TC2: Optimization for Machine Learning

Master of Science in AI and Master of Science in Data Science

@ UPSaclay

2025/2026.

Faïcel Chamroukhi

université PARIS-SACLAY

SystemX INSTITUT DE RECHERCHE TECHNOLOGIQUE

chamroukhi.com

week 2 : November 13, 2025.

- **Continuous Optimization ;**

↪ **preparing to (Gradient) Descent methods**

Outline

université PARIS-SACLAY

1. Continuous Optimization
   - Formulation and basic concepts
   - Examples in Machine Learning

2. Optimization Concepts in $\mathbb{R}^n$
   - General form of convexity in $\mathbb{R}^n$
   - Example : Convexity of the Least Squares
   - Differentiability and Gradient Concepts in $\mathbb{R}^n$
   - Convexity and the Hessian
   - Least Squares Convexity (Exercise)

Continuous Optimization

université PARIS-SACLAY

Problem Formulation :

- Consider **unconstrained** optimization (minimization) problems, where we seek to minimize a function $f(x)$ defined over a domain $\mathcal{D} = \text{dom}\, f$, without any explicit constraints on x :

$$\min_{x \in \mathcal{D}} f(x)$$

- Then for uncosntrained problems, the feasible set is $\mathcal{D}$ the dom$f$.
- If there are restrictions on x (e.g., x must satisfy $g_i(x) \leq 0$ for certain constraint functions $g_i(x)$), the problem is said to be **constrained**.

Continuous Optimization Definition :

- Continuous optimization refers to optimization problems where the objective function $f(x)$ is defined over a continuous domain $\mathcal{D} \subseteq \mathbb{R}^n$.
- In these problems, the feasible set is typically **uncountably infinite** since x can take any real values within $\mathcal{D}$.

Continuous vs. Discrete Optimization

université PARIS-SACLAY

- In **continuous optimization**, the variables x can take any value within the continuous domain $\mathcal{D} \subseteq \mathbb{R}^n$.
- In **discrete optimization**, the variables are restricted to discrete sets (e.g., integers or binary values).
- Continuous optimization is often more analytically tractable and can be solved using numerical methods (e.g., gradient descent), while discrete optimization often requires combinatorial approaches.
- Many problems in machine learning are naturally formulated as continuous optimization problems (eg. training models using gradient-based methods).

Advantages in Continuous Optimization

- **Use of numerical methods :** The ability to compute gradients (and Hessians) when the function is (twice) differentiable facilitates the application of numerical methods, such as :
  - **Gradient Descent :** Uses first-order derivative information (Gradient) to move in the direction of steepest descent.
  - **Newton's method :** Utilizes second-order derivative information (Hessian) to refine the search direction, potentially leading to faster convergence.
- **Comparison to Discrete Optimization :** Unlike continuous optimization, discrete optimization problems have feasible sets that are finite or countably infinite, making them less amenable to smooth methods. It often requires combinatorial or exhaustive search techniques, which are typically more computationally intensive.
- Many problems in machine learning are naturally formulated as continuous optimization problems (such as training models using gradient-based methods).

Advantages in Continuous Optimization

- **Use of numerical methods :** The ability to compute gradients (and Hessians) when the function is (twice) differentiable facilitates the application of numerical methods, such as :
  - **Gradient Descent :** Uses first-order derivative information (Gradient) to move in the direction of steepest descent.
  - **Newton's method :** Utilizes second-order derivative information (Hessian) to refine the search direction, potentially leading to faster convergence.
- **Comparison to Discrete Optimization :** Unlike continuous optimization, discrete optimization problems have feasible sets that are finite or countably infinite, making them less amenable to smooth methods. It often requires combinatorial or exhaustive search techniques, which are typically more computationally intensive.
- Many problems in machine learning are naturally formulated as continuous optimization problems (such as training models using gradient-based methods).

