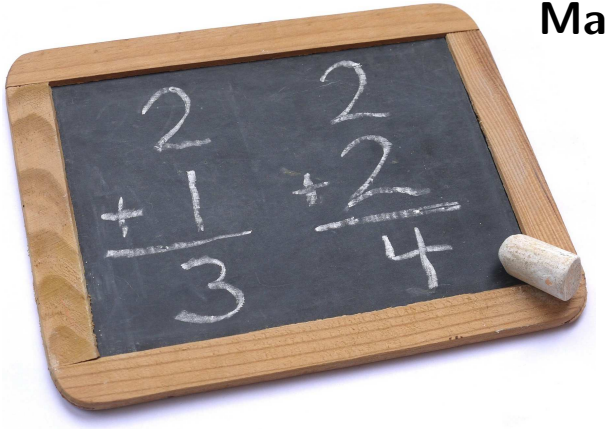


Image by Julia S from Pixabay

A Refresher On

Machine Learning Maths

CAS Machine Learning
Module 1: Introduction

Michael Calonder
Hochschule Luzern HSLU
Oct 20/21, 2023

Outline

Linear Algebra

Introduction

What is linear algebra and where is it used?

- Branch of mathematics concerned with linear equations such as

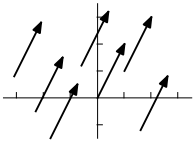
$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$$

- Of central important to most areas of science, mathematics, and engineering: ranging from differential equations, to modeling electrical circuits and control systems that fly aircraft, to econometry and genetics
- Not just "linear": Non-linear systems may well be treated through a series of linear operations, e.g. first-order Taylor expansion
- At the heart of most machine learning algorithms
- Modern computers can process them very efficiently, still active research (DeepMind's AlphaTensor)

Vectors

Properties:

- A vector is an object in $\mathbb{R}^n$ that is comprised of a magnitude and a direction
- Often visualized as an "arrow" for $n = 2$ or $n = 3$



Note that all above arrow represent the same vector

- Notation:

$$\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Vectors

Linear Dependence

Given $\boldsymbol{x}_1, \ldots \boldsymbol{x}_m$ be vectors in $\mathbb{R}^n$.
The set $\{\boldsymbol{x}_1, \ldots \boldsymbol{x}_m\}$ of these vectors *linearly independent* if

$$c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \ldots + c_m\boldsymbol{x}_m \quad \text{implies} \quad c_1 = c_2 = \ldots = c_m = 0$$

Likewise, any two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are *linearly dependent* if $\boldsymbol{x}_1 = c\boldsymbol{x}_2$ for some real $c \neq 0$

Determine whether or not the two vectors $(6, 2, 8, 4)^\top$ and $(21, 7, 28, 14)^\top$ are linearly dependent

Answer:

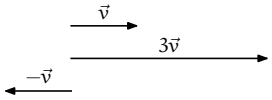
They are linearly dependent, $3.5 \times$ the first vector yields the second

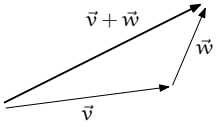
Vectors

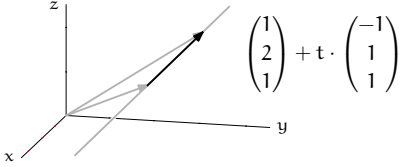
Algebraic properties:

- Additive identity: $\forall \boldsymbol{x} : \exists \mathbf{0}$ s.t. $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$
- Multiplicative identity: $1\,\boldsymbol{x} = \boldsymbol{x}$
- Commutativity: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
- Associativity: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{y} + (\mathbf{x} + \mathbf{z})$
- Additive inverse: $\forall \boldsymbol{v} : \exists \bar{\boldsymbol{x}} := -\boldsymbol{x}$ s.t. $\boldsymbol{x} + \bar{\boldsymbol{x}} = \boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$
- Distributivity:
  - $\forall \boldsymbol{v}, r$: $r(\boldsymbol{x} + \boldsymbol{y}) = r\boldsymbol{x} + r\boldsymbol{y}$
  - $\forall \boldsymbol{v}, r, s$: $(r + s)\boldsymbol{x} = r\boldsymbol{x} + s\boldsymbol{x}$

For $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ and r, s scalars







Matrices

Basic properties:

- A matrix is an object in $\mathbb{R}^{m\times n}$, i.e. a rectangular array of numbers.
- Central object in linear algebra, enable actual operations, e.g. addition or multiplication
- Widely used, even outside linear algebra e.g. in graph theory (adjacency matrices)
- Notation:

$$\boldsymbol{A} := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

- Note that for now we choose a_{ij} to be in $\mathbb{R}$, however, $a_{ij} \in \mathbb{C}$ in general
- Note that algebraic properties and identities differ form those above ones for vectors (see below)
- Important special case: *square* matrix, $m = n$

Matrix Multiplication (MM)

Matrix product $\boldsymbol{C} = \boldsymbol{AB} \in \mathbb{R}^{m \times p}$ with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times p}$,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mp} \end{pmatrix}$$

is defined s.t.

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

For instance,

$$\begin{pmatrix} a_{11} & a_{12} \\ \cdot & \cdot \\ a_{31} & a_{32} \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & b_{12} & b_{13} \\ \cdot & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} \cdot & c_{12} & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & c_{33} \\ \cdot & \cdot & \cdot \end{pmatrix}$$

MM Is Non-Commutative

MM is undefined if the inner dimensions differ.
But even when the MM is defined, changing the order of the factors will change the result.

For two square matrices $\boldsymbol{A} \in \mathbb{R}^n, \boldsymbol{B} \in \mathbb{R}^n$, and $n > 1$, in general

$$\boldsymbol{A}\,\boldsymbol{B} \neq \boldsymbol{B}\,\boldsymbol{A}$$

Example:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \neq \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

MM Is Distributive

Distributivity: MM is distributive with respect to matrix addition. If $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ are matrices of size $m \times n$, $n \times p$, $n \times p$, , $p \times q$, respectively, we have

- left distributivity: $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$
- right distributivity: $(\boldsymbol{B} + \boldsymbol{C})\boldsymbol{D} = \boldsymbol{BD} + \boldsymbol{CD}$

Scalar multiplication:

- For any $c \in \mathbb{R}$ we have $c\boldsymbol{A} = \boldsymbol{A}c$
- Also, $c(\boldsymbol{AB}) = (c\boldsymbol{A})\boldsymbol{B}$ and $(\boldsymbol{AB})c = \boldsymbol{A}(\boldsymbol{B}c)$

Matrix Transpose

Transposition
For any matrix $\boldsymbol{A} = [a_{ij}]$, its transpose is defined to be

$$\boldsymbol{A}^\top = [a_{ji}]$$

Example:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}^\top = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$$

Matrix Transpose: Identities

Some useful matrix identities related to its transpose:

- $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$
- $(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$
- $(\boldsymbol{A}_1 \boldsymbol{A}_2 \dots \boldsymbol{A}_{k-1} \boldsymbol{A}_k)^\top = \boldsymbol{A}_k^\top \boldsymbol{A}_{k-1}^\top \dots \boldsymbol{A}_2^\top \boldsymbol{A}_1^\top$ e.g., $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$
- $(c\boldsymbol{A})^\top = c\boldsymbol{A}^\top$
- $\det \boldsymbol{A} = \det \boldsymbol{A}^\top$
- If $\forall i, j : a_{ij} \in \mathbb{R} \Rightarrow \boldsymbol{A}^\top \boldsymbol{A} \succeq 0$ i.e., $\boldsymbol{A}$ PSD
- $(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$
- If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ then eigenvalues unchanged under transposition

Matrix Rank

Various definitions exists, all equivalent.

We define

$$\operatorname{rank} \boldsymbol{A} := \text{number of linearly independent rows (or columns)}$$

Example: The matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{pmatrix}$$

is of rank 2 (do you see why?)

Determinants

The determinant of a *square* matrix $\boldsymbol{A}$ is scalar that characterizes certain properties of $\boldsymbol{A}$. Most importantly, the determinant is nonzero iff the matrix is invertible

Due to being expensive to compute, explicit use of determinants in machine learning applications is rare

Computation of the determinant of $\boldsymbol{A} \in \mathbb{R}^n$:

- $n = 1$: $\det a = a$
- $n = 2$: $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$
- $n = 3$: $\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + cdh - ceg - bdi - afh$
- $n > 3$: Usually computed via decomposition (see below)

Definiteness

For a *symmetric* matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ consider

$$s := \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}.$$

Then, $\boldsymbol{A}$ is called

- *positive-definite* (PSD) if $s > 0 \ \forall \boldsymbol{x}$
- *positive-semidefinite* (PD) if $s \geq 0 \ \forall \boldsymbol{x}$
- *negative-semidefinite* (NSD) if $s \leq 0 \ \forall \boldsymbol{x}$
- *negative-definite* (ND) if $s < 0 \ \forall \boldsymbol{x}$
- and *indefinite* otherwise

Example: The identity matrix $\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is positive-definite. To see this, let $\boldsymbol{x} = \begin{pmatrix} a \\ b \end{pmatrix}$ and consider

$$\boldsymbol{x}^\top \boldsymbol{I} \boldsymbol{x} = \begin{pmatrix} a & b \end{pmatrix}^\top \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = a^2 + b^2$$

which is strictly positive for all $\boldsymbol{x} \neq \boldsymbol{0}$.

