Euclidean Norm

CE40282-1: Linear Algebra

Hamid R. Rabiee and Maryam Ramezani
Sharif University of Technology

The reason to use norms

- Machine learning uses vectors, matrices, and tensors as the basic units of representation
- Two reasons to use norms
  - To estimate how **big** a vector/matrix/tensor is
    - How big is the difference between two tensors is
  - To estimate how **close** one tensor is to another
    - How close is one image to another

Root-mean-square value

- Mean-square (MS) value of n-vector x is:

$$\frac{x_1^2 + \cdots + x_n^2}{n} = \frac{\|x\|^2}{n}$$

- Root-mean-square value (RMS)

$$\mathbf{rms}(x) = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}} = \frac{\|x\|}{\sqrt{n}}$$

- The RMS value of a vector x is useful when comparing norms of vectors with different dimensions
- rms(x) gives 'typical' value of $|x_i|$
  - e.g., rms(**1**) = 1 (independent of n)
  - if all the entries of a vector are the same, (a) then the RMS value of the vector is |a|

Chebyshev inequality

- suppose that k of the numbers $|x_1|, \ldots, |x_n|$ are $\geq a$

  then k of the numbers $x_1^2, \ldots, x_n^2$ are $\geq a^2$

  so $\|x\|^2 = x_1^2 + \cdots + x_n^2 \geq ka^2$

  so we have $k \leq \|x\|^2 / a^2$

  number of x_i with $|x_i| \geq a$ is no more than $\|x\|^2 / a^2$

  this is the *Chebyshev inequality*
- What happens when $\|x\|^2/a^2 \geq n$?
- No entry of a vector can be larger in magnitude than the norm of the vector

Chebyshev inequality

- Chebyshev inequality is easier to interpret in terms of the RMS value of a vector.

$$\frac{k}{n} \leq \left(\frac{\mathbf{rms}(x)}{a}\right)^2$$

- How many entries of x can have value more than 5rms(x)?
- The Chebyshev inequality partially justifies the idea that the RMS value of a vector gives an idea of the size of a typical entry: It states that not too many of the entries of a vector can be much bigger (in absolute value) than its RMS value

Standard deviation

- for n-vector x, $\mathbf{avg}(x) = \mathbf{1}^T x/n$
- *de-meaned vector* is $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ (so $\mathbf{avg}(\tilde{x}) = 0$)
- *standard deviation* of x is

$$\mathbf{std}(x) = \mathbf{rms}(\tilde{x}) = \frac{\|x - (\mathbf{1}^T x/n)\mathbf{1}\|}{\sqrt{n}}$$

- $\mathbf{std}(x)$ gives 'typical' amount x_i vary from $\mathbf{avg}(x)$
- $\mathbf{std}(x) = 0$ only if $x = \alpha\mathbf{1}$ for some α
- greek letters μ, σ commonly used for mean, standard deviation
- a basic formula:

$$\mathbf{rms}(x)^2 = \mathbf{avg}(x)^2 + \mathbf{std}(x)^2$$

Chebyshev inequality for standard deviation

x is an n-vector with mean $\mathbf{avg}(x)$, standard deviation $\mathbf{std}(x)$

rough idea: most entries of x are not too far from the mean

by Chebyshev inequality, fraction of entries of x with

$$|x_i - \mathbf{avg}(x)| \geq \alpha \ \mathbf{std}(x)$$

is no more than $1/\alpha^2$ (for $\alpha > 1$)

- The fraction of entries of x within θ standard deviations of avg(x) is at least $(1 - \frac{1}{\theta^2})$ for $\theta > 1$

Properties of standard deviation

- *Adding a constant.* For any vector x and any number a, we have $\mathbf{std}(x + a\mathbf{1}) = \mathbf{std}(x)$. Adding a constant to every entry of a vector does not change its standard deviation.

- *Multiplying by a scalar.* For any vector x and any number a, we have $\mathbf{std}(ax) = |a| \, \mathbf{std}(x)$. Multiplying a vector by a scalar multiplies the standard deviation by the absolute value of the scalar.

Vector Standardization

$$z = \frac{1}{\mathbf{std}(x)}(x - \mathbf{avg}(x)\mathbf{1}).$$

- It has mean zero, and standard deviation one.
- Its entries are sometimes called the z-scores associated with the original entries of x.
- The standardized values for a vector give a simple way to interpret the original values in the vectors.

Cauchy–Schwarz inequality

- - for two n-vectors a and b, $|a^T b| \leq \|a\| \|b\|$
  - written out,

$$|a_1 b_1 + \cdots + a_n b_n| \leq \left(a_1^2 + \cdots + a_n^2\right)^{1/2} \left(b_1^2 + \cdots + b_n^2\right)^{1/2}$$

Derivation of Cauchy–Schwarz inequality

- it's clearly true if either a or b is 0

  so assume $\alpha = \|a\|$ and $\beta = \|b\|$ are nonzero

  we have

$$
\begin{aligned}
0 &\leq \|\beta a - \alpha b\|^2 \\
&= \|\beta a\|^2 - 2(\beta a)^T(\alpha b) + \|\alpha b\|^2 \\
&= \beta^2\|a\|^2 - 2\beta\alpha(a^T b) + \alpha^2\|b\|^2 \\
&= 2\|a\|^2\|b\|^2 - 2\|a\|\,\|b\|(a^T b)
\end{aligned}
$$

