{O(d)}}$ more volume.

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

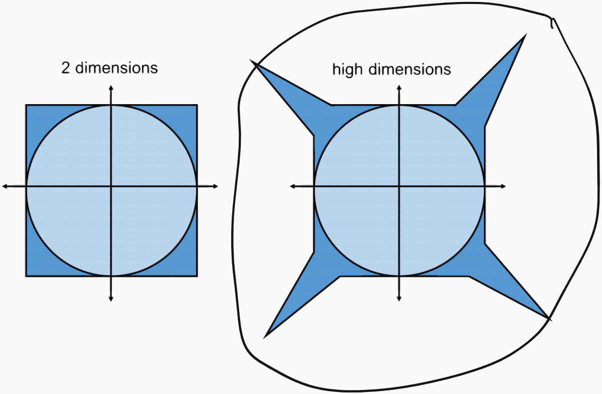
Some other ways to see these shapes are very different:

- $\mathbb{E}_{\mathbf{x}\sim\mathcal{B}_d}\|\mathbf{x}\|_2^2$ ≤ 1
- $\mathbb{E}_{\mathbf{x}\sim\mathcal{C}_d}\|\mathbf{x}\|_2^2 =$ $\sum_{i=1}^{d} \mathbb{E}\left(\text{unif}(-1,1)^2\right) = d/3$

$$\frac{1}{2}\int_{-1}^{1} x^2 = \frac{1}{2} \left.\frac{x^3}{3}\right|_{-1}^{1} = \frac{1}{3} = O(1)$$

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 NEXT LECTURE: THE JOHNSON-LINDENSTRAUSS LEMMA

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

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

CONNECTION TO NEXT LECTURE: THE JOHNSON-LINDENSTRAUSS LEMMA

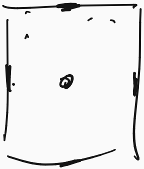
Despite **all this** warning that low-dimensional space looks nothing like high-dimensional space, next lecture we are going to learn about an algorithm to **compress high dimensional vectors to low dimensions.**

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

ADMINISTRATIVE THINGS

$$\mathrm{Var}[X] = \mathbb{E}\left[\left(X - \underset{0}{\cancel{\mathbb{E}[X]}}\right)^2\right] = \mathbb{E}[X^2] = \underline{1}$$

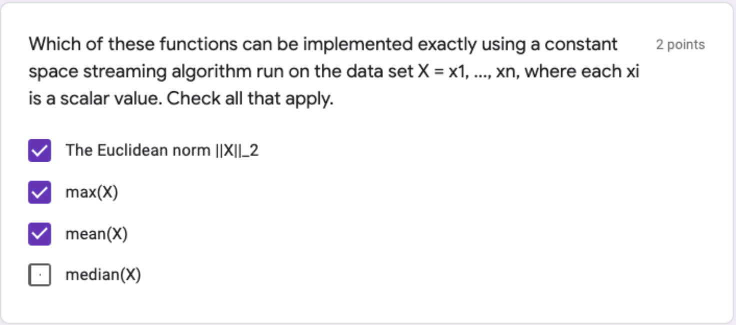
- Second problem set posted. I will release sample solutions to the first so you have a better sense of expected length/rigor.
- Remember to complete **poll for reading group time.**
- I will be posting a document with guidance on finding a project topic, and some suggested papers shortly.



$\underline{X(i)}$

$$X = [\underset{\hat{X}(1)}{\_\_}\ \underset{\bar{X}(2)}{\_\_}\ \_\_\ \_\ \underset{X(d)}{\_\_}]$$

WEEKLY QUIZ SOLUTIONS

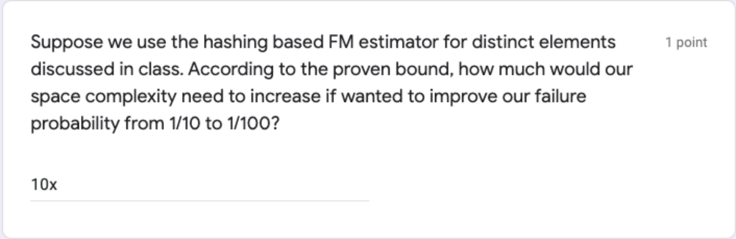


Which of these functions can be implemented exactly using a constant space streaming algorithm run on the data set X = x1, ..., xn, where each xi is a scalar value. Check all that apply.

2 points

- ☑ The Euclidean norm ||X||_2
- ☑ max(X)
- ☑ mean(X)
- □ median(X)

WEEKLY QUIZ SOLUTIONS

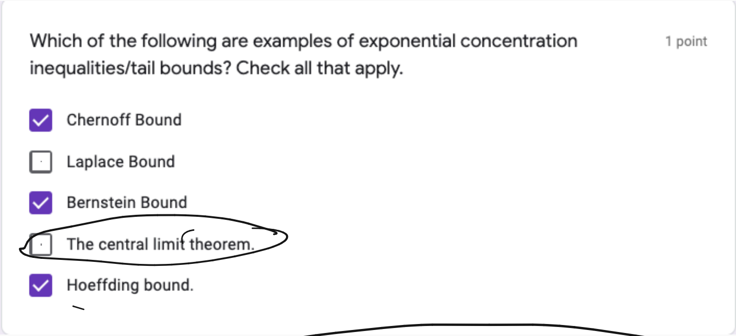


Suppose we use the hashing based FM estimator for distinct elements discussed in class. According to the proven bound, how much would our space complexity need to increase if wanted to improve our failure probability from 1/10 to 1/100?

