Foundations of Data Science, Fall 2020

2. Mathematics Basics

Prof. Dan Olteanu

DaST Data•(Systems+Theory)

University of Zurich

Sept 18, 2020

https://lms.uzh.ch/url/RepositoryEntry/16830890400

https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09

Today's Lecture

- No Machine Learning without rigorous mathematics
- Serves as reference for notation used throughout the course
- If there are any holes make sure to fill them sooner than later
- Attempt Exercise Sheet 1 to see where you are standing
- Good reference: Maths4ML document in OLAT
- Specific maths topics will be discussed when needed

Lecture topics

- Linear algebra
- Calculus
- Probability theory

Linear Algebra

Vectors

We will mostly work in the real vector space

- Scalar: single number $r \in \mathbb{R}$
- Vector: array of numbers $\mathbf{v} \in \mathbb{R}^D$ of dimension D arranged in a **column**

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_D \end{bmatrix}$$

- $\mathbf{v}^\top = (v_1, \ldots, v_D)$ is the transpose of $\mathbf{v}$
- $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^D$ are linearly independent if

$$\nexists r_1, \ldots, r_n \in \mathbb{R} \setminus \{0\} \text{ such that } \sum_{i \in [n]} r_i \mathbf{v}_i = \mathbf{0}$$

- The span of $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ for a vector space V is the set of all vectors that can be expressed as a linear combination of them:

$$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\mathbf{v} \in V : \exists \alpha_1, \ldots, \alpha_n \text{ such that } \alpha_1\mathbf{v}_1 + \ldots + \alpha_n\mathbf{v}_n = \mathbf{v}\}$$

Vector Norms

Vector norms allow us to talk about the length of vectors

- The L^p norm of $\mathbf{v} \in \mathbb{R}^D$ is given by

$$\|\mathbf{v}\|_p = \left(\sum_{i \in [D]} |v_i|^p \right)^{1/p}$$

- Properties of L^p (which actually hold for any norm):
  - $\|\mathbf{v}\|_p = 0$ implies $\mathbf{v} = \mathbf{0}$
  - $\|\mathbf{v} + \mathbf{w}\|_p \leq \|\mathbf{v}\|_p + \|\mathbf{w}\|_p$
  - $\|r\,\mathbf{v}\|_p = |r|\;\|\mathbf{v}\|_p$ for all $r \in \mathbb{R}$
- Popular norms:
  - Manhattan norm L^1
  - Eucledian norm L^2
  - Maximum norm L^∞ where $\|\mathbf{v}\|_\infty = \max_{i \in [D]} |v_i|$

Inner Product Spaces

An **inner product** on a real vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ with:

- $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{0}$
- Linearity: $\langle \mathbf{x} + \mathbf{y}, \mathbf{v} \rangle = \langle \mathbf{x}, \mathbf{v} \rangle + \langle \mathbf{y}, \mathbf{v} \rangle$ and $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle$
- Commutativity: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

(for $\alpha \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y}, \mathbf{v} \in V$)

Any inner product on V induces a norm on V: $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

- Check out the Pythagorean theorem and the Cauchy-Schwarz inequality

Standard inner product on $\mathbb{R}^D$ is given by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i \in [D]} x_i y_i = \mathbf{x}^\top \mathbf{y}$

- $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Also denoted by $\mathbf{x} \perp \mathbf{y}$.
- $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ are orthonormal if they are orthogonal and $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$.

Matrices

Matrix: two-dimensional array $\mathbf{A} \in \mathbb{R}^{m \times n}$ written as

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} = [\mathbf{a}_1 \ldots \mathbf{a}_n]$$

- Vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^m$ are $\mathbb{R}^{m \times 1}$ matrices
- $\mathbf{A}_{i,:} = (a_{i,1}, \ldots, a_{i,n})$ denotes i-th row
- $\mathbf{A}_{:,i} = \mathbf{a}_i$ denotes i-th column
- $\mathbf{A}^\intercal$ is the transpose of $\mathbf{A}$ such that $(\mathbf{A}^\intercal)_{i,j} = \mathbf{A}_{j,i}$