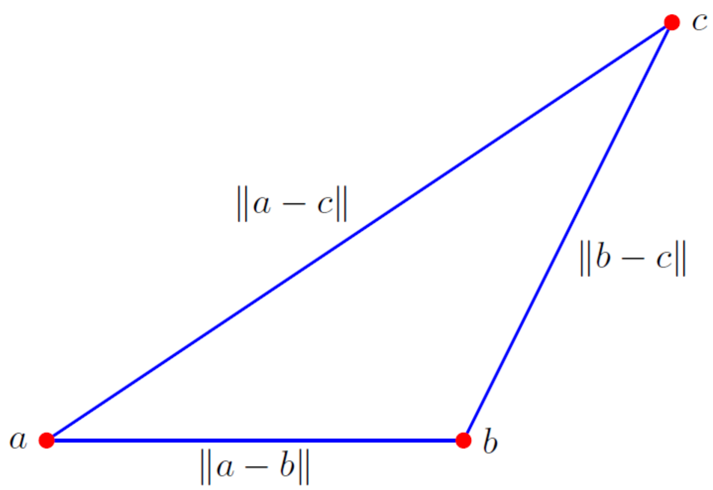
  divide by $2\|a\|\,\|b\|$ to get $a^T b \leq \|a\|\,\|b\|$

  apply to $-a$, b to get other half of Cauchy–Schwarz inequality

- Cauchy-Schwarz inequality holds with equality when one of the vectors is a multiple of the other

Triangle inequality

- Consider a triangle in two or three dimensions, whose vertices have coordinates a, b, and c.



Cauchy–Schwarz inequality

- Verification of triangle inequality.

$$
\begin{aligned}
\|a+b\|^2 &= \|a\|^2 + 2a^Tb + \|b\|^2 \\
&\leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \\
&= (\|a\| + \|b\|)^2
\end{aligned}
$$

Euclidean Norm

- Euclidean Norm (2-norm, l_2 norm, length)
  - A vector whose length is 1 is called a unit vector
  - Normalizing: divide a nonzero vector by its length which is a unit vector in the same direction of original vector

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x^T x}$$

  - It is a nonnegative scalar
  - In R^2 follows from the Pythagorean Theorem.
  - What about R^3?
  - What is the shape of $\|x\|_2 = 1$?

Vector Norms Properties

- Absolute homogeneity/Linearity:
  - $||\alpha x|| = |\alpha|||x||$
- Subadditivity/Triangle inequality
  - $||x + y|| \leq ||x|| + ||y||$
- Positive definiteness/Point separating
  - If $||x|| = 0$ then $x = 0$
  - (1&3): For every x, $||x|| = 0$ if and only if $x = 0$
- Non-negativity
  - $||x|| \geq 0$

Nom of sum

- If x and y are vectors:

$$\|x+y\| = \sqrt{\|x\|^2 + 2x^T y + \|y\|^2}.$$

  - Proof:

$$\begin{aligned} \|x+y\|^2 &= (x+y)^T(x+y) \\ &= x^T x + x^T y + y^T x + y^T y \\ &= \|x\|^2 + 2x^T y + \|y\|^2. \end{aligned}$$

Norm of block vectors

- suppose a, b, c are vectors
- $\|(a, b, c)\|^2 = a^T a + b^T b + c^T c = \|a\|^2 + \|b\|^2 + \|c\|^2$
- so we have

$$\|(a, b, c)\| = \sqrt{\|a\|^2 + \|b\|^2 + \|c\|^2} = \|(\|a\|, \|b\|, \|c\|)\|$$

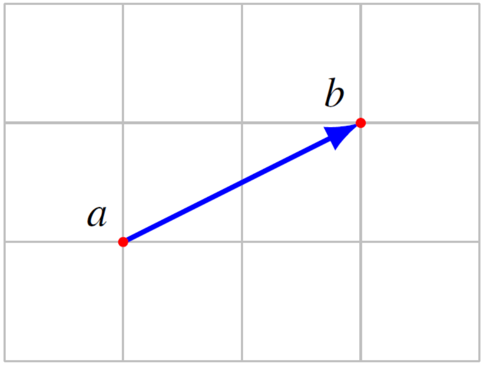
(parse RHS very carefully!)

- The norm of a stacked vector is the norm of the vector formed from the norms of the sub vectors.

Euclidean distance

- Distance

$$\mathbf{dist}(a, b) = \|a - b\|$$



- RMS deviation between the two vectors

$$\mathbf{rms}(a - b) \qquad \|a - b\| / \sqrt{n}$$

Euclidean distance

- Distance between two n-vectors shows the vectors are “close” or “nearby” or “far”.

As an example, consider

$$u = \begin{bmatrix} 1.8 \\ 2.0 \\ -3.7 \\ 4.7 \end{bmatrix}, \qquad v = \begin{bmatrix} 0.6 \\ 2.1 \\ 1.9 \\ -1.4 \end{bmatrix}, \qquad w = \begin{bmatrix} 2.0 \\ 1.9 \\ -4.0 \\ 4.6 \end{bmatrix}.$$

The distances between pairs of them are

$$\|u - v\| = 8.368, \qquad \|u - w\| = 0.387, \qquad \|v - w\| = 8.533,$$

Compare norm and distance

Norm* *(Normed Linear Space)	***Distance function* *(Metric Space)***
1. $\|\|x-y\|\| \geq 0$	1. $d(x,y) \geq 0$
2. $\|\|x-y\|\| = 0 \implies x = y$	2. $d(x,y) = 0 \implies x = y$
3. $\|\|\lambda(x-y)\|\| = \|\lambda\| \, \|\|x-y\|\|$	3. $d(x,y) = d(y,x)$

