Numerical optimization : theory and applications

Ammar Mian
Associate professor, LISTIC, Université Savoie Mont Blanc

Outline

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

Online ressources

The syllabus, course monograph and slides are available at:

Book ressources

Main book

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999

Additional in convex optimization

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004

For reminders

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019

Part I - Fundamentals

Oranisation of first week

Session	Duration	Content	Date	Room
CM1	1.5h	Introduction, Linear algebra and Differentiation reminders, and exercices	2 June 2025 10am	B-120
CM2	1.5h	Steepest descent algorithm, Newton method and convexity	2 June 2025 1.15pm	B-120
TD1	1.5h	Application to linear regression	2 June 2025 3pm	C-213
CM3	1.5h	Linesearch algorithms and their convergence	3 June 2025 10am	B-120
CM4	1.5h	Constrained optimization : linear programming and lagrangian methods	3 June 2025 1.15pm	B-120
TD2	1.5h	Implementation of Linesearch methods	3 June 2025 3pm	C-213

Then on 5 June 2025 at 1pm, a project on Implementation of inverse problems for image processing, by *Yassine Mhiri*.

Numerical optimization

What is this course about ?

Numerical optimization

What is this course about ?

Numerical optimization

Numerical optimization is the computational process of finding the best solution to a mathematical problem when analytical (exact) methods are impractical or impossible.

What problem ?

Numerical optimization

What is this course about ?

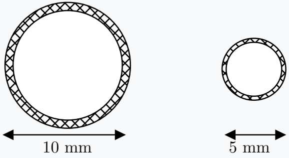
Numerical optimization

Numerical optimization is the computational process of finding the best solution to a mathematical problem when analytical (exact) methods are impractical or impossible.

What problem ?

- **Variables :** $\mathbf{x}_1, \ldots, \mathbf{x}_d$ organised as $\mathbf{x} \in \mathbb{R}^d$
- **Objective function:** $f : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R}$
- **Constraints :** $\mathcal{S} = \{\mathbf{x} \in \mathcal{X} : h_{1,\ldots,p}(\mathbf{x}) = 0,\ g_{1,\ldots,q}(\mathbf{x}) \geq 0\}$

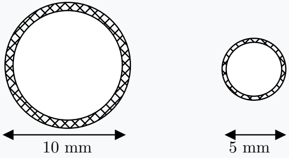
Practical examples (1/3)



Cable factory

A factory produces copper cables of 5mm and 10mm diameter, on which the profit is respectively 2 and 7 euros per meter. The copper available to the factory allows for the production of 20 km of 5mm diameter cable per week. The production of 10mm cable requires 4 times more copper than that of 5mm cable. For demand reasons, the weekly production of 5mm cable must not exceed 15 km, and for logistical reasons, the production of 10mm cable must not represent more than 40% of the total production.

Practical examples (1/3)



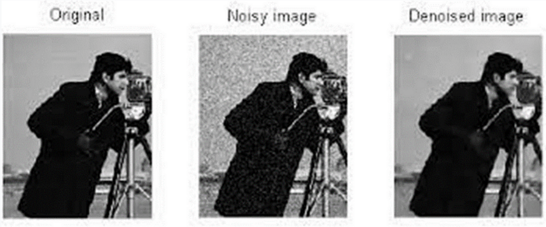
Cable factory

A factory produces copper cables of 5mm and 10mm diameter, on which the profit is respectively 2 and 7 euros per meter. The copper available to the factory allows for the production of 20 km of 5mm diameter cable per week. The production of 10mm cable requires 4 times more copper than that of 5mm cable. For demand reasons, the weekly production of 5mm cable must not exceed 15 km, and for logistical reasons, the production of 10mm cable must not represent more than 40% of the total production.

→ How to know what is the most profitable setup ?

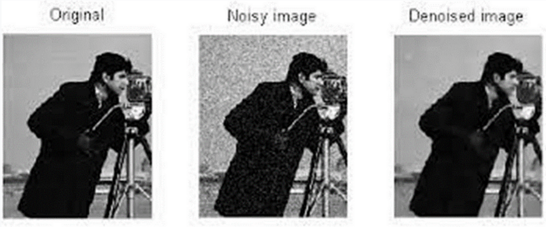
Practical exmaples (2/3)

Image denoising



Practical exmaples (2/3)

Image denoising



$\rightarrow$ How to model the wanted signal and then find the best one among all possible signals ?

Practical examples (3/3)

Portfolio optimization

An investor has \$1M to allocate between 3 assets: stocks (expected return 8%, risk 15%), bonds (expected return 4%, risk 5%), and real estate (expected return 6%, risk 10%). The correlations between assets are: stocks-bonds $= 0.2$, stocks-real estate $= 0.3$, bonds-real estate $= 0.1$. The investor wants to maximize expected return while keeping portfolio risk below 8%.

Practical examples (3/3)

Portfolio optimization

An investor has \$1M to allocate between 3 assets: stocks (expected return 8%, risk 15%), bonds (expected return 4%, risk 5%), and real estate (expected return 6%, risk 10%). The correlations between assets are: stocks-bonds = 0.2, stocks-real estate = 0.3, bonds-real estate = 0.1. The investor wants to maximize expected return while keeping portfolio risk below 8%.