Special Matrices

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

- $\mathbf{A}$ is symmetric if $\mathbf{A} = \mathbf{A}^\intercal$
- $\mathbf{A} \in \mathbb{R}^{n \times n}$ is diagonal if $\mathbf{A}_{i,j} = 0$ for all $i \neq j$
- The identity matrix $\mathbf{I}_n$ is the $n \times n$ diagonal matrix s.t. $(\mathbf{I}_n)_{i,i} = 1$

Operations on Matrices: Addition and Multiplication

- Addition: $\mathbf{C} = \mathbf{A} + \mathbf{B}$ s.t. $\mathbf{C}_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$
  - associative: $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
  - commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Scalar multiplication: $\mathbf{B} = r\,\mathbf{A}$ s.t. $\mathbf{B}_{i,j} = r\,\mathbf{A}_{i,j}$
- Multiplication: $\underbrace{\mathbf{C}}_{m \times p} = \underbrace{\mathbf{A}}_{m \times n}\,\underbrace{\mathbf{B}}_{n \times p}$ s.t. $\mathbf{C}_{i,j} = \sum_{k \in [n]} \mathbf{A}_{i,k}\,\mathbf{B}_{k,j}$
  - associative: $\mathbf{A}\,(\mathbf{B}\,\mathbf{C}) = (\mathbf{A}\,\mathbf{B})\,\mathbf{C}$
  - not commutative in general: $\mathbf{A}\,\mathbf{B} \neq \mathbf{B}\,\mathbf{A}$
  - distributive wrt. addition: $\mathbf{A}\,(\mathbf{B} + \mathbf{C}) = \mathbf{A}\,\mathbf{B} + \mathbf{A}\,\mathbf{C}$
  - $(\mathbf{A}\,\mathbf{B})^\intercal = \mathbf{B}^\intercal\,\mathbf{A}^\intercal$

$$\begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad i \in [n] \equiv 1 \le i \le n.$$

$$\begin{bmatrix} \boxed{A_{1,1} \;\cdots\; A_{1,n}} \\ \vdots \\ \boxed{A_{m,1} \;\cdots\; A_{m,n}} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \overset{a_1}{A_{1,1}} \cdot x_1 + \cdots + \overset{a_n}{A_{1,n}} \cdot x_n \\ \vdots \\ A_{m,1} \cdot x_1 + \cdots + A_{m,n} \cdot x_n \end{bmatrix} = \sum_{i \in [n]} x_i \cdot a_i$$

quadratic form: $\underbrace{\underbrace{x^T}_{1 \times n}\,\underbrace{\underbrace{A}_{n \times n}\,\underbrace{x}_{n \times 1}}_{n \times 1}}_{1 \times 1}.$

$$\begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix} \begin{bmatrix} \sum_{i \in [n]} A_{1,i} \cdot x_i \\ \vdots \\ \sum_{i \in [n]} A_{m,i} \cdot x_i \end{bmatrix} = x_1 \cdot \sum_{i \in [n]} A_{1,i} \cdot x_i + \cdots + x_n \cdot \sum_{i \in [n]} A_{m,i} \cdot x_i$$

$$= \sum_{j \in [n]} \sum_{i \in [n]} A_{j,i} \cdot x_i \cdot x_j.$$

Operations on Matrices: Inversion

Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ invertible if there is $\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$ s.t. $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\,\mathbf{A} = \mathbf{I}_n$.

- $\mathbf{A}$ is invertible if and only if rows of $\mathbf{A}$ are linearly independent
- $\mathbf{A}$ invertible if and only if $\det(\mathbf{A}) = |\mathbf{A}| \neq 0$
- If $\mathbf{A}$ invertible then $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ has solution $\mathbf{x} = \mathbf{A}^{-1}\,\mathbf{b}$

Examples on how to compute the determinant of a square matrix:

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}$$

$$\begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} = a_{1,1} \begin{vmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{vmatrix} - a_{1,2} \begin{vmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{vmatrix} + a_{1,3} \begin{vmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,2} \end{vmatrix}$$

Properties of determinants: $\det(\mathbf{A}^\intercal) = \det(\mathbf{A})$ $\qquad \det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$

