2. Vectors

- vector notation
- vector operations
- linear, affine functions
- norm, distance, angle
- standard deviation, correlation
- complexity

Vector

a *vector* is a collection of elements written as

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \quad \text{or} \quad a = (a_1, a_2, \ldots, a_n)$$

- a_i is the ith *element* (*entry, coefficient, component*) of vector a
- i is the *index* of the ith element a_i
- number of elements n is the *size* (*length, dimension*) of the vector
- a vector of size n is called an n*-vector*
- example of a 4-vector:

$$a = \begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ 7.2 \end{bmatrix} = (-1.1, 0.0, 3.6, 7.2), \qquad a_3 = 3.6$$

Notes and conventions

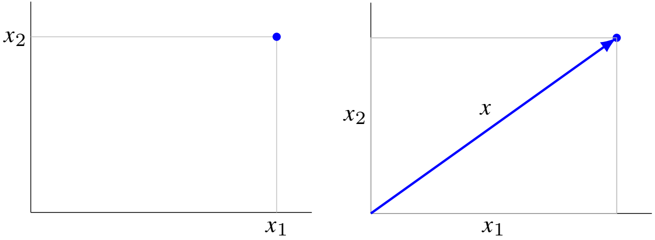
- $\mathbb{R}^n$ is the set of n-vectors with real entries
- $a \in \mathbb{R}^n$ means a is n-vector with real entries
- two n-vectors a and b are equal, denoted as $a = b$, if $a_i = b_i$ for all i
- warning: a_i can refer to an ith vector in a collection of vectors
  - in this case, we use $(a_i)_j$ to denote the jth entry of vector a_i
  - example: if $a_2 = (-1, 2, -5)$, then $(a_2)_3 = -5$

Conventions

- parentheses are also used instead of rectangular brackets to represent a vector
- other notations exist to distinguish vectors from numbers (*e.g.*, $\boldsymbol{a}, \vec{a}, \mathbf{a}$)
- conventions vary; be prepared to distinguish scalars from vectors

Geometric interpretation: location and displacement

- location (position): coordinates of a point in 2-D (plane) or 3-D space
- displacement: vector represents the change in position from one point to another (shown as an arrow in plane or 3-D space)



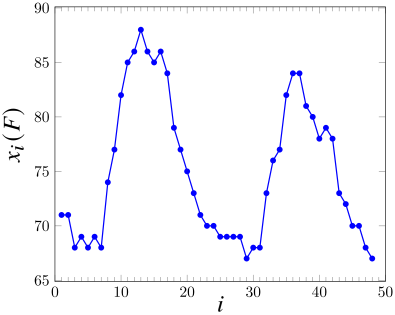
- other quantities that have direction and magnitude (velocity, force vector, ...)

Examples of vectors

Time series or signal

elements of n-vector are values of some quantity at n different times

- hourly temperature over a period of n hours



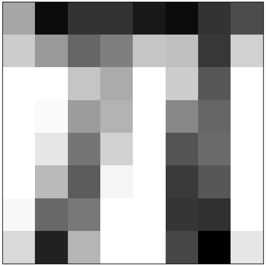
- audio signal: entries give the value of acoustic pressure at equally spaced times

Examples of vectors

Color: 3-vector can represent a color, with RGB intensity values

Monochrome (black and white) image

grayscale values of $M \times N$ pixels stored as MN-vector (row-wise or column-wise)



$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{62} \\ x_{63} \\ x_{64} \end{bmatrix} = \begin{bmatrix} 0.65 \\ 0.05 \\ 0.20 \\ \vdots \\ 0.28 \\ 0.00 \\ 0.90 \end{bmatrix}$$

Color image: $3MN$-vectors with R, G, B values of the MN pixels

Video: vector of size KMN represents K monochrome images of $M \times N$ pixels

Examples of vectors

Quantities

- elements of n-vector represent quantities of n resources or products
- sign indicates whether quantity is held or owed, produced or consumed, ...
- example: *bill of materials* is the list of resources (items) required to build a product represented as an n-vector whose entries give the amounts resources required

Portfolio vector

- n-vector s can represent stock portfolio (*e.g.*, investment in n assets)
  - assets can be stocks, bonds, cash, commodities (*e.g.*, gold), real estate ...
- s_i is the number of shares of asset i held (or invested in asset i)
- elements can be the no. of shares, dollar values, fractions of total dollar amount
- shares you owe another party (short positions) are represented by negative values

Examples of vectors

Daily return

- daily fractional return of a stock for a period of n trading days
- example: return time series vector $(-0.022, +0.014, +0.004)$ means stock price
  - went down 2.2% on the first day
  - then up 1.4% the next day
  - and up again 0.4% on the third day

Cash flow

- cash flow: payments into and out of an entity over n periods
- example: vector $(1000, -10, -10, -10, -1010)$ represents
  - a one year loan of 1000
  - with 1% interest only payments made each period (*e.g.*, quarter)
  - and the principal and last interest payment at the end

Examples of vectors

Word count vectors

- vector represents a document
- size of vector is the number of words in a dictionary
- word *count vector:* entry i is the number of times word i occurs in document
- word *histogram:* entry i is frequency of word i in document (in percentage)

Example: *word count vectors are used in computer-based document analysis; each entry of the word count vector represents the number of times the associated dictionary word appears in the document*

$$\begin{array}{l} \text{word} \\ \text{in} \\ \text{number} \\ \text{horse} \\ \text{document} \end{array} \qquad \begin{bmatrix} 3 \\ 2 \\ 1 \\ 0 \\ 2 \end{bmatrix}$$

Examples of vectors

Feature vector

- collects together n different quantities that relate to a single object
- entries are called the *features* or *attributes*