Matrix Trace

The trace of a *square* matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is commonly defined as the sum of its diagonal elements:

$$\operatorname{tr} \boldsymbol{A} = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

Example:

$$\operatorname{tr} \begin{pmatrix} 1 & 0 & 3 \\ 11 & 5 & 2 \\ 6 & 12 & -5 \end{pmatrix} = 1 + 5 + (-5) = 1$$

Properties:

- Linearity: $\operatorname{tr}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B})$ and $\operatorname{tr}(c\boldsymbol{A}) = c \operatorname{tr}(\boldsymbol{A})$
- $\operatorname{tr} \boldsymbol{A} = \operatorname{tr} \boldsymbol{A}^{\top}$
- $\operatorname{tr}(\boldsymbol{ABCD}) = \operatorname{tr}(\boldsymbol{BCDA}) = \operatorname{tr}(\boldsymbol{CDAB}) = \operatorname{tr}(\boldsymbol{DABC})$ and as a special case $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA})$
- Trace as the sum of eigenvalues λ_i of $\boldsymbol{A}$: $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$

Determine whether or not the real symmetric matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

is positive-definite.

Answer:

$\boldsymbol{A}$ is PD since for any non-zero column vector $\boldsymbol{z} = \begin{pmatrix} a & b & c \end{pmatrix}^\top$:

$$\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} = \left(\boldsymbol{z}^\top \boldsymbol{M}\right)\boldsymbol{z} = \begin{pmatrix} (2a-b) & (-a+2b-c) & (-b+2c) \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \ldots = a^2+(a-b)^2+(b-c)^2+c^2 > 0$$

Matrix Inverse

The inverse matrix $\boldsymbol{A}^{-1}$ to a regular matrix $\boldsymbol{A}$ is defined s.t.

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

with $\boldsymbol{I}$ then identity matrix
Example: The inverse of the real matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix} \quad \text{is} \quad \boldsymbol{A}^{-1} = \begin{pmatrix} 2 & -\frac{1}{2} \\ -3 & 1 \end{pmatrix}$$

Verify this:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \begin{pmatrix} 2 & 1 \\ 6 & 4 \end{pmatrix} \begin{pmatrix} 2 & -\frac{1}{2} \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 4-3 & -1+1 \\ 12-12 & -3+4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \boldsymbol{I}$$

Matrix Inverse

When is a matrix $\boldsymbol{A}$ invertible? Many equivalent conditions exist ("Invertible Matrix Theorem"):

- $\boldsymbol{A}$ is row- or column-equivalent to the $n \times n$ identity matrix $\boldsymbol{I}_n \Rightarrow$ Gaussian Elimination
- The columns of A are linearly independent
- $\boldsymbol{A}$ has full rank, rank $\boldsymbol{A} = n$
- 0 is *not* an eigenvalue of $\boldsymbol{A} \Rightarrow$ Definiteness
- $\det A \neq 0$

Inversion plays an important part in the solution of linear systems of equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \quad \Longleftrightarrow \quad \boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

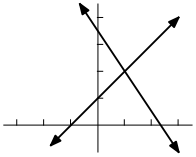
Closely related to rank: For instance, if rank $\boldsymbol{A} = n$ (full rank), the system

- $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ has only the trivial solution $\boldsymbol{x} = 0$
- $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has exactly one solution for each $\boldsymbol{b} \in \mathbb{R}^n$

In practice, computing the inverse is costly and numerically delicate—avoid whenever possible

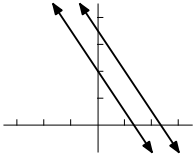
Linear Systems

Unique solution



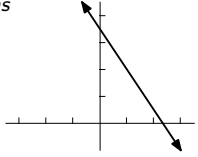
$$\begin{aligned} 3x &+ 2y &= 7 \\ x &- y &= -1 \end{aligned}$$

No solutions



$$\begin{aligned} 3x &+ 2y &= 7 \\ 3x &+ 2y &= 4 \end{aligned}$$

Infinitely many solutions



$$\begin{aligned} 3x &+ 2y &= 7 \\ 6x &+ 4y &= 14 \end{aligned}$$

Find the coefficients a, b, and c so that the graph of $f(x) = ax^2 + bx + c$ passes through the points $(1, 2)$, $(-1, 6)$, and $(2, 3)$

Answer:

$a = 1$, $b = -2$, $c = 3$

Matrix Inverse: Identities

Some useful matrix identities to inversion:

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(k\mathbf{A})^{-1} = k^{-1}\mathbf{A}^{-1}$ for any scalar $k \neq 0$
- $(\mathbf{A}^{\top})^{-1} = (\mathbf{A}^{-1})^{\top}$
- For invertible matrices $\boldsymbol{A}_i \in \mathbb{R}^{n \times n}$, $(\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_{k-1}\mathbf{A}_k)^{-1} = \mathbf{A}_k^{-1}\mathbf{A}_{k-1}^{-1} \cdots \mathbf{A}_2^{-1}\mathbf{A}_1^{-1}$
- $\det \mathbf{A}^{-1} = (\det \mathbf{A})^{-1}$

Matrix Decompositions

Eigendecomposition

The Eigendecomposition factorizes a diagonalizable matrix into a its canonical form, represented by *eigenvalues* and *eigenvectors*. Particularly important special case in machine learning: real, symmetric matrices

A vector $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is an eigenvector of a square $N \times N$ matrix $\boldsymbol{A}$, if it satisfies

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

for some scalar λ. Then λ is called the eigenvalue corresponding to eigenvector $\boldsymbol{x}$.

Computation

- Characteristic polynomial $p(\lambda) := \det(\boldsymbol{A}\boldsymbol{I} - \lambda) = 0$ yields up to N distinct eigenvalues λ_i
- Each λ_i gives rise to a specific eigenvalue equation $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{v}_i = \mathbf{0}$, yielding the eigenvectors $\boldsymbol{v}_i$

Matrix Decompositions

Lower-upper (LU) Decomposition

The LU decomposition factors a square matrix $\boldsymbol{A}$ into a lower triangular and an upper triangular matrix:

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$$

Example: An invertible 3×3 matrix $\boldsymbol{A}$ can be factored as

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}.$$

Remarks:

- Typical strategy used by computers to solve square systems of equations
- Often "LU" actually refers to "LUP" where $\boldsymbol{P}$ is a permutation matrix ("partial pivoting")
- Any square matrix $\boldsymbol{A}$ admits LUP and PLU factorizations. If $\boldsymbol{A}$ is invertible, then it admits LU factorization

Matrix Decompositions

Lower-upper (LU) Decomposition

LU decompositions find wide-spread application from solving linear equations, to inverting a matrix, to computation of the determinant.

Example: Solve a system of linear equations with LUP
Given the system $\boldsymbol{Ax} = \boldsymbol{b}$ and the decomposition $\boldsymbol{PA} = \boldsymbol{LU}$, rewrite the problem

$$\boldsymbol{Ax} = \boldsymbol{b} \iff \boldsymbol{PAx} = \boldsymbol{Pb} \iff \boldsymbol{LUx} = \boldsymbol{Pb}$$

Then $\boldsymbol{x}$ can be found by

1. first solving $\boldsymbol{Ly} = \boldsymbol{Pb}$ for $\boldsymbol{y}$ (forward substitution)
2. second solving $\boldsymbol{Ux} = \boldsymbol{y}$ for $\boldsymbol{x}$ (backward substitution)

Matrix Decompositions

Cholesky Decomposition

Why Cholesky? As a special case of LU for PD matrices, about twice as efficient to compute

The Cholesky decomposition of an invertible matrix $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\top}$$

where $\boldsymbol{L}$ is a real lower triangular matrix with positive diagonal entries

Wide-spread applications:

- Linear least squares: efficient solutions of linear system arising from partial differential equations
- Non-linear optimization: numerical stability in Quasi-Newton by avoiding direct updates of the Hessian
- Monte Carlo simulations: produce correlated samples from uniform ones
- Kalman filters: forcing the system covariance matrix to remain positive semi-definite

Matrix Decompositions

Singular Value Decomposition (SVD)

SVD generalizes the eigendecomposition of a square matrix to any $m \times n$ matrix.

The SVD decomposes a real $m \times n$ matrix $\boldsymbol{A}$ as follows,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$$

with $\boldsymbol{U}$ and $\boldsymbol{V}$ are real orthogonal matrices and $\boldsymbol{\Sigma}$ a $m \times n$ rectangular diagonal matrix w

If M is real, then U and V can be guaranteed to be real orthogonal matrices; in such contexts, the SVD is often denoted $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$

Paper exercises

- Does $(1, 0, 2, 1)$ lie on the line through $(-2, 1, 1, 0)$ and $(5, 10, -1, 4)$?
- For what value(s) of a does the system have nontrivial solutions?

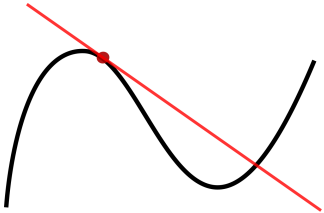
$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0 \\ -x_1 - x_2 + x_3 &= 0 \\ 3x_1 + 4x_2 + a\,x_3 &= 0 \end{aligned}$$

Exercises with Python

- Implement your own matrix multiplication in pure Python; measure speed for a 1000×1000 matrix; then use numpy for the multiplication and compare
- Compute the SVD of a random $m \times n$ matrix $\boldsymbol{A}$ using your implementation of choice and numerically compare the reconstructed matrix $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ with $\boldsymbol{A}$
- Verify numerically the identity $(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$
- Implement your own solver for linear systems using LU decomposition; use `scipy.linalg.lu` for the decomposition and implement the foward- and backward substitution yourself
- Given a set of m vectors in $\mathbb{R}^n$, determine whether it is linearly independent

Intuition

Geometrically: The derivative is the tangent line to the curve at the point $(x0, y0)$ is a line passing through $(x0, y0)$ and 'flat against' the curve



Intuition: The derivative measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value)

Definition

Roughly speaking, a function $f(x)$ is differentiable at a point a of its domain, if the limit

$$L = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

exists.