Eigenvectors and Eigenvalues

- $\mathbf{v} \in \mathbb{R}^n$ is an eigenvector of $\mathbf{A} \in \mathbb{R}^{n \times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $\mathbf{A}\,\mathbf{v} = \lambda\,\mathbf{v}$
- If $\mathbf{A} \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, then the determinant of $\mathbf{A}$ is

$$\det(\mathbf{A}) = |\mathbf{A}| = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$$

A recipe to compute the eigenvalues and eigenvectors of $\mathbf{A}$:

- Compute the determinant of $\mathbf{A} - \lambda\mathbf{I}_n$. With λ subtracted along the diagonal, this determinant is a polynomial of degree n. It starts with $(-\lambda)^n$.
- Find the roots of this polynomial. The n roots are the eigenvalues of $\mathbf{A}$.
- For each eigenvalue λ solve the the equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$. The solution $\mathbf{x} \neq \mathbf{0}$ is the eigenvector corresponding to λ.

A has eigenvector x and corresp. eigenvalue λ. : $Ax = \lambda x$

Q1: $(A + \mu I)x = \underbrace{Ax}_{\lambda x} + \mu I x = \lambda \cdot x + \mu x$ $\quad \mu \in \mathbb{R}$.

$= (\lambda + \mu) \cdot x$

Q2: if A is invertible, then x is an eigenvector of A with λ^{-1}.

$A^{-1} A = I \Leftrightarrow A^{-1} \underbrace{Ax}_{\lambda x} = \underbrace{Ix}_{x} \Leftrightarrow A^{-1} \lambda x = x$

$\lambda \neq 0$. $\quad A^{-1} x = \lambda^{-1} x$

Q3: $A^k x$ $\quad k \in \mathbb{Z}$. $\quad \boxed{A^k x = \lambda^k x}$

$k > 0$: $\quad A^k x = A^{k-1} \underbrace{Ax}_{\lambda x} = \lambda A^{k-1} x = \dots = \lambda^k x$

$k < 0$

$k = 0$: $\quad A^0 x = x = \underbrace{\lambda^0}_{1} \cdot x$.

Positive (Semi-)Definite Matrices

A symmetric matrix $\mathbf{A}$ is:

- positive semi-definite (PSD) if for all $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$.
- positive definite (PD) if for all non-zero $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$.

Properties:

- PD $\equiv$ all eigenvalues are strictly positive
  $\Rightarrow$ non-zero determinant $\Rightarrow$ invertible
- PSD $\equiv$ all eigenvalues are nonnegative

Exercises:

1. If $\mathbf{A}$ is PSD then $(\mathbf{A} + \epsilon\,\mathbf{I})$ is PD
2. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, is $\mathbf{A}^\top \mathbf{A}$ PSD?

$A \in \mathbb{R}^{m \times n}$, is $A^T A$ PSD?

Def: $\forall x \in \mathbb{R}^n$: $x^T \underbrace{(A^T A)} \cdot x \geq 0$.

$\underbrace{x^T A^T} A x = (Ax)^T Ax = \langle Ax, Ax \rangle = \| Ax \|_2^2 \geq 0$.

if A is PSD $\Rightarrow$ is $(A + \varepsilon I)$ PD? for $\varepsilon > 0$.

$x \neq 0$: $\quad x^T (A + \varepsilon I) x = \underbrace{x^T A x}_{\geq 0} + \underbrace{x^T \varepsilon I x}_{\varepsilon \cdot x^T x}$

$\varepsilon > 0$, $\langle x, x \rangle = \|x\|_2^2 > 0$

> 0

> 0.

Calculus

Minimising Objective Functions

Function $f : \mathbb{R}^D \to \mathbb{R}$

Extrema

- $\mathbf{x}$ is local minimum for f if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y}$ in some neighbourhood of $\mathbf{x}$
- $\mathbf{x}$ is global minimum for f if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y}$

How to find extrema?