Examples

- age, height, weight, blood pressure, gender, etc., of patients
- square footage, number of bedrooms, list price, etc., of houses in an inventory

Notes

- vector elements can represent very different quantities, in different units
- can contain categorical features (*e.g.*, 1/0 for house/condo)
- ordering has no particular meaning

Row vector and transpose

an *row* vector b of size n with entries $b_1, \ldots, b_n$ has the form:

$$b = \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix}$$

- all vectors are column vectors unless otherwise stated
- other notation exists, *e.g.*, $b = [b_1, b_2, \ldots, b_n]$ (we will not use)

Transpose: the *transpose* of an n-column vector a is the row vector a^T:

$$a^T = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}$$

- $(\cdot)^T$ is transpose operation
- $(a^T)^T = a$ (transpose of row vector is a column vector)

Block vectors, subvectors

Stacking

- vectors can be *stacked* (*concatenated*) to create larger vectors
- stacking vectors b, c, d of size m, n, p gives an $(m+n+p)$-vector

$$a = \begin{bmatrix} b \\ c \\ d \end{bmatrix} = (b, c, d) = \left(b_1, \ldots, b_m, c_1, \ldots, c_n, d_1, \ldots, d_p\right)$$

- b, c, and d are called *subvectors* or *slices* of a
- example: if $a = 1$, $b = (2, -1)$, $c = (4, 2, 7)$, then $(a, b, c) = (1, 2, -1, 4, 2, 7)$

Subvectors slicing

- colon (:) notation is used to define subvectors (slices) of a vector
- for vector a, we define $a_{r:s} = (a_r, \ldots, a_s)$
- example: if $a = (1, -1, 2, 0, 3)$, then $a_{2:4} = (-1, 2, 0)$

Special vectors

Zero vector and ones vector

$$0 = (0, 0, \ldots, 0), \quad \mathbf{1} = (1, 1, \ldots, 1)$$

size follows from context (if not, we add a subscript and write $0_n, \mathbf{1}_n$)

Unit vectors

- there are n *unit vectors* of size n, denoted by $e_1, e_2, \ldots, e_n$:

$$(e_i)_j = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}$$

- the ith unit vector is zero except its ith element which is 1
- example: for $n = 3$,

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

- the size of e_i follows from context (or should be specified explicitly)

Sparsity

- a vector is *sparse* if many of its entries are 0
- can be stored and manipulated efficiently on a computer
- $\mathbf{nnz}(x)$ is number of entries that are nonzero
- examples:
  - $x = 0$ with $\mathbf{nnz}(x) = 0$
  - $x = e_i$ (unit vectors), $\mathbf{nnz}(x) = 1$
  - $x = (0, 0, 1, 0, 0, 0, -2, 0, 5, 0, 0)$, $\mathbf{nnz}(x) = 3$
- sparse vectors arise in many applications

Outline

- vector notation
- **vector operations**
- linear, affine functions
- norm, distance, angle
- standard deviation, correlation
- complexity

Addition and subtraction

for n-vectors a and b,

$$a + b = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{bmatrix}, \quad a - b = \begin{bmatrix} a_1 - b_1 \\ a_2 - b_2 \\ \vdots \\ a_n - b_n \end{bmatrix}$$

Example

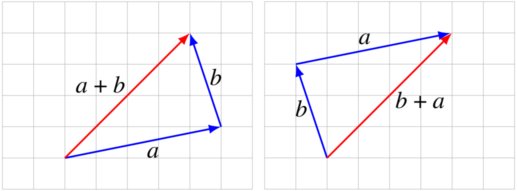
$$\begin{bmatrix} 0 \\ 7 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 9 \\ 3 \end{bmatrix}$$

Properties: for vectors a, b of equal size

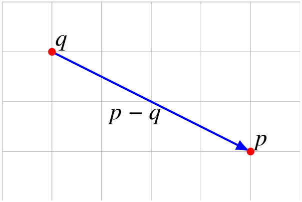
- commutative: $a + b = b + a$
- associative: $a + (b + c) = (a + b) + c$

Geometric interpretation: displacements addition

- if a and b are displacements, $a + b$ is the net displacement



- displacement from point q to point p is $p - q$



Scalar-vector multiplication

for scalar β and n-vector a,

$$\beta \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \beta a_1 \\ \beta a_2 \\ \vdots \\ \beta a_n \end{bmatrix}$$

example:

$$(-2) \begin{bmatrix} 1 \\ 9 \\ 6 \end{bmatrix} = \begin{bmatrix} -2 \\ -18 \\ -12 \end{bmatrix}$$

Properties: for vectors a, b of equal size, scalars β, γ

- commutative: $\beta a = a\beta$
- associative: $(\beta\gamma)a = \beta(\gamma a)$, we write as $\beta\gamma a$
- distributive with scalar addition: $(\beta + \gamma)a = \beta a + \gamma a$
- distributive with vector addition: $\beta(a + b) = \beta a + \beta b$

Linear combination

a *linear combination* of vectors $a_1, \ldots, a_m$ is a sum of scalar-vector products:

$$\beta_1 a_1 + \beta_2 a_2 + \cdots + \beta_m a_m$$

- scalars $\beta_1, \ldots, \beta_m$ are the *coefficients* of the linear combination
- example: any n-vector b can be written as

$$b = b_1 e_1 + \cdots + b_n e_n$$

Special linear combinations

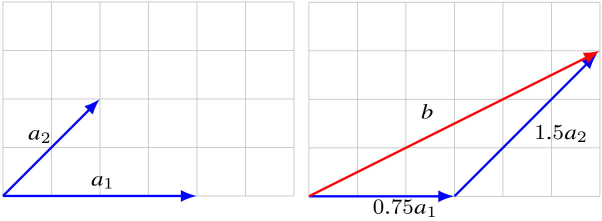
- *affine combination*: when $\beta_1 + \cdots + \beta_m = 1$
- *convex combination* (*weighted average*): when $\beta_1 + \cdots + \beta_m = 1$ and $\beta_i \geq 0$