1 point

10x

WEEKLY QUIZ SOLUTIONS



Which of the following are examples of exponential concentration inequalities/tail bounds? Check all that apply.

1 point

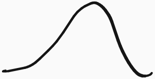
☑ Chernoff Bound

□ Laplace Bound

☑ Bernstein Bound

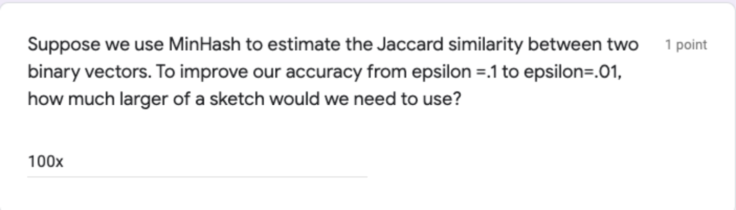
□ The central limit theorem.

☑ Hoeffding bound.



$$\Pr\{|X-\mathbb{E}X] \geq \delta\} \leq \Delta$$

WEEKLY QUIZ SOLUTIONS

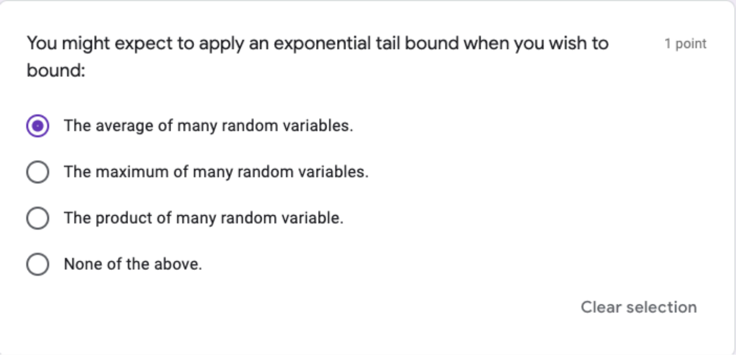


Suppose we use MinHash to estimate the Jaccard similarity between two binary vectors. To improve our accuracy from epsilon =.1 to epsilon=.01, how much larger of a sketch would we need to use?

1 point

100x

WEEKLY QUIZ SOLUTIONS

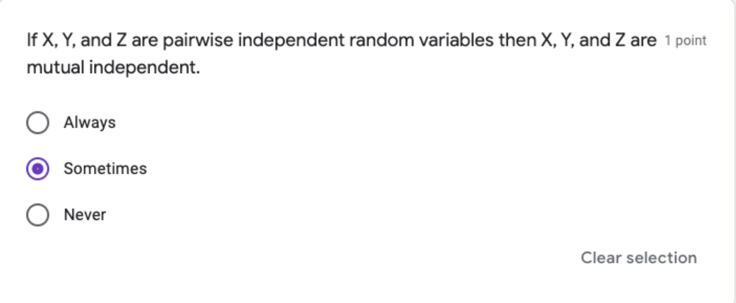


You might expect to apply an exponential tail bound when you wish to bound:

1 point

- ● The average of many random variables.
- ○ The maximum of many random variables.
- ○ The product of many random variable.
- ○ None of the above.

WEEKLY QUIZ SOLUTIONS



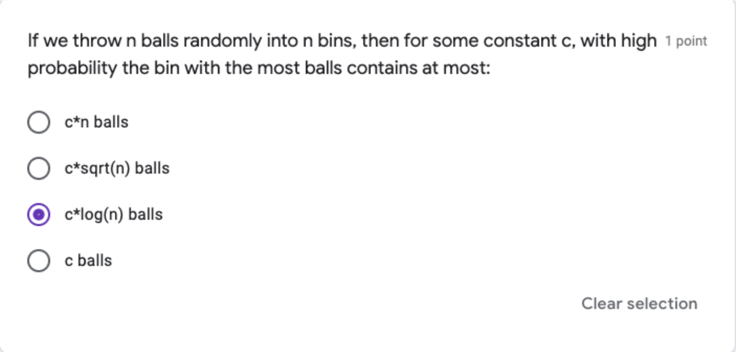
If X, Y, and Z are pairwise independent random variables then X, Y, and Z are mutual independent. 1 point

○ Always

● Sometimes

○ Never

WEEKLY QUIZ SOLUTIONS



If we throw n balls randomly into n bins, then for some constant c, with high probability the bin with the most balls contains at most: 1 point

○ c*n balls

○ c*sqrt(n) balls

● c*log(n) balls

○ c balls