Angle

- *angle* between two nonzero vectors a, b defined as

$$\angle(a,b) = \arccos\left(\frac{a^T b}{\|a\| \, \|b\|}\right)$$

- $\angle(a,b)$ is the number in $[0,\pi]$ that satisfies

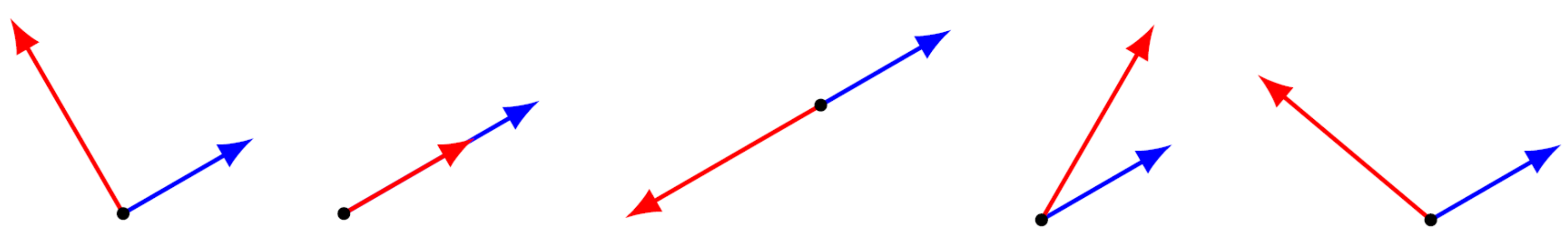
$$a^T b = \|a\| \, \|b\| \cos\left(\angle(a,b)\right)$$

- coincides with ordinary angle between vectors in 2-D and 3-D

Classification of angles

$\theta = \angle(a, b)$

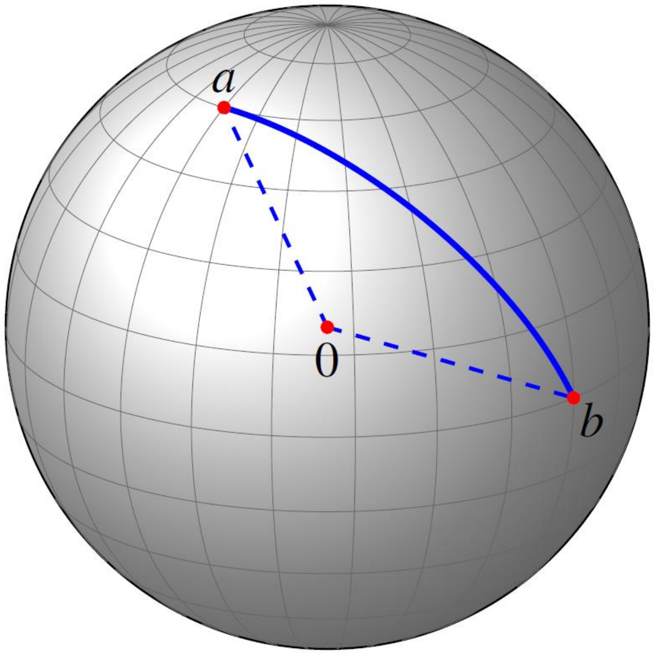
- $\theta = \pi/2 = 90^\circ$: a and b are *orthogonal*, written $a \perp b$ ($a^T b = 0$)
- $\theta = 0$: a and b are *aligned* ($a^T b = \|a\| \|b\|$)
- $\theta = \pi = 180^\circ$: a and b are *anti-aligned* ($a^T b = -\|a\| \, \|b\|$)
- $\theta \leq \pi/2 = 90^\circ$: a and b make an *acute angle* ($a^T b \geq 0$)
- $\theta \geq \pi/2 = 90^\circ$: a and b make an *obtuse angle* ($a^T b \leq 0$)



Applications

Spherical distance

if a, b are on sphere of radius R, distance *along the sphere* is $R\angle(a,b)$



Applications

Correlation coefficient

- vectors a and b, and de-meaned vectors

$$\tilde{a} = a - \mathbf{avg}(a)\mathbf{1}, \qquad \tilde{b} = b - \mathbf{avg}(b)\mathbf{1}$$

- *correlation coefficient* (between a and b, with $\tilde{a} \neq 0$, $\tilde{b} \neq 0$)

$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\| \, \|\tilde{b}\|}$$

- $\rho = \cos \angle(\tilde{a}, \tilde{b})$
  - $\rho = 0$: a and b are *uncorrelated*
  - $\rho > 0.8$ (or so): a and b are *highly correlated*
  - $\rho < -0.8$ (or so): a and b are *highly anti-correlated*
- very roughly: highly correlated means a_i and b_i are typically both above (below) their means together

Applications

Document dissimilarity by angles

- measure dissimilarity by angle of word count histogram vectors
- pairwise angles (in degrees) for 5 Wikipedia pages shown below

	Veterans Day	Memorial Day	Academy Awards	Golden Globe Awards	Super Bowl
Veterans Day	0	60.6	85.7	87.0	87.7
Memorial Day	60.6	0	85.6	87.5	87.5
Academy A.	85.7	85.6	0	58.7	85.7
Golden Globe A.	87.0	87.5	58.7	0	86.0
Super Bowl	87.7	87.5	86.1	86.0	0

Vector Norms

- p-norm:

$$||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

$$p \geq 1$$

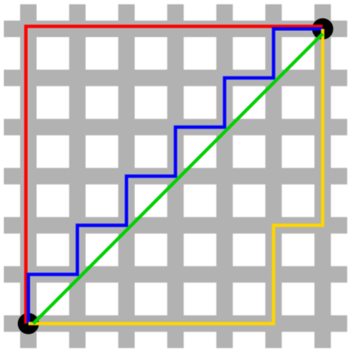
- What is the shape of $||x||_p = 1$?

Vector Norms

- 1-norm: $(\boldsymbol{l_1})$

$$||x||_1 = (|x_1| + |x_2| + \cdots + |x_n|)$$

- What is the shape of $||x||_1 = 1$?
- The distance between two vectors under the L1 norm is also referred to as the Manhattan distance
- Example:
  - L1 distance between (0,1) and (1,0)?



Vector Norms

- ∞-norm: ($\boldsymbol{l}_\infty$) (max norm)

$$L_\infty = max(|x_1|, |x_2|, \dots, |x_n|)$$

- What is the shape of $||x||_\infty = 1$?