Example: combination of displacements

- vector a represents a displacement
- for $\beta > 0$, βa is displacement in same direction of a, with magnitude scaled by β
- for $\beta < 0$, βa is displac. in the opposite direction of a, with mag. scaled by $|\beta|$

Example

$$b = 0.75a_1 + 1.5a_2$$

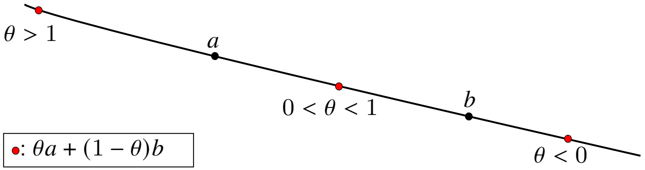


Line segment

any point on the line passing through distinct a and b can be written as

$$c = \theta a + (1 - \theta)b$$

- θ is a scalar
- an affine combination
- for $0 \leq \theta \leq 1$, point c lie on the segment between a and b



Addition and multiplication examples

Word count

- a and b are word count vectors (using the same dictionary) for two documents
- $a + b$ is the word count vector of the document combining the original two
- $a - b$ how many more times each word appears in 1st document compared to 2nd

Audio mixing

- $a_1, \ldots, a_m$ are vectors representing audio signals over the same period of time
- βa_i is the same audio signal, but changed in volume (loudness) by the factor $|\beta_i|$
- linear combination $\beta_1 a_1 + \cdots + \beta_m a_m$ is a mixture of the audio tracks

Portfolio trading

- s is n-vector giving no. of shares of n assets in a portfolio
- b is n-vector giving no. of shares of assets that we buy ($b_i > 0$) or sell ($b_i < 0$)
- after trading, our portfolio is $s + b$, which is called the *trade vector* or *trade list*

Inner product

the *inner product* (or *dot product*) of two n-vectors a, b is

$$a^T b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

- a scalar
- example:

$$\begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix}^T \begin{bmatrix} 1 \\ 0 \\ -3 \end{bmatrix} = (-1)(1) + (2)(0) + (2)(-3) = -7$$

- other notation exists: $\langle a, b \rangle$, $\langle a \mid b \rangle$, $a \cdot b$

Properties of inner product

for vectors a, b, c of equal size, scalar γ

- nonnegativity: $a^T a \geq 0$, and $a^T a = 0$ if and only if $a = 0$.
- commutative: $a^T b = b^T a$
- associative with scalar multiplication: $(\gamma a)^T b = \gamma(a^T b)$
- distributive with vector addition: $(a + b)^T c = a^T c + b^T c$

Useful combination: for vectors a, b, c, d

$$(a + b)^T (c + d) = a^T c + a^T d + b^T c + b^T d$$

Block vectors: if vectors a, b are block vectors, and corresponding blocks $a_i, b_i \in \mathbb{R}^{n_i}$ have the same sizes (they conform),

$$a^T b = \left[\begin{array}{c} a_1 \\ \vdots \\ a_k \end{array}\right]^T \left[\begin{array}{c} b_1 \\ \vdots \\ b_k \end{array}\right] = a_1^T b_1 + \cdots + a_k^T b_k$$

Simple examples

Inner product with unit vector

$$e_i^T a = a_i$$

Differencing

$$(e_i - e_j)^T a = a_i - a_j$$

Sum and average

$$\mathbf{1}^T a = a_1 + a_2 + \cdots + a_n$$

$$\mathrm{avg}(a) = \frac{a_1 + a_2 + \cdots + a_n}{n} = \left(\frac{1}{n}\mathbf{1}\right)^T a$$

Inner product examples

Weights, features, scores

- vectors of features f and weights w
- $w^T f = w_1 f_1 + w_2 f_2 + \cdots + w_n f_n$ is the total score
- example: features are associated with a loan applicant (*e.g.*, age, income, . . .)
  - we can interpret $s = w^T f$ as a credit score
  - we can interpret w_i as the weight given to feature i in forming the score

Price quantity (cost)

- vectors of prices p and quantities q of n goods
- $p^T q = p_1 q_1 + p_2 q_2 + \cdots + p_n q_n$ is the total cost

Speed time

- vehicle travels over n segments with constant speed in each segment
- n-vector s gives the speed in the segments
- n-vector t gives the times taken to traverse the segments
- $s^T t$ is the total distance traveled

Inner product examples

Polynomial evaluation

- n-vector c represents the coefficients of a polynomial p of degree $n-1$ or less:

$$p(x) = c_1 + c_2x + \cdots + c_{n-1}x^{n-2} + c_nx^{n-1}$$

- let $z = (1, t, t^2, \ldots, t^{n-1})$ be the n-vector of powers of a scalar t
- $c^Tz = p(t)$ is the value of the polynomial p at the point t

Discounted total

- cash flow vector c where c_i is value at period i
- r is interest rate and $d = (1, 1/(1+r), \ldots, 1/(1+r)^{n-1})$
- $d^Tc = c_1 + c_2/(1+r) + \ldots, c_n/(1+r)^{n-1}$ is the discounted total of cash flow
  - money received in the future is worth less than money received today
- called *net present value* (NPV) with interest rate r