If the limit L exists, then this limit is called the *derivative* of f at a, and denoted

$$f'(a) \quad \text{or} \quad \frac{df}{dx}(a)$$

Derivatives of polynomials

There's just four simple facts which suffice to take the derivative of any polynomial plussomewhat more general things.
First, there is the rule for taking the derivative of a *power function* $f(x) = x^n$:

$$\frac{d}{dx} x^n = n\, x^{n-1}$$

Second, a special case of the rule above, the derivative of any constant c is zero:

$$\frac{d}{dx}\, c = 0$$

Third, for any function $f(x)$, multiplicative constants do not affect the result,

$$\frac{d}{dx}\, cf = c\, \frac{d}{dx}\, f$$

Fourth, for any two functions $f(x)$ and $g(x)$, the derivative of the sum is the sum of the derivatives:

$$\frac{d}{dx}(f + g) = \frac{d}{dx} f + \frac{d}{dx} g$$

Putting these four things together, we can write general formulas like

$$\frac{d}{dx}(ax^m + bx^n + cx^p) = a\,m\,x^{m-1} + b\,n\,x^{n-1} + c\,p\,x^{p-1}$$

and so on, with any number of summands.
As a refresher, here are some examples:

$$\frac{d}{dx}5x^3 = 15x^2$$

$$\frac{d}{dx}(3x^7 + 5x^3 - 11) = 21x^6 + 15x^2$$

$$\frac{d}{dx}(2 - 3x^2 - 2x^3) = -6x - 6x^2$$

$$\frac{d}{dx}(-x^4 + 2x^5 + 1) = -4x^3 + 10x^4$$

More general power functions

Remember some of the other possibilities for the exponential notation x^n. For example

$$x^{\frac{1}{2}} = \sqrt{x}$$

$$x^{-1} = \frac{1}{x}$$

$$x^{-\frac{1}{2}} = \frac{1}{\sqrt{x}}$$

and so on.
Good news: Above rule for taking the derivative of powers of x is still correct here, even for exponents which are negative or fractions or even real numbers:

$$\frac{d}{dx} x^r = r\, x^{r-1}$$

Thus, in particular,

$$\frac{d}{dx}\sqrt{x} = \frac{d}{dx}x^{\frac{1}{2}} = \frac{1}{2}x^{-\frac{1}{2}}$$

$$\frac{d}{dx}\frac{1}{x} = \frac{d}{dx}x^{-1} = -1 \cdot x^{-2} = \frac{-1}{x^2}$$

When combined with the sum rule from above, we have the obvious possibilities:

$$\frac{d}{dx}(3x^2 - 7\sqrt{x} + \frac{5}{x^2}) = \frac{d}{dx}(3x^2 - 7x^{\frac{1}{2}} + 5x^{-2}) = 6x - \frac{7}{2}x^{-\frac{1}{2}} - 10x^{-3}$$

Expressing square roots, cube roots, inverses, etc., in terms of exponents often simplifies the rearranging of algebraic expressions

Quotient rule

Quotient rule: Given two differentiable functions $f(x)$ and $g(x)$, the derivative of the ratio of the two is given by:

$$\frac{d}{dx}\frac{f}{g} = \frac{f'g - g'f}{g^2}$$

Example: For $f(x) = x - 1$ and $g(x) = x - 2$ we get

$$\frac{d}{dx}\frac{x-1}{x-2} = \frac{(x-1)'(x-2) - (x-1)(x-2)'}{(x-2)^2} = \frac{1\cdot(x-2) - (x-1)\cdot 1}{(x-2)^2}$$

$$= \frac{(x-2) - (x-1)}{(x-2)^2} = \frac{-1}{(x-2)^2}$$

Special case of the quotient rule: Applying the quotient rule to the expression $1/g(x)$ yields

Reciprocal rule:

$$\frac{d}{dx}\frac{1}{g(x)} = -\frac{g'(x)}{g(x)^2}$$

Because

$$\frac{d}{dx}\frac{1}{g(x)} = \frac{0 \cdot g(x) - 1 \cdot g'(x)}{g(x)^2} = \frac{-g'(x)}{g(x)^2}$$

Product rule

We have:

Product rule:

$$\frac{d}{dx}(fg) = f'g + fg'$$

Note the this rule behaves differently from the simple rule for sums above
Example: Differentiate $f(x) = x^2 \sin x$. We get $f'(x) = 2x \sin x + x^2 \cos x$
Note: "constant multiple rule" is a special case: $(cf(x))' = cf'(x)$ since $\frac{d}{dx}c = 0$

Find the derivative of the sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

and express it using the sigmoid function itself. Hint: The identity $\frac{a}{1+a} = \frac{a+1-1}{1+a} = 1 - \frac{1}{1+a}$ should prove useful

Answer:

$$\frac{d}{d\sigma}\frac{1}{1+e^{-x}} = \ldots = \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} = \frac{1}{1+e^{-x}}\frac{1+e^{-x}-1}{1+e^{-x}} = \frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right) = \sigma(x)\,(1-\sigma(x))$$

Chain rule

Function composition operator "$\circ$":
Operator that takes two functions f and g and produces a function $h = g \circ f$ such that $h(x) = g(f(x))$. In other words, g is applied to the result of $f(x)$.

The chain rule states how the derivative of the composition of two differentiable functions f and g behaves:

Chain rule:

$$h' = (f \circ g)' = (f' \circ g) \cdot g'$$

or less formally

$$\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$$

Note that the chain rule is a core utility in machine learning, e.g. forming the base of how neural networks are learning, among countless other uses.

Use of chain rule is often straightforward. Just be clear what $f(x)$ and what $g(x)$ are. Here are some examples.

Example 1:

$$\frac{d}{dx}(1+x^2)^{100} = f'(g(x)) \cdot g'(x) = 100\, g^{99}(x) \cdot 2x = 100\,(1+x^2)^{99} \cdot 2x$$

Example 2:

$$\frac{d}{dx}\sqrt{3x+2} = \frac{d}{dx}(3x+2)^{1/2} = \frac{1}{2}(3x+2)^{-1/2} \cdot 3$$

Example 3:

$$\frac{d}{dx}(3x^5 - x + 14)^{11} = 11\,(3x^5 - x + 14)^{10} \cdot (15x^4 - 1)$$

Similarly the chain rule applies to composites of more than two functions:

$$(f \circ g \circ h)' = f' \circ (g \circ h) \cdot (g \circ h)' = f' \circ (g \circ h) \cdot g'(h) \cdot h' = (f' \circ g \circ h) \cdot (g' \circ h) \cdot h'$$

For example, compute $\frac{d}{dx} e^{\sin x^2}$:

- Define individual functions and compute their derivatives:

$$\begin{aligned} y := f(u) = e^u \quad &\Longrightarrow \quad \frac{dy}{du} = f'(u) = e^u = e^{\sin x^2} \\ u := g(v) = \sin v = \sin x^2 \quad &\Longrightarrow \quad \frac{du}{dv} = g'(v) = \cos v = \cos x^2 \\ v := h(x) = x^2 \quad &\Longrightarrow \quad \frac{dv}{dx} = h'(x) = 2x \end{aligned}$$

- Therefore:

$$\frac{d}{dx} e^{\sin x^2} = e^{\sin x^2} \cos(2x)\, 2x$$

Multivariate Chain Rule

Multivariate Chain Rule: Let

- $z = f(x_1, x_2, \ldots, x_m)$ be differentiable in m independent variables, and let
- $\forall i \in 1, \ldots, m : x_i = x_i(t_1, t_2, \ldots, t_n)$ be differentiable in n independent variables

Then, $\forall j \in 1, 2, \ldots, n$:

$$\frac{\partial z}{\partial t_j} = \sum_{i=1}^{m} \frac{\partial z}{\partial x_i} \frac{\partial x_i}{\partial t_j} = \frac{\partial z}{\partial x_1} \frac{\partial x_1}{\partial t_j} + \frac{\partial z}{\partial x_2} \frac{\partial x_2}{\partial t_j} + \cdots + \frac{\partial z}{\partial x_m} \frac{\partial x_m}{\partial t_j} = \nabla z \cdot \frac{\partial \boldsymbol{x}}{\partial t_j}$$

For example, in the case of $x(t)$, $y(t)$, and $z = f(x, y)$, we get

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \frac{dx}{dt} + \frac{\partial z}{\partial y} \frac{dy}{dt}$$

with $z = f(x(t), y(t))$ being differentiable at t.

Example: Multivariate Chain Rule

Given $u(x, y) = x^2 + 2y$ with $x(r, t) = r\sin(t)$ and $y(r, t) = \sin^2(t)$, we are interested in finding $\partial u/\partial r$ and $\partial u/\partial t$.

$$\frac{\partial u}{\partial r} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial r} = 2x\sin t + 2 \times 0 = 2r\sin^2 t$$

and

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial t} \\
&= 2x \times r\cos t + 2 \times 2\sin t\cos t \\
&= 2r\sin t \times r\cos t + 4\sin t\cos t \\
&= 2(r^2 + 2)\sin t\cos t \\
&= (r^2 + 2)\sin(2t)
\end{aligned}$$

You are given 200 m of wire to make a fence around some rectangular area of side length a and b, respectively. How should you choose a and b s.t. the area which you are covering is maximized?