Vector Norms

- $\frac{\mathbf{1}}{\mathbf{2}}$-norm: $(\boldsymbol{l}_{\frac{\mathbf{1}}{\mathbf{2}}})$

- What is the shape of $||x||_{\frac{\mathbf{1}}{\mathbf{2}}} = 1$?

Vector Norms

- $\boldsymbol{zero}$-norm: $(\boldsymbol{l_0})$

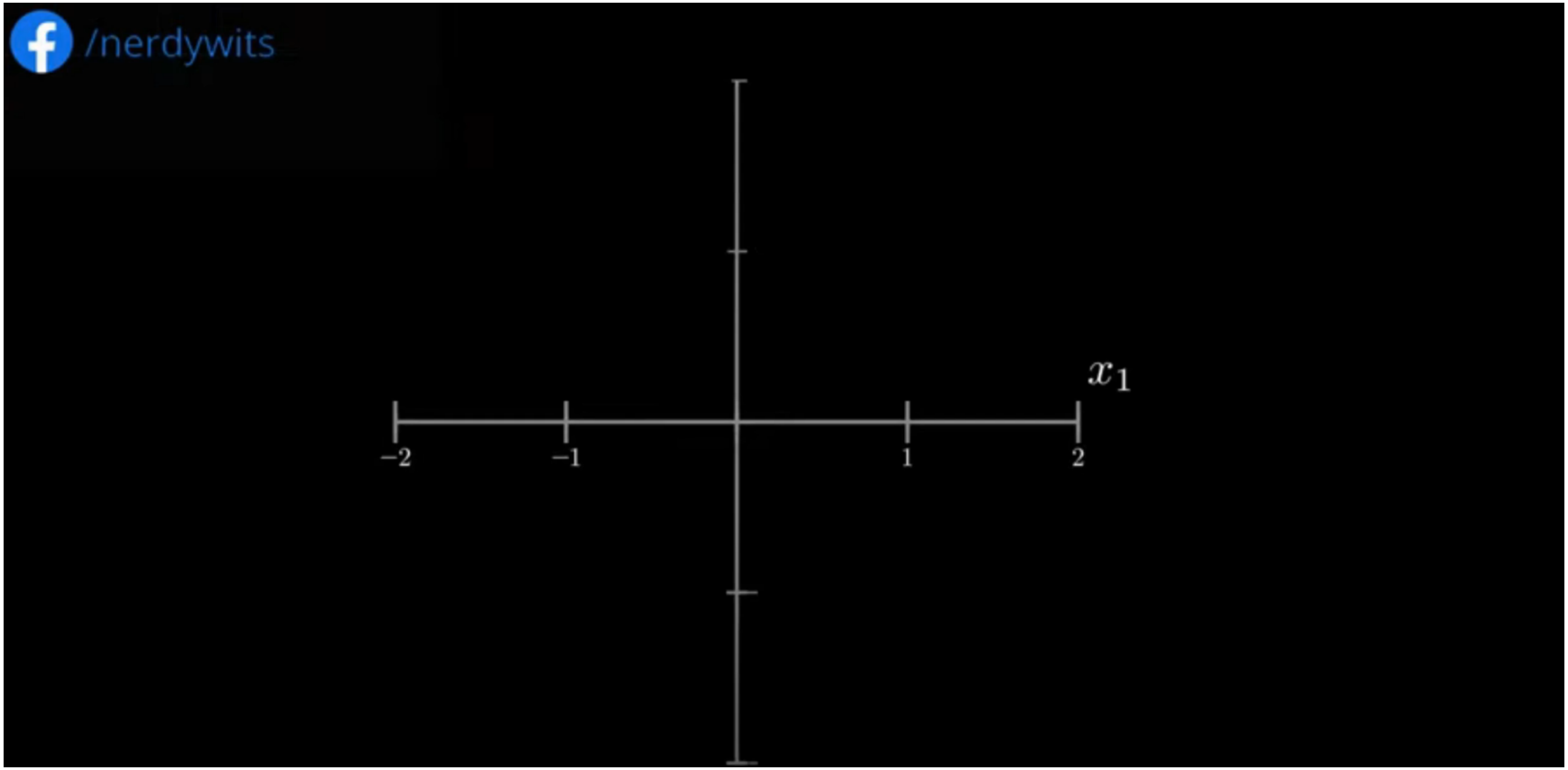
$$||x||_0 = \lim_{\alpha \to 0^+} ||x||_\alpha = \left(\sum_{k=1}^{n} |x|^\alpha \right)^{1/\alpha} = \sum_{k=1}^{n} 1_{(0,\infty)}(|x|)$$

- Zero-norm, defined as **the number of non-zero elements in a vector**, is an ideal quantity for feature selection. However, minimization of zero-norm is generally regarded as a combinatorically difficult optimization
- $||x||_0 = \sum_{x_i \neq 0} 1$

