### 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
- Conditioning
- Exercices

- 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
- Conditioning
- Exercices

- 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
- Conditioning
- Exercices

- 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*.

### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

- 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

# 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 :  $x_1, \ldots, x_d$  organised as  $x \in \mathbb{R}^d$
- Objective function:  $f : \mathcal{X} \subset \mathbb{R}^d \mapsto \mathbb{R}$
- Constraints :  $S = \{ \mathbf{x} \in \mathcal{X} : h_{1,\dots,p}(\mathbf{x}) = 0, g_{1,\dots,q}(\mathbf{x}) \ge 0 \}$

# 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.

 $\rightarrow$  How to know what is the most profitable setup ?

# Practical exmaples (2/3)



# Practical exmaples (2/3)



 $\rightarrow$  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:

$$\max_{\mathbf{w}} \quad \sum_{i=1}^{3} w_{i} \mu_{i} s.t \quad \sqrt{\mathbf{w}^{\mathsf{T}} \mathbf{\Sigma} \mathbf{w}} \leq 0.08, \ \sum_{i=1}^{3} w_{i} = 1, \ w_{i} \geq 0, \quad i = 1, 2, 3$$
(1)

where  $w_i$  = weight in asset *i*,  $\mu_i$  = expected return,  $\Sigma$  = 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:

$$\max_{\mathbf{w}} \quad \sum_{i=1}^{3} w_{i} \mu_{i} s.t \quad \sqrt{\mathbf{w}^{\mathsf{T}} \mathbf{\Sigma} \mathbf{w}} \leq 0.08, \ \sum_{i=1}^{3} w_{i} = 1, \ w_{i} \geq 0, \quad i = 1, 2, 3$$
(1)

where  $w_i$  = weight in asset *i*,  $\mu_i$  = expected return,  $\Sigma$  = 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
- Conditioning
- Exercices

- 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
- Conditioning
- Exercices

- 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:  $lpha, eta, \gamma$
- Vectors  $\in \mathbb{R}^d$  are lowercase bold letters: **x**, **y**, **z** or greek letters:  $oldsymbol{ heta}$
- Matrices  $\in \mathbb{R}^{m,n}$  are uppercase bold letters: X, Y, 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.  $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$
- $g: \mathbb{R}^{m,n} \to \mathbb{R}$  is a function from  $\mathbb{R}^{m,n}$  to  $\mathbb{R}$ , e.g.  $g(\mathsf{X}) = \|\mathsf{X}\|_F^2$
- $h : \mathbb{R}^d \to \mathbb{R}^p$  is a function from  $\mathbb{R}^d$  to  $\mathbb{R}^p$ , e.g.  $h(\mathbf{x}) = [x_1^2, x_2^2, \dots, x_d^2]^T$

• 
$$I : \mathbb{R}^{m,n} \to \mathbb{R}^{q}$$
 is a function from  $\mathbb{R}^{m,n}$  to  $\mathbb{R}^{q}$ , e.g.  $I(\mathbf{X}) = \begin{pmatrix} X_{11}^{2} \\ X_{12}^{2} \\ \vdots \\ X_{mn}^{2} \end{pmatrix}$ 

## 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:  $\mathbf{x}^{\mathsf{T}} \mathbf{y} = x_1 y_1 + x_2 y_2 + \dots + x_d y_d$   
• *p*-norm:  $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ , with  $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^{\mathsf{T}} \mathbf{x}}$ 

### Vector spaces

## Span and Subspace

For a set of vectors  $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ : span $(\{\mathbf{v}_1, \dots, \mathbf{v}_k\}) = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_k\mathbf{v}_k : c_i \in \mathbb{R}\}.$ 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  $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$  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, ..., m and j = 1, ..., 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, ..., m and j = 1, ..., 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, ..., m and j = 1, ..., p

• Frobenius norm: 
$$\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j |X_{ij}|^2}$$
, with  $\|\mathbf{X}\|_F^2 = \text{Tr}(\mathbf{X} \circ \mathbf{X})$ 

Matrix/vector multiplication: for  $\mathbf{A} \in \mathbb{R}^{m,n}$  and  $\mathbf{x} \in \mathbb{R}^n$ , we have  $\mathbf{A}\mathbf{x} = \mathbf{y} \in \mathbb{R}^m$  with  $y_i = \sum_{j=1}^n A_{ij}x_j$ , for i = 1, ..., 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  $\mathbf{A}^{\mathsf{T}} \in \mathbb{R}^{n,m}$  and is defined as  $(\mathbf{A}^{\mathsf{T}})_{ij} = A_{ji}$ .

#### Symmetric matrices

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

#### 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  $\mathbf{x}^{\mathsf{T}} \mathbf{A} \mathbf{x} > 0$ .

## 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  $\text{Tr}(\mathbf{A})$  and is defined as the sum of the diagonal elements:  $\text{Tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$ .

# Exercices

TODO

#### 1. Introduction

- Course organization
- The setup

## 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

- 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  $\lambda$  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 **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 **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 **A**, then  $\mathbf{u}_i^{\mathsf{T}}\mathbf{u}_j = 0$  for  $i \neq j$ ).

# Eigenvalue decomposition

#### Eigenvalue decomposition

If A is a symmetric matrix, it can be decomposed as:

 $\mathbf{A} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{\mathsf{T}}$ ,

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

#### Properties

- The eigenvalues of A are the diagonal elements of  $\Lambda$ .
- The columns of **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  $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\mathsf{T}}$ ,

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

#### Properties

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

#### 1. Introduction

- Course organization
- The setup

## 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

- 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

# Matrix Conditioning: Definition

### Definition: Condition Number

For a nonsingular matrix  $A \in \mathbb{R}^{n \times n}$  and matrix norm  $\|\cdot\|$ , the condition number is:

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$

#### **Basic Properties**

- $\kappa(A) \geq 1$  for any nonsingular A
- $\kappa(\alpha A) = \kappa(A)$  for any  $\alpha \neq 0$
- $\kappa(A) = \infty$  if A is singular

# Well vs Ill-Conditioned Matrices

#### Classification

- Well-conditioned:  $\kappa(A)$  is small (close to 1)
- Ill-conditioned:  $\kappa(A)$  is large ( $\gg 1$ )
- Singular:  $\kappa(A) = \infty$  (no inverse exists)

### **Relation to Linear Systems**

For Ax = b, if we perturb  $b \to b + \Delta b$ , then  $x \to x + \Delta x$  where:  $\frac{\|\Delta x\|}{\|x\|} \le \kappa(A) \frac{\|\Delta b\|}{\|b\|}$ Interpretation:  $\kappa(A)$  amplifies relative errors by factor  $\kappa(A)$ .

Numerical Stability

- Rule of thumb: Lose  $\approx \log_{10}(\kappa(A))$  decimal digits
- If  $\kappa(A) \approx 10^k$ , expect (d k) correct digits in *d*-digit precision
- **Example**:  $\kappa(A) = 10^6$  in double precision  $\Rightarrow$  lose 6 digits

## Relationship to Eigenvalues

Symmetric Matrices

For symmetric A with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ :

 $\kappa_2(A) = rac{|\lambda_1|}{|\lambda_n|} = rac{|\lambda_{\max}|}{|\lambda_{\min}|}$ 

Geometric meaning: Ratio of largest to smallest stretching factors.

**Geometric Interpretation** 

Matrix A transforms unit sphere  $\{x : ||x||_2 = 1\}$  into ellipsoid:

• Semi-axes have lengths  $\sigma_1, \sigma_2, \ldots, \sigma_n$  (singular values)

• 
$$\kappa_2(A) = \frac{\sigma_{\max}}{\sigma_{\min}} =$$
aspect ratio of ellipsoid

High condition number ⇔ very elongated ellipsoid
#### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

### 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

#### Exercices

### Factorization

Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  such that  $\mathbf{A}^3 - 4\mathbf{A}^2 + 5\mathbf{A} - 2\mathbf{I}_n = \mathbf{0}$ , where  $\mathbf{I}_n$  is the identity matrix of size *n*. Show that  $\mathbf{A}$  is invertible and find its inverse  $\mathbf{A}^{-1}$  as a function of  $\mathbf{A}$ .

# Classification of Quadratic Forms

#### Problem

For the quadratic form  $q(x) = x^T A x$ , consider the following matrices:

$$A_1 = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}$$

#### Tasks

- 1. Classify each quadratic form as:
  - Positive definite
  - Negative definite
  - Indefinite
- 2. Find the minimum/maximum values of q(x) subject to  $||x||_2 = 1$  (if they exist)
- 3. State the geometric interpretation of each quadratic form

### Hint

## Matrix Inversion Formulas

#### Problem

Consider the matrices:

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & k \\ k & 1 \end{pmatrix}$$

### Tasks

1. Find the inverse of A using the formula:

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

where adj(A) is the adjugate matrix.

- 2. Verify your result by computing  $AA^{-1} = I$ .
- 3. For which values of k is matrix B invertible?
- **4.** Find  $B^{-1}$  in terms of k (when it exists).

24

# **Orthogonal Projections**

#### Problem

Consider the vectors:

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$
,  $u_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ 

#### Tasks

1. Find the projection matrix P that projects onto the subspace  $W = \text{span}\{v_1, v_2\}.$ 

#### 2. Verify that *P* satisfies the projection properties:

- $P^2 = P$  (idempotent)
- $P^T = P$  (symmetric)
- 3. What is the matrix I P? What subspace does it project onto?

4. Find the projection of the vector 
$$b = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$
 onto  $W$ .

## Exercise: Condition Number Analysis

#### Problem

Consider the symmetric matrix:  $A_{\epsilon} = \begin{pmatrix} 4 & 2 \\ 2 & 1+\epsilon \end{pmatrix}$ , where  $\epsilon > 0$  is a small parameter. This matrix arises as the Hessian of a quadratic function.

### Part A: Eigenvalue Analysis

- 1. Find the eigenvalues  $\lambda_1$  and  $\lambda_2$  of  $A_{\epsilon}$  in terms of  $\epsilon$ .
- 2. Compute the 2-norm condition number  $\kappa_2(A_{\epsilon}) = \frac{\lambda_{\max}}{\lambda_{\min}}$ .
- 3. What happens to  $\kappa_2(A_\epsilon)$  as  $\epsilon 
  ightarrow 0^+?$

## Part B: Numerical Computation

- For  $\epsilon \in \{1, 0.1, 0.01, 0.001\}$ :
  - **1**. Compute  $A_{\epsilon}^{-1}$  explicitly.
  - 2. Calculate  $\kappa_2(A_\epsilon)$  using  $\kappa_2(A) = ||A||_2 ||A^{-1}||_2.$
  - **3.** Verify your results match the eigenvalue computation from Part A.

#### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

## 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
- Conditioning
- Exercices

## 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
- Conditioning
- Exercices

## 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:

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

### 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:

$$abla f(\mathbf{x}_0) = \left( Df(\mathbf{x}_0) [\mathbf{e}_1], Df(\mathbf{x}_0) \mathbf{e}_1, \dots, Df(\mathbf{x}_0) \mathbf{e}_1 
ight)^\mathsf{T}$$

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:

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

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

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

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

 $\rightarrow$  The gradient is a vector containing all the partial derivatives of the function at the point  $\textbf{x}_{0}.$ 

## 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:

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

#### Derivative of a composition

Two cases:

- f = h ∘ g with: h : ℝ → ℝ and g : ℝ<sup>d</sup> → ℝ. The gradient of f can be computed using the chain rule: ∇f(x) = h'(g(x))∇g(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:

$$\mathbf{H}(\mathbf{x}_{0}) = \begin{pmatrix} \frac{\partial^{2} f}{\partial x_{1}^{2}}(\mathbf{x}_{0}) & \frac{\partial^{2} f}{\partial x_{1} \partial x_{2}}(\mathbf{x}_{0}) & \cdots & \frac{\partial^{2} f}{\partial x_{1} \partial x_{d}}(\mathbf{x}_{0}) \\ \frac{\partial^{2} f}{\partial x_{2} \partial x_{1}}(\mathbf{x}_{0}) & \frac{\partial^{2} f}{\partial x_{2}^{2}}(\mathbf{x}_{0}) & \cdots & \frac{\partial^{2} f}{\partial x_{2} \partial x_{d}}(\mathbf{x}_{0}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^{2} f}{\partial x_{d} \partial x_{1}}(\mathbf{x}_{0}) & \frac{\partial^{2} f}{\partial x_{d} \partial x_{2}}(\mathbf{x}_{0}) & \cdots & \frac{\partial^{2} f}{\partial x_{d}^{2}}(\mathbf{x}_{0}) \end{pmatrix}$$

 $\rightarrow$  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:  $\mathbf{H}(\mathbf{x}_0) = \mathbf{H}(\mathbf{x}_0)^{\mathsf{T}}$ .
- 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 x<sub>0</sub>.
  - 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 **H**(*x*, *y*). Analyse local curvature at the stationary point(s).

#### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

## 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

# 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:

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

### Jacobian matrix

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

$$\mathbf{J}_{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_{1}}{\partial x_{1}} & \frac{\partial f_{1}}{\partial x_{2}} & \dots & \frac{\partial f_{1}}{\partial x_{d}} \\ \frac{\partial f_{2}}{\partial x_{1}} & \frac{\partial f_{2}}{\partial x_{2}} & \dots & \frac{\partial f_{2}}{\partial x_{d}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_{p}}{\partial x_{1}} & \frac{\partial f_{p}}{\partial x_{2}} & \dots & \frac{\partial f_{p}}{\partial x_{d}} \end{pmatrix} \in \mathbb{R}^{p \times d}$$

 $\rightarrow$  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} = egin{pmatrix} 
abla f_1(\mathbf{x})^T \mathbf{v} \\

abla f_2(\mathbf{x})^T \mathbf{v} \\
\vdots \\

abla 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 **v** at the point **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:

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

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  $\mathbf{J}_{g}(\mathbf{y}) \in \mathbb{R}^{d \times m}$ .

Introduction Linear Algebra Differentiation

### 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:

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

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

Introduction Linear Algebra Differentiation

## Exercise 1: Basic Chain Rule

### Problem

Given:

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

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

**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 $J_g(y)$

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

### Step 3: Apply chain rule

$$\nabla h(\mathbf{y}) = \mathbf{J}_g(\mathbf{y})^T \nabla f(g(\mathbf{y})) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot 2g(\mathbf{y})$$
$$= 2 \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} y_1 + y_2 \\ y_1 - y_{24} \end{pmatrix} = 2 \begin{pmatrix} 2y_1 \\ 2y_2 \end{pmatrix} = \begin{pmatrix} 4y_1 \\ 4y_2 \end{pmatrix}$$

# 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:

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

# 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  $f : \mathbb{R}^n \to \mathbb{R}$ 

- $\mathbf{A} \in \mathbb{R}^{n imes n}$  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 **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 $J_g(y)$

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

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

### Step 3: Apply chain rule

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

## 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:

$$\mathbf{C}^{T}\mathbf{A}\mathbf{C} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 7 & 6 \\ 6 & 9 \end{pmatrix}$$
$$\mathbf{C}^{T}\mathbf{b} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ {}_{48} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

Thoughouse

# 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  $\nabla f$  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

#### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

## 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

## Fréchet derivative – Definition

#### Fréchet differentiability

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

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

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

### Gateaux derivative (alternative characterization)

If f is Fréchet differentiable at **X**, then for any **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  $\frac{d}{dt}\Big|_{t=0} f(\mathbf{X} + t\mathbf{V})$  is not linear in  $\boldsymbol{\mathbb{V}}$ , 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 **X** in direction **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  $Tr(\cdot)$  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:**  $f(\mathbf{X}) = \|\mathbf{X}\|_F^2 = \text{Tr}(\mathbf{X}^T\mathbf{X})$ Using Gateaux derivative:

$$Df(\mathbf{X})[\mathbf{V}] = \frac{d}{dt} \Big|_{t=0} \operatorname{Tr}((\mathbf{X} + t\mathbf{V})^{\mathsf{T}}(\mathbf{X} + t\mathbf{V}))$$
$$= \frac{d}{dt} \Big|_{t=0} [\operatorname{Tr}(\mathbf{X}^{\mathsf{T}}\mathbf{X}) + 2t\operatorname{Tr}(\mathbf{X}^{\mathsf{T}}\mathbf{V}) + t^{2}\operatorname{Tr}(\mathbf{V}^{\mathsf{T}}\mathbf{V})]$$
$$= 2\operatorname{Tr}(\mathbf{X}^{\mathsf{T}}\mathbf{V})$$

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

Tł

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

$$Df(\mathbf{X})[\mathbf{V}] = \left. \frac{d}{dt} \right|_{t=0} \log \det(\mathbf{X} + t\mathbf{V}) = \operatorname{Tr}(\mathbf{X}^{-1}\mathbf{V})$$
  
herefore:  $\nabla f(\mathbf{X}) = \mathbf{X}^{-\mathsf{T}}$ 

#### 1. Introduction

- Course organization
- The setup

### 2. Linear Algebra

- Vectors and matrices
- Matrix decompositions
- Conditioning
- Exercices

# 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:  $f : \mathbb{R}^{m,n} \to \mathbb{R}^{p,q}$

# 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:

```
\operatorname{vec}(Df(\mathbf{X})[\mathbf{V}]) = \mathbf{M}_{\mathbf{X}}\operatorname{vec}(\mathbf{V})
```

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

# Vectorization identities

## Key identities for matrix calculus

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

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:

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

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

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

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

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

Introduction Linear Algebra Differentiation

## 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:

- $Dg(\mathbf{Y})[\mathbf{W}] = \mathbf{Y}\mathbf{W} + \mathbf{W}\mathbf{Y}$
- $Df(\mathbf{X})[\mathbf{V}] = \mathbf{V}$  (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  $Df(\mathbf{X})[\mathbf{V}] = \text{Tr}(\nabla f(\mathbf{X})^{\mathsf{T}}\mathbf{V})$
- Matrix-to-matrix: Linear operator Df(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