Answer:

$$\max_{a,b} a\,b \quad \text{s.t.} \quad 2a + 2b = 200 \iff \max_{a} a\,(100 - a) \iff a^{\star} = 50 \quad \text{(square area)}$$

Extreme Value Problems

Systematic procedure to find the minimum and maximum values of a function $f(x)$ on an interval $[a, b]$:

1. Solve $f'(x) = 0$ to find the list of critical points of f
2. Exclude any critical points not inside the interval $[a, b]$
3. Add to the list the *endpoints* a, b of the interval and any points of discontinuity or non-differentiability
4. At each point on the list, evaluate f: the biggest number that occurs is the maximum, and the littlest number that occurs is the minimum.

Example: Find the minima and maxima of the function $f(x) = x^4 - 8x^2 + 5$ on the interval $[-1, 3]$.

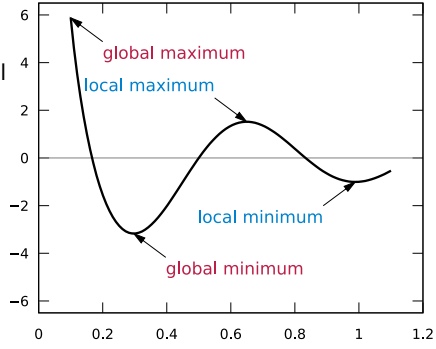
1. $f'(x) = 4x^3 - 16x = 0 \iff x(x^2 - 4) = 0 \iff x(x + 2)(x - 2) = 0$
   So critical points are $-2, 0, 2$
2. Remove -2 from list since outside interval
3. Add endpoints $-1, 3$ to list. Now list is $-1, 0, 2, 3$
4. Evaluate f at $-1, 0, 2, 3$ which yields $-2, 5, -11, 14$, respectively

Hence maximum is 14, which occurs at $x = 3$, and the minimum is -11, which occurs at $x = 2$.

Global and Local Extrema

Use the *derivative test* to determine the type of each cricitcal point.
In the scalar case:

- $f''(x) < 0$: local maximum
- $f''(x) > 0$: local minimum
- $f''(x) = 0$: test inconclusive $\Rightarrow$ higher-order derivative test



Remarks:

- There is other tests to determine the nature of a critical point, e.g. first derivative test
- The above holds only for "reasonably well-behaved" functions
- The rules and intuition readily extend to the multi-dimensional case

Mathematical Optimization

General optimization problem: Given: a function $f : A \to \mathbb{R}$ from some set A to the real numbers
Objective: find an element $\boldsymbol{x}^\star \in A$ such that

- $\forall \boldsymbol{x} \in A : f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ "minimization"
- $\forall \boldsymbol{x} \in A : f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x})$ "maximization"

Optimization problems

- Two major classes: discrete and continuous optimization (focus on latter)
- We can always use *minimization* because

$$f(\mathbf{x}^\star) \geq f(\mathbf{x}) \quad \Longleftrightarrow \quad -f(\mathbf{x}^\star) \leq -f(\mathbf{x})$$

Optimization in Practice

In practice, algorithms that solve optimization problems fall into three broad classes:

- those that terminate in a finite number of steps, e.g. Simplex, combinatorial algos
- iterative methods that converge to a solution, e.g. Gradient descent, Newton's method, SQP
- heuristics that provide approx. solutions, e.g. genetic and evolutionary algos, PSO

Relevance for machine learning:

- Many machine learning problems are formulated as the minimization of some (so-called) loss function over a training set of examples
- Finding a solution may require to employ method from any of the three classes above

Nabla Operator: Gradients in $\mathbb{R}^n$ are conveniently expressed with the *Nabla operator*:

$$\nabla \equiv \begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} \quad \text{or applied to some } f(\boldsymbol{x}) \quad \nabla f(\boldsymbol{x}) = \begin{pmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$$

Hessian Matrix: The Hessian is a square matrix of 2nd order partial derivatives of a scalar-valued function and a measure of local curvature:

$$H_f = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\, \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\, \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\, \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\, \partial x_1} & \dfrac{\partial^2 f}{\partial x_n\, \partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

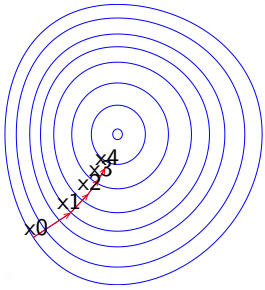
Gradient Descent

Gradient Descent (GD) is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function:

- Assume $F(\boldsymbol{x})$ is defined and locally differentiable around $\boldsymbol{a}$
- Then $F(\boldsymbol{x})$ decreases fastest in the direction of the *negative* gradient of F at a, which is $-\nabla F(\boldsymbol{a})$
- For a sufficiently small step size (learning rate) $\gamma \in \mathbb{R}^+$, we have

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_n - \gamma \nabla F(\boldsymbol{a}_n) \quad \text{and} \quad F(\mathbf{a_n}) \geq F(\mathbf{a_{n+1}})$$

Iterating the above we get the



Gradient Descent Update Rule:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_n \nabla F(\boldsymbol{x}_n), \ n \geq 0$$

where $F(\boldsymbol{x}_0) \geq F(\boldsymbol{x}_1) \geq F(\boldsymbol{x}_2) \geq \ldots$ is a monotonic sequence

Note: Convergence is *guaranteed* under certain reasonable assumptions and a proper choice of γ (e.g. via line search or the Barzilai Borwein method)

Wide range of applications: Linear and non-linear systems, least squares problems, anything "minimizable"

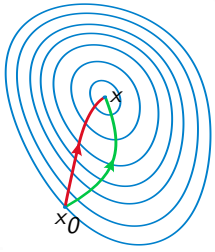
Newton's Method

Iterative prodecure for finding roots $\boldsymbol{x}^\star$ of a differentiable function F:

$$\{\boldsymbol{x}_i^\star : F(\boldsymbol{x}_i^\star) = 0\}$$

Intuitively, does so by repeatedly intersecting tangents with x-Axis:

$$y = F(x_k) + F'(x_k)(x - x_k) \stackrel{!}{=} 0 \quad \Longrightarrow \quad x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}$$



Applicable to minimization problem $\min_{x \in \mathbb{R}} f(x)$: Simply let $F(x) = f'(x)$ which yields:

Newton's Method (for optimization):

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \quad \text{(univariate case)}$$

Corresponds to a second-order approxmiation of $f \Rightarrow$ fast convergence

Really Newton is just a second-order Taylor expansion of f around x_k:

$$f(x_k + t) \approx f(x_k) + f'(x_k)t + \frac{1}{2}f''(x_k)t^2$$

and minmizing its first derivative:

$$\frac{d}{dt}f(x_k + t) \approx f'(x_k) + f''(x_k)t \stackrel{!}{=} 0 \quad \Longrightarrow \quad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

Similarly in the multi-variate case, with $\boldsymbol{H}_f$ the Hessian of f:

$$f(\boldsymbol{x}_k + \boldsymbol{a}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top \boldsymbol{a} + \frac{1}{2}\boldsymbol{a}^\top \boldsymbol{H}\boldsymbol{a} \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{a}} f(\boldsymbol{x}_k + \boldsymbol{a}) \approx \nabla f(\boldsymbol{x}_k) + \boldsymbol{H}\boldsymbol{a} \stackrel{!}{=} \boldsymbol{0}$$

Newton's Method (for optimization):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}_f^{-1} \nabla f(\boldsymbol{x}_k) \quad \text{(multivariate case)}$$

Note that this yields the exact solution for any quadratic objective function in a *single* step.

Stochastic Gradient Descent (SGD)

SGD is:

- an iterative optimization method, stochastic approximation of gradient descent
- the fundamental idea behind many machine learning algorithms such as linear support vector machines, logistic regression, graphical models
- the de facto standard training algorithm for neural networks

Given an objective function $Q_i(\boldsymbol{w})$ which computes some sort of error for sample i, e.g. $Q_i(\boldsymbol{w}) = (\hat{y}_i - y_i)^2$

Classical SGD:

Initialize parameter vector $\boldsymbol{w}$ and set learning rate η
while termination criterion not satisfied **do**
  randomly shuffle samples in training set
  for $i \leftarrow 1, n$ **do**
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla Q_i(\boldsymbol{w})$
  end for
end while

Example: SGD to compute a least squares fit of a straight line
Given training data $(x_i, y_i)_{i=1}^N$, fit the model

$$\hat{y} = w_1 + w_2 x$$

Objective function (minimize):

$$Q(w) = \sum_{i=1}^{n} Q_i(w) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (w_1 + w_2 x_i - y_i)^2$$

The update step of the Basic SGD algorithm above becomes

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \leftarrow \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \eta \begin{bmatrix} \frac{\partial}{\partial w_1}(w_1 + w_2 x_i - y_i)^2 \\ \frac{\partial}{\partial w_2}(w_1 + w_2 x_i - y_i)^2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \eta \begin{bmatrix} 2(w_1 + w_2 x_i - y_i) \\ 2x_i(w_1 + w_2 x_i - y_i) \end{bmatrix}.$$

SGD Extensions

Extension 1: Batch updates

Issue: Each iteration of classical SGD evaluates the gradient at a single point x_i only