- First and second order derivative tests

Maximising f is the same as minimising $-f \Rightarrow$ OK to focus on minimisation

Continuous and Differentiable Functions of One Variable

Functions of one variable $f : \mathbb{R} \to \mathbb{R}$

- f is differentiable at x_0 if

$$f'(x_0) = \frac{d}{dx} f(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \quad \text{exists}$$

- Differentiation rules:

$$\frac{d}{dx} x^n = n \cdot x^{n-1} \qquad \frac{d}{dx} a^x = a^x \cdot \ln(a) \qquad \frac{d}{dx} \log_a(x) = \frac{1}{x \cdot \ln(a)}$$

$$(f + g)' = f' + g' \qquad (f \cdot g)' = f' \cdot g + f \cdot g'$$

- Chain rule: if $f = h(g)$ then $f' = h'(g) \cdot g'$

1. $f(x) = |x|$. Q: is f diff. at 0?

$$f'(0) = \lim_{h \to 0} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0} \frac{f(h)}{h} = \lim_{h \to 0} \frac{|h|}{h}.$$

$$\lim_{h \to 0^+} \frac{|h|}{h} = 1 \neq \lim_{h \to 0^-} \frac{|h|}{h} = -1.$$

2. $f(x) = \max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0. \end{cases}$

$$f'(0) = \lim_{h \to 0} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0} \frac{f(h)}{h}.$$

$$\lim_{h \to 0^-} \frac{f(h)}{h} = 0 . \neq \lim_{h \to 0^+} \frac{f(h)}{h} = \frac{h}{h} = 1$$

3. $f(x) = [\max(0, 1-x)]^2$. TODO.

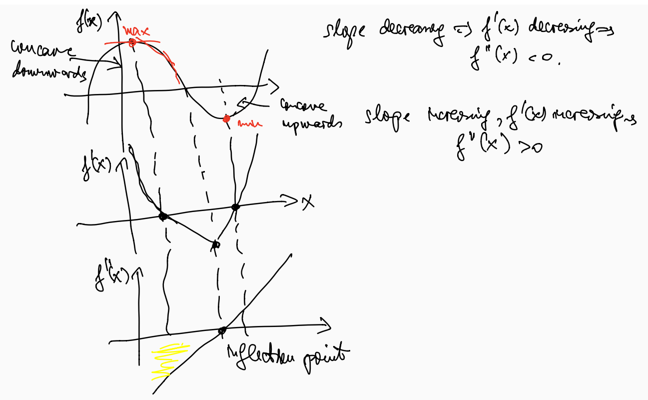
Testing for Extrema

First derivative test:

- $f'(x^*) = 0$ means that x^* is a critical or stationary point for f
  - Can be a local minimum, a local maximum, or a saddle point

Second derivative test to (partially) decide nature of critical point:

- $f'(x^*) = 0$ and $f''(x^*) > 0$ means that f has local minimum at x^*
- $f'(x^*) = 0$ and $f''(x^*) < 0$ means that f has local maximum at x^*
- $f'(x^*) = f''(x^*) = 0$ and $f'''(x^*) \neq 0$ means that f has a saddle point at x^*
- Otherwise, higher order derivative tests necessary



Functions of Multiple Variables

Functions of multiple variables $f : \mathbb{R}^m \to \mathbb{R}$

- Partial derivative of $f(x_1, \ldots, x_m)$ in direction x_i at $\mathbf{a} = (a_1, \ldots, a_m)$:

$$\frac{\partial}{\partial x_i} f(\mathbf{a}) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_m) - f(a_1, \ldots, a_i, \ldots, a_m)}{h}$$

- Gradient (assuming f is differentiable everywhere):

$$\nabla_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_m} \end{bmatrix} \qquad \text{This means: } [\nabla_{\mathbf{x}} f]_i = \frac{\partial f}{\partial x_i}$$

  - $\nabla_{\mathbf{x}} f$ points in direction of steepest ascent
    $\Rightarrow -\nabla_{\mathbf{x}} f$ points in direction of steepest descent
  - Critical point if $\nabla_{\mathbf{x}} f(\mathbf{a}) = \mathbf{0}$

Functions of Multiple Variables

Functions of multiple variables $f : \mathbb{R}^m \to \mathbb{R}$

- Hessian $\nabla^2_{\mathbf{x}} f$ is a matrix of second-order partial derivatives

$$\nabla^2_{\mathbf{x}} f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix} \qquad \text{This means: } [\nabla^2_{\mathbf{x}} f]_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

- If the partial derivatives are continuous, the order of differentiation does not matter $\Rightarrow$ the Hessian matrix is symmetric

Functions of Multiple Variables

Functions of multiple variables to vectors $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$:

- $\mathbf{f}$ given as $\mathbf{f} = (f_1, \ldots, f_n)$ with $f_i : \mathbb{R}^m \to \mathbb{R}$
- Jacobian $\mathbf{J}$ of $\mathbf{f}$ is an $n \times m$ matrix:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_m} \end{bmatrix} \qquad \text{This means: } [\mathbf{J}_f]_{i,j} = \frac{\partial f_i}{\partial x_j}$$

Matrix Calculus: Useful Differentiation Rules

$$\nabla_{\mathbf{x}}(\mathbf{c}^{\top}\mathbf{x}) = \mathbf{c}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^{\top}\mathbf{x}) = 2\mathbf{x}$$

$$\nabla_{\mathbf{x}}(\mathbf{A}\,\mathbf{x}) = \mathbf{A}^{\top}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^{\top}\mathbf{A}\,\mathbf{x}) = \mathbf{A}\,\mathbf{x} + \mathbf{A}^{\top}\mathbf{x} \qquad (= 2\mathbf{A}\,\mathbf{x} \text{ for symmetric } \mathbf{A})$$

$$\nabla_{\mathbf{x}}(f + g) = \nabla_{\mathbf{x}}f + \nabla_{\mathbf{x}}g$$

$$\nabla_{\mathbf{x}}(f\,g) = f\,\nabla_{\mathbf{x}}g + g\,\nabla_{\mathbf{x}}f$$

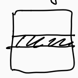
See `http://en.wikipedia.org/wiki/Matrix_calculus` for many more useful rules, and use them!

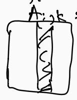
1. $c^T x = \underline{\sum_{i\in[n]} c_i x_i}$ $\qquad \frac{\partial(c^T x)}{\partial x_i} = c_i, \quad \nabla_x(c^T x) = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = c.$

2. $\boxed{\nabla_x(x^T A x) = A x + A^T x.}$ $\quad = 2Ax$ if A is symmetric,

$x^T A x = \sum_{i\in[n]} \sum_{j\in[n]} x_i \cdot x_j \cdot A_{i,j}.$

$\frac{\partial(x^T A x)}{\partial x_k} = \sum_{j\in[n]} x_j A_{k,j} + \sum_{i\in[n]} x_i A_{i,k} = A_{k,:}x + A_{:,k}^T x$

$A_{k,:}$ 

$A_{:,k} = a_k$ 

Chain Rule in Higher Dimensions

Let $\mathbf{y} = g(\mathbf{x})$, $z = f(\mathbf{y})$ for $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$:

$$\frac{\partial z}{\partial x_i} = \sum_{j\in[n]} \frac{\partial z}{\partial y_j} \cdot \frac{\partial y_j}{\partial x_i}$$

$$\nabla_{\mathbf{x}} z = \mathbf{J}_g \cdot \nabla_{\mathbf{y}} z = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \cdot \nabla_{\mathbf{y}} z$$

Let $g(x, y) = (x^2, y)$, $f(s, t) = (s + t)^2$ and $z = f(g(x, y))$. Then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial s} \cdot \frac{\partial s}{\partial x} + \frac{\partial z}{\partial t} \cdot \frac{\partial t}{\partial x} = 2 \cdot (x^2 + y) \cdot 1 \cdot 2 \cdot x + 2 \cdot (x^2 + y) \cdot 1 \cdot 0 = 4x(x^2 + y)$$

$$\mathbf{J}_g = \begin{bmatrix} 2 \cdot x & 0 \\ 0 & 1 \end{bmatrix}$$

$$\nabla_{\mathbf{y}} z = (2 \cdot (x^2 + y), 2 \cdot (x^2 + y))$$

$$\nabla_{\mathbf{x}} z = (4 \cdot x \cdot (x^2 + y), 2 \cdot (x^2 + y))$$

Optima with Side Conditions: Lagrange Multipliers

We will often encounter constrained optimisation problems:

$$\begin{aligned} &\text{maximise} \quad f(\mathbf{x}) \\ &\text{subject to} \quad g_i(\mathbf{x}) = 0 \qquad \text{for all } i \in [n] \end{aligned}$$

- Optimal points of f lie tangential to the g_i
- For $n = 1$, optimum should fulfil:

$$\nabla_{\mathbf{x}} f = \lambda \nabla_{\mathbf{x}} g$$

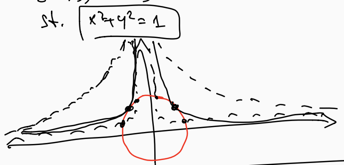
- Optimum of the original optimisation problem will be critical point of the Lagrangian:

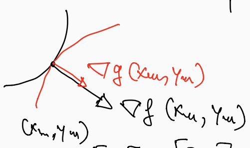
$$\Lambda(\mathbf{x}, \lambda) := f(\mathbf{x}) - \lambda \cdot g(\mathbf{x})$$

- Generalises to any $n > 0$ and inequality constraints

max $f(x, y) = x^2 y$

st. $\boxed{x^2 + y^2 = 1}$





$\boxed{\lambda \cdot \nabla g(x_m, y_m) = \nabla f(x_m, y_m)}$

$\nabla g = \nabla(x^2 + y^2) = \begin{bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$

$\nabla f = \nabla(x^2 y) = \begin{bmatrix} 2xy \\ x^2 \end{bmatrix}$

$\lambda \cdot \begin{bmatrix} 2x \\ 2y \end{bmatrix} = \begin{bmatrix} 2xy \\ x^2 \end{bmatrix}$

$\begin{cases} 2x\lambda = 2xy \\ 2y\lambda = x^2 \\ x^2 + y^2 = 1. \end{cases} \Rightarrow$ FIND x, y, λ.

Probability theory

Probability Space

- Consists of sample space S and a probability function $p : \mathcal{P}(S) \to [0, 1]$ assigning a probability to every event
- Satisfies axioms of probability:
  - $p(\emptyset) = 0$ and $p(S) = 1$
  - For mutually exclusive events $A_1, A_2, \ldots$

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

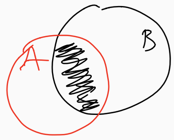
Trivial properties:

- $p(\overline{A}) = 1 - p(A)$
- If $A \subseteq B$ then $p(A) \leq p(B)$
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Conditional Probability

Given events A, B with $p(B) > 0$, conditional probability of A given B is

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

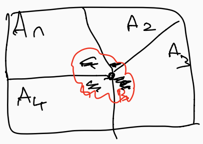


$$p(A \cap B) = p(A|B) \cdot p(B)$$
$$= p(B|A) \cdot p(A).$$

Conditional Probability

Law of total probability: Given partition $A_1, \ldots, A_n$ of S with $p(A_i) > 0$,

$$p(B) = \sum_{i=1}^{n} p(B|A_i) \cdot p(A_i)$$



$$p(B) = p(B|A_n) \cdot p(A_n) + p(B|A_2) \cdot p(A_2) + \ldots$$
$$p(B|A_4) \cdot p(A_4).$$

Conditional Probability

Bayes' rule:

likelihood

prior

$$\underbrace{p(A|B)}_{\text{posterior}} = \frac{p(B|A) \cdot p(A)}{\boxed{p(B)}}$$

$$p(A \cap B) = p(A|B) \cdot \boxed{p(B)} = p(B|A) \cdot p(A)$$

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}.$$

Random Variables

- Function from sample space to some numeric domain (usually $\mathbb{R}$)
- $p(X = x)$ denotes probability of event $\{s \in S : X(s) = x\}$
- Write $X \sim p(x)$ to specify probability distribution of X

Discrete random variables:

- Discrete if there are countably many $a_1, a_2, \ldots$ such that $\sum_{a_j} p(X = a_j) = 1$
- Probability mass function (PMF) p_X giving distribution of X

$$p_X(x) = p(X = x)$$

- Cumulative distribution function (CDF) maps x to $p(X \leq x)$

Continuous random variables:

- Probability density function (PDF) $p(x)$ is derivative of CDF giving distribution of X

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad P(a \leq X \leq b) = \int_a^b p(x)dx$$

Joint Probability Distributions

- Natural generalisation to vectors of random variables giving joint probability distributions, e.g., $p(X = x, Y = y)$
- Marginal probability distribution: Given $p(X, Y)$, obtain $p(X)$ via

$$p(X = x) = \sum_y p(X = x, Y = y) \qquad \text{resp.} \qquad p(x) = \int p(x, y)\, dy$$

- Conditional probabilities: Assuming $p(X = x) > 0$,

$$p(Y = y \mid X = x) = \frac{p(Y = y, X = x)}{p(X = x)}$$

- Chain rule of conditional probability:

$$p(X^{(1)}, \ldots, X^{(n)}) = p(X^{(1)}) \cdot \prod_{i=2}^{n} p(X^{(i)} \mid X^{(1)}, \ldots, X^{(i-1)})$$

Expectation and Variance

Expected value of random variable

- Discrete random variables: $\mathbb{E}[X] = \sum_{x \in \mathrm{dom}(X)} x \cdot p(x)$
- Continuous random variables: $\mathbb{E}[X] = \int x \cdot p(x)\, dx$
- Linearity of expectation:

$$\mathbb{E}[\alpha \cdot X + \beta \cdot Y] = \alpha \cdot \mathbb{E}[X] + \beta \cdot \mathbb{E}[Y]$$

Variance of a random variable

- Captures how much values of probability distribution vary on average if randomly drawn:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Properties of variance:
  - $Var(\alpha \cdot X + \beta) = \alpha^2 \cdot Var(X)$
  - If X and Y are independent: $Var(X + Y) = Var(X) + Var(Y)$

Standard Deviation and Covariance

- Standard deviation is square root of variance

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

- Covariance generalises variance to two random variables

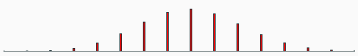
$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- Covariance matrix $\mathbf{\Sigma}$ generalises covariance to multiple random variables X_i

$$\mathbf{\Sigma}_{i,j} = \mathrm{Cov}(X_i, X_j)$$

Well-known Discrete Probability Distributions

- Bernoulli:
  - Parameter: $\phi \in [0, 1]$
  - PMF: $p(X = 1) = \phi$, $p(X = 0) = 1 - \phi$;
  - $\mathbb{E}[X] = \phi$; $\mathrm{Var}(X) = \phi \cdot (1 - \phi)$
- Binomial distribution:
  - Parameters: $\phi \in [0, 1]$, $n \in \mathbb{N} \setminus \{0\}$
  - PMF: $p(X = k) = \binom{n}{k} \cdot \phi^k \cdot (1 - \phi)^{n-k}$
  - $\mathbb{E}[X] = n \cdot \phi$; $\mathrm{Var}(X) = n \cdot \phi \cdot (1 - \phi)$

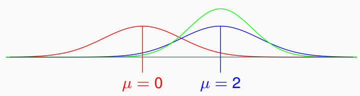


Well-known Continuous Probability Distributions

- Normal (Gaussian) distribution:
  - Parameters: μ, σ^2
  - PDF:

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

  - $\mathbb{E}[X] = \mu$; $\mathrm{Var}(X) = \sigma^2$

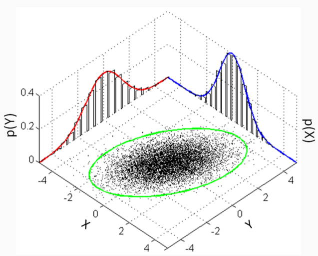


Well-known Continuous Probability Distributions

- Multivariate normal (Gaussian) distribution:
  - Parameters: $k, \mu, \mathbf{\Sigma}$ positive semi-definite
  - PDF:

$$\mathcal{N}(\mathbf{x}; \mu, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^k \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$

  - $\mathbb{E}[\mathbf{X}] = \mu$; $\mathrm{Var}(\mathbf{X}) = \mathbf{\Sigma}$



Well-known Continuous Probability Distributions

- Laplace distribution:
  - Parameters: μ (location) γ^2 (scale)
  - PDF:

$$\mathrm{Lap}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right)$$

  - $\mathbb{E}[X] = \mu$; $\mathrm{Var}(X) = 2\gamma^2$