Vector Norms

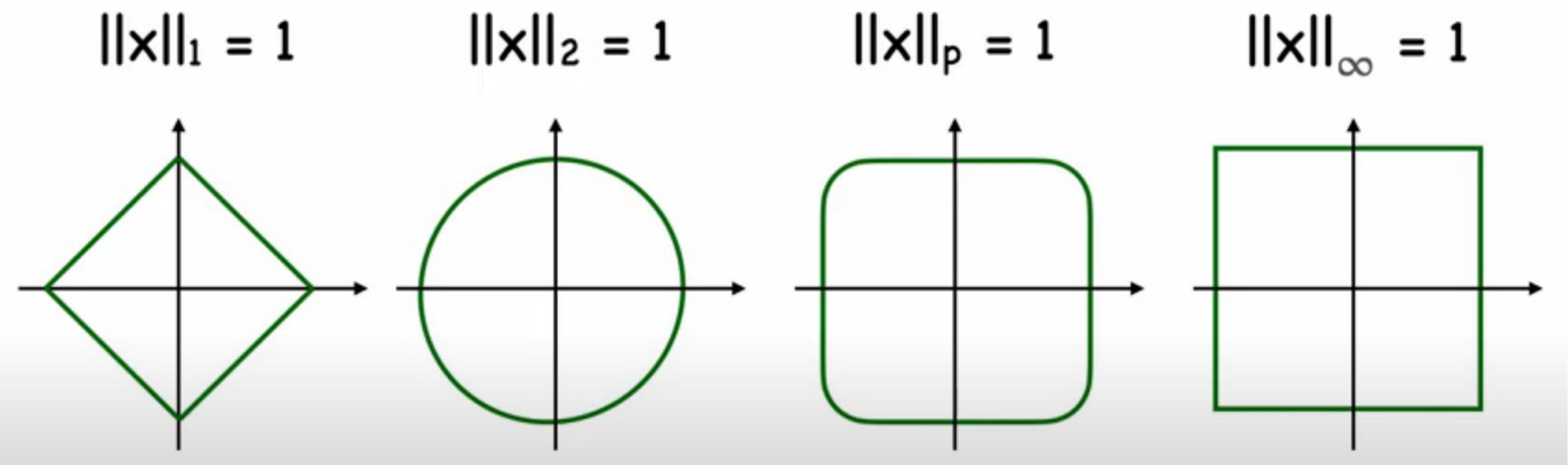
- Is zero-norm a norm??
- What is the shape of $||x||_0 = 1$?
- Examples:
  - L0 distance between (0,0) and (0,5)?
  - L0 distance between (1,1) and (2,2)?
  - (username,password)

Vector Norms Shapes



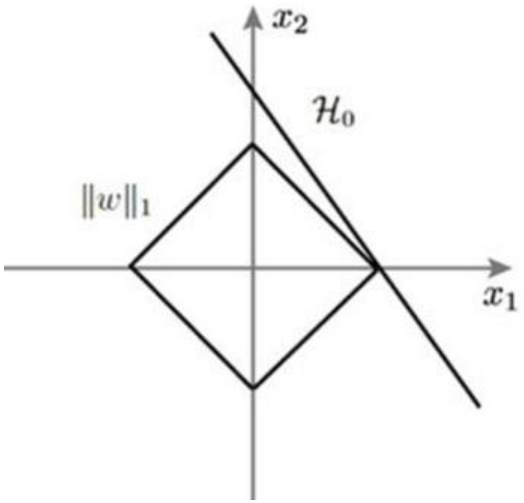
Norms and Convexity

- For $p \geq 1$, l_p norm is convex



Norms and Convexity

***Theorem:** If A is a convex subset of a normed linear space B whose norm is strictly convex, then, for every f in B, there exists a unique best approximation a* in A to f*



Norm Derivations

- Square of l_2

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{bmatrix} \quad ||u||_2 = u_1^2 + u_2^2 + \cdots + u_n^2$$

$$\begin{cases} \dfrac{d||u||_2}{du_1} = 2u_1 \\ \dfrac{d||u||_2}{du_2} = 2u_2 \\ \cdots \\ \dfrac{d||u||_2}{du_n} = 2u_n \end{cases}$$

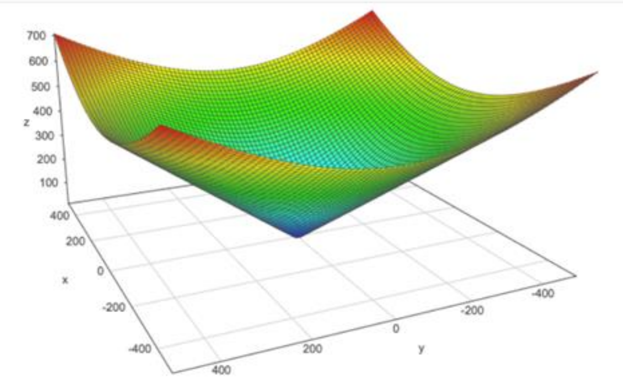
- l_2

$$||u||_2 = \sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)} = (u_1^2 + u_2^2 + \cdots + u_n^2)^{\frac{1}{2}}$$

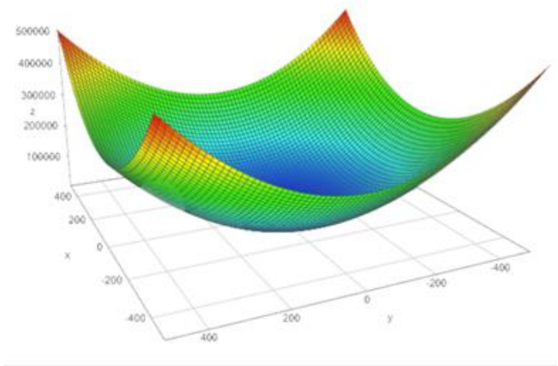
$$\begin{aligned} \frac{d||u||_2}{du_1} &= \frac{1}{2}(u_1^2 + u_2^2 + \cdots + u_n^2)^{\frac{1}{2}-1} \cdot \frac{d}{du_1}(u_1^2 + u_2^2 + \cdots + u_n^2) \\ &= \frac{1}{2}(u_1^2 + u_2^2 + \cdots + u_n^2)^{-\frac{1}{2}} \cdot \frac{d}{du_1}(u_1^2 + u_2^2 + \cdots + u_n^2) \\ &= \frac{1}{2} \cdot \frac{1}{(u_1^2 + u_2^2 + \cdots + u_n^2)^{\frac{1}{2}}} \cdot \frac{d}{du_1}(u_1^2 + u_2^2 + \cdots + u_n^2) \\ &= \frac{1}{2} \cdot \frac{1}{(u_1^2 + u_2^2 + \cdots + u_n^2)^{\frac{1}{2}}} \cdot 2 \cdot u_1 \\ &= \frac{u_1}{\sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)}} \end{aligned}$$

$$\begin{cases} \dfrac{d||u||_2}{du_1} = \dfrac{u_1}{\sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)}} \\ \dfrac{d||u||_2}{du_2} = \dfrac{u_2}{\sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)}} \\ \cdots \\ \dfrac{d||u||_2}{du_n} = \dfrac{u_n}{\sqrt{(u_1^2 + u_2^2 + \cdots + u_n^2)}} \end{cases}$$