Mathematical formulation:

$$\max_{\mathbf{w}} \quad \sum_{i=1}^{3} w_i \mu_i s.t \quad \sqrt{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}} \leq 0.08, \sum_{i=1}^{3} w_i = 1, \; w_i \geq 0, \quad i = 1, 2, 3 \tag{1}$$

where w_i = weight in asset i, μ_i = expected return, Σ = covariance matrix

Practical examples (3/3)

Portfolio optimization

An investor has \$1M to allocate between 3 assets: stocks (expected return 8%, risk 15%), bonds (expected return 4%, risk 5%), and real estate (expected return 6%, risk 10%). The correlations between assets are: stocks-bonds = 0.2, stocks-real estate = 0.3, bonds-real estate = 0.1. The investor wants to maximize expected return while keeping portfolio risk below 8%.

Mathematical formulation:

$$\max_{\mathbf{w}} \quad \sum_{i=1}^{3} w_i \mu_i s.t \quad \sqrt{\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}} \leq 0.08, \sum_{i=1}^{3} w_i = 1, \ w_i \geq 0, \quad i = 1, 2, 3 \tag{1}$$

where w_i = weight in asset i, μ_i = expected return, Σ = covariance matrix
$\rightarrow$ How to find the optimal balance between risk and return?

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

Notations

Matrix vectors, and scalars

- Scalars $\in \mathbb{R}$ are lowercase letters: x, y, z or greek letters: α, β, γ
- Vectors $\in \mathbb{R}^d$ are lowercase bold letters: $\mathbf{x}, \mathbf{y}, \mathbf{z}$ or greek letters: $\boldsymbol{\theta}$
- Matrices $\in \mathbb{R}^{m,n}$ are uppercase bold letters: $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$

> We don't consider data in $\mathbb{C}^d$ but it could be treated with equivalence $\mathbb{C}^d \equiv \mathbb{R}^{2d}$.

We also consider functions:

- $f : \mathbb{R}^d \to \mathbb{R}$ is a function from $\mathbb{R}^d$ to $\mathbb{R}$, e.g. $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$
- $g : \mathbb{R}^{m,n} \to \mathbb{R}$ is a function from $\mathbb{R}^{m,n}$ to $\mathbb{R}$, e.g. $g(\mathbf{X}) = \|\mathbf{X}\|_F^2$
- $h : \mathbb{R}^d \to \mathbb{R}^p$ is a function from $\mathbb{R}^d$ to $\mathbb{R}^p$, e.g. $h(\mathbf{x}) = [x_1^2, x_2^2, \dots, x_d^2]^\top$
- $l : \mathbb{R}^{m,n} \to \mathbb{R}^q$ is a function from $\mathbb{R}^{m,n}$ to $\mathbb{R}^q$, e.g. $l(\mathbf{X}) = \begin{pmatrix} X_{11}^2 \\ X_{12}^2 \\ \vdots \\ X_{mn}^2 \end{pmatrix}$

Vectors

Usual operations

- Sum: $\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_d + y_d \end{pmatrix}$
- Scalar product: $\mathbf{x}^\top \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_d y_d$
- p-norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}$, with $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$

Vector spaces

Span and Subspace

For a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$: $\text{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}) = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_k\mathbf{v}_k : c_i \in \mathbb{R}\}$.
A subspace is a subset of a vector space that is closed under addition and scalar multiplication.

Linear independence

A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is said to be **linearly independent** if the only solution to the equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_k\mathbf{v}_k = \mathbf{0}$ is $c_1 = c_2 = \ldots = c_k = 0$.

Basis: A set of vectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ is a **basis** of a vector space if they are linearly independent and span the space.

Matrices

Usual operations

- Sum for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m,n}$: $\mathbf{X} + \mathbf{Y} = \mathbf{Z} \in \mathbb{R}^{m,n}$ with $Z_{ij} = X_{ij} + Y_{ij}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$
- Hadamard product for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m,n}$: $\mathbf{X} \circ \mathbf{Y} = \mathbf{Z} \in \mathbb{R}^{m,n}$ with $Z_{ij} = X_{ij} Y_{ij}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$
- Matrix multiplication for $\mathbf{A} \in \mathbb{R}^{m,n}$, $\mathbf{B} \in \mathbb{R}^{n,p}$: $\mathbf{AB} = \mathbf{C} \in \mathbb{R}^{m,p}$ with $C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$, for $i = 1, \ldots, m$ and $j = 1, \ldots, p$
- Frobenius norm: $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j |X_{ij}|^2}$, with $\|\mathbf{X}\|_F^2 = \mathrm{Tr}\left(\mathbf{X} \circ \mathbf{X}\right)$

Matrix/vector multiplication: for $\mathbf{A} \in \mathbb{R}^{m,n}$ and $\mathbf{x} \in \mathbb{R}^{n}$, we have $\mathbf{Ax} = \mathbf{y} \in \mathbb{R}^{m}$ with $y_i = \sum_{j=1}^{n} A_{ij} x_j$, for $i = 1, \ldots, m$.