Examples of Continuous Optimization in Machine Learning

université PARIS-SACLAY

- **Linear Regression :**
  - **Objective Function :** the mean squared error between predicted and actual values.
  - $\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{2}\|y - Xw\|^2$, where $y \in \mathbb{R}^m$ is the target variable, $X \in \mathbb{R}^{m \times n}$ is the design matrix, and $w \in \mathbb{R}^n$ is the continuous vector of parameters we aim to optimize.
- **Logistic Regression :**
  - **Objective Function :** (log)-likelihood of observations in maximization, or by equivalence the "negative" log-likelihood, in minimization
  - $\min_{w \in \mathbb{R}^n} f(w) = -\sum_{i=1}^{m} \left(y_i \log(\sigma(X_i w)) + (1 - y_i) \log(1 - \sigma(X_i w))\right)$, where $y_i \in \{0, 1\}$ represents binary labels, X_i is the i-th row of the design matrix X, and σ is the sigmoid function. The vector of continuous variables $w \in \mathbb{R}^n$ represents model parameters.
- **Neural Network Training : Objective Function :** Minimize a loss function, such as cross-entropy or MSE, over the network parameters w (the weights).

  $\min_{w \in \mathbb{R}^n} f(w) = \sum_{i=1}^{m} L(y_i, \hat{y}_i(w))$, where L is the loss function, y_i are the true outputs, $\hat{y}_i(w)$ the predictions, and $w \in \mathbb{R}^n$ represents all network parameters.

Examples of Continuous Optimization in Machine Learning

université PARIS-SACLAY

- **Linear Regression :**
  - **Objective Function :** the mean squared error between predicted and actual values.
  - $\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{2}\|y - Xw\|^2$, where $y \in \mathbb{R}^m$ is the target variable, $X \in \mathbb{R}^{m \times n}$ is the design matrix, and $w \in \mathbb{R}^n$ is the continuous vector of parameters we aim to optimize.
- **Logistic Regression :**
  - **Objective Function :** (log)-likelihood of observations in maximization, or by equivalence the "negative" log-likelihood, in minimization
  - $\min_{w \in \mathbb{R}^n} f(w) = -\sum_{i=1}^{m} \left(y_i \log(\sigma(X_i w)) + (1 - y_i) \log(1 - \sigma(X_i w))\right)$, where $y_i \in \{0, 1\}$ represents binary labels, X_i is the i-th row of the design matrix X, and σ is the sigmoid function. The vector of continuous variables $w \in \mathbb{R}^n$ represents model parameters.
- **Neural Network Training : Objective Function :** Minimize a loss function, such as cross-entropy or MSE, over the network parameters w (the weights).

  $\min_{w \in \mathbb{R}^n} f(w) = \sum_{i=1}^{m} L(y_i, \hat{y}_i(w))$, where L is the loss function, y_i are the true outputs, $\hat{y}_i(w)$ the predictions, and $w \in \mathbb{R}^n$ represents all network parameters.

Examples of Continuous Optimization in Machine Learning

université PARIS-SACLAY

- **Linear Regression :**
  - **Objective Function :** the mean squared error between predicted and actual values.
  - $\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{2}\|y - Xw\|^2$, where $y \in \mathbb{R}^m$ is the target variable, $X \in \mathbb{R}^{m \times n}$ is the design matrix, and $w \in \mathbb{R}^n$ is the continuous vector of parameters we aim to optimize.
- **Logistic Regression :**
  - **Objective Function :** (log)-likelihood of observations in maximization, or by equivalence the "negative" log-likelihood, in minimization
  - $\min_{w \in \mathbb{R}^n} f(w) = -\sum_{i=1}^{m} \left(y_i \log(\sigma(X_i w)) + (1 - y_i) \log(1 - \sigma(X_i w))\right)$, where $y_i \in \{0, 1\}$ represents binary labels, X_i is the i-th row of the design matrix X, and σ is the sigmoid function. The vector of continuous variables $w \in \mathbb{R}^n$ represents model parameters.
- **Neural Network Training : Objective Function :** Minimize a loss function, such as cross-entropy or MSE, over the network parameters w (the weights).

  $\min_{w \in \mathbb{R}^n} f(w) = \sum_{i=1}^{m} L(y_i, \hat{y}_i(w))$, where L is the loss function, y_i are the true outputs, $\hat{y}_i(w)$ the predictions, and $w \in \mathbb{R}^n$ represents all network parameters.