Inner product examples

Portfolio value

- s is an n-vector of holdings in shares of a portfolio of n assets
- p is an n-vector for the prices of the assets
- $p^T s$ is the total (or net) value of the portfolio

Portfolio return

- portfolio vector x with x_i representing dollar value of asset i
- r_i is rate (fraction) of return of asset i over the investment period:

$$p_i^{\text{final}} = (1 + r_i) p_i^{\text{init}}, \qquad r_i = \frac{p_i^{\text{final}} - p_i^{\text{init}}}{p_i^{\text{init}}}$$

  p_i^{init} and p_i^{final} are the prices of asset i at the beginning and end of the period

- $r^T x = r_1 x_1 + \cdots + r_n x_n$ is total return in dollars over the period
- if w is the fractional (dollar) holdings of our portfolio, then $r^T w$ is rate of return
  - example: if $r^T w = 0.09$, then our portfolio return is 9%; if we had invested 10000 initially, we would have earned $\$900$

Outline

- vector notation
- vector operations
- **linear, affine functions**
- norm, distance, angle
- standard deviation, correlation
- complexity

Linear functions

- $f : \mathbb{R}^n \to \mathbb{R}$ means f is a *function* mapping n-vectors to numbers
- example: $f(x) = x_1 + x_2 - x_4^2$ ($f : \mathbb{R}^4 \to \mathbb{R}$)

Linear functions: f is *linear* if it satisfies the superposition property

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

for all scalars α, β, and all n-vectors x, y

Extension: if f is linear, then

$$f\left(\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_m u_m\right) = \alpha_1 f\left(u_1\right) + \alpha_2 f\left(u_2\right) + \cdots + \alpha_m f\left(u_m\right)$$

for all n-vectors $u_1, \ldots, u_m$ and all scalars $\alpha_1, \ldots, \alpha_m$

Inner product function

$$f(x) = a^T x = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

the inner product function is linear:

$$\begin{aligned} f(\alpha x + \beta y) &= a^T(\alpha x + \beta y) \\ &= a^T(\alpha x) + a^T(\beta y) \\ &= \alpha(a^T x) + \beta(a^T y) \\ &= \alpha f(x) + \beta f(y) \end{aligned}$$

All linear functions are inner products

- if $f : \mathbb{R}^n \to \mathbb{R}$ is linear, then $f(x) = a^T x$ for some (unique) a
- this follows from

$$\begin{aligned} f(x) &= f\left(x_1 e_1 + x_2 e_2 + \cdots + x_n e_n\right) \\ &= x_1 f\left(e_1\right) + x_2 f\left(e_2\right) + \cdots + x_n f\left(e_n\right) = a^T x \end{aligned}$$

  with $a = (f(e_1), \ldots, f(e_n))$

Example

- mean or average value of an n-vector is linear

$$f(x) = \text{avg}(x) = (x_1 + x_2 + \cdots + x_n)\,/n = a^T x$$

  where $a = (1/n, \ldots, 1/n) = (1/n)\mathbf{1}$ (sometimes denoted $\bar{x}$ or μ_x)

- maximum element func. $f(x) = \max\{x_1, \ldots, x_n\}$, is not linear (unless $n = 1$)
  - we can show this by a counterexample for $n = 2$
  - take $x = (1, -1), y = (-1, 1), \alpha = 1/2, \beta = 1/2$
  - then

$$f(\alpha x + \beta y) = 0 \neq \alpha f(x) + \beta f(y) = 1$$

Affine functions

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is *affine* if

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

for all n-vectors and scalars $\alpha + \beta = 1$

- extension: if f is affine, then

$$f\,(\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_m u_m) = \alpha_1 f\,(u_1) + \alpha_2 f\,(u_2) + \cdots + \alpha_m f\,(u_m)$$

  for all n-vectors $u_1, \ldots, u_m$ and all scalars $\alpha_1, \ldots, \alpha_m$ with $\alpha_1 + \cdots + \alpha_m = 1$

- every affine function f can be expressed as $f(x) = a^T x + b$ with

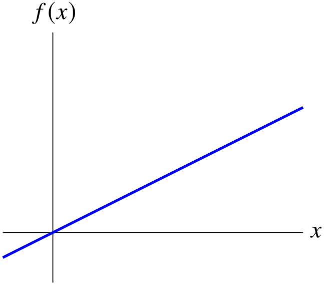
$$\begin{aligned} a &= \big(f\,(e_1) - f(0), f\,(e_2) - f(0), \ldots, f\,(e_n) - f(0)\big) \\ b &= f(0) \end{aligned}$$

- an affine function is a linear function plus a constant

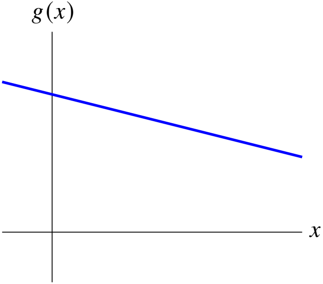
- often affine functions are called linear (which is mathematically not true)

Linear versus affine functions

f is linear

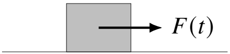


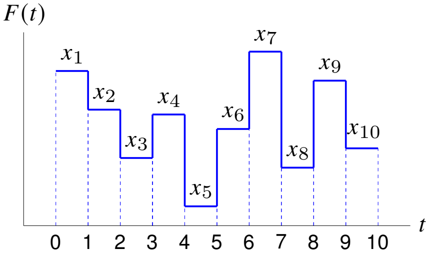
g is affine



linear functions pass through zero since $f(0) = 0$

Example: motion of a mass





- a unit mass with zero initial position and velocity
- we apply piecewise-constant force $F(t)$ during interval $[0, 10)$:

$$F(t) = x_j \quad \text{for } t \in [j-1, j), \quad j = 1, \ldots, 10$$

- define $f(x)$ as position at $t = 10$, $g(x)$ as velocity at $t = 10$

find f and g and determine whether they are linear or affine in x?

