Cookbook for data scientists

Charles Deledalle

Linear algebra I

Notations

$x, y, z, \ldots$: vectors of $\mathbb{C}^n$
$a, b, c, \ldots$: scalars of $\mathbb{C}$
A, B, C : matrices of $\mathbb{C}^{m \times n}$
Id : identity matrix
$i = 1, \ldots, m$ and $j = 1, \ldots, n$

Matrix vector product

$$(Ax)_i = \sum_{k=1}^n A_{i,k} x_k$$

$$(AB)_{i,j} = \sum_{k=1}^n A_{i,k} B_{k,j}$$

Basic properties

$$A(ax + by) = aAx + bAy$$

$$A\mathrm{Id} = \mathrm{Id}A = A$$

Inverse $(m = n)$

A is said invertible, if it exists B st

$$AB = BA = \mathrm{Id}.$$

B is unique and called inverse of A.
We write $B = A^{-1}$.

Adjoint and transpose

$$(A^t)_{j,i} = A_{i,j}, \quad A^t \in \mathbb{C}^{m \times n}$$

$$(A^*)_{j,i} = (A_{i,j})^*, \quad A^* \in \mathbb{C}^{m \times n}$$

$$\langle Ax,\, y \rangle = \langle x,\, A^* y \rangle$$

Trace and determinant $(m = n)$

$$\mathrm{tr}\, A = \sum_{i=1}^n A_{i,i} = \sum_{i=1}^n \lambda_i$$

$$\mathrm{tr}\, A = \mathrm{tr}\, A^*$$

$$\mathrm{tr}\, AB = \mathrm{tr}\, BA$$

$$\det A^* = \det A$$

$$\det A = \prod_{i=1}^n \lambda_i$$

$$\det A^{-1} = (\det A)^{-1}$$

$$\det AB = \det A \det B$$

A is invertible $\Leftrightarrow \det A \neq 0 \Leftrightarrow \lambda_i \neq 0, \forall i$

Scalar products, angles and norms

$$\langle x,\, y \rangle = x \cdot y = x^* y = \sum_{k=1}^n x_k y_k \quad \text{(dot product)}$$

$$\|x\|^2 = \langle x,\, x \rangle = \sum_{k=1}^n x_k^2 \quad (\ell_2 \text{ norm})$$

$$|\langle x,\, y \rangle| \leqslant \|x\| \|y\| \quad \text{(Cauchy-Schwartz inequality)}$$

$$\cos(\angle(x, y)) = \frac{\langle x,\, y \rangle}{\|x\| \|y\|} \quad \text{(angle and cosine)}$$

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x,\, y \rangle \quad \text{(law of cosines)}$$

$$\|x\|_p^p = \sum_{k=1}^n |x_k|^p, \quad p \geqslant 1 \quad (\ell_p \text{ norm})$$

$$\|x + y\|_p \leqslant \|x\|_p + \|y\|_p \quad \text{(triangular inequality)}$$

Orthogonality, vector space, basis, dimension

$$x \perp y \Leftrightarrow \langle x,\, y \rangle = 0 \quad \text{(Orthogonality)}$$

$$x \perp y \Leftrightarrow \|x + y\|^2 = \|x\|^2 + \|y\|^2 \quad \text{(Pythagorean)}$$

Let d vectors x_i be st $x_i \perp x_j$, $\|x_i\| = 1$. Define

$$V = \mathrm{Span}(\{x_i\}) = \left\{ y \setminus \exists \alpha \in \mathbb{C}^d, y = \sum_{i=1}^d \alpha_i x_i \right\}$$

V is a vector space, $\{x_i\}$ is an orthonormal basis of V and

$$\forall y \in V, \quad y = \sum_{i=1}^d \langle y,\, x_i \rangle x_i$$

and $d = \dim V$ is called the dimensionality of V. We have

$$\dim(V \cup W) = \dim V + \dim W - \dim(V \cap W)$$

Column/Range/Image and Kernel/Null spaces

$$\mathrm{Im}[A] = \{ y \in \mathbb{R}^m \setminus \exists x \in \mathbb{R}^n \text{ such that } y = Ax \} \quad \text{(image)}$$

$$\mathrm{Ker}[A] = \{ x \in \mathbb{R}^n \setminus Ax = 0 \} \quad \text{(kernel)}$$

$\mathrm{Im}[A]$ and $\mathrm{Ker}[A]$ are vector spaces satisfying

$$\mathrm{Im}[A] = \mathrm{Ker}[A^*]^\perp \quad \text{and} \quad \mathrm{Ker}[A] = \mathrm{Im}[A^*]^\perp$$

$$\mathrm{rank}\, A + \dim(\mathrm{Ker}[A]) = n \quad \text{(rank-nullity theorem)}$$

$$\text{where} \quad \mathrm{rank}\, A = \dim(\mathrm{Im}[A]) \quad \text{(matrix rank)}$$

Note also

$$\mathrm{rank}\, A = \mathrm{rank}\, A^*$$

$$\mathrm{rank}\, A + \dim(\mathrm{Ker}[A^*]) = m$$

Linear algebra II

Eigenvalues / eigenvectors

If $\lambda \in \mathbb{C}$ and $e \in \mathbb{C}^n (\neq 0)$ satisfy

$$Ae = \lambda e$$

λ is called the eigenvalue associated to the eigenvector e of A. There are at most n distinct eigenvalues λ_i and at least n linearly independent eigenvectors e_i (with norm 1). The set λ_i of n (non necessarily distinct) eigenvalues is called the spectrum of A (for a proper definition see characteristic polynomial, multiplicity, eigenspace). This set has exactly $r = \operatorname{rank} A$ non zero values.

Eigendecomposition ($m = n$)

If it exists $E \in \mathbb{C}^{n \times n}$, and a diagonal matrix $\Lambda \in \mathbb{C}^{n \times n}$ st

$$A = E \Lambda E^{-1}$$

A is said diagonalizable and the columns of E are the n eigenvectors e_i of A with corresponding eigenvalues $\Lambda_{i,i} = \lambda_i$.

Properties of eigendecomposition ($m = n$)

- If, for all i, $\Lambda_{i,i} \neq 0$, then A is invertible and

$$A^{-1} = E \Lambda^{-1} E^{-1} \quad \text{with} \quad \Lambda^{-1}_{i,i} = (\Lambda_{i,i})^{-1}$$

- If A is Hermitian ($A = A^*$), such decomposition always exists, the eigenvectors of E can be chosen orthonormal such that E is unitary ($E^{-1} = E^*$), and λ_i are real.
- If A is Hermitian ($A = A^*$) and $\lambda_i > 0$, A is said positive definite, and for all $x \neq 0$, $xAx^* > 0$.

Singular value decomposition (SVD)

For **all** matrices A there exists two unitary matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$, and a real non-negative diagonal matrix $\Sigma \in \mathbb{R}^{m \times n}$ st

$$A = U \Sigma V^* \quad \text{and} \quad A = \sum_{k=1}^{r} \sigma_k u_k v_k^*$$

with $r = \operatorname{rank} A$ non zero singular values $\Sigma_{k,k} = \sigma_k$.

Eigendecomposition and SVD

- If A is Hermitian, the two decompositions coincide with $V = U = E$ and $\Sigma = \Lambda$.
- Let $A = U \Sigma V^*$ be the SVD of A, then the eigendecomposition of AA^* is $E = U$ and $\Lambda = \Sigma^2$.

SVD, image and kernel

Let $A = U \Sigma V^*$ be the SVD of A, and assume $\Sigma_{i,i}$ are ordered in decreasing order then

$$\operatorname{Im}[A] = \operatorname{Span}(\{u_i \in \mathbb{R}^m \setminus i \in (1 \ldots r)\})$$
$$\operatorname{Ker}[A] = \operatorname{Span}(\{v_i \in \mathbb{R}^n \setminus i \in (r+1 \ldots n)\})$$

Moore-Penrose pseudo-inverse

The Moore-Penrose pseudo-inverse reads

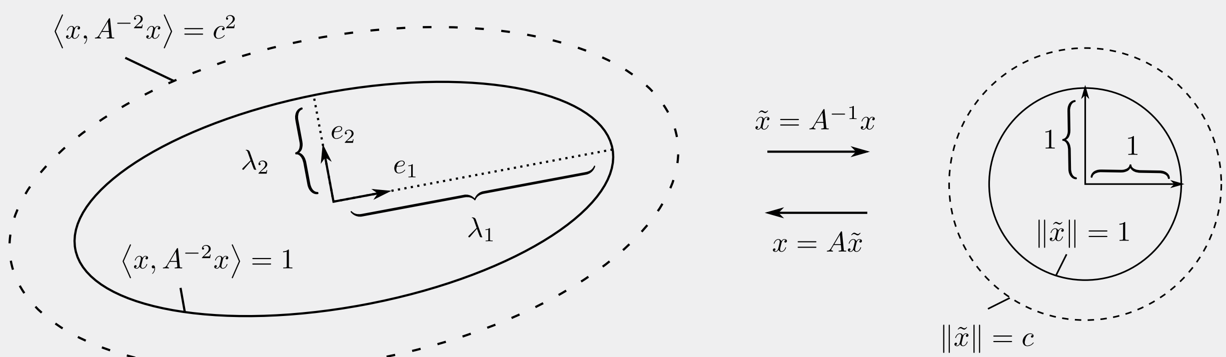
$$A^+ = V \Sigma^+ U^* \quad \text{with} \quad \Sigma^+_{i,i} = \begin{cases} (\Sigma_{i,i})^{-1} & \text{if } \Sigma_{ii} > 0, \\ 0 & \text{otherwise} \end{cases}$$

and is the unique matrix satisfying $A^+ A A^+ = A^+$ and $AA^+A = A$ with A^+A and AA^+ Hermitian.
If A is invertible, $A^+ = A^{-1}$.

Matrix norms

$$\|A\|_p = \sup_{x; \|x\|_p = 1} \|Ax\|_p, \; \|A\|_2 = \max_k \sigma_k, \; \|A\|_* = \sum_k \sigma_k,$$

$$\|A\|_F^2 = \sum_{i,j} |a_{i,j}|^2 = \operatorname{tr} A^* A = \sum_k \sigma_k^2$$



Cookbook for data scientists

Charles Deledalle

Fourier analysis

Fourier Transform (FT)

Let $x : \mathbb{R} \to \mathbb{C}$ such that $\int_{-\infty}^{+\infty} |x(t)| \, \mathrm{d}t < \infty$. Its Fourier transform $X : \mathbb{R} \to \mathbb{C}$ is defined as

$$X(u) = \mathcal{F}[x](u) = \int_{-\infty}^{+\infty} x(t) e^{-i2\pi ut} \, \mathrm{d}t$$

$$x(t) = \mathcal{F}^{-1}[X](t) = \int_{-\infty}^{+\infty} X(u) e^{i2\pi ut} \, \mathrm{d}u$$

where u is referred to as the frequency.

Properties of continuous FT

$$\mathcal{F}[ax + by] = a\mathcal{F}[x] + b\mathcal{F}[y] \quad \text{(Linearity)}$$

$$\mathcal{F}[x(t - a)] = e^{-i2\pi au}\mathcal{F}[x] \quad \text{(Shift)}$$

$$\mathcal{F}[x(at)](u) = \frac{1}{|a|}\mathcal{F}[x](u/a) \quad \text{(Modulation)}$$

$$\mathcal{F}[x^*](u) = \mathcal{F}[x](-u)^* \quad \text{(Conjugation)}$$

$$\mathcal{F}[x](0) = \int_{-\infty}^{+\infty} x(t) \, \mathrm{d}t \quad \text{(Integration)}$$

$$\int_{-\infty}^{+\infty} |x(t)|^2 \, \mathrm{d}t = \int_{-\infty}^{+\infty} |X(u)|^2 \, \mathrm{d}u \quad \text{(Parseval)}$$

$$\mathcal{F}[x^{(n)}](u) = (2\pi iu)^n \mathcal{F}[x](u) \quad \text{(Derivation)}$$

$$\mathcal{F}[e^{-\pi^2 a t^2}](u) = \frac{1}{\sqrt{\pi a}} e^{-u^2/a} \quad \text{(Gaussian)}$$

$$x \text{ is real} \Leftrightarrow X(\varepsilon) = X(-\varepsilon)^* \quad \text{(Real} \leftrightarrow \text{Hermitian)}$$

Properties with convolutions

$$(x \star y)(t) = \int_{-\infty}^{\infty} x(s) y(t - s) \, \mathrm{d}s \quad \text{(Convolution)}$$

$$\mathcal{F}[x \star y] = \mathcal{F}[x]\mathcal{F}[y] \quad \text{(Convolution theorem)}$$

Multidimensional Fourier Transform

Fourier transform is separable over the different d dimensions, hence can be defined recursively as

$$\mathcal{F}[x] = (\mathcal{F}_1 \circ \mathcal{F}_2 \circ \ldots \circ \mathcal{F}_d)[x]$$

$$\text{where} \quad \mathcal{F}_k[x](t_1 \ldots, \varepsilon_k, \ldots, t_d) = \mathcal{F}[t_k \mapsto x(t_1, \ldots, t_k, \ldots, t_d)](\varepsilon_k)$$

and inherits from above properties (same for DFT).

Discrete Fourier Transform (DFT)

$$X_u = \mathcal{F}[x]_u = \sum_{t=0}^{n-1} x_t e^{-i2\pi ut/n}$$

$$x_t = \mathcal{F}^{-1}[X]_t = \frac{1}{n}\sum_{u=0}^{n-1} X_k e^{i2\pi ut/n}$$

Or in a matrix-vector form $X = Fx$ and $x = F^{-1}X$ where $F_{u,k} = e^{-i2\pi uk/n}$. We have

$$F^* = nF^{-1} \quad \text{and} \quad U = n^{-1/2}F \quad \text{is unitary}$$

Properties of discrete FT

$$\mathcal{F}[ax + by] = a\mathcal{F}[x] + b\mathcal{F}[y] \quad \text{(Linearity)}$$

$$\mathcal{F}[x_{t-a}] = e^{-i2\pi au/n}\mathcal{F}[x] \quad \text{(Shift)}$$

$$\mathcal{F}[x^*]_u = \mathcal{F}[x]^*_{n-u \bmod n} \quad \text{(Conjugation)}$$

$$\mathcal{F}[x]_0 = \sum_{t=0}^{n-1} x_t \quad \text{(Integration)}$$

$$\|x\|_2^2 = \frac{1}{n}\|X\|_2^2 \quad \text{(Parseval)}$$

$$\|x\|_1 \leqslant \|X\|_1 \leqslant n\|x\|_1$$

$$\|X\|_\infty \leqslant \|x\|_1 \quad \text{and} \quad \|x\|_\infty \leqslant \frac{1}{n}\|X\|_1$$

$$x \text{ is real} \Leftrightarrow X_u = X^*_{n-u \bmod n} \quad \text{(Real} \leftrightarrow \text{Hermitian)}$$

Discrete circular convolution

$$(x * y)_t = \sum_{s=1}^{n} x_s y_{(t-s \bmod n)+1} \quad \text{or} \quad x * y = \Phi_y x$$

where $(\Phi_y)_{t,s} = y_{(t-s \bmod n)+1}$ is a circulant matrix diagonalizable in the discrete Fourier basis, thus

$$\mathcal{F}[x * y]_u = \mathcal{F}[x]_u \mathcal{F}[y]_u$$

Fast Fourier Transform (FFT)

The matrix-by-vector product Fx can be computed in $\mathcal{O}(n \log n)$ operations (much faster than the general matrix-by-vector product that required $\mathcal{O}(n^2)$ operations). Same for F^{-1} and same for multi-dimensional signals.

Probability and Statistics

Kolmogorov's probability axioms

Let Ω be a sample set and A an event

$$\mathbb{P}[\Omega] = 1, \quad \mathbb{P}[A] \geqslant 0$$

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbb{P}[A_i] \quad \text{with} \quad A_i \cap A_j = \emptyset$$

Basic properties

$$\mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[A] \in [0,1], \quad \mathbb{P}[A^c] = 1 - \mathbb{P}[A]$$

$$\mathbb{P}[A] \leqslant \mathbb{P}[B] \quad \text{if} \quad A \subseteq B$$

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Conditional probability

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \text{subject to} \quad \mathbb{P}[B] > 0$$

Bayes' rule

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

Independence

Let A and B be two events, X and Y be two rv

$$A \perp B \quad \text{if} \quad \mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

$$X \perp Y \quad \text{if} \quad (X \leqslant x) \perp (Y \leqslant y)$$

If X and Y admit a density, then

$$X \perp Y \quad \text{if} \quad f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Properties of Independence and uncorrelation

$$\mathbb{P}[A|B] = \mathbb{P}[A] \Rightarrow A \perp B$$

$$X \perp Y \Rightarrow (\mathbb{E}[XY^*] = \mathbb{E}[X]\mathbb{E}[Y^*] \Leftrightarrow \text{Cov}[X,Y] = 0)$$

Independence $\Rightarrow$ uncorrelation

correlation $\Rightarrow$ dependence

uncorrelation $\not\Rightarrow$ Independence

dependence $\not\Rightarrow$ correlation

Discrete random vectors

Let X be a discrete random vector defined on $\mathbb{N}^n$

$$\mathbb{E}[X]_i = \sum_{k=0}^{\infty} k \mathbb{P}[X_i = k]$$

The function $f_X : k \to \mathbb{P}[X = k]$ is called the probability mass function (pmf) of X.

Continuous random vectors

Let X be a continuous random vector on $\mathbb{C}^n$. Assume there exist f_X such that, for all $A \subseteq \mathbb{C}^n$,

$$\mathbb{P}[X \in A] = \int_A f_X(x) \, \mathrm{d}x.$$

Then f_X is called the probability density function (pdf) of X, and

$$\mathbb{E}[X] = \int_{\mathbb{C}^n} x f_X(x) \, \mathrm{d}x.$$

Variance / Covariance

Let X and Y be two random vectors. The covariance matrix between X and Y is defined as

$$\text{Cov}[X,Y] = \mathbb{E}[XY^*] - \mathbb{E}[X]\mathbb{E}[Y]^*.$$

X and Y are said uncorrelated if $\text{Cov}[X,Y] = 0$. The variance-covariance matrix is

$$\text{Var}[X] = \text{Cov}[X,X] = \mathbb{E}[XX^*] - \mathbb{E}[X]\mathbb{E}[X]^*.$$

Basic properties

- The expectation is linear

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

- If X and Y are independent

$$\text{Var}[aX + bY + c] = a^2\text{Var}[X] + b^2\text{Var}[Y]$$

- $\text{Var}[X]$ is always Hermitian positive definite

Cookbook for data scientists

Charles Deledalle

Multi-variate differential calculus

Partial and directional derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}^m$. The (i, j)-th partial derivative of f, if it exists, is

$$\frac{\partial f_i}{\partial x_j}(x) = \lim_{\varepsilon \to 0} \frac{f_i(x + \varepsilon e_j) - f_i(x)}{\varepsilon}$$

where $e_i \in \mathbb{R}^n$, $(e_j)_j = 1$ and $(e_j)_k = 0$ for $k \neq j$. The directional derivative in the dir. $d \in \mathbb{R}^n$ is

$$\mathcal{D}_d f(x) = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon d) - f(x)}{\varepsilon} \in \mathbb{R}^m$$

Jacobian and total derivative

$$J_f = \frac{\partial f}{\partial x} = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j} \qquad (m \times n \text{ Jacobian matrix})$$

$$\mathrm{d}f(x) = \mathrm{tr}\left[\frac{\partial f}{\partial x}(x)\, \mathrm{d}x\right] \qquad \text{(total derivative)}$$

Gradient, Hessian, divergence, Laplacian

$$\nabla f = \left(\frac{\partial f}{\partial x_i}\right)_i \qquad \text{(Gradient)}$$

$$H_f = \nabla\nabla f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j} \qquad \text{(Hessian)}$$

$$\mathrm{div}\, f = \nabla^t f = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} = \mathrm{tr}\, J_f \qquad \text{(Divergence)}$$

$$\Delta f = \mathrm{div}\, \nabla f = \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2} = \mathrm{tr}\, H_f \qquad \text{(Laplacian)}$$

Properties and generalizations

$$\nabla f = J_f^t \qquad \text{(Jacobian} \leftrightarrow \text{gradient)}$$

$$\mathrm{div} = -\nabla^* \qquad \text{(Integration by part)}$$

$$\mathrm{d}f(x) = \mathrm{tr}\left[J_f\, \mathrm{d}x\right] \qquad \text{(Jacob. character. I)}$$

$$\mathcal{D}_d f(x) = J_f(x) \times d \qquad \text{(II)}$$

$$f(x+h) = f(x) + \mathcal{D}_h f(x) + o(\|h\|) \qquad \text{(1st order exp.)}$$

$$f(x+h) = f(x) + \mathcal{D}_h f(x) + \tfrac{1}{2} h^* H_f(x) h + o(\|h\|^2)$$

$$\frac{\partial (f \circ g)}{\partial x} = \left(\frac{\partial f}{\partial x} \circ g\right) \frac{\partial g}{\partial x} \qquad \text{(Chain rule)}$$

Elementary calculation rules

$$\mathrm{d}A = 0$$

$$\mathrm{d}[aX + bY] = a\mathrm{d}X + b\mathrm{d}Y \qquad \text{(Linearity)}$$

$$\mathrm{d}[XY] = (\mathrm{d}X)Y + X(\mathrm{d}Y) \qquad \text{(Product rule)}$$

$$\mathrm{d}[X^*] = (\mathrm{d}X)^*$$

$$\mathrm{d}[X^{-1}] = -X^{-1}(\mathrm{d}X)X^{-1}$$

$$\mathrm{d}\,\mathrm{tr}[X] = \mathrm{tr}[\mathrm{d}X]$$

$$\frac{\mathrm{d}Z}{\mathrm{d}X} = \frac{\mathrm{d}Z}{\mathrm{d}Y}\frac{\mathrm{d}Y}{\mathrm{d}X} \qquad \text{(Leibniz's chain rule)}$$

Classical identities

$$\mathrm{d}\,\mathrm{tr}[AXB] = \mathrm{tr}[BA\, \mathrm{d}X]$$

$$\mathrm{d}\,\mathrm{tr}[X^*AX] = \mathrm{tr}[X^*(A^* + A)\, \mathrm{d}X]$$

$$\mathrm{d}\,\mathrm{tr}[X^{-1}A] = \mathrm{tr}[-X^{-1}AX^{-1}\, \mathrm{d}X]$$

$$\mathrm{d}\,\mathrm{tr}[X^n] = \mathrm{tr}[nX^{n-1}\, \mathrm{d}X]$$

$$\mathrm{d}\,\mathrm{tr}[e^X] = \mathrm{tr}[e^X\, \mathrm{d}X]$$

$$\mathrm{d}|AXB| = \mathrm{tr}[|AXB|X^{-1}\, \mathrm{d}X]$$

$$\mathrm{d}|X^*AX| = \mathrm{tr}[2|X^*AX|X^{-1}\, \mathrm{d}X]$$

$$\mathrm{d}|X^n| = \mathrm{tr}[n|X^n|X^{-1}\, \mathrm{d}X]$$

$$\mathrm{d}\log|aX| = \mathrm{tr}[X^{-1}\, \mathrm{d}X]$$

$$\mathrm{d}\log|X^*X| = \mathrm{tr}[2X^+\, \mathrm{d}X]$$

Implicit function theorem

Let $f : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be continuously differentiable and $f(a, b) = 0$ for $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. If $\frac{\partial f}{\partial y}(a, b)$ is invertible, then there exist g such that $g(a) = b$ and for all $x \in \mathbb{R}^n$ in the neighborhood of a

$$f(x, g(x)) = 0$$

$$\frac{\partial g}{\partial x_i}(x) = -\left(\frac{\partial f}{\partial y}(x, g(x))\right)^{-1} \frac{\partial f}{\partial x_i}(x, g(x))$$

In a system of equations $f(x, y) = 0$ with an infinite number of solutions (x, y), IFT tells us about the relative variations of x with respect to y, even in situations where we cannot write down explicit solutions (*i.e.*, $y = g(x)$). For instance, without solving the system, it shows that the solutions (x, y) of $x^2 + y^2 = 1$ satisfies $\frac{\partial y}{\partial x} = -x/y$.

Cookbook for data scientists

Charles Deledalle

Convex optimization

Conjugate gradient

Let $A \in \mathbb{C}^{n \times n}$ be Hermitian positive definite The sequence x_k defined as, $r_0 = p_0 = b$, and

$$x_{k+1} = x_k + \alpha_k p_k \qquad r_{k+1} = r_k - \alpha_k A p_k \quad \text{with} \quad \alpha_k = \frac{r_k^* r_k}{p_k^* A p_k}$$

$$p_{k+1} = r_{k+1} + \beta_k p_k \quad \text{with} \quad \beta_k = \frac{r_{k+1}^* r_{k+1}}{r_k^* r_k}$$

converges towards $A^{-1}b$ in at most n steps.

Lipschitz gradient

$f : \mathbb{R}^n \to \mathbb{R}$ has a L Lipschitz gradient if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L\|x - y\|_2$$

If $\nabla f(x) = Ax$, $L = \|A\|_2$. If f is twice differentiable $L = \sup_x \|H_f(x)\|_2$, *i.e.*, the highest eigenvalue of $H_f(x)$ among all possible x.

Convexity

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if for all x, y and $\lambda \in (0, 1)$

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y)$$

f is strictly convex if the inequality is strict. f is convex and twice differentiable iif $H_f(x)$ is Hermitian non-negative definite. f is strictly convex and twice differentiable iif $H_f(x)$ is Hermitian positive definite. If f is convex, f has only global minima if any. We write the set of minima as

$$\operatorname*{argmin}_x f(x) = \{x \setminus \text{for all } z \in \mathbb{R}^n f(x) \leqslant f(z)\}$$

Gradient descent

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable with L Lipschitz gradient then, for $0 < \gamma \leqslant 1/L$, the sequence

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

converges towards a stationary point $x^\star$ in $O(1/k)$

$$\nabla f(x^\star) = 0$$

If f is moreover convex then

$$x^\star \in \operatorname*{argmin}_x f(x).$$

Newton's method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and twice continuously differentiable then, the sequence

$$x_{k+1} = x_k - H_f(x_k)^{-1} \nabla f(x_k)$$

converges towards a minimizer of f in $O(1/k^2)$.

Subdifferential / subgradient

The subdifferential of a convex† function f is

$$\partial f(x) = \{p \setminus \forall x',\ f(x) - f(x') \geqslant \langle p,\, x - x' \rangle\}.$$

$p \in \partial f(x)$ is called a subgradient of f at x.
A point $x^\star$ is a global minimizer of f iif

$$0 \in \partial f(x^\star).$$

If f is differentiable then $\partial f(x) = \{\nabla f(x)\}$.

Proximal gradient method

Let $f = g + h$ with g convex and differentiable with Lip. gradient and h convex†. Then, for $0 < \gamma \leqslant 1/L$,

$$x_{k+1} = \text{prox}_{\gamma h}(x_k - \gamma \nabla g(x_k))$$

converges towards a global minimizer of f where

$$\begin{aligned} \text{prox}_{\gamma h}(x) &= (\text{Id} + \gamma \partial h)^{-1}(x) \\ &= \operatorname*{argmin}_z \frac{1}{2}\|x - z\|^2 + \gamma h(z) \end{aligned}$$

is called proximal operator of f.

Convex conjugate and primal dual problem

The convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$f^*(z) = \sup_x \langle z,\, x \rangle - f(x)$$

if f is convex (and lower semi-continuous) $f = f^{\star\star}$. Moreover, if $f(x) = g(x) + h(Lx)$, then minimizers $x^\star$ of f are solutions of the saddle point problem

$$(x^\star, z^\star) \in \text{args} \min_x \max_z g(x) + \langle Lx,\, z \rangle - h^*(z)$$

$z^\star$ is called dual of $x^\star$ and satisfies $\begin{cases} Lx^\star \in \partial h^*(z^\star) \\ L^* z \in \partial g(x^\star) \end{cases}$