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.

$$||v||_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{v^\top v}$$

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

Manhattan Norm (L_1 norm)

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

$$||v||_1 = |v_1| + |v_2| + \cdots + |v_n|$$

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

General L_p Norm

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

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

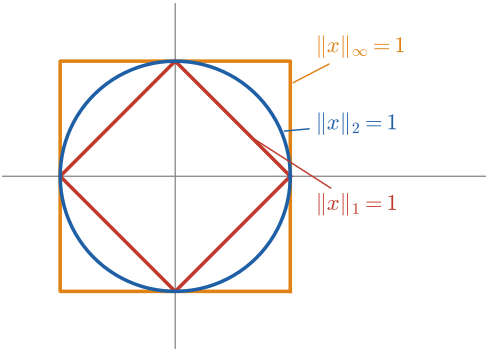
Special cases: $p = 1$ (Manhattan), $p = 2$ (Euclidean), $p = \infty$ (max norm)

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

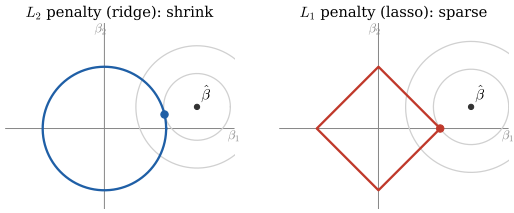
Losses measure prediction error:

- $||y - \hat{y}||_2^2$: 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.

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

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 $\to$ vectors point in similar directions
- Zero $\to$ vectors are perpendicular
- Negative $\to$ vectors point in opposite directions

Inner Products: Examples and Key Results

Standard Dot Product

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

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

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

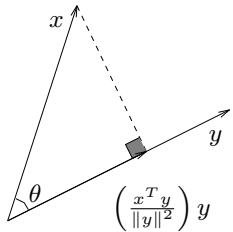
Weighted Inner Product

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

where $w_i > 0$ are positive weights.

Fundamental Results:

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



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: $||v||_2 = \sqrt{v^\top v}$
- 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

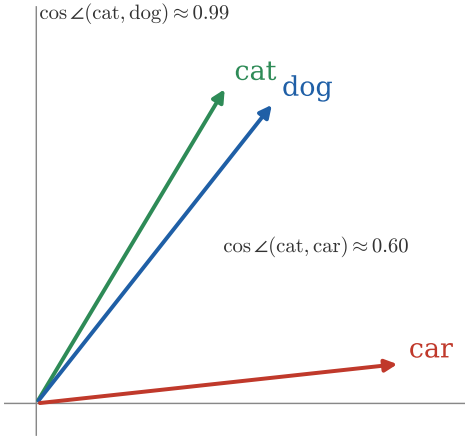
Embeddings represent words, sentences, images, or users as vectors in $\mathbb{R}^d$, so that *similar items map to nearby vectors*

similarity is the cosine of the angle between them:

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

- 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 $x^\mathsf{T} y$



Linear functionals

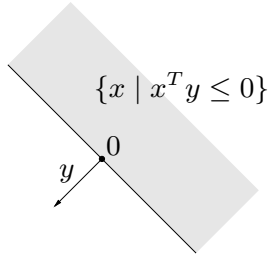
- a function $f : \mathbb{R}^n \to \mathbb{R}$ is called a *functional*
- for $a \in \mathbb{R}^n$, the function $f(x) = a^\mathsf{T} x$ is a *linear functional*
- every linear functional has this form
- if $a \neq 0$ then for any $\alpha \in \mathbb{R}$ the set $H_\alpha = \{x \in \mathbb{R}^n \mid a^\mathsf{T} x = \alpha\}$ is called a *hyperplane*
- hyperplanes H_α and H_β are parallel, H_0 passes through the origin

Halfspaces

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

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

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

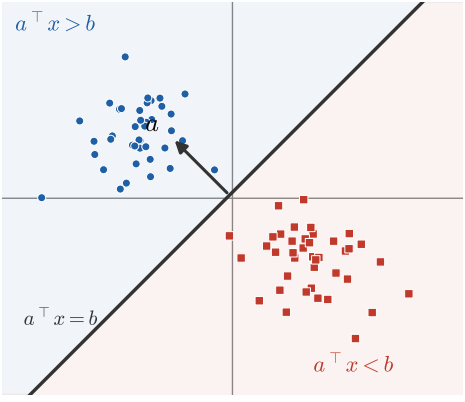


Hyperplanes in practice: linear classifiers

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

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

- 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

- $S_1 = \mathbb{R}^n$, *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 $v_1, v_2, \ldots, v_k \in \mathbb{R}^n$ is a subspace

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

  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

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

- 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

Independent set of vectors

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

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = 0 \quad \Longrightarrow \quad \alpha_1 = \alpha_2 = \cdots = 0$$

some equivalent conditions:

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

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

  implies $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \ldots, \alpha_k = \beta_k$
- no vector v_i can be expressed as a linear combination of the other vectors $v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k$
- no one vector v_i is in the span of the others

Basis and dimension

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

$$S = \textbf{span}(v_1, v_2, \ldots, v_k)$$
and
$$\{v_1, v_2, \ldots, v_k\} \text{ is independent}$$

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

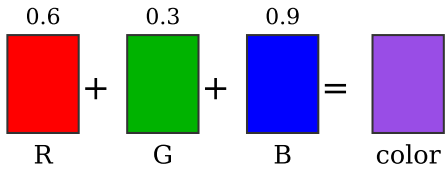
$$v = \alpha_1 v_1 + \cdots + \alpha_k v_k$$

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

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

$$BA = I$$

- 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

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

- standard basis vectors in $\mathbb{R}^n$ are $e_1, e_2, \ldots, e_n$, where $e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ has a 1 in ith component
- for any x we have
$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$
where x_i are called the *coordinates* of x (in the standard basis)
- if $t_1, t_2, \ldots, t_n$ is another basis for $\mathbb{R}^n$, we have
$$x = \tilde{x}_1 t_1 + \tilde{x}_2 t_2 + \cdots + \tilde{x}_n t_n$$
where $\tilde{x}_i$ are the coordinates of x in the basis $t_1, t_2, \ldots, t_n$
- then $x = T\tilde{x}$ and $\tilde{x} = T^{-1}x$

Similarity transformation

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

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

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

- $A \longrightarrow T^{-1}AT$ is called *similarity transformation*
- similarity transformation by T expresses linear transformation $y = Ax$ in coordinates $t_1, t_2, \ldots, t_n$

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