Solution

- from Newton's law $p''(t) = F(t)$ where $p(t)$ is the position at time t
- integrate to get final velocity and position

$$
\begin{aligned}
g(x) = p'(10) &= \int_0^{10} F(t)dt \\
&= x_1 + x_2 + \cdots + x_{10} \\
f(x) = p(10) &= \int_0^{10} p'(t)dt \\
&= \frac{19}{2}x_1 + \frac{17}{2}x_2 + \frac{15}{2}x_3 + \cdots + \frac{1}{2}x_{10}
\end{aligned}
$$

- the two functions are linear: $f(x) = a^T x$ and $g(x) = b^T x$ with

$$a = \left(\frac{19}{2}, \frac{17}{2}, \ldots, \frac{3}{2}, \frac{1}{2}\right), \quad b = (1, 1, \ldots, 1)$$

First-order Taylor (affine) approximation

first-order *Taylor approximation* of $f : \mathbb{R}^n \to \mathbb{R}$, near point z:

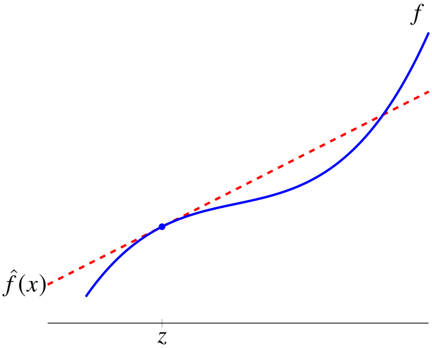
$$\begin{aligned}\hat{f}(x) &= f(z) + \frac{\partial f}{\partial x_1}(z)\,(x_1 - z_1) + \cdots + \frac{\partial f}{\partial x_n}(z)\,(x_n - z_n) \\ &= f(z) + \nabla f(z)^T(x - z)\end{aligned}$$

- n-vector $\nabla f(z)$ is the *gradient* of f at z,

$$\nabla f(z) = \left(\frac{\partial f}{\partial x_1}(z), \ldots, \frac{\partial f}{\partial x_n}(z)\right)$$

- $\hat{f}(x)$ is very close to $f(x)$ when x_i are all near z_i
- sometimes written $\hat{f}(x; z)$, to indicate that z where the approximation appear
- $\hat{f}$ is an affine function of x
- often called *linear approximation* of f near z, even though it is in general affine

Example with one variable



$$\hat{f}(x) = f(z) + f'(z)(x - z)$$

Example with two variables

$$f(x_1, x_2) = x_1 - 3x_2 + e^{2x_1+x_2-1}$$

- gradient:

$$\nabla f(x) = \begin{bmatrix} 1 + 2e^{2x_1+x_2-1} \\ -3 + e^{2x_1+x_2-1} \end{bmatrix}$$

- Taylor approximation around $z = 0$:

$$\begin{aligned} \hat{f}(x) &= f(0) + \nabla f(0)^T(x - 0) \\ &= e^{-1} + (1 + 2e^{-1})x_1 + (-3 + e^{-1})x_2 \end{aligned}$$

Outline

- vector notation
- vector operations
- linear, affine functions
- **norm, distance, angle**
- standard deviation, correlation
- complexity

Euclidean norm

Euclidean norm of vector $a \in \mathbb{R}^n$:

$$\|a\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} = \sqrt{a^T a}$$

- reduces to absolute value $|a|$ when $n = 1$
- measures the magnitude of a
- examples

$$\left\| \begin{bmatrix} 2 \\ -1 \\ 2 \end{bmatrix} \right\| = \sqrt{9} = 3, \quad \left\| \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\| = 1$$

Properties

Positive definiteness

$$\|a\| \geq 0 \quad \text{for all } a, \quad \|a\| = 0 \quad \text{only if } a = 0$$

Homogeneity

$$\|\beta a\| = |\beta| \|a\| \quad \text{for all vectors } a \text{ and scalars } \beta$$

Triangle inequality

$$\|a + b\| \leq \|a\| + \|b\| \text{ for all vectors } a \text{ and } b \text{ of equal length}$$

- any real function that satisfies these properties is called a (general) *norm*
- Euclidean norm is often written as $\|a\|_2$ to distinguish from other norms
  - also called 2-norm or ℓ_2-norm
- examples of other norms are the one-norm and infinity-norm:

$$\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n|$$
$$\|a\|_\infty = \max\{|a_1|, |a_2|, \ldots, |a_n|\}$$

Norm of block vector and norm of sum

Norm of block vector: for vectors a, b, c,

$$\left\| \begin{bmatrix} a \\ b \\ c \end{bmatrix} \right\| = \sqrt{\|a\|^2 + \|b\|^2 + \|c\|^2} = \|(\|a\|, \|b\|, \|c\|)\|$$

Norm of sum: for vectors a, b,

$$\|a + b\| = \sqrt{\|a\|^2 + 2a^Tb + \|b\|^2}$$

Cauchy-Schwarz inequality

$$|a^T b| \leq \|a\|\|b\| \quad \text{for all } a, b \in \mathbb{R}^n$$

moreover, equality $|a^T b| = \|a\|\|b\|$ holds if:

- $a = 0$ or $b = 0$; in this case $a^T b = 0 = \|a\|\|b\|$
- $b = \gamma a$ for some $\gamma > 0$; in this case

$$0 < a^T b = \gamma\|a\|^2 = \|a\|\|b\|$$

- $b = -\gamma a$ for some $\gamma > 0$; in this case

$$0 > a^T b = -\gamma\|a\|^2 = -\|a\|\|b\|$$

Proof of Cauchy-Schwarz inequality

1. trivial if $a = 0$ or $b = 0$

2. assume $\|a\| = \|b\| = 1$; we show that $-1 \leq a^T b \leq 1$

$$\begin{aligned} 0 &\leq \|a - b\|^2 \\ &= (a-b)^T(a-b) \\ &= \|a\|^2 - 2a^T b + \|b\|^2 \\ &= 2(1 - a^T b) \end{aligned}$$

with equality only if $a = b$

$$\begin{aligned} 0 &\leq \|a + b\|^2 \\ &= (a+b)^T(a+b) \\ &= \|a\|^2 + 2a^T b + \|b\|^2 \\ &= 2(1 + a^T b) \end{aligned}$$

with equality only if $a = -b$

3. for general nonzero a, b, apply case 2 to the unit-norm vectors

$$\frac{1}{\|a\|}a, \quad \frac{1}{\|b\|}b$$

Triangle inequality from Cauchy-Schwarz inequality

for vectors a, b of equal size

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

- taking square roots gives the triangle inequality
- triangle inequality is an equality if and only if $a^Tb = \|a\|\|b\|$
- also note from line 3 that $\|a+b\|^2 = \|a\|^2 + \|b\|^2$ if $a^Tb = 0$

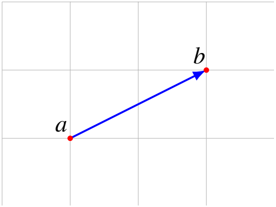
Euclidean distance

Euclidean distance between two vectors a and b,

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

- agrees with ordinary distance for $n = 1, 2, 3$

2-D illustration



- when the distance between two vectors is small, we say they are ‘close’ or ‘nearby’, and when the distance is large, we say they are ‘far’

Feature distance and nearest neighbors

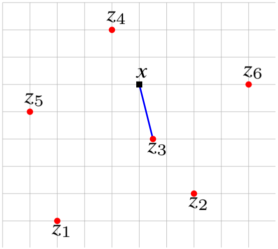
Feature distance

- let x and y be feature vectors for two entities
- $\|x - y\|$ is the *feature distance*; gives a measure of how different the objects are
  - example: features associated with patients in a hospital (weight, age, results of tests)
  - feature vector distance gives similarity between one patient case and another one

Nearest neighbor

- $z_1, \ldots, z_m$ is a list of vectors
- z_j is the nearest neighbor of x if

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



Document dissimilarity

- if x_i represent histogram of word occurrence in document i
- $\|x_i - x_j\|$ measures the dissimilarity between documents

Example

- 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

Units for heterogeneous vector entries

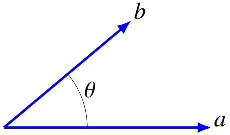
$$\|a - b\|^2 = (a_1 - b_1)^2 + \cdots + (a_n - b_n)^2$$

- suppose entries of vectors a_i, b_i represent different types of quantities
- choice of units for each entry affects the distance between a and b
- general rule: choose units so typical vector entries have similar ranges of values

Angle between vectors

the *angle* between nonzero real vectors a, b is defined as

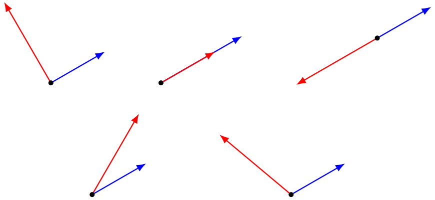
$$\theta = \angle(a, b) = \arccos\left(\frac{a^T b}{\|a\|\|b\|}\right)$$



- this is the unique value of $\theta \in [0, \pi]$ that satisfies $a^T b = \|a\|\|b\| \cos\theta$
- coincides with ordinary angle between vectors in 2-D and 3-D
- symmetric: $\angle(a, b) = \angle(b, a)$
- unaffected by positive scaling: $\angle(\alpha a, \beta b) = \angle(a, b)$ for +ve α, β

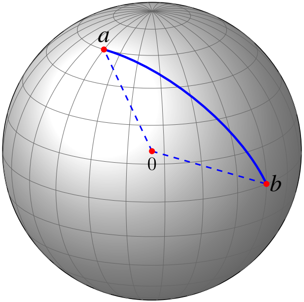
Classification of angles

$\theta = 0$	$a^T b = \|a\|\|b\|$	aligned/parallel
$0 \leq \theta < \pi/2$	$a^T b > 0$	acute angle
$\theta = \pi/2$	$a^T b = 0$	orthogonal ($a \perp b$)
$\pi/2 < \theta \leq \pi$	$a^T b < 0$	obtuse angle
$\theta = \pi$	$a^T b = -\|a\|\|b\|$	anti-aligned/opposed



Example: spherical distance

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



Document dissimilarity by angles

- if n-vectors x_i are word counts for documents, their angle $\angle(x_i, x_j)$ can be used as a measure of document dissimilarity
- example: 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	86.1
Golden Globe A.	87.0	87.5	58.7	0	86.0
Super Bowl	87.7	87.5	86.1	86.0	0

Norm of sum via angles

for vectors a and b we have

$$\begin{aligned}\|a+b\|^2 &= \|a\|^2 + 2a^Tb + \|b\|^2 \\ &= \|a\|^2 + 2\|a\|\|b\|\cos\theta + \|b\|^2\end{aligned}$$

- if a and b are aligned $(\theta = 0)$, then $\|a+b\| = \|a\| + \|b\|$
- if a and b are orthogonal $(\theta = 90^\circ)$, then