$\Rightarrow$ often results in noisy estimates of the true gradient

$\Rightarrow$ computationally expensive (blocking vectorization)

Idea: Combine multiple gradient estimates of a so-called "mini-batch", then do a single update

$\Rightarrow$ smoother convergence thanks to better estimates

$\Rightarrow$ speed-up owing to vectorization

Extension 2: Implicit updates (ISGD)

Issues:

- Classical SGD sensitive to choice of η and may easily diverge
- Better convergence with some decay schedule of η over iterations, $\eta = \eta(\text{iter})$

Idea: Evaluate stochastic gradient at the next iteration rather than the current one:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla Q_i(\boldsymbol{w}_{k+1})$$

To see the effect, assuming an ordinary least squares objective

$$Q_i(\boldsymbol{w}) = \frac{1}{2}(\hat{y}_i - y_i)^2$$

in the usual samples $\{(x_i, y_i)\}_{i=1}^N$, compare the resulting update rules:

$$\text{Classical SGD:} \quad \boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \eta(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}_k)\boldsymbol{x}_i$$

$$\text{Implicit SGD (closed-form!):} \quad \boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \frac{\eta}{1 + \eta \left\|\boldsymbol{x}_i\right\|^2} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}_k\right) \boldsymbol{x}_i$$

Note the normalizing effect on η, resulting in numerically stable procedure for virtually all η

You are in the middle of a minimization with ISGD:

- Iteration $k = 348$, number of samples $n = 50'000$
- Current weight vector: $\boldsymbol{w}_{348} = \begin{pmatrix} 5 & -3 \end{pmatrix}^\top$
- Current sample: $\boldsymbol{x}_{1283} = \begin{pmatrix} 3 & -4 \end{pmatrix}^\top, y_{1283} = 57$
- Learning rate: $\eta = 0.04$

Continue the iteration for one more step to find the next weight vector $\boldsymbol{w}_{349}$.

Answer:

$$\begin{aligned} \boldsymbol{w}_{349} &= \boldsymbol{w}_{348} + \frac{\eta}{1 + \eta \left\| \boldsymbol{x}_i \right\|^2} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}_k \right) \boldsymbol{x}_i \\ &= \begin{pmatrix} 5 \\ -3 \end{pmatrix} + \frac{0.04}{1 + 0.04 \times 25} (57 - 27) \begin{pmatrix} 3 \\ -4 \end{pmatrix} \\ &= \begin{pmatrix} 6.8 \\ -5.4 \end{pmatrix} \end{aligned}$$

Extension 3: Momentum

Issue: Classical SGD frequently shows oscillations, hampering convergence

Idea: Introduce some "memory" $\Delta \boldsymbol{w}_k$ of the current direction, i.e., do not take "hard turns" at every step:

$$\Delta \boldsymbol{w}_{k+1} = -\eta \nabla Q_i(\boldsymbol{w}_k) + \alpha \Delta \boldsymbol{w}_k$$

resulting in the update rule

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \Delta \boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla Q_i(\boldsymbol{w}_k) + \alpha \Delta \boldsymbol{w}_k$$

with $\alpha \in [0, 1)$ an exponential decay factor (hyperparameter)

More Extensions

More sophisticated extensions that adapt the learning rate exists:

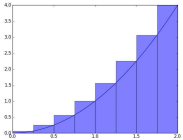
- AdaGrad (adaptive gradient): offers per-parameter learning rate
- Adam (adaptive momentum): also per-parameter learning rate; basis for most modern algorithms (AdaMax, AMSGrad)
- Second-order methods: stochastic analogue of the standard (deterministic) Newton algorithm

Integration

Definite integrals of real-valued functions $f(x)$ w.r.t. a real variable x on an interval $[a, b]$ is written as

$$\int_a^b f(x)\,dx$$

with limits a and b. A function f is *integrable* if its integral over its domain is finite

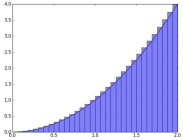


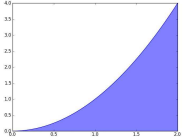
Indefinite integrals (antiderivative), written

$$F(x) := \int f(x)\,dx$$

represent a class of functions whose derivative is the integrand, hence

$$F' = f$$





Please note that all statemets around *integrals* and *integrability* are meant in the *Riemann* sense

Property 1 (linearity): The integral of a linear combination is the linear combination of the integrals:

$$\int_a^b (\alpha f + \beta g)(x)\, dx = \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx$$

Property 2: Flipping integration bounds flips the sign of the result,

$$\int_a^b f(x)\ dx = -\int_b^a f(x)\ dx$$

Property 3: Given $\int_a^b f(x)\, dx$ exists with $a \leq b$, then for any $c \in [a, b]$, we have

$$\int_a^b f(x)\ dx = \int_a^c f(x)\ dx + \int_c^b f(x)\ dx$$

Closed-form Integration

Refresher by example:

- $\int a\,dx = ax + C$
- $\int x^n\,dx = \frac{x^{n+1}}{n+1} + C, \quad n \neq -1$
- $\int (ax+b)^n\,dx = \frac{(ax+b)^{n+1}}{a(n+1)} + C, \quad n \neq -1$
- $\int \frac{1}{x}\,dx = \ln|x| + C$
- $\int e^{ax}\,dx = \frac{1}{a}e^{ax} + C$
- $\int f'(x)e^{f(x)}\,dx = e^{f(x)} + C$
- $\int a^x\,dx = \frac{a^x}{\ln a} + C$
- $\int \ln x\,dx = x\ln x - x + C$
- $\int \log_a x\,dx = \frac{x\ln x - x}{\ln a} + C$
- $\int \sin x\,dx = -\cos x + C$
- $\int \cos x\,dx = \sin x + C$
- $\int \tan x\,dx = \ln|\sec x| + C = -\ln|\cos x| + C$
- $\int_0^\infty \sqrt{x}\,e^{-x}\,dx = \frac{1}{2}\sqrt{\pi}$ Gamma function
- $\int_0^\infty e^{-ax^2}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}$ Gaussian integral, $a > 0$

Integration By Parts

If $y(x) = u(x)\,v(x)$, then

$$y' = (uv)' = u'v + uv' \quad \Longleftrightarrow \quad uv' = (uv)' - u'v$$

Therefore we get:

Integration by parts (helpful to integrate products of functions):

$$\int uv'\,dx = \int (uv)'\,dx - \int u'v\,dx = uv - \int vu'\,dx$$

Example: To find the antiderivative $\int x\sin(x)dx$, we *choose* $u := x$ and $v := \sin x$,

$$\int x\sin x\,dx = -x\cos x - \int(-\cos x)\,dx = \sin x - x\cos x + C$$

Note that the "right" choice of u and v is crucial

Integration By Substitution

Transforms one integral into another integral that is easier to compute

Integration by substitution (helpful to integrate compositions of functions):
To transform $\int f(g(x))\,dx$

- choose a *suitable* substitution $u = g(x)$
- compute $du = g'(x)\,dx$ (not obvious)
- substitute integrand and differential, then solve the (simpler) integral
- back-substitute

Example: To find the antiderivative of $x\cos\left(x^2+1\right)$, choose $u := x^2+1$, $du = 2\,dx$. Then

$$\int x\cos\left(x^2+1\right)dx = \frac{1}{2}\int 2x\cos\left(x^2+1\right)dx = \frac{1}{2}\int \cos u\,du = \frac{1}{2}\sin u + C = \frac{1}{2}\sin\left(x^2+1\right) + C$$

Numerical Integration

Issue: In practice, many integrals possess no closed-form solution
Resort: Quadrature (numerical integration) for 1-dimensional integrals works well, e.g.

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} f(x_i^*)\,\Delta x$$

Multi-dimensional integrals (arising e.g. as normalizers in Bayesian stats) require more sophisticated techniques such as Monte Carlo integration. Popular algorithms in this class include Metropolis-Hastings and Gibbs sampling (both instances of Markov Chain Monte Carlo)

Practice!

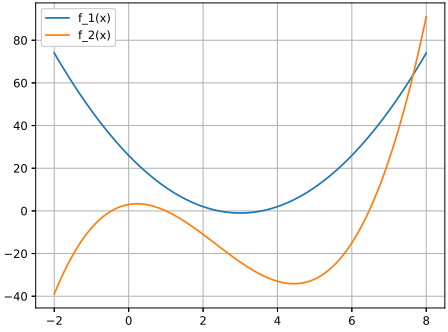
Implement the 1-dimensional version of

- Gradient Descent
- Basic SGD

Then compare convergence of the two algorithms on one or both test functions

$$f_1(x) = 3(x-3)^2 - 1 \quad \text{and} \quad f_2(x) = (x-1)^3 - 4x^2 + 4$$

Also vary the initial guess x_0 and the learning rate η. What happens e.g. for f_2 with $x_0 = 0$?