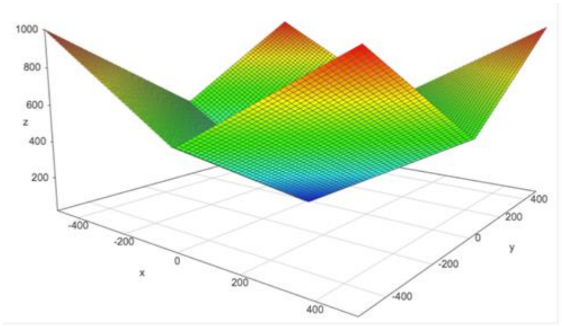
Norm Comparisons



l_2 norm

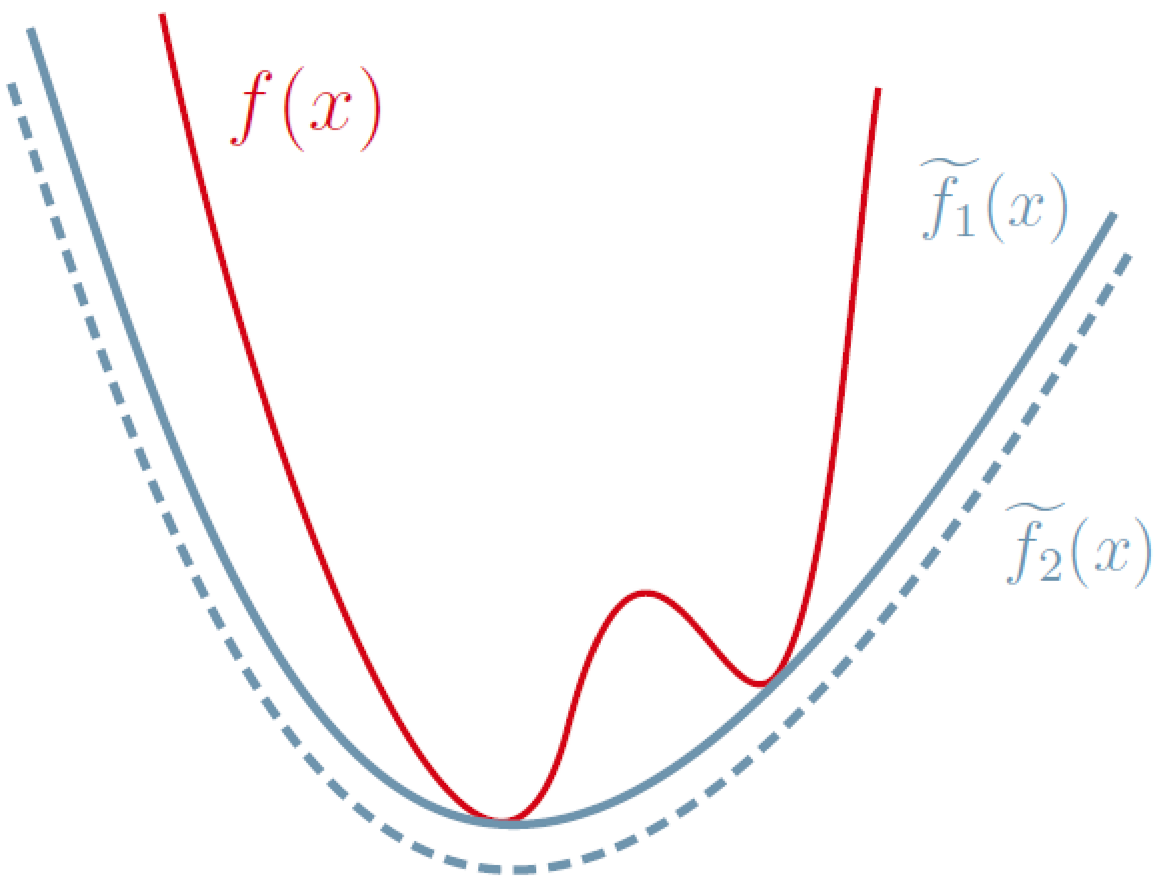


Square l_2 norm



l_1 norm

Convex Relaxation

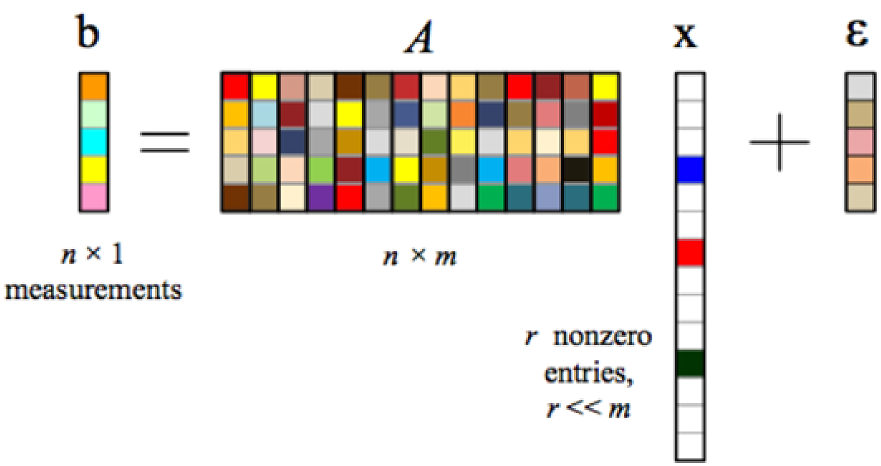


Sparse applications

- **Alternative viewpoint:** We try to find the sparsest solution which explains our noisy measurements

$$\min_{x} \|\mathbf{x}\|_0 \qquad \text{subject to } \|\mathbf{Ax}-\mathbf{b}\|_2 < \varepsilon$$

- Here, the l_0-norm is a shorthand notation for *counting the number of non-zero elements in* x.



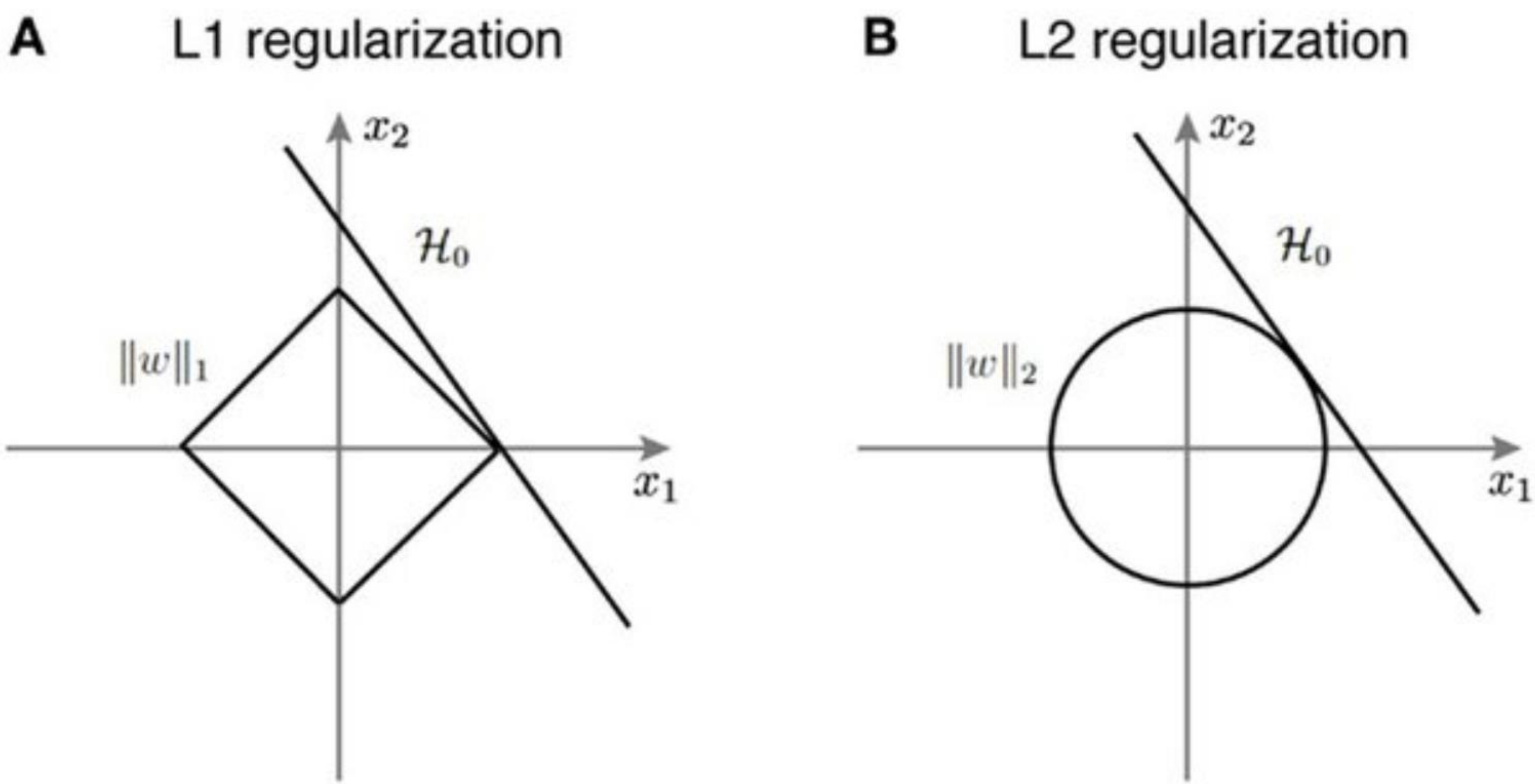
Sparse Solution

- l_0 optimization is np-hard
- Convex relaxation for solving the problem

$$\min_x \|x\|_1 \quad \text{subject to } \|Ax-b\|_2 < \varepsilon$$

$$\min_x \|x\|_0 \quad \text{subject to } \|Ax-b\|_2 < \varepsilon$$

Why is L1 supposed to lead to sparsity than L2?