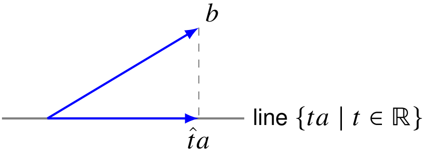
$$\|a+b\|^2 = \|a\|^2 + \|b\|^2$$

  and $\|a+b\| = \sqrt{\|a\|^2 + \|b\|^2}$ (called the *Pythagorean theorem*)

Projection onto a vector

given two vectors $a, b \in \mathbb{R}^n$, with $a \neq 0$, the vector multiple ta closest to b has

$$\hat{t} = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|^2}$$



Proof: minimize squared distance between ta and b:

$$\|ta - b\|^2 = (ta - b)^T(ta - b) = t^2 a^T a - 2ta^T b + b^T b$$

setting derivative with respect to t to zero gives

$$2ta^T a - 2a^T b = 0 \quad \Longrightarrow \quad \hat{t} = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|^2}$$

Geometric interpretation: $b - \hat{t}a \perp a$:

$$(b - \hat{t}a)^T a = 0 \Longrightarrow \hat{t} = \frac{a^T b}{\|a\|^2}$$

Outline

- vector notation
- vector operations
- linear, affine functions
- norm, distance, angle
- **standard deviation, correlation**
- complexity

RMS value

the *root-mean-square* value of $a \in \mathbb{R}^n$ is the root of the average squared entry

$$\mathrm{rms}(x) = \sqrt{\frac{a_1^2 + \cdots + a_n^2}{n}} = \frac{\|a\|}{\sqrt{n}}$$

- it is root of *mean-square value*: $\mathrm{ms} = (a_1^2 + \cdots + a_n^2)/n$
- RMS value useful for comparing sizes of vectors of different lengths
- $\mathrm{rms}(a)$ gives 'typical' value of $|a_i|$
- *e.g.*, $\mathrm{rms}(\alpha \mathbf{1}) = |\alpha|$ (independent of n)
- $\mathrm{rms}(a - b)$ is called the RMS *deviation* between a and b

Standard deviation

the *standard deviation* of $a \in \mathbb{R}^n$ is

$$\mathrm{std}(a) = \mathrm{rms}(a - \mathrm{avg}(a)\mathbf{1}) = \left\|a - ((\mathbf{1}^T a)/n)\mathbf{1}\right\| / \sqrt{n}$$

- std is RMS deviation from the average
- std gives us the 'typical' amount a vector entries deviate from their average (mean)
- $\tilde{a} = a - \mathrm{avg}(a)\mathbf{1}$ is called *de-meaned* vector (since $\mathrm{avg}(\tilde{a}) = 0$)
- other notation: μ and σ are often used for mean and standard deviation

Standard deviation formula

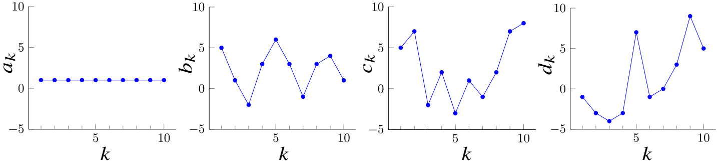
$$\mathrm{rms}(a)^2 = \mathrm{avg}(a)^2 + \mathrm{std}(a)^2$$

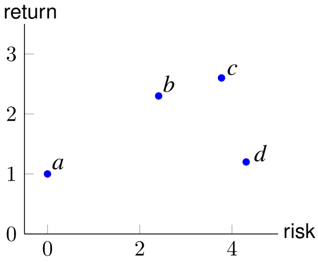
Proof

$$\begin{aligned}
\mathrm{std}(a)^2 &= \frac{\|a - \mathrm{avg}(a)\mathbf{1}\|^2}{n} \\
&= \frac{1}{n}\left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right)^T \left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right) \\
&= \frac{1}{n}\left(a^T a - \frac{\left(\mathbf{1}^T a\right)^2}{n} - \frac{\left(\mathbf{1}^T a\right)^2}{n} + \left(\frac{\mathbf{1}^T a}{n}\right)^2 n\right) \\
&= \frac{1}{n}\left(a^T a - \frac{\left(\mathbf{1}^T a\right)^2}{n}\right) \\
&= \mathrm{rms}(a)^2 - \mathrm{avg}(a)^2
\end{aligned}$$

Mean return and risk of investment

- vectors represent time series of returns on an investment (as a percentage)
- average value is (mean) return of the investment
- standard deviation measures variation around the mean (called risk)





Correlation coefficient

correlation coefficient (between a and b)

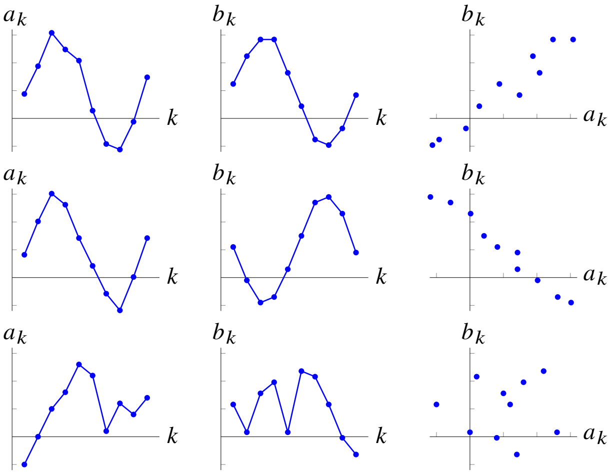
$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\| \|\tilde{b}\|}$$

where vectors $\tilde{a}$ and $\tilde{b}$ are de-meaned vectors ($\tilde{a} \neq 0, \tilde{b} \neq 0$):