Introduction

Background

- Probability is the mathematical field concerned with reasoning under uncertainty
- Probabilistic models enable us to reason about the *likelihood* of various events
- Use of probabilities to describe the frequencies of repeatable events (like coin tosses) is fairly uncontroversial

Schools of thoughts

- Frequentists scholars adhere to an interpretation of probability that applies only to such repeatable events
- Bayesian scholars use the language of probability more broadly to formalize our reasoning under uncertainty

Bayesian probability is characterized by two unique features:

- assigning prior belief e.g., P["moon is made out of cheese"], opions may differ
- evidence, that is hard facts that we measure

Bayesian probability provides unambiguous rules for how we should *update* our belief in light of new evidence, while allows for different individuals to start off with different prior beliefs

Terminology

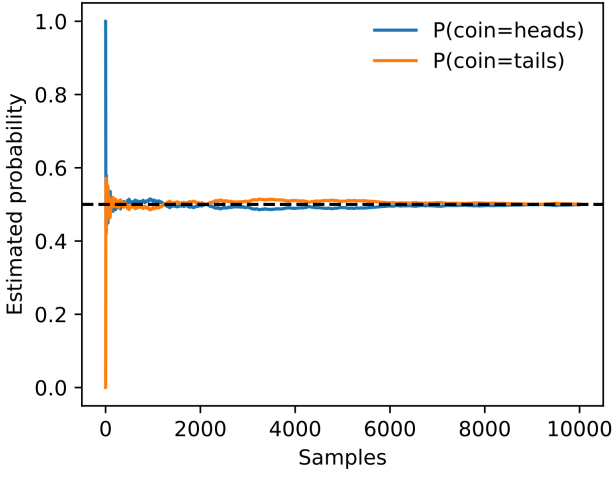
Imagine the experiment:

- Toss a coin n times. The probability for heads is $p_h = 1/2$.
- We observe n_h heads and $n_t = n - n_h$ tails
- Fraction of heads that we *expect* to see should exactly match the expected fraction of tails
- The larger n the less likely $n_h = n_t$ becomes

Terminology:

- p_h is called a *probability*, captures the certainty with which any given toss will come up heads
- Probabilities assign scores between 0 and 1 to *events* (i.e. the outcomes of interest)
- Here the event of interest is "heads" and we denote the corresponding probability $P(\text{heads})$
- $P(\cdot) = 1$ indicates absolute certainty, and $P(\cdot) = 0$ impossibility
- n_h/n and n_t/n are called *statistics* (not probabilities)—empirical quantities computed based on the observed data
- *Estimators* are special statistics that use the data to produce estimates of model parameters (like probabilities). *Consistent estimators* converge to the corresponding probability (cf. next slide)

Estimators and Consistency



Formal Definition

Denote the set of possible outcomes $\mathcal{S}$, and call it the *sample space* or *outcome space*, e.g.:

- Rolling a single coin: $\mathcal{S} = \{\text{heads}, \text{tails}\}$
- Rolling a single die, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$
- flipping two coins, possible outcomes are $\{(\text{heads}, \text{heads}), (\text{heads}, \text{tails}), (\text{tails}, \text{heads}), (\text{tails}, \text{tails})\}$

Events are subsets of the sample space.

Axioms of Probability: A *probability function* maps events onto real values $P : \mathcal{A} \subseteq \mathcal{S} \to [0, 1]$. The probability of an event $\mathcal{A}$ in the given sample space $\mathcal{S}$, denoted $P(\mathcal{A})$, satisfies the following properties (Kolmogorov, 1933):

- The probability of any event $\mathcal{A}$ is a non-negative real number, i.e., $P(\mathcal{A}) \geq 0$;
- The probability of the entire sample space is 1, i.e., $P(\mathcal{S}) = 1$;
- For any countable sequence of events $\mathcal{A}_1, \mathcal{A}_2, \ldots$ that are *mutually exclusive* ($\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $i \neq j$), the probability that any of them happens is equal to the sum of their individual probabilities, i.e., $P(\bigcup_{i=1}^{\infty} \mathcal{A}_i) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$.

Random Variables

A *random variable* (RV) is a mapping from an underlying sample space to a set of values

- Random variables can be much coarser than the raw sample space. E.g. define RV "greater than 0.5" is binary but has an infinite sample space
- Multiple RVs may share the same underlying sample space. E.g. for RVs "home alarm goes off" and "my house was burglarized", knowing the value taken by one random variable can tell us something about the likely value of another random variable

Every value taken by a random variable corresponds to a subset of the underlying sample space. Thus the occurrence where the random variable X takes value v, denoted by $X = v$, is an *event* and $P(X = v)$ denotes its probability.

Notation in practice

- $P(X)$ may refer broadly to the *distribution* of X
- $P(X, Y) = P(X)P(Y)$, may be shorthand to express a statement that is true for all of the values that the random variables X and Y, i.e., $\forall i, j : P(X = i \text{ and } Y = j) = P(X = i)P(Y = j)$
- $P(v)$ may be used when the random variable is clear from the context
- $P(1 \leq X \leq 3)$ may denote the probability of the event $\{1 \leq X \leq 3\}$

In general, there's two types of random variables:

- Discrete, e.g. flips of a coin or tosses of a die
- Continuous, e.g. height H of a person. Note that in this case, asking e.g. for $P(H = 1.8796549851915233234)$ does not make sense. Instead, we work with probability *densities* p

Probability Density Functions (PDFs)

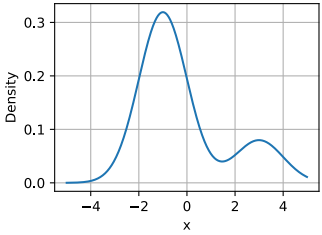
PDF: Informally, a PDF $p(x)$ is characterized by the following properties:

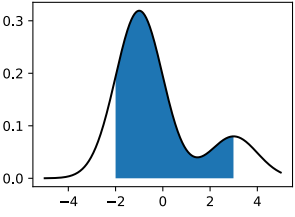
- $p(x) \geq 0$
- $\int_{-\infty}^{\infty} p(x)\,dx = 1$

From this it follows that the probability of any single value x is 0, $P(X = x) = 0$ and that $p(x) > 1$ is admissible

To compute actual probabilities we integrate $p(x)$ over an *interval*:

$$P(X \in (a, b]) = \int_a^b p(x)\,dx$$





Cumulative Distribution Functions (CDFs)

The CDF $F(x)$ of a random variable X with density $p(x)$ is given by

$$F(x) = \int_{-\infty}^{x} p(x)\,dx = P(X \leq x)$$

Note that $F(x)$ represents an actual probability

Properties:

- $F(x) \to 0$ as $x \to -\infty$
- $F(x) \to 1$ as $x \to \infty$
- $F(x)$ is non-decreasing, i.e. $y > x \implies F(y) \geq F(x)$
- $F(x)$ is continuous (has no jumps) if X is a continuous random variable

Characterizing Distributions

Expected value

Expected value of a RV (mean):

- $\mu_X = \mathrm{E}[X] = \sum_i x_i \, p_i \quad$ (for discrete RV X)
- $\mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \, p(x) \, dx \quad$ (for continuous RV X)

Properties:

- For any RV X and numbers a and b, we have that $\mu_{aX+b} = a\mu_X + b$
- For any two random variables X and Y, we have $\mu_{X+Y} = \mu_X + \mu_Y$

Means characterize average behavior of a RV, which is not sufficient however: A profit of $CHF10 \pm CHF1$ per sale is very different from making $CHF10 \pm CHF15$ per sale.

One possibility to quantify risk r as expected loss. That is, by weighting the loss l (e.g., in CHF) which some event causes with the probability p_e that this event actually occurs:

$$r := p_e \, l$$

Now assume you want to park your car in the city of Zurich. The parking garage will charge you CHF 24. If you park in some quiet corner, you may get fined and pay CHF 40. The chance that you'll get caught is 2/3. What's the strictly economic choice here, parking garage or taking your chances?

Answer:

The parking garage as it has the lower expected loss: $24 < \frac{2}{3} \times 40 \approx 26.67$

Characterizing Distributions

Variance and Standard Deviation

Variance: The variance measures the variation of a RV about its mean:

$$\sigma_X^2 = \mathrm{Var}(X) = E\left[(X - \mu_X)^2\right]$$

Properties:

- For any RV X, $\mathrm{Var}(X) \geq 0$, with $\mathrm{Var}(X) = 0$ iff X constant
- For any RV X and numbers a and b, we have that $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
- For any two independent RV X and Y, we have $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
- $E\left[(X - \mu_X)^2\right] = E[X^2] - \mu_X^2$

Standard Deviation (SD): The SD also measures the variation of a RV about its mean, but has the same units as the mean:

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

Analogous properties as in the case of variance apply

Continuous Distributions

A Well Behaved Example

Task: Compute the first two moments of the uniform distribution

$$p(x) = \begin{cases} 1 & x \in [0,1], \\ 0 & \text{otherwise} \end{cases}$$

Find expected value:

$$\mu_X = \int_{-\infty}^{\infty} x p(x)\, dx = \int_0^1 x\, dx = \frac{1}{2}$$

Find variance:

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 p(x)\, dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

where we used above identity $\mathrm{E}\left[(X-\mu_X)^2\right] = \mathrm{E}\left[X^2\right] - \mu_X^2$

Continuous Distributions

An Less Well-Behaved Example

Task: Compute the first two moments of a Cauchy-like distribution with the PDF given by

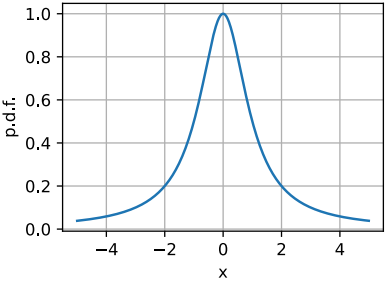
$$p(x) = \frac{1}{1+x^2}$$

Trying to compute the variance yields

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 p(x)\, dx = \int_{-\infty}^{\infty} \frac{x^2}{1+x^2}\, dx = \infty$$

Hence the variance of $p(x)$ is undefined.

A similar problem arises for the mean, which is undefined too.



