

Linear Algebra Review

Stephen Boyd and Sanjay Lall
Revision by Amirhossein Afsharrad

EE263
Stanford University

Norms: Measuring Vector "Size"

A **norm** is a function that assigns a non-negative number to each vector, representing its "size" or "length."

$$||v|| \geq 0 \text{ for all vectors } v$$

Key Properties of Norms:

- ▶ **Non-negativity:** $||v|| \geq 0$, and $||v|| = 0$ only when $v = 0$
- ▶ **Scaling:** $||cv|| = |c| \cdot ||v||$ for any scalar c
- ▶ **Triangle inequality:** $||u + v|| \leq ||u|| + ||v||$

Intuition: Think of a norm as a generalized way to measure distance from the origin.

Famous Norms

Euclidean Norm (L_2 norm)

The most familiar norm – the "straight-line" distance.

$$\|v\|_2 = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2} = \sqrt{v^T v}$$

Example: For $v = (3, 4)$, $\|v\|_2 = \sqrt{3^2 + 4^2} = 5$

$$\|v\|_\infty = \lim_{p \rightarrow \infty} \|v\|_p = \max_i |v_i| \quad \sqrt[p]{3^{1000} + 4^{1000}} \approx 4$$

Manhattan Norm (L_1 norm)

Named after Manhattan's grid system – you can only move horizontally or vertically.

$$\|v\|_1 = |v_1| + |v_2| + \dots + |v_n| \quad \|v_\infty\| = 1$$

Example: For $v = (3, 4)$, $\|v\|_1 = |3| + |4| = 7$

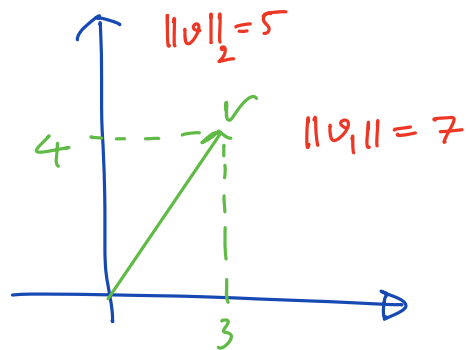
General L_p Norm

A family of norms parameterized by $p \geq 1$:

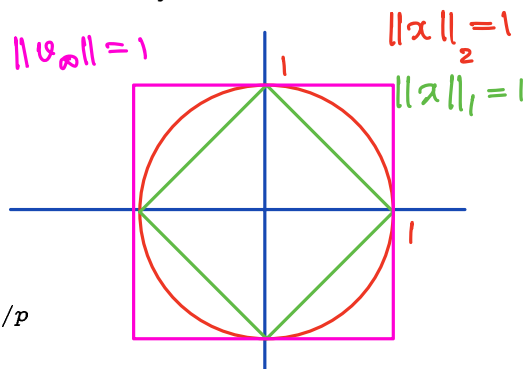
$$\|v\|_p = (|v_1|^p + |v_2|^p + \dots + |v_n|^p)^{1/p}$$

Special cases: $p = 1$ (Manhattan), $p = 2$ (Euclidean), $p = \infty$ (max norm)

$$v = \begin{bmatrix} 1 \\ 2 \\ 5 \\ -3 \end{bmatrix} \rightarrow \|v\|_\infty = 5$$

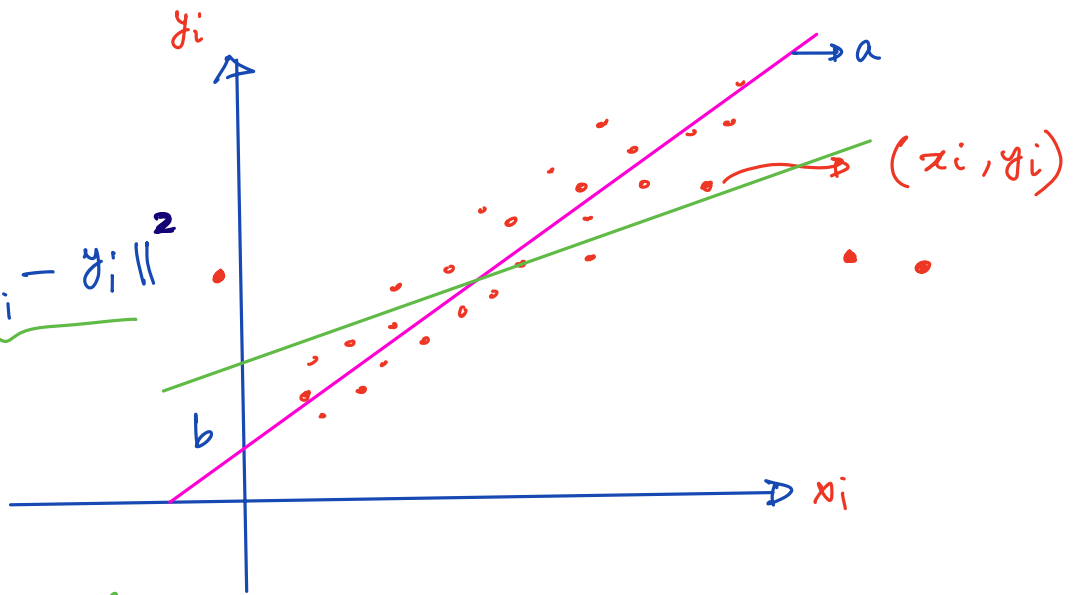


$$\|v\|_3 = \sqrt[3]{64 + 27}$$



$$\hat{y}_i = ax_i + b$$

$$\text{minimize } \sum_i \underbrace{\|\hat{y}_i - y_i\|}_{}^2$$



$$\|\cdot\|_2 \rightarrow (y - \hat{y})^2$$

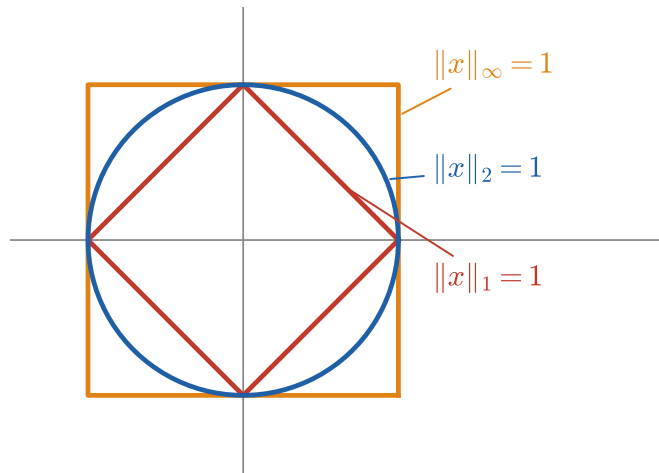
$$\|\cdot\|_1 \rightarrow |y - \hat{y}|$$

Visualizing norms: the unit ball

the *unit ball* of a norm is the set of vectors of norm at most one,

$$\{v \mid \|v\| \leq 1\}$$

- ▶ its shape is different for each norm
- ▶ ℓ_1 is a diamond, ℓ_2 a circle, ℓ_∞ a square
- ▶ as p grows from 1 to ∞ the ball grows from the diamond to the square



Norms in practice: loss and regularization

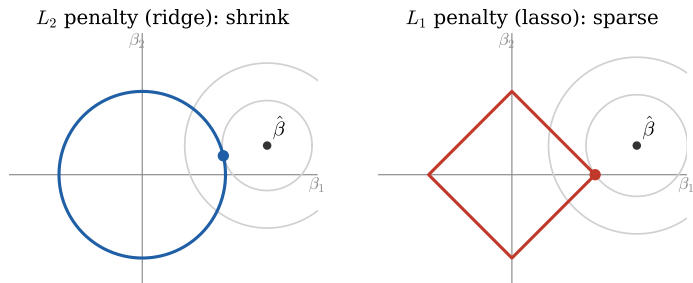
Losses measure prediction error:

- ▶ $\|y - \hat{y}\|_2^2$: mean-squared error
- ▶ $\|y - \hat{y}\|_1$: mean-absolute error

Regularizers penalize large parameters β :

- ▶ ridge / weight decay $\|\beta\|_2^2$: shrinks all coefficients
- ▶ lasso $\|\beta\|_1$: drives some to *exactly zero* (sparsity)

the ℓ_1 ball has corners on the axes, so the penalized optimum tends to land there



Inner Products: Measuring Vector "Similarity"

An **inner product** is a function that takes two vectors and returns a scalar, measuring how "similar" or "aligned" they are.

$$\langle u, v \rangle \rightarrow \text{real number}$$

Key Properties of Inner Products:

- ▶ **Symmetry:** $\langle u, v \rangle = \langle v, u \rangle$
- ▶ **Linearity:** $\langle au + bv, w \rangle = a\langle u, w \rangle + b\langle v, w \rangle$
- ▶ **Positive definite:** $\langle v, v \rangle \geq 0$, and equals 0 only when $v = 0$

Geometric Intuition: The inner product tells us about the angle between vectors:

- ▶ Positive $\rightarrow$ vectors point in similar directions
- ▶ Zero $\rightarrow$ vectors are perpendicular
- ▶ Negative $\rightarrow$ vectors point in opposite directions

Inner Products: Examples and Key Results

Standard Dot Product

$$x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Also called the *dot product* or *scalar product*.

Example: $\langle (2, 3), (4, 1) \rangle = 2 \times 4 + 3 \times 1 = 11$

$$e_i = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 < \sqrt{10} \\ \pm 1 \end{bmatrix} \rightarrow \begin{matrix} \text{age} \\ \text{gender} \end{matrix}$$

$$\langle w, v \rangle_A = w^T A v$$

Weighted Inner Product

$$\langle u, v \rangle_w = w_1 u_1 v_1 + w_2 u_2 v_2 + \cdots + w_n u_n v_n$$

$$\rightarrow \langle e_i, e_i \rangle = w_i < 0 \quad \times$$

where $w_i > 0$ are positive weights.

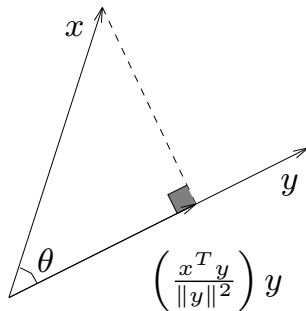
$$\downarrow \\ < 0$$

Fundamental Results:

► **Cauchy-Schwarz inequality:** $|\langle x, y \rangle| \leq \|x\| \|y\|$

► **Angle between vectors:** $\theta = \angle(x, y) = \cos^{-1} \frac{\langle x, y \rangle}{\|x\| \|y\|}$

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$



The Big Connection: Inner Products Create Norms

Key Insight: Every inner product creates a norm, but not every norm comes from an inner product!

$$||v|| = \sqrt{\langle v, v \rangle}$$

Examples of the Connection:

- ▶ Standard dot product $\rightarrow$ Euclidean norm: $||v||_2 = \sqrt{v^T v}$
- ▶ Weighted inner product $\rightarrow$ Weighted norm

Counter-Example: The Manhattan norm (L_1) does **not** come from any inner product!

Why this matters: Inner products are more restrictive than norms, but they give us powerful geometric tools like angles and orthogonality.

Applications: This connection is fundamental in machine learning, signal processing, and optimization.

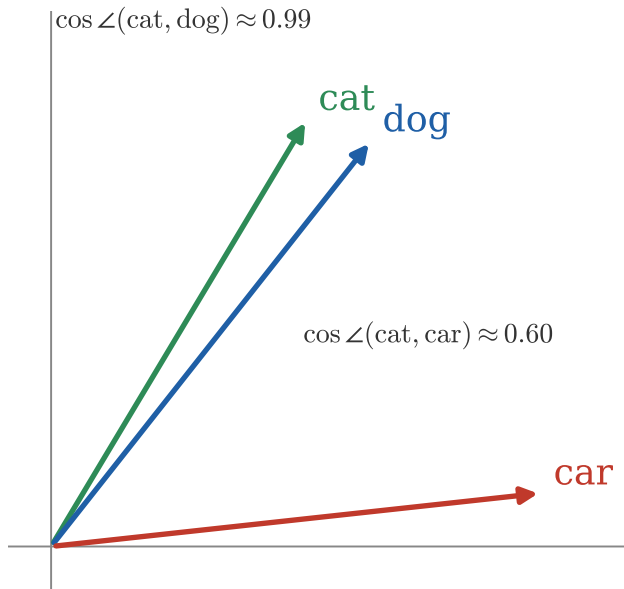
Inner products in practice: embeddings

Embeddings represent words, sentences, images, or users as vectors in $\mathbb{R}^d$, so that *similar items map to nearby vectors* similarity is the cosine of the angle between them:

$$\text{sim}(x, y) = \frac{x^\top y}{\|x\| \|y\|}$$

- ▶ word and sentence embeddings (word2vec, LLM token vectors)
- ▶ image and multimodal embeddings (e.g. CLIP)

used in semantic search, retrieval (RAG), and recommendation; for unit-norm embeddings the similarity is just $x^\top y$



Linear functionals

$$a^T x \in \mathbb{R}$$

$$f: \mathbb{R}^n \rightarrow \mathbb{R}^m$$

linear $\Rightarrow f(x) = Ax$
 $A \in \mathbb{R}^{m \times n}$

► a function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called a *functional*

► for $a \in \mathbb{R}^n$, the function $f(x) = a^T x$ is a *linear functional*

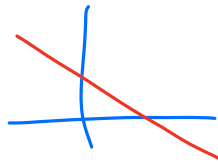
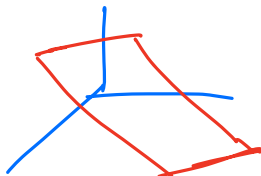
► every linear functional has this form

► if $a \neq 0$ then for any $\alpha \in \mathbb{R}$ the set $H_\alpha = \{x \in \mathbb{R}^n \mid a^T x = \alpha\}$ is called a *hyperplane*

► hyperplanes H_α and H_β are parallel, H_0 passes through the origin

$m=1$

2D: line
3D: plane



Halfspaces

$$y = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$
$$H = \{x \mid x^T y \leq 0\}$$

- define a *halfspace* with outward normal vector y , and boundary passing through 0 by

$$H = \{x \mid x^T y \leq 0\}$$

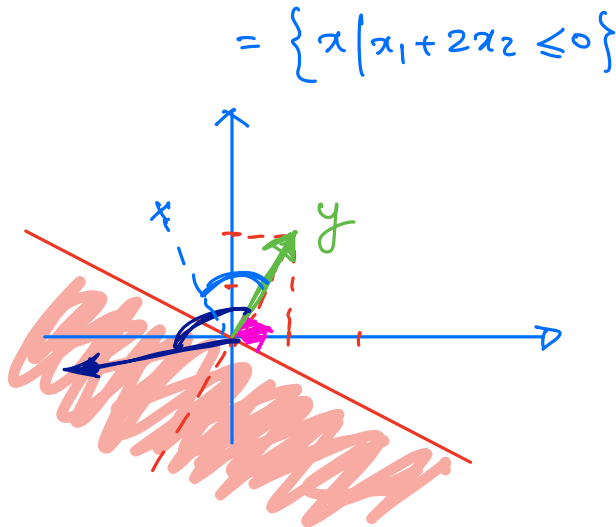
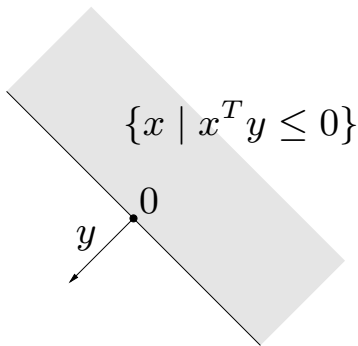
- H is all vectors x that make an obtuse or right angle with y

$$x^T y < 0$$

boundary: $x^T y = 0$



all the vectors x
that are orthogonal
to y .

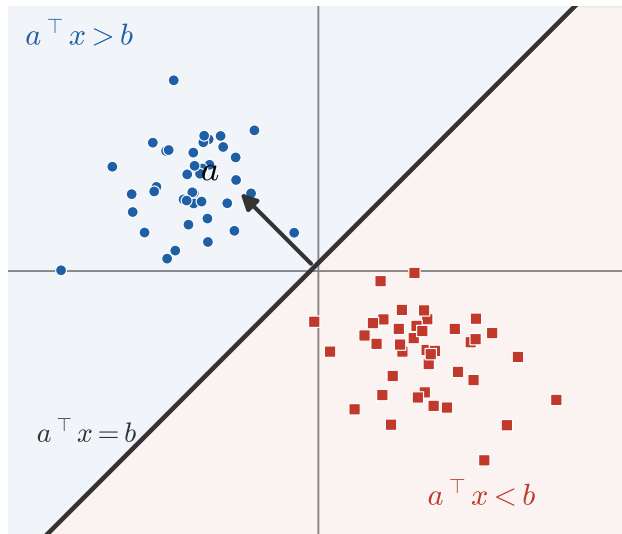


Hyperplanes in practice: linear classifiers

a *linear classifier* predicts a label from the sign of $a^\top x - b$

- ▶ the boundary $a^\top x = b$ is a hyperplane
- ▶ the two halfspaces are the two predicted classes
- ▶ the normal a points toward the positive class

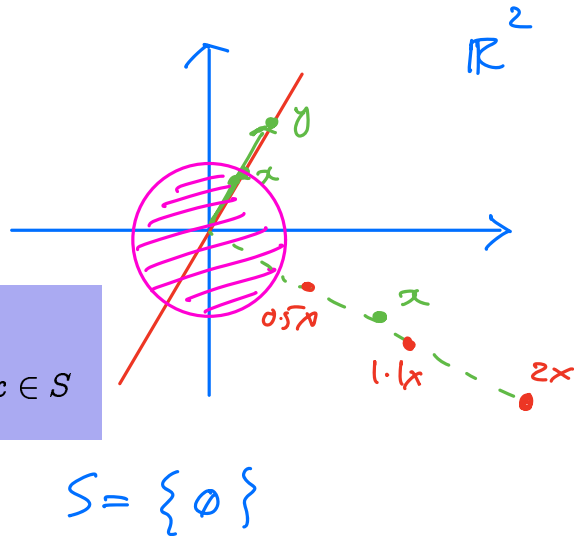
the perceptron, logistic regression, and linear support vector machines all learn such a hyperplane; a single neuron computes $a^\top x - b$ followed by a nonlinearity



Subspaces

A set $S \subset \mathbb{R}^n$ is called a subspace if

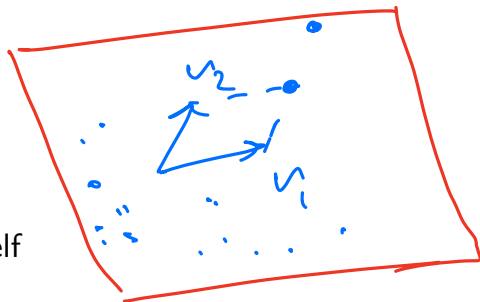
$$\begin{array}{ll} x + y \in S & \text{for all } x, y \in S \\ \alpha x \in S & \text{for all } \alpha \in \mathbb{R} \text{ and } x \in S \end{array}$$



- ▶ we say S is *closed under addition* and *closed under scalar multiplication*
- ▶ geometrically, S is a flat set which passes through the origin

Examples of subspaces

$$Ax = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1 a_1 + \dots + x_n a_n$$
$$\{y \mid y = Ax\} = \text{span}(\{a_1, \dots, a_n\})$$



- ▶ $S_1 = \mathbb{R}^n$, i.e., the entire vector space is considered a subspace of itself
- ▶ $S_2 = \{0\}$, the origin is the smallest subspace of $\mathbb{R}^n$
- ▶ the *span* of a set of vectors $v_1, v_2, \dots, v_k \in \mathbb{R}^n$ is a subspace

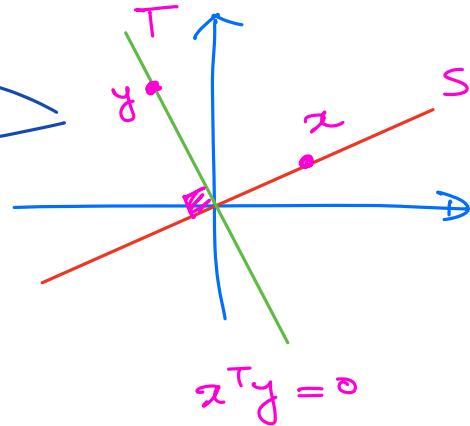
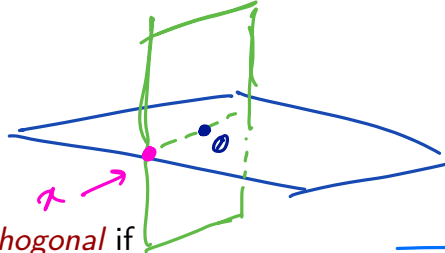
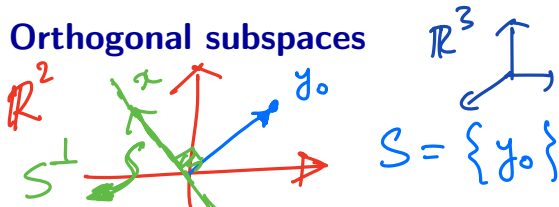
$$\text{span}(v_1, v_2, \dots, v_k) = \{\alpha_1 v_1 + \dots + \alpha_k v_k \mid \alpha_i \in \mathbb{R}\}$$

the set of all *linear combinations* of the vectors

- ▶ the *sum of two subspaces* is a subspace

$$S + T = \{x + y \mid x \in S, y \in T\}$$

Orthogonal subspaces



- ▶ two subspaces $S, T \subset \mathbb{R}^n$ are called *orthogonal* if

$$x \in S, y = x \in T \rightarrow x^T y = x^T x = \|x\|^2 > 0$$

$$x^T y = 0 \text{ for all } x \in S, y \in T$$

- ▶ for any set $S \subset \mathbb{R}^n$, the *orthogonal complement* is

$$S^\perp = \{x \mid x^T y = 0 \text{ for all } y \in S\}$$

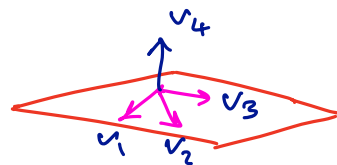
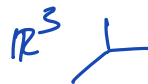
- ▶ $S^\perp$ is always a subspace, even if S is not
- ▶ $S^\perp$ is the set of all vectors x , each of which is orthogonal to every vector in S

$$x \in \mathbb{R}^4$$

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ 0 \\ 0 \end{bmatrix} \right\}$$

$$T = \left\{ \begin{bmatrix} 0 \\ 0 \\ x_3 \\ x_4 \end{bmatrix} \right\}$$

Independent set of vectors



a set of vectors $\{v_1, v_2, \dots, v_k\}$ is *independent* if

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_k v_k = 0 \implies \alpha_1 = \alpha_2 = \dots = 0$$

$$\{v_1, v_2, v_3, v_4\}$$

some equivalent conditions:

$$\underbrace{(\alpha_1 - \beta_1)}_0 v_1 + \underbrace{(\alpha_2 - \beta_2)}_0 v_2 + \dots + \underbrace{(\alpha_k - \beta_k)}_0 v_k = 0$$

- coefficients of $\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_k v_k$ are uniquely determined, i.e.,

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_k v_k = \beta_1 v_1 + \beta_2 v_2 + \dots + \beta_k v_k$$



implies $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \dots, \alpha_k = \beta_k$

- no vector v_i can be expressed as a linear combination of the other vectors $v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_k$
- no one vector v_i is in the span of the others

$$0 \cancel{v_k} = \alpha_1 v_1 + \dots + \alpha_{k-1} v_{k-1} - v_k$$

Basis and dimension

set of vectors $\{v_1, v_2, \dots, v_k\}$ is called a **basis** for a subspace S if

$$(v_3 = -v_1)$$

$$S = \{v_1, v_2, v_3\}$$

$$x = 3v_1 + 2v_2$$

$$= -3v_3 + 2v_2$$

$$= v_1 + 2v_2 - 2v_3$$

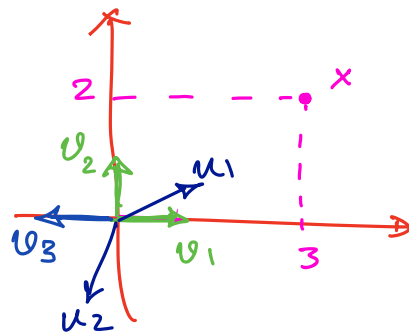
► equivalently, every $v \in S$ **can be uniquely** expressed as

$$v = \alpha_1 v_1 + \dots + \alpha_k v_k$$

$$\textcircled{1} S = \text{span}(v_1, v_2, \dots, v_k)$$

and

$$\textcircled{2} \{v_1, v_2, \dots, v_k\} \text{ is independent}$$



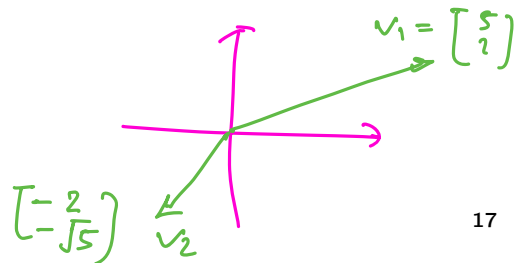
$$x = 3v_1 + 2v_2 = \alpha_1 u_1 + \alpha_2 u_2$$

✓ $\{v_1, v_2\}$
 $\{u_1, u_2\}$

► for a given subspace S , the number of vectors in any basis is the same, called the **dimension** of S , denoted $d = \dim S$