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

- $\rho = \cos \angle(\tilde{a}, \tilde{b})$ hence $-1 \leq \rho \leq 1$
- $\rho = 0$, 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
- highly correlated “means” many a_i, b_i are both above (below) their means

Example



$\rho_{ab} = 0.968$

$\rho_{ab} = -0.988$

$\rho_{ab} = 0.004$

Examples

highly correlated vectors:

- rainfall time series at nearby locations
- daily returns of similar companies in same industry
- word count vectors of closely related documents (*e.g.*, same author, topic, ...)
- sales of shoes and socks (at different locations or periods)

approximately uncorrelated vectors

- unrelated vectors
- audio signals (even different tracks in multi-track recording)

(somewhat) negatively correlated vectors

- daily temperatures in Palo Alto and Melbourne

Properties and standardization

Properties of standard deviation

- *adding a constant:* $\mathrm{std}(a+\beta\mathbf{1}) = \mathrm{std}(a)$ for vector a and number β
- *multiplying by a scalar:* $\mathrm{std}(\beta a) = |\beta|\,\mathrm{std}(a)$ for vector a and number β
- *sum:* $\mathrm{std}(a+b) = \sqrt{\mathrm{std}(a)^2 + 2\rho\,\mathrm{std}(a)\,\mathrm{std}(b) + \mathrm{std}(b)^2}$ for vectors a, b

Standardization

- de-meaned vector of a in standard units is

$$z = \frac{1}{\mathrm{std}(a)}(a - \mathrm{avg}(a)\mathbf{1})$$

- z is called *standardized* or z*-scored* version of a ($\mathrm{avg}(z) = 0$ and $\mathrm{std}(z) = 1$)
- $z_4 = 1.4$ means a_4 is 1.4 standard deviations above the mean of a

Example: hedging investments

- a and b are time series of returns for two assets with the same return (average) μ, risk (standard deviation) σ, and correlation coefficient ρ

- $c = (a + b)/2$ is time series of returns for an investment with 50% in each asset

- this blended investment has the same return as the original assets, since

$$\mathrm{avg}(c) = \mathrm{avg}((a + b)/2) = (\mathrm{avg}(a) + \mathrm{avg}(b))/2 = \mu$$

- the risk (standard deviation) of this blended investment is

$$\mathrm{std}(c) = \sqrt{2\sigma^2 + 2\rho\sigma^2}/2 = \sigma\sqrt{(1 + \rho)/2}$$

- risk of the blended investment is never more than the risk of the original assets, and is smaller when the correlation of the original asset returns is smaller

- investing in two uncorrelated or -ve correlated assets is called *hedging*

Outline

- vector notation
- vector operations
- linear, affine functions
- norm, distance, angle
- standard deviation, correlation
- **complexity**

Floating point operation (FLOP)

Computer representation of numbers

- computers store (real) numbers in *floating-point format*
- number represented as 64 bits (0s and 1s), or 8 bytes (group of bits)
- each of 2^{64} sequences of bits corresponds to a specific number

Floating point operations

- 1 flop = one basic arithmetic operation ($+, -, *, /, \surd, \ldots$) in $\mathbb{R}$ (or complex $\mathbb{C}$)
- speed with which a computer can carry out flops is typically in 1-10 Gflop/s
- *complexity* of an operation is the number of flops required to carry it out
- flop is the unit of complexity when comparing algorithms; run time of the algorithm:

$$\text{run time} \approx \frac{\text{number of operations (flops)}}{\text{computer speed (flops per second)}}$$

  this is a very crude and simplified model of complexity of algorithms

Dominant terms

- typically, complexity is highly simplified, dropping small or negligible terms
- dominant term: the highest-order term in the flop count

$$\frac{1}{3}n^3 + 100n^2 + 10n + 5 \approx \frac{1}{3}n^3$$

- order: the power in the dominant term

$$\frac{1}{3}n^3 + 10n^2 + 100 = \text{order } n^3 = \mathcal{O}(n^3)$$

- order is useful in understanding how the time to execute the computation will scale when the size of the operands changes

Complexity of vector operations

for vectors of size n

- $x + y$ needs n additions, so n flops
- scalar multiplication: n flops
- inner product: $2n - 1 \approx 2n$ flops
  - we simplify this to $2n$ (or even n) flops
- these operations are all order n

Sparse vectors: when x and/or y is sparse

- αx requires **nnz**(x) flops
- $x + y$ requires $\min\{\textbf{nnz}(x), \textbf{nnz}(y)\}$ flops
- if sparsity pattern do not overlap, $x + y$ requires zero flops
- $x^T y$ requires no more than $2\min\{\textbf{nnz}(x), \textbf{nnz}(y)\}$ flops

Complexity of norms

for n-vectors

- $\|x\|$ requires $2n$ flops
  - n multiplications (to square each entry)
  - $n-1$ additions (to add the squares)
  - one squareroot
- RMS value costs $2n$ (ignore two flops from division of $\sqrt{n}$)
- distance between two vectors costs $3n$ flops
- angle between them costs $6n$ flops
- de-meaning an n-vector requires $2n$ flops
  - n for forming the average
  - n flops for subtracting the average from each entry
- standard deviation costs $4n$ flops
  - $2n$ for computing the de-meaned vector
  - $2n$ for computing its RMS value
- correlation coefficient costs $10n$ flops to compute

References and further readings

- S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares.* Cambridge University Press, 2018.

- L. Vandenberghe, *EE133A Lecture Notes,* University of California, Los Angeles.