Matrices as linear applications

Given a basis in a vector space $\mathcal{V}$, a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ can be seen as a linear application $\mathcal{A} : \mathcal{V} \to \mathbb{R}^m$ such that for any vector $\mathbf{x} \in \mathcal{V}$, we have:
$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i \Rightarrow$

$$\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i \mathcal{A}(\mathbf{e}_i) = \sum_{i=1}^{n} x_i \mathbf{a}_i,$$

where $\mathbf{a}_i$ is the i-th column of $\mathbf{A}$.

Matrix properties (1/2)

Transpose

The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{m,n}$ is denoted $\mathbf{A}^\top \in \mathbb{R}^{n,m}$ and is defined as $(\mathbf{A}^\top)_{ij} = A_{ji}$.

Symmetric matrices

A matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is symmetric if $\mathbf{A} = \mathbf{A}^\top$.

Positive definite matrices

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is positive definite if for all non-zero vectors $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$.

Matrix properties (2/2)

Inverse

The inverse of a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is denoted $\mathbf{A}^{-1}$ and is defined such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of size n.

Determinant

The determinant of a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is denoted $\det(\mathbf{A})$ or $|\mathbf{A}|$ and is a scalar value that provides information about the matrix, such as whether it is invertible.

Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$ is denoted $\mathrm{Tr}(\mathbf{A})$ and is defined as the sum of the diagonal elements: $\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$.

Exercices

TODO

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

Eigenvalues and eigenvectors

Eigenvalues and eigenvectors

For a square matrix $\mathbf{A} \in \mathbb{R}^{n,n}$, a scalar λ is an **eigenvalue** and a non-zero vector $\mathbf{v} \in \mathbb{R}^{n}$ is an **eigenvector** if they satisfy the equation:

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

Computing eigenvalues and eigenvectors: The eigenvalues of a matrix $\mathbf{A}$ are the roots of the characteristic polynomial $\det(\mathbf{A} - \lambda \mathbf{I}_n) = 0$, where $\mathbf{I}_n$ is the identity matrix of size n.

Spectral theorem

If $\mathbf{A}$ is symmetric, then it has n real eigenvalues and n orthogonal eigenvectors (i.e if $\mathbf{u}_i$ and $\mathbf{u}_j$ are eigenvectors of $\mathbf{A}$, then $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for $i \neq j$).

Eigenvalue decomposition

Eigenvalue decomposition

If $\mathbf{A}$ is a symmetric matrix, it can be decomposed as:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top,$$

where $\mathbf{U}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}$.

Properties

- The eigenvalues of $\mathbf{A}$ are the diagonal elements of $\boldsymbol{\Lambda}$.
- The columns of $\mathbf{U}$ form an orthonormal basis for $\mathbb{R}^n$.
- The eigenvalue decomposition is unique up to the order of the eigenvalues and eigenvectors.

Singular Value Decomposition (SVD)

Singular Value Decomposition

Any matrix $\mathbf{X} \in \mathbb{R}^{m,n}$ can be decomposed as $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$,

- $\mathbf{U} \in \mathbb{R}^{m,m}$ is an orthogonal matrix whose columns are the left singular vectors of $\mathbf{X}$.
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m,n}$ is a diagonal matrix with non-negative entries (the singular values).
- $\mathbf{V} \in \mathbb{R}^{n,n}$ is an orthogonal matrix whose columns are the right singular vectors of $\mathbf{X}$.

Properties

- The singular values are the square roots of the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ or $\mathbf{X}\mathbf{X}^\top$.
- The SVD is unique up to the order of the singular values and the signs of the singular vectors.

Exercices

TODO

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

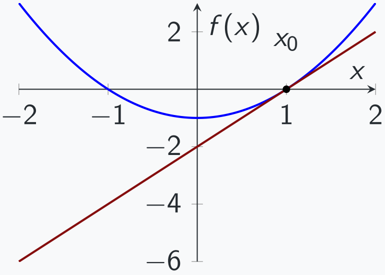
3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. **Differentiation**
   - **Monovariate reminders**
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

Derivative of a function



Definition

The derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at a point x_0 is defined as:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

$\rightarrow$ The derivative represents the slope of the tangent line to the curve at the point $(x_0, f(x_0))$.

Usually computed thanks to product and chain rules:

- Product rule: $(uv)' = u'v + uv'$
- Chain rule: $(f(g(x)))' = f'(g(x))g'(x)$

1. Introduction
   - Course organization
   - The setup

2. Linear Algebra
   - Vectors and matrices
   - Matrix decompositions

3. Differentiation
   - Monovariate reminders
   - Extension to Multivariate setup: $f : \mathbb{R}^d \to \mathbb{R}$
   - Multivariate case: $f : \mathbb{R}^d \to \mathbb{R}^p$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}$
   - Matrix functions: $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

Limits and Continuity

Open disk

An open disk of radius $\epsilon > 0$ centered at a point $\mathbf{x}_0 \in \mathbb{R}^d$ is defined as:

$$\mathcal{B}(\mathbf{x}_0, \epsilon) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\|_2 < \epsilon\}$$

Limit

The limit of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{x}_0$ is defined as:

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = L$$

if $\forall \epsilon > 0, \exists \delta > 0$ such that if $\|\mathbf{x} - \mathbf{x}_0\|_2 < \delta$, then $|f(\mathbf{x}) - L| < \epsilon$.