Joint Density

Similar to the univariate case, a PDF can be defined for two or more variables.
The two RVs X and Y can be defined to have a *joint density* $p(x, y)$ or $p_{X,Y}$. Then properties similar to the univariable case hold:

- $p(x, y) \geq 0$
- $\int_{\mathbb{R}^2} p(x, y)\, dx\, dy = 1$
- $P((X, Y) \in \mathcal{D}) = \int_{\mathcal{D}} p(x, y)\, dx\, dy$

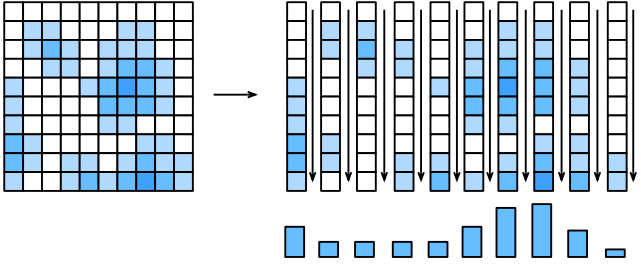
This generalizes to any number of RVs.

Marginal Distributions

When dealing with multiple variables, we often ask "how is this one variable distributed?". The distribution that answers this questions this is the *marginal distribution*.
Mathematically (often referred to as *integrating out* or *marginalizing out* unneeded variables):

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)\, dy$$

Visually:



Joint Probability Marginal Probability

Conditional Probability and Statistical Independence

The conditional distribution quantifies the probability of an event occurring, given that another event (e.g. by assertion or by evidence) has already occurred.

Conditional Probability: We define the conditional probability of "$\mathcal{A}$ given $\mathcal{B}$", written $P(\mathcal{A} \mid \mathcal{B})$ to be:

$$P(\mathcal{A} \mid \mathcal{B}) = \frac{P(\mathcal{A}, \mathcal{B})}{P(\mathcal{B})} \quad \text{or equivalently} \quad P(\mathcal{A}, \mathcal{B}) = P(\mathcal{A} \mid \mathcal{B})\, P(\mathcal{B})$$

Based on the above we define independence:

Statistical Independence: Events $\mathcal{A}$ and $\mathcal{B}$ are defined to be *statistically independent* if

$$P(\mathcal{A}, \mathcal{B}) = P(\mathcal{A})\, P(\mathcal{B})$$

Therefore independence implies both $P(\mathcal{A} \mid \mathcal{B}) = P(\mathcal{A})$ and $P(\mathcal{B} \mid \mathcal{A}) = P(\mathcal{B})$

Conditional Independence

Conditional probabilities are proper probabilities $\Rightarrow$ concepts of (in)dependence also apply to them.

Conditional Independence: Two RVs A and B are *conditionally independent* given a third RV C iff

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

Two independent RV can become dependent when conditioning on a third RV C. Occurs when RVs A and B correspond to causes of RV C.
Example:

- A broken bone and lung cancer might be independent in the general population
- Now condition on being in the hospital
- Conditionally, the broken bone is negatively correlated with lung cancer
- We say the broken bone *explained away* why some person is in the hospital, in turn lowering the probability that the person has lung cancer

Bayes' Theorem

By definition of conditional probabilities, $P(A, B) = P(B \mid A)\, P(A)$ and $P(A, B) = P(A \mid B)\, P(B)$, which is equivalent to

Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Please note the profound implications: It allows us to reverse the order of conditioning. If we know how to estimate $P(B \mid A)$, $P(A)$, and $P(B)$, then we can estimate $P(A \mid B)$.
Example:
Given the prevalence of symptoms for a given disease, and the overall prevalences of the disease and symptoms, respectively, we can determine how likely someone is to have the disease based on their symptoms.
Special case: Without direct access to $P(B)$, we still know that

$$P(A \mid B) \propto P(B \mid A) P(A)$$

and use the normalization constraint $\sum_a P(A = a \mid B) = 1$.

Marginalization

Marginalization ("sum rule") is key ingredient to make Bayes theorem work.
It simply sums or integrates out, respectively, other variables from the joint distribution:

$$P(B) = \sum_a P(A, B)$$

$$p_B = \int_{-\infty}^{\infty} p(a, b)\, da$$

The result is called marginal probability.
In practice, where joint probabilities consist of many variables, this "normalizer" term in Bayes' theorem often becomes a computational bottleneck.

It's 1963 and you participate in "Let's Make a Deal" (TV Show). You play this game:

- There's three doors: behind two of them there's a goat, behind the third a Ferrari
- First you pick a door
- Then the host opens one of the two other doors with a goat behind
- Then you can optionally switch to the other closed door or stay with your initial one
- Then your door will be opened and if it's the Ferrari, the car is yours

If you want the car, what's best to do? a) stay, b) switch, c) makes no difference
Hint: Use Python or Bayes' theorem

Answer:

With loss of generality, you pick door 1 and the host opens door 2 or 3. Let
$D :=$ "number of the door with the car behind" $\in \{1, 2, 3\}$, and likewise
$H :=$ number of door which host opens. At the end of the game there's two options, $d = 1$ and $d = 3$:

$$P(D = d \,|\, H = 2) = \frac{P(H = 2 \,|\, D = d)\, P(D = d)}{P(H = 2)}$$

$$P(D = 1 \,|\, H = 2) = \frac{1/2 \times 1/3}{1/3 + 1/6} = \frac{1}{3} \qquad P(D = 3 \,|\, H = 2) = \frac{1 \times 1/3}{1/3 + 1/6} = \frac{2}{3}$$

Covariance

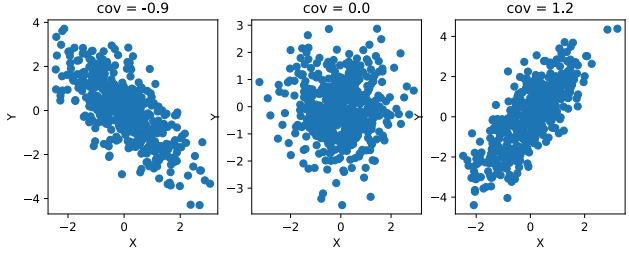
Covariance: Covariance measures the degree that two random variables fluctuate together:

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

Spelled out this means for the discrete and continuous case, respectively,

$$\sigma_{XY} = \sum_{i,j}(x_i - \mu_X)(y_j - \mu_Y)p_{ij} \quad \text{and} \quad \sigma_{XY} = \int_{\mathbb{R}^2}(x - \mu_X)(y - \mu_y)p(x, y)\,dx\,dy$$

Visually:



Summarizing the main properties of covariance:

- For any RV X, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- For any RVs X and Y and numbers a and b, $\mathrm{Cov}(aX + b, Y) = \mathrm{Cov}(X, aY + b) = a\mathrm{Cov}(X, Y)$
- If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$

Correlation

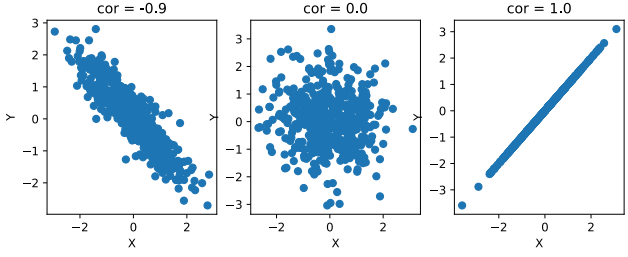
Correlation coefficient: Define the *correlation coefficient* to be

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Properties:

- For any RV X, $\rho(X, X) = 1$
- For any RVs X and Y and numbers a and b, $\rho(aX + b, Y) = \rho(X, aY + b) = \rho(X, Y)$
- If X and Y are independent with non-zero variance then $\rho(X, Y) = 0$

Visually:



Example

Let X be any random variable, and $Y = aX + b$ as linear deterministic function of X.
Then, one can compute that

$$\sigma_Y = \sigma_{aX+b} = |a|\,\sigma_X$$

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, aX + b) = a\mathrm{Cov}(X, X) = a\mathrm{Var}(X),$$

and by definition

$$\rho(X, Y) = \frac{a\mathrm{Var}(X)}{|a|\sigma_X^2} = \frac{a}{|a|} = \mathrm{sign}(a)$$

Thus the correlation is $+1$ for any $a > 0$, and -1 for any $a < 0$.
This shows that correlation measures the *degree and directionality* of variation of the two RV, not the scale that the variation takes.

Maximum Likelihood

Maximum Likelihood:

- A very common way of thinking in machine learning. Really it is a "point of view"
- Given a probabilistic model with unknown parameters, the most likely parameters are those under which the chance that the model generated the data are maximized

Technically, given our data X, we want

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta} \mid X)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ the parameter vector of interest.
Using Bayes' theorem the Maximum Likelihood Estimator (MLE) for $\boldsymbol{\theta}$ is

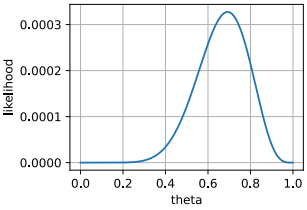
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \frac{P(X \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(X)} = \arg\max_{\boldsymbol{\theta}} P(X \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})$$

since $P(X)$ is just scaling the objective w.r.t. $\boldsymbol{\theta}$.
Exact same arguments apply to the densities of the continuous case.

Maximum Likelihood: An Example

Setting:

- Given a single, unknown parameter θ: Probability that a coin flip is heads
- Assume "uniformative prior" $\Rightarrow$ drop $P(\theta)$ from Bayes' theorem
- Data: "HHHTHTTHHHHHT", i.e. $n_H = 9$ heads, $n_T = 4$ tails



Question: What is the probability that the coin comes up heads?

Answer: Assume samples independent $\Rightarrow$ probabilities multiply. Likelihood becomes

$$P(X \mid \theta) = \theta^{n_H}(1-\theta)^{n_T} = P(X \mid \theta) = \theta^9(1-\theta)^4$$

Then maximize the likelihood: The first-order condition

$$\frac{d}{d\theta}P(X \mid \theta) = \frac{d}{d\theta}\theta^9(1-\theta)^4 = 9\theta^8(1-\theta)^4 - 4\theta^9(1-\theta)^3 = \theta^8(1-\theta)^3(9-13\theta) = 0$$

has three solutions: 0, 1 and 9/13. First two are minima since $P(X \mid \theta = 0) = P(X \mid \theta = 1) = 0$. Last solution assigns non-zero probability to our data, hence the maximum likelihood estimate is $\hat{\theta} = 9/13$.

Note that despite everyone would correctly guess 9/13, MLE provides a structured approach which equally applies to vastly more complex situations

Numerical Optimization and the Negative Log-Likelihood

In practice, we may have billions of parameters and examples.

- Independence assumption results in long product of small values, say, e.g. $\prod_{i=1}^{10^9} \frac{1}{2}$, something far below machine precision
- However, the log of this product is a sum, $-\sum_{i=0}^{10^9} \log_{10} 2 = -10^9 \times \log_{10} 2 \approx -301029995.664$, which fits into a 32-bit floating point value

Hence should work with log-likelihood,

$$\log P(X \mid \boldsymbol{\theta})$$

which will not affect the result since log is monotonically increasing.
Often working with loss functions, to be *minimized*, hence use negative log-likelihood:

$$-\log P(X \mid \boldsymbol{\theta})$$

Example: For the the coin flipping problem from before, this means minimizing

$$-\log(P(X \mid \boldsymbol{\theta})) = -\log(\theta^{n_H}(1-\theta)^{n_T}) = -n_H \log(\theta) - n_T \log(1-\theta)$$

which again leads to $\hat{\theta} = n_H/(n_H + n_T)$

Distributions

Bernoulli

Simple RV encoding e.g. a coin flip which comes up 1 (heads) with probability p and 0 (tails) with probability $1 - p$.

Bernoulli Distribution:

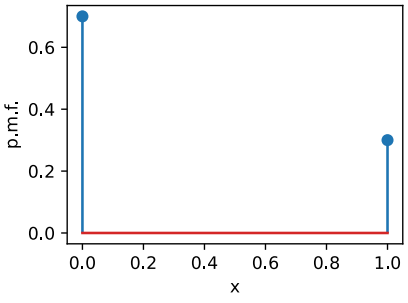
$$X \sim \text{Bernoulli}(p)$$

Cumulative distribution function:

$$F(x) = \begin{cases} 0 & x < 0, \\ 1 - p & 0 \leq x < 1, \\ 1 & x \geq 1. \end{cases}$$

Moments:

- $\mu_X = p$
- $\sigma_X^2 = p(1 - p)$



Distributions

Discrete Uniform

Every value out of given set is equally likely has a probability of $p_i = \frac{1}{n}$ to occur

Discrete Uniform Distribution: Assume the set of possible outcomes is $\{1, 2, \ldots, n\}$.

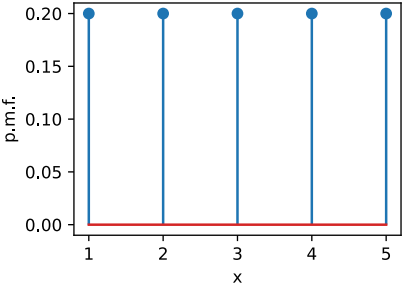
$$X \sim U(n)$$

Cumulative distribution function:

$$F(x) = \begin{cases} 0 & x < 1, \\ \frac{k}{n} & k \leq x < k+1 \text{ with } 1 \leq k < n, \\ 1 & x >= n. \end{cases}$$

Moments:

- $\mu_X = \frac{1+n}{2}$
- $\sigma_X^2 = \frac{n^2-1}{12}$



Distributions

Continuous Uniform

Similar to discrete uniform distribution: Starting from $U(n)$ with $n \to \infty$ and linearly scaling it to the interval $[a, b]$ will approach a continuous random variable that just picks an arbitrary value in $[a, b]$ with equal probability

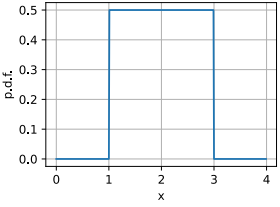
Continuous Uniform Distribution:

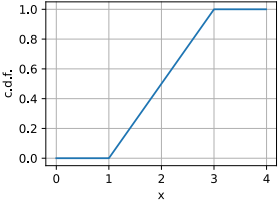
$$X \sim U(a, b)$$

PDF: $$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

CDF: $$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x >= b \end{cases}$$

Moments: $$\mu_X = \frac{a+b}{2} \quad \sigma_X^2 = \frac{(b-a)^2}{12}$$





Distributions

Binomial

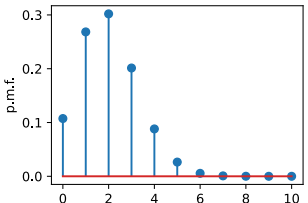
Given n independent experiments, each with success probability p. Interested in number of successes X:

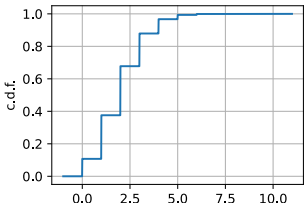
$$X = \sum_{i=1}^{n} X_i \quad \text{with} \quad X_i \sim \text{Bernoulli}(p) \qquad \Longleftrightarrow \qquad X \sim \text{Binomial}(n, p)$$

Binomial Distribution:

$$\text{CDF:} \qquad F(x) = \begin{cases} 0 & x < 0 \\ \sum_{m \leq k} \binom{n}{m} p^m (1-p)^{n-m} & k \leq x < k+1 \text{ with } 0 \leq k < n \\ 1 & x >= n \end{cases}$$

$$\text{Moments:} \qquad \mu_X = np \quad \sigma_X^2 = np(1-p)$$





Distributions

Poisson

Model for arrival rates, e.g. number of buses at a given stop per time interval. Call this rate λ (with units [1/s])

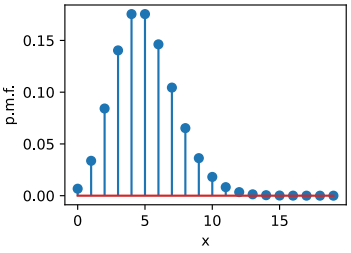
Poisson Distribution:

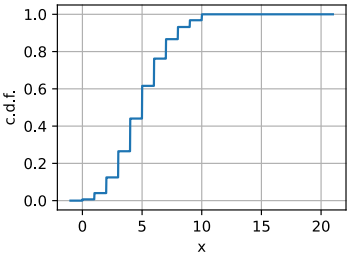
$$X \sim U(a, b)$$

$$\text{PMF:} \qquad p_k = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\text{CDF:} \qquad F(x) = \begin{cases} 0 & x < 0, \\ e^{-\lambda} \sum_{m=0}^{k} \frac{\lambda^m}{m!} & k \leq x < k+1 \text{ with } 0 \leq k \end{cases}$$

$$\text{Moments:} \qquad \mu_X = \lambda \quad \sigma_X^2 = \lambda$$





Distributions

Gaussian

The central distribution in probability theory with many unique properties because of the *central limit theorem*.

Informally, given N arbitrarily distributed, i.i.d. RVs X_i and

$$X^{(N)} = \sum_{i=1}^{N} X_i$$

their sum. Then, under mild conditions,

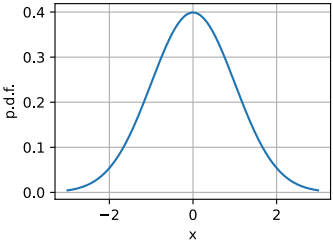
$$\frac{X^{(N)} - \mu_{X^{(N)}}}{\sigma_{X^{(N)}}} \quad \sim \quad \text{approximately Gaussian}$$

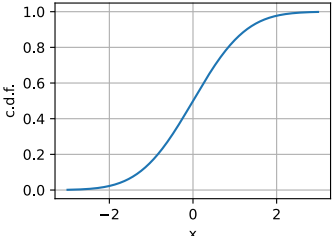
Gaussian Distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

PDF: $p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Moments: $\mu_X = \mu \quad \sigma_X^2 = \sigma^2$





Distributions: Summary

Things to remember about distributions:

- Bernoulli RVs can be used to model events with a yes/no outcome
- Discrete uniform distributions model selects from a finite set of possibilities
- Continuous uniform distributions select from an interval
- Binomial distributions model a series of Bernoulli RVs, and count the number of successes
- Poisson RVs model the arrival of rare events
- Gaussian RVs model the result of adding a large number of independent RVs together
- All the above distributions belong to the exponential family

License

This work is based on the following works:

- Jim Hefferson, *Linear Algebra*, 4th Edition, freely available at `https://hefferon.net/linearalgebra/`
- Wikipedia, The Free Encyclopedia, `https://www.wikipedia.org/`
- D2L.ai: Interactive Deep Learning Book with Multi-Framework Code, Math, and Discussions, freely available at `https://github.com/d2l-ai/d2l-en`