► any set of independent vectors in S has no more than d elements

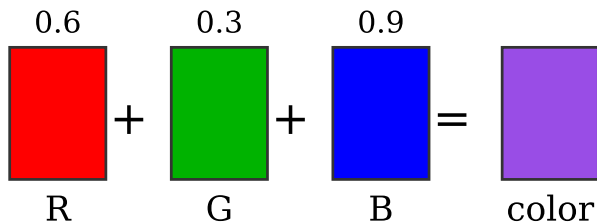
► any set of vectors that span S has at least d elements



Example: color is a three-dimensional space

any color on a screen is a combination of three primaries:

$$\text{color} = c_R R + c_G G + c_B B$$



- ▶ $\{R, G, B\}$ *span* all colors and are *independent*, so they form a *basis*: color space has *dimension* three
- ▶ the coordinates (c_R, c_G, c_B) are exactly how a color is stored
- ▶ a redundant primary (a mix of the others) would be like *collinear features* in regression, whose separate effects cannot be recovered

Invertibility

$$\begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = A\alpha \\ = x_1 a_1 + x_2 a_2 + \dots + x_n a_n$$

$$AB \neq BA \quad (\text{in general}) \quad \left\{ \begin{array}{l} BA = I \Rightarrow AB = I \\ B = A^{-1} \end{array} \right.$$

A **square** matrix $A \in \mathbb{R}^{n \times n}$ is called **invertible** if there is a matrix $B \in \mathbb{R}^{n \times n}$ such that

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

indep. columns:

$$Ax = 0 \Rightarrow x = 0$$

$$BA = I$$

$$\rightarrow AB = I ?$$

$$\rightarrow CA = I \Rightarrow C = B ?$$

► B is called the **inverse** of A , written A^{-1}

► we have $AA^{-1} = A^{-1}A = I$

► A is invertible iff it has linearly independent columns

$$\Rightarrow Ax = 0 \xLeftrightarrow{\text{}} \underbrace{A^{-1}}_I A x = A^{-1} 0 = 0 \Rightarrow x = 0 \quad \checkmark$$

B : left inverse

C : right inverse

$$BA = I \rightarrow BAC = C \Rightarrow \boxed{B = C}$$

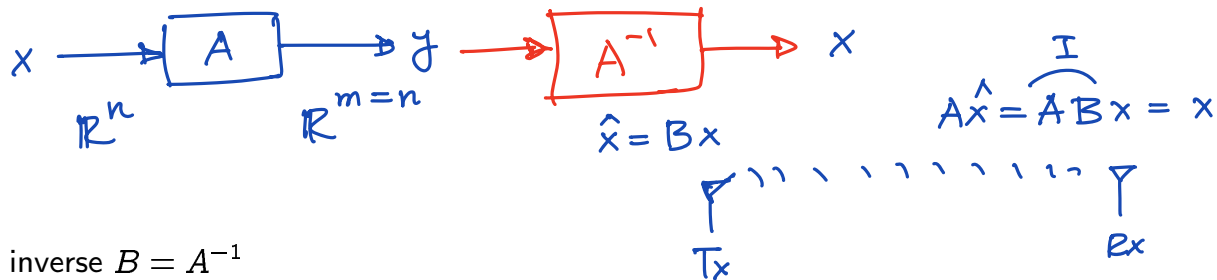
for any y we have some x s.t. $Ax = y$
 $y = e_i \rightarrow b_i$ exists s.t. $Ab_i = e_i$

for any y we have some x
s.t. $Ax = y$
 $y = \underline{e_i} \rightarrow b_i$ exists s.t. $Ab_i = e_i$

$$B = [b_1 \ b_2 \ \dots \ b_n]$$

$$\begin{aligned} AB &= \begin{bmatrix} | & | & \dots & | \\ Ab_1 & Ab_2 & \dots & Ab_n \\ | & | & & | \end{bmatrix} \\ &= \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix} \\ &= I \quad \Rightarrow \quad B = A^{-1} \end{aligned}$$

Interpretations of inverse



suppose $A \in \mathbb{R}^{n \times n}$ has inverse $B = A^{-1}$

- ▶ mapping associated with B undoes mapping associated with A (applied either before or after!)
- ▶ $x = By$ is a perfect (pre- or post-) *equalizer* for the *channel* $y = Ax$
- ▶ $x = By$ is unique solution of $Ax = y$

Change of coordinates

$$T = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- standard basis vectors in $\mathbb{R}^n$ are $e_1, e_2, \dots, e_n$, where $e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ has a 1 in i th component

- for any x we have

$$x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n$$

where x_i are called the *coordinates* of x (in the standard basis)

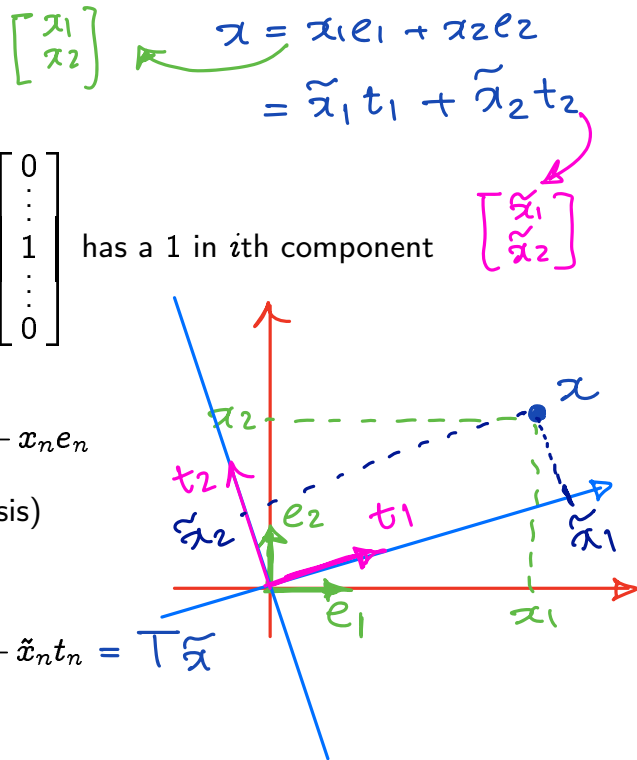
- if $t_1, t_2, \dots, t_n$ is another basis for $\mathbb{R}^n$, we have

$$T = [t_1 \dots t_n] \quad \tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix} \quad x = \tilde{x}_1 t_1 + \tilde{x}_2 t_2 + \dots + \tilde{x}_n t_n = T \tilde{x}$$

where $\tilde{x}_i$ are the coordinates of x in the basis $t_1, t_2, \dots, t_n$

- then $x = T \tilde{x}$ and $\tilde{x} = T^{-1} x$

$$x = T \tilde{x} \rightarrow \tilde{x} = T^{-1} x$$



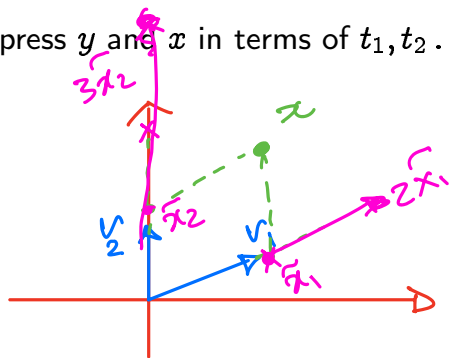
Similarity transformation

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \rightarrow \boxed{A} \rightarrow \begin{pmatrix} 2v_1 \\ 3v_2 \end{pmatrix}$$

$$T = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \quad T^{-1}AT = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

consider linear transformation $y = Ax$, $A \in \mathbb{R}^{n \times n}$

express y and x in terms of $t_1, t_2, \dots, t_n$, so $x = T\tilde{x}$ and $y = T\tilde{y}$, then



$$\tilde{y} = (T^{-1}AT)\tilde{x}$$

$$A = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$$

► $A \rightarrow T^{-1}AT$ is called *similarity transformation*

► similarity transformation by T expresses linear transformation $y = Ax$ in coordinates $t_1, t_2, \dots, t_n$

$$e_i \rightarrow \boxed{A} \rightarrow a_i$$

$$y = Ax$$

$$T\tilde{y} = A T\tilde{x}$$

$$\tilde{y} = (T^{-1}AT)\tilde{x}$$

$$T^{-1}AT = I$$

$$A = T^{-1}IT$$

$$= T^{-1}T$$

$$= I$$

Choosing a good basis: compression

the same vector has different coordinates in different bases;
a *good* basis makes the data simple

- ▶ in the right basis most coordinates are nearly zero, so a few numbers describe the data
- ▶ JPEG (a cosine basis) and MP3 keep only the large coordinates
- ▶ PCA picks the basis that captures the most variation first

here the cloud is nearly one-dimensional in the t_1, t_2 basis:
keep t_1 , drop t_2