Optimization Concepts in $\mathbb{R}^n$

université PARIS-SACLAY

In multivariate optimization, we consider functions $f : \mathbb{R}^n \to \mathbb{R}$ with variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$.

Problem Formulation :

- Objective : Minimize a function $f(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^n$.
- Common form :

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{subject to constraints}$$

- Applications in machine learning, such as least squares regression and logistic regression, often use unconstrained multivariate optimization.

Example : Least Squares

- Given a design matrix $X \in \mathbb{R}^{m \times n}$ and a target vector $\mathbf{y} \in \mathbb{R}^m$, least squares seeks to find $\mathbf{w} \in \mathbb{R}^n$ that minimizes :

$$f(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \left(y_i - \mathbf{x}_i^\top \mathbf{w} \right)^2$$

Optimization Concepts in $\mathbb{R}^n$

In multivariate optimization, we consider functions $f : \mathbb{R}^n \to \mathbb{R}$ with variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$.

Problem Formulation :

- Objective : Minimize a function $f(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^n$.
- Common form :

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \quad \text{subject to constraints}$$

- Applications in machine learning, such as least squares regression and logistic regression, often use unconstrained multivariate optimization.

Example : Least Squares

- Given a design matrix $X \in \mathbb{R}^{m \times n}$ and a target vector $\mathbf{y} \in \mathbb{R}^m$, least squares seeks to find $\mathbf{w} \in \mathbb{R}^n$ that minimizes :

$$f(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \left(y_i - \mathbf{x}_i^\top \mathbf{w} \right)^2$$

Convexity in $\mathbb{R}^n$

université PARIS-SACLAY

A set $S \subset \mathbb{R}^n$ is convex if, for any $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in S$$

Convex Functions : A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Example : Least Squares Convexity

Show the least squares function :

$$f(w) = \frac{1}{2}\|y - Xw\|^2$$

where $y \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$, and $w \in \mathbb{R}^n$, is convex.

Convexity in $\mathbb{R}^n$

A set $S \subset \mathbb{R}^n$ is convex if, for any $\mathbf{x}, \mathbf{y} \in S$ and $\lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in S$$

Convex Functions : A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$

Example : Least Squares Convexity

Show the least squares function :

$$f(w) = \frac{1}{2}\|y - Xw\|^2$$

where $y \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$, and $w \in \mathbb{R}^n$, is convex.

Example : Convexity of the Least Squares Function

université PARIS-SACLAY

Let $f(w) = \frac{1}{2}\|y - Xw\|^2$, where $y \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$, and $w \in \mathbb{R}^n$.
To prove $f(w)$ is convex, we show :

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2), \quad \forall w_1, w_2 \in \mathbb{R}^n,\ \lambda \in [0,1].$$

Proof :

We have : $f(\lambda w_1 + (1-\lambda)w_2) = \frac{1}{2}\|y - X(\lambda w_1 + (1-\lambda)w_2)\|^2$

$$= \frac{1}{2}\|\lambda(y - Xw_1) + (1-\lambda)(y - Xw_2)\|^2 \quad \text{(by linearity of } X\text{)}$$

Using the property of the squared norm : $\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2u^T v$,

$$= \frac{1}{2}\left(\lambda^2\|y - Xw_1\|^2 + (1-\lambda)^2\|y - Xw_2\|^2 + 2\lambda(1-\lambda)(y - Xw_1)^T(y - Xw_2)\right).$$

Compared with : $\lambda f(w_1) + (1-\lambda) f(w_2) = \frac{1}{2}\left(\lambda\|y - Xw_1\|^2 + (1-\lambda)\|y - Xw_2\|^2\right)$.
The difference is (after some simple calculations and factoring out $\lambda(1-\lambda)$) :

$$f(\lambda w_1 + (1-\lambda)w_2) - (\lambda f(w_1) + (1-\lambda) f(w_2)) = -\frac{\lambda(1-\lambda)}{2}\|Xw_1 - Xw_2\|^2.$$

Since $\|Xw_1 - Xw_2\|^2 \geq 0$ and $\lambda(1-\lambda) \geq 0$ for $\lambda \in [0,1]$, the difference is negative.
Hence :

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2).$$

CQFD.