A function is continuous at a point $\mathbf{x}_0$ if $\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0)$.

Directional derivative

Directional derivative

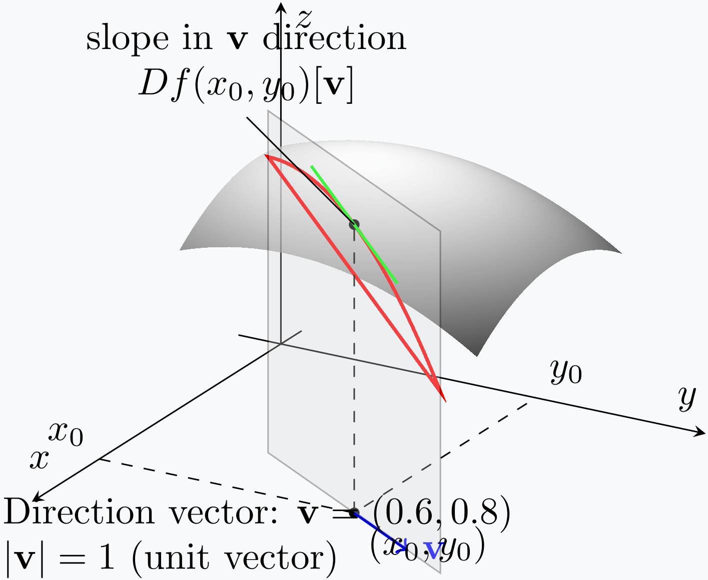
The directional derivative of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{x}_0$ in the direction of a vector $\mathbf{v} \in \mathbb{R}^d$ is defined as:

$$Df(\mathbf{x}_0)[\mathbf{v}] = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h}$$

If $\|\mathbf{v}\|_2 = 1$, then $Df(\mathbf{x}_0)[\mathbf{v}]$ represents the rate of change of f in the direction of $\mathbf{v}$ at the point $\mathbf{x}_0$.

Note : We use alternatively the notation $\nabla_{\mathbf{v}} f(\mathbf{x}_0)$ for the directional derivative.

Illustration of directional derivative



Gradient

Gradient

The gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{x}_0$ is defined as the vector of all directional derivatives in the standard basis directions:

$$\nabla f(\mathbf{x}_0) = (Df(\mathbf{x}_0)[\mathbf{e}_1], Df(\mathbf{x}_0)\mathbf{e}_1, \ldots, Df(\mathbf{x}_0)\mathbf{e}_1)^\mathsf{T}$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$ is the standard basis of $\mathbb{R}^d$.

$\rightarrow$ the direction of the steepest ascent of the function f at the point $\mathbf{x}_0$.

Relationship with directional derivative

For any vector $\mathbf{v} \in \mathbb{R}^d$, the directional derivative can be expressed as:

$$Df(\mathbf{x}_0)[\mathbf{v}] = \nabla f(\mathbf{x}_0)^\mathsf{T}\mathbf{v}$$

Graident and partial derivatives

Partial derivatives

The partial derivative of a function $f : \mathbb{R}^d \to \mathbb{R}$ with respect to the i-th variable is defined as:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}_0) = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h}$$

where $\mathbf{e}_i$ is the i-th standard basis vector.

The gradient can be expressed in terms of partial derivatives as:

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}_0), \frac{\partial f}{\partial x_2}(\mathbf{x}_0), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x}_0) \right)^\mathsf{T}$$

$\rightarrow$ The gradient is a vector containing all the partial derivatives of the function at the point $\mathbf{x}_0$.

Gradient properties and practical computation

Derivative of a prodct

Let $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$ be two functions. Then the gradient of their product $f(\mathbf{x}) = g(\mathbf{x})h(\mathbf{x})$ can be computed using the product rule:

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

Derivative of a composition

Two cases:

- $f = h \circ g$ with: $h : \mathbb{R} \mapsto \mathbb{R}$ and $g : \mathbb{R}^d \mapsto \mathbb{R}$. The gradient of f can be computed using the chain rule: $\nabla f(\mathbf{x}) = h'(g(\mathbf{x}))\nabla g(\mathbf{x})$, where h' is the derivative of h.
- $f = h \circ g$ with: $h : \mathbb{R}^d \mapsto \mathbb{R}$ and $g : \mathbb{R}^{d'} \mapsto \mathbb{R}^d$. (We look at that later)

Hessian matrix

Hessian matrix

The Hessian matrix of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point $\mathbf{x}_0$ is defined as the square matrix of second-order partial derivatives:

$$\mathbf{H}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\mathbf{x}_0) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(\mathbf{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\mathbf{x}_0) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(\mathbf{x}_0) \end{pmatrix}$$

$\to$ The Hessian matrix provides information about the curvature of the function f at the point $\mathbf{x}_0$.

Exercices

Exercices

- Compute the gradient and Hessian matrix of the function $f(x, y) = x^2 + 3xy + y^2$ at the point $(1, 2)$.
- Using chain-rule compute gradient of $f(\mathbf{x}) = (\sum_i^d x_i^2)^(1/2)$.

Hessian matrix properties

Properties of the Hessian matrix

- The Hessian is symmetric: $\mathbf{H}(\mathbf{x}_0) = \mathbf{H}(\mathbf{x}_0)^T$.
- If f is twice continuously differentiable, then the mixed partial derivatives are equal: $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.
- The eigenvalues of the Hessian provide information about the local curvature of the function:
  - If all eigenvalues are positive, f is locally convex at $\mathbf{x}_0$.
  - If all eigenvalues are negative, f is locally concave at $\mathbf{x}_0$.
  - If some eigenvalues are positive and others are negative, f has a saddle point at $\mathbf{x}_0$.

Exercice

Rosenbrock function

The Rosenbrock function is defined as:

$$f(x, y) = (a - x)^2 + b(y - x^2)^2$$

where a and b are constants (commonly $a = 1$ and $b = 100$).

- Compute the gradient $\nabla f(x, y)$. And find stationary point(s).
- Compute the Hessian matrix $\mathbf{H}(x, y)$. Analyse local curvature at the stationary point(s).

Multivariate functions

Multivariate function

A function $f : \mathbb{R}^d \to \mathbb{R}^p$ maps a vector $\mathbf{x} \in \mathbb{R}^d$ to a vector $\mathbf{y} \in \mathbb{R}^p$. We can write:

$$f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_p(\mathbf{x}) \end{pmatrix}$$

The function f is said to be vector-valued, and each component $f_i : \mathbb{R}^d \to \mathbb{R}$ is a scalar function.

Gradient and Jacobian

Gradient

The gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right)^\mathsf{T} \in \mathbb{R}^d$$

Jacobian matrix

The Jacobian matrix of a function $f : \mathbb{R}^d \to \mathbb{R}^p$ is defined as:

$$\mathbf{J}_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \ldots & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \ldots & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \frac{\partial f_p}{\partial x_2} & \ldots & \frac{\partial f_p}{\partial x_d} \end{pmatrix} \in \mathbb{R}^{p \times d}$$

$\rightarrow$ The Jacobian matrix generalizes the concept of gradient to vector-valued functions.

Jacobian and directional derivative

Directional derivative

The directional derivative of a function $f : \mathbb{R}^d \to \mathbb{R}^p$ in the direction of a vector $\mathbf{v} \in \mathbb{R}^d$ is defined as:

$$Df(\mathbf{x})[\mathbf{v}] = \mathbf{J}_f(\mathbf{x})\mathbf{v} = \begin{pmatrix} \nabla f_1(\mathbf{x})^T\mathbf{v} \\ \nabla f_2(\mathbf{x})^T\mathbf{v} \\ \vdots \\ \nabla f_p(\mathbf{x})^T\mathbf{v} \end{pmatrix} \in \mathbb{R}^p$$

The directional derivative gives the rate of change of the function f for all components in the direction of $\mathbf{v}$ at the point $\mathbf{x}$.

Chain rule for composition of functions

General chain rule

If $f : \mathbb{R}^d \to \mathbb{R}^p$ and $g : \mathbb{R}^m \to \mathbb{R}^d$, then the composition $h = f \circ g : \mathbb{R}^m \to \mathbb{R}^p$ is defined as:

$$h(\mathbf{y}) = f(g(\mathbf{y}))$$

The Jacobian of h can be computed using the chain rule:

$$\mathbf{J}_h(\mathbf{y}) = \mathbf{J}_f(g(\mathbf{y}))\mathbf{J}_g(\mathbf{y})$$

where $\mathbf{J}_h(\mathbf{y}) \in \mathbb{R}^{p \times m}$, $\mathbf{J}_f(g(\mathbf{y})) \in \mathbb{R}^{p \times d}$, and $\mathbf{J}_g(\mathbf{y}) \in \mathbb{R}^{d \times m}$.

Chain rule: Special cases

Case 1: $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^d$

For $h = f \circ g : \mathbb{R}^m \to \mathbb{R}$, we have:

$$\nabla h(\mathbf{y}) = \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y}))$$

where $\mathbf{J}_g(\mathbf{y}) \in \mathbb{R}^{d \times m}$ and $\nabla f(g(\mathbf{y})) \in \mathbb{R}^d$.

Case 2: $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$

For $h = f \circ g : \mathbb{R}^m \to \mathbb{R}$, we have:

$$\nabla h(\mathbf{y}) = f'(g(\mathbf{y}))\nabla g(\mathbf{y})$$

where $f'(g(\mathbf{y})) \in \mathbb{R}$ is a scalar and $\nabla g(\mathbf{y}) \in \mathbb{R}^m$.

Exercise 1: Basic Chain Rule

Problem

Given:

- $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ where $f : \mathbb{R}^2 \to \mathbb{R}$
- $g(\mathbf{y}) = \begin{pmatrix} y_1 + y_2 \\ y_1 - y_2 \end{pmatrix}$ where $g : \mathbb{R}^2 \to \mathbb{R}^2$
- $h = f \circ g$

Task: Find $\nabla h(\mathbf{y})$ using the chain rule.

Chain Rule Formula (Case 1)

For $h = f \circ g$ where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^n$:

$$\nabla h(\mathbf{y}) = \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y}))$$

Exercise 1: Solution

Step 1: Compute $\nabla f(\mathbf{x})$

Using matrix calculus: $f(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$

$$\nabla f(\mathbf{x}) = 2\mathbf{x}$$

Step 2: Compute Jacobian $\mathbf{J}_g(\mathbf{y})$

$$g(\mathbf{y}) = \begin{pmatrix} y_1 + y_2 \\ y_1 - y_2 \end{pmatrix} \Rightarrow \mathbf{J}_g(\mathbf{y}) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Step 3: Apply chain rule

$$\nabla h(\mathbf{y}) = \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y})) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot 2g(\mathbf{y})$$

$$= 2\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} y_1 + y_2 \\ y_1 - y_2 \end{pmatrix} = 2\begin{pmatrix} 2y_1 \\ 2y_2 \end{pmatrix} = \begin{pmatrix} 4y_1 \\ 4y_2 \end{pmatrix}$$

Exercise 1: Verification

Direct computation

$$h(\mathbf{y}) = f(g(\mathbf{y})) = (y_1 + y_2)^2 + (y_1 - y_2)^2$$
$$= y_1^2 + 2y_1y_2 + y_2^2 + y_1^2 - 2y_1y_2 + y_2^2 = 2y_1^2 + 2y_2^2$$

Therefore:

$$\nabla h(\mathbf{y}) = \begin{pmatrix} 4y_1 \\ 4y_2 \end{pmatrix} \quad \checkmark$$

Exercise 2: General Quadratic Forms

Problem (General Case)

Given:

- $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}$ where $f : \mathbb{R}^n \to \mathbb{R}$
- $\mathbf{A} \in \mathbb{R}^{n\times n}$ symmetric, $\mathbf{b} \in \mathbb{R}^n$
- $g(\mathbf{y}) = \mathbf{C}\mathbf{y}$ where $g : \mathbb{R}^m \to \mathbb{R}^n$
- $\mathbf{C} \in \mathbb{R}^{n\times m}$
- $h = f \circ g$

Task: Find $\nabla h(\mathbf{y})$ using matrix calculus and the chain rule.

Matrix Calculus Rules

- $\nabla(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$ (when $\mathbf{A}$ symmetric)
- $\nabla(\mathbf{b}^T\mathbf{x}) = \mathbf{b}$
- For linear $g(\mathbf{y}) = \mathbf{C}\mathbf{y}$: $\mathbf{J}_g(\mathbf{y}) = \mathbf{C}$

Exercise 2: General Solution

Step 1: Compute $\nabla f(\mathbf{x})$

Using matrix calculus rules:

$$\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x} + \mathbf{b}$$

Step 2: Compute Jacobian $\mathbf{J}_g(\mathbf{y})$

Since $g(\mathbf{y}) = \mathbf{C}\mathbf{y}$ (linear transformation):

$$\mathbf{J}_g(\mathbf{y}) = \mathbf{C} \in \mathbb{R}^{n \times m}$$

Step 3: Apply chain rule

$$\begin{aligned}
\nabla h(\mathbf{y}) &= \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y})) \\
&= \mathbf{C}^T[2\mathbf{A}(\mathbf{C}\mathbf{y}) + \mathbf{b}] \\
&= 2\mathbf{C}^T\mathbf{A}\mathbf{C}\mathbf{y} + \mathbf{C}^T\mathbf{b}
\end{aligned}$$

Exercise 2: Numerical Example

Specific case: $n = 3, m = 2$

- $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$
- $\mathbf{C} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$

Computing the required matrices:

$$\mathbf{C}^T\mathbf{A}\mathbf{C} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 7 & 6 \\ 6 & 9 \end{pmatrix}$$

$$\mathbf{C}^T\mathbf{b} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Therefore:

Key Takeaways

Advantages of Matrix Calculus Approach

- **Scalability:** Works for arbitrary dimensions without modification
- **Efficiency:** No component-wise partial derivatives needed
- **Clarity:** Clean matrix operations instead of summations
- **Generality:** Same formulas apply regardless of problem size

General Chain Rule Pattern

For compositions $h = f \circ g$:

1. Identify the structure of f and g
2. Use appropriate matrix calculus rules for ∇f and $\mathbf{J}_g$
3. Apply: $\nabla h(\mathbf{y}) = \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y}))$
4. Verify using direct computation when possible

Fréchet derivative – Definition

Fréchet differentiability

A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is differentiable at $\mathbf{X}$ if there exists a linear mapping $Df(\mathbf{X}) : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ such that

$$\lim_{\|\mathbf{V}\|_F \to 0} \frac{\|f(\mathbf{X}+\mathbf{V}) - f(\mathbf{X}) - Df(\mathbf{X})[\mathbf{V}]\|_F}{\|\mathbf{V}\|_F} = 0$$

$$\mathbf{D}f(\mathbf{X})[\boldsymbol{\xi}] = f(\mathbf{X}+\xi) - f(\mathbf{X}) + o(\|\boldsymbol{\xi}\|)$$

Gateaux derivative (alternative characterization)

If f is Fréchet differentiable at $\mathbf{X}$, then for any $\mathbf{V}$:

$$Df(\mathbf{X})[\mathbf{V}] = \left.\frac{d}{dt}\right|_{t=0} f(\mathbf{X}+t\mathbf{V}) = \lim_{t\to 0} \frac{f(\mathbf{X}+t\mathbf{V}) - f(\mathbf{X})}{t}$$

Note: If $\left.\frac{d}{dt}\right|_{t=0} f(\mathbf{X}+t\mathbf{V})$ is not linear in $\mathbf{V}$, then f is not Fréchet differentiable.

What it means to derive a matrix function?

Matrix function

A matrix function $f : \mathbb{R}^{m,n} \to \mathbb{R}$ maps a matrix $\mathbf{X} \in \mathbb{R}^{m,n}$ to a scalar value. For example, $f(\mathbf{X}) = \|\mathbf{X}\|_F^2$.

Directional derivative

The directional derivative of f at $\mathbf{X}$ in direction $\mathbf{V}$ is:

$$Df(\mathbf{X})[\mathbf{V}] = \lim_{h \to 0} \frac{f(\mathbf{X} + h\mathbf{V}) - f(\mathbf{X})}{h}$$

where $\mathbf{V} \in \mathbb{R}^{m,n}$ is a perturbation matrix.

Since $Df(\mathbf{X})$ is linear, it can be represented using the gradient matrix.

Gradient in the matrix to scalar case

Gradient definition

For $f : \mathbb{R}^{m,n} \to \mathbb{R}$, the gradient $\nabla f(\mathbf{X}) \in \mathbb{R}^{m,n}$ satisfies:

$$Df(\mathbf{X})[\mathbf{V}] = \mathrm{Tr}(\nabla f(\mathbf{X})^\mathsf{T} \mathbf{V})$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

Link to partial derivatives

The gradient can be computed using partial derivatives:

$$\nabla f(\mathbf{X}) = \begin{pmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{m1}} & \frac{\partial f}{\partial X_{m2}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{pmatrix}$$

Examples: Matrix to scalar functions

Example 1: $f(\mathbf{X}) = \|\mathbf{X}\|_F^2 = \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X})$
Using Gateaux derivative:

$$
\begin{aligned}
Df(\mathbf{X})[\mathbf{V}] &= \left.\frac{d}{dt}\right|_{t=0} \mathrm{Tr}((\mathbf{X}+t\mathbf{V})^\mathsf{T}(\mathbf{X}+t\mathbf{V})) \\
&= \left.\frac{d}{dt}\right|_{t=0} [\mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) + 2t\mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{V}) + t^2\mathrm{Tr}(\mathbf{V}^\mathsf{T}\mathbf{V})] \\
&= 2\mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{V})
\end{aligned}
$$

Therefore: $\boxed{\nabla f(\mathbf{X}) = 2\mathbf{X}}$

Example 2: $f(\mathbf{X}) = \log\det(\mathbf{X})$ (for invertible $\mathbf{X}$)

$$
Df(\mathbf{X})[\mathbf{V}] = \left.\frac{d}{dt}\right|_{t=0} \log\det(\mathbf{X}+t\mathbf{V}) = \mathrm{Tr}(\mathbf{X}^{-1}\mathbf{V})
$$

Therefore: $\boxed{\nabla f(\mathbf{X}) = \mathbf{X}^{-\mathsf{T}}}$

Matrix to matrix functions

Matrix-valued function

A function $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$ maps a matrix $\mathbf{X} \in \mathbb{R}^{m,n}$ to a matrix $\mathbf{Y} \in \mathbb{R}^{p,q}$.

Directional derivative

The directional derivative $Df(\mathbf{X})[\mathbf{V}]$ is a linear mapping from $\mathbb{R}^{m,n}$ to $\mathbb{R}^{p,q}$:

$$Df(\mathbf{X})[\mathbf{V}] = \lim_{t \to 0} \frac{f(\mathbf{X} + t\mathbf{V}) - f(\mathbf{X})}{t}$$

Vectorization representation

Since $Df(\mathbf{X})$ is linear, there exists a matrix $\mathbf{M}_{\mathbf{X}} \in \mathbb{R}^{pq \times mn}$ such that:

$$\text{vec}(Df(\mathbf{X})[\mathbf{V}]) = \mathbf{M}_{\mathbf{X}} \text{vec}(\mathbf{V})$$

where $\text{vec}(\cdot)$ stacks matrix columns into a vector.

Vectorization identities

Key identities for matrix calculus

- $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\text{vec}(\mathbf{B})$
- $\text{Tr}(\mathbf{AB}) = \text{vec}(\mathbf{A})^\top\text{vec}(\mathbf{B})$
- $\text{Tr}(\mathbf{A}^\top\mathbf{B}) = \text{vec}(\mathbf{A})^\top\text{vec}(\mathbf{B})$

where $\otimes$ denotes the Kronecker product.

These identities allow us to:

- Convert matrix operations to vector operations
- Find the matrix $\mathbf{M}_{\mathbf{X}}$ in the vectorization representation
- Compute derivatives efficiently

Examples: Matrix to matrix functions

Example 1: $f(\mathbf{X}) = \mathbf{X}^2$
Using Gateaux derivative:

$$Df(\mathbf{X})[\mathbf{V}] = \left.\frac{d}{dt}\right|_{t=0} (\mathbf{X} + t\mathbf{V})^2$$

$$= \left.\frac{d}{dt}\right|_{t=0} [\mathbf{X}^2 + t(\mathbf{XV} + \mathbf{VX}) + t^2\mathbf{V}^2]$$

$$= \mathbf{XV} + \mathbf{VX}$$

Therefore: $\boxed{Df(\mathbf{X})[\mathbf{V}] = \mathbf{XV} + \mathbf{VX}}$

Example 2: $f(\mathbf{X}) = \mathbf{X}^{-1}$ (for invertible $\mathbf{X}$)
From the identity $\mathbf{XX}^{-1} = \mathbf{I}$ and product rule:

$$\mathbf{VX}^{-1} + \mathbf{X}\, Df(\mathbf{X})[\mathbf{V}] = \mathbf{0}$$

Therefore: $\boxed{Df(\mathbf{X})[\mathbf{V}] = -\mathbf{X}^{-1}\mathbf{VX}^{-1}}$

Properties of matrix function derivatives

Linearity

For $f = \alpha g + \beta h$:

$$Df(\mathbf{X})[\mathbf{V}] = \alpha\, Dg(\mathbf{X})[\mathbf{V}] + \beta\, Dh(\mathbf{X})[\mathbf{V}]$$

Product rule

For $f(\mathbf{X}) = g(\mathbf{X}) \cdot h(\mathbf{X})$ (matrix multiplication):

$$Df(\mathbf{X})[\mathbf{V}] = Dg(\mathbf{X})[\mathbf{V}] \cdot h(\mathbf{X}) + g(\mathbf{X}) \cdot Dh(\mathbf{X})[\mathbf{V}]$$

Chain rule

For $f(\mathbf{X}) = g(h(\mathbf{X}))$:

$$Df(\mathbf{X})[\mathbf{V}] = Dg(h(\mathbf{X}))[Dh(\mathbf{X})[\mathbf{V}]]$$

Chain rule example

Example: $f(\mathbf{X}) = (\mathbf{X}^{1/2})^2 = \mathbf{X}$
Let $g(\mathbf{Y}) = \mathbf{Y}^2$ and $h(\mathbf{X}) = \mathbf{X}^{1/2}$, so $f(\mathbf{X}) = g(h(\mathbf{X}))$.
We know:

- $Dg(\mathbf{Y})[\mathbf{W}] = \mathbf{YW} + \mathbf{WY}$
- $Df(\mathbf{X})[\mathbf{V}] = \mathbf{V}$ (since $f(\mathbf{X}) = \mathbf{X}$)

Using the chain rule:

$$Df(\mathbf{X})[\mathbf{V}] = Dg(h(\mathbf{X}))[Dh(\mathbf{X})[\mathbf{V}]]$$

$$\mathbf{V} = \mathbf{X}^{1/2}\, Dh(\mathbf{X})[\mathbf{V}] + Dh(\mathbf{X})[\mathbf{V}]\, \mathbf{X}^{1/2}$$

This gives us a Sylvester equation for $Dh(\mathbf{X})[\mathbf{V}]$:

$$\mathbf{X}^{1/2}\, Dh(\mathbf{X})[\mathbf{V}] + Dh(\mathbf{X})[\mathbf{V}]\, \mathbf{X}^{1/2} = \mathbf{V}$$

Summary: Matrix function differentiation

Key concepts

- **Fréchet derivative**: Captures linear approximation of matrix functions
- **Matrix-to-scalar**: Gradient $\nabla f(\mathbf{X})$ with $Df(\mathbf{X})[\mathbf{V}] = \mathrm{Tr}(\nabla f(\mathbf{X})^\top \mathbf{V})$
- **Matrix-to-matrix**: Linear operator $Df(\mathbf{X})$ with vectorization representation
- **Vectorization**: Converts matrix operations to linear algebra on vectors

Applications

- Optimization on matrix manifolds
- Machine learning (gradient descent for matrix parameters)
- Sensitivity analysis
- Perturbation theory