Vector Norms

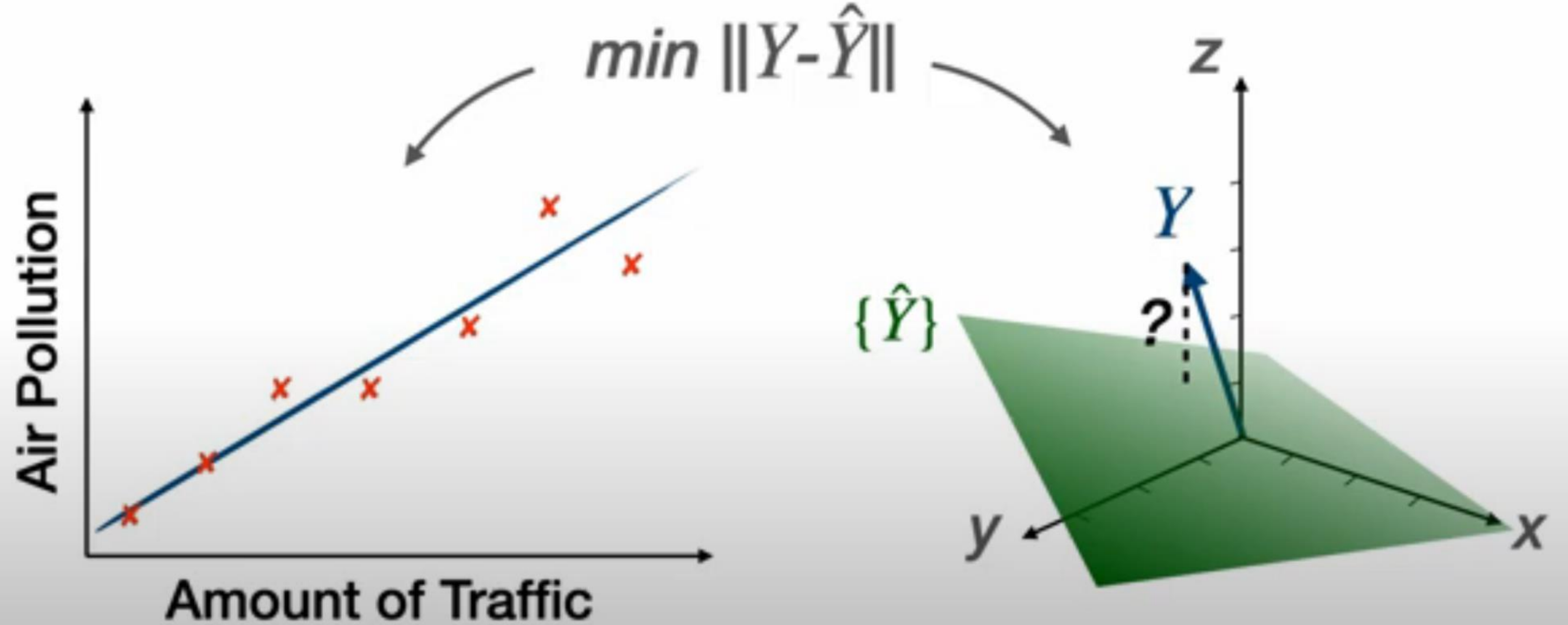
- Which norm is the convex hull of the intersection between the L0 norm ball and L2 norm ball?

- Any valid norm $||.||$ is a convex function.
  - Proof?

- The L0 norm is not convex.
  - Proof?

ML application

The best linear regression model comes from choosing the closest $\hat{Y}$ to Y based on

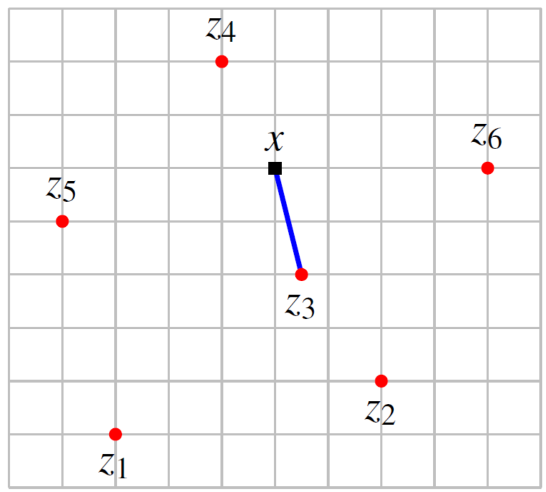


ML Application

Feature distance and nearest neighbors

- if x and y are feature vectors for two entities, $\|x - y\|$ is the *feature distance*
- if $z_1, \ldots, z_m$ is a list of vectors, z_j is the *nearest neighbor* of x if

$$\|x - z_j\| \leq \|x - z_i\|, \quad i = 1, \ldots, m$$



- these simple ideas are very widely used

- Number of flops and order?

ML Application

Document dissimilarity

- 5 Wikipedia articles: ‘Veterans Day’, ‘Memorial Day’, ‘Academy Awards’, ‘Golden Globe Awards’, ‘Super Bowl’
- word count histograms, dictionary of 4423 words
- pairwise distances shown below

	Veterans Day	Memorial Day	Academy Awards	Golden Globe Awards	Super Bowl
Veterans Day	0	0.095	0.130	0.153	0.170
Memorial Day	0.095	0	0.122	0.147	0.164
Academy A.	0.130	0.122	0	0.108	0.164
Golden Globe A.	0.153	0.147	0.108	0	0.181
Super Bowl	0.170	0.164	0.164	0.181	0

Complexity

- Norm: 2n flops. O(n)
- RMS: 2n flops. O(n)
- Distance: 3n flops. O(n)
- Angle: 6n flops. O(n)
- Standard deviation: 4n flops. O(n) can reduce to 3n flops $\mathbf{std}(x)^2 = \mathbf{rms}(x)^2 - \mathbf{avg}(x)^2,$
- Standardizing: 5n flops. O(n)
- Correlation coefficient: 10n flops. O(n)

Conclusion

By a normed linear space (briefly normed space) is meant a real or complex vector space E in which every vector x is associated with a real number $|x|$, called its absolute value or norm, in such a manner **that the properties** $(\mathbf{a}') - (\mathbf{c}')$ of §9 hold. That is, for any vectors $x, y \in E$ and scalar a, we have

(i) $|x| \geq 0$;

(i′) $|x| = 0$ iff $x = \overrightarrow{0}$;

(ii) $|ax| = |a||x|$; and

(iii) $|x + y| \leq |x| + |y|$ (triangle inequality).

Conclusion

A metric space is a set $S \neq \emptyset$ together with a function

$$\rho : S \times S \to E^1$$

(called a metric for S) satisfying the metric laws (axioms):

For any x, y, and z in S, we have

i. $\rho(x, y) \geq 0$, and (i') $\rho(x, y) = 0$ iff $x = y$;
ii. $\rho(x, y) = \rho(y, x)$ (symmetry law); and
iii. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$(triangle law).

Reference

- Linear Algebra and Its Applications David C. Lay
- Introduction to Applied Linear Algebra Vectors, Matrices, and Least Squares
- https://www.youtube.com/watch?v=76B5cMEZA4Y