Example : Convexity of the Least Squares Function

université PARIS-SACLAY

Let $f(w) = \frac{1}{2}\|y - Xw\|^2$, where $y \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$, and $w \in \mathbb{R}^n$.
To prove $f(w)$ is convex, we show :

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2), \quad \forall w_1, w_2 \in \mathbb{R}^n,\ \lambda \in [0,1].$$

Proof :

We have : $f(\lambda w_1 + (1-\lambda)w_2) = \frac{1}{2}\|y - X(\lambda w_1 + (1-\lambda)w_2)\|^2$

$$= \frac{1}{2}\|\lambda(y - Xw_1) + (1-\lambda)(y - Xw_2)\|^2 \quad \text{(by linearity of } X\text{)}$$

Using the property of the squared norm : $\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2u^T v$,

$$= \frac{1}{2}\left(\lambda^2\|y - Xw_1\|^2 + (1-\lambda)^2\|y - Xw_2\|^2 + 2\lambda(1-\lambda)(y - Xw_1)^T(y - Xw_2)\right).$$

Compared with : $\lambda f(w_1) + (1-\lambda) f(w_2) = \frac{1}{2}\left(\lambda\|y - Xw_1\|^2 + (1-\lambda)\|y - Xw_2\|^2\right)$.
The difference is (after some simple calculations and factoring out $\lambda(1-\lambda)$) :

$$f(\lambda w_1 + (1-\lambda)w_2) - (\lambda f(w_1) + (1-\lambda) f(w_2)) = -\frac{\lambda(1-\lambda)}{2}\|Xw_1 - Xw_2\|^2.$$

Since $\|Xw_1 - Xw_2\|^2 \geq 0$ and $\lambda(1-\lambda) \geq 0$ for $\lambda \in [0,1]$, the difference is negative.
Hence :

$$f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2).$$

CQFD.

Gradient in the Multivariate Case :

- The gradient $\nabla f(\mathbf{x})$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the vector of partial derivatives :

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- The gradient vector serves as a fundamental tool in optimization, providing both the direction of steepest descent and the magnitude of change needed to iteratively adjust $\mathbf{w}$ for minimizing the objective function.

Example : Gradient of Least Squares :

- For a least squares function $f(\mathbf{w}) = \frac{1}{2}\|y - X\mathbf{w}\|^2$, the gradient with respect to $\mathbf{w}$ is [proof left as an Exercise] :

$$\nabla f(\mathbf{w}) = -X^\top (y - X\mathbf{w})$$

Gradient in the Multivariate Case :

- The gradient $\nabla f(\mathbf{x})$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the vector of partial derivatives :

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- The gradient vector serves as a fundamental tool in optimization, providing both the direction of steepest descent and the magnitude of change needed to iteratively adjust $\mathbf{w}$ for minimizing the objective function.

Example : Gradient of Least Squares :

- For a least squares function $f(\mathbf{w}) = \frac{1}{2}\|y - X\mathbf{w}\|^2$, the gradient with respect to $\mathbf{w}$ is [proof left as an Exercise] :

$$\nabla f(\mathbf{w}) = -X^\top (y - X\mathbf{w})$$

Hessian Matrix in the Multivariate Case :

- The Hessian matrix $H_f(\mathbf{x})$ of a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is the matrix of second-order partial derivatives :

$$H_f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- The Hessian matrix provides information about the curvature of f at $\mathbf{x}$ and is essential in analyzing convexity and optimizing the function more effectively.
- The Hessian matrix also enables second-order optimization methods, such as Newton's method, to achieve faster convergence by incorporating information about how the gradient changes.

Hessian of Least Squares : [Proof left as an Exercice]

- For the least squares function $f(\mathbf{w}) = \frac{1}{2}\|y - X\mathbf{w}\|^2$, the Hessian with respect to $\mathbf{w}$ is :

$$H_f(\mathbf{w}) = X^\top X$$

- This Hessian captures the curvature of the least squares objective and is constant, as f is a quadratic function in $\mathbf{w}$.

Convexity Condition for Twice-Differentiable Functions :

- A twice-differentiable function $f(\mathbf{x})$ is convex iff its Hessian $\nabla^2 f(\mathbf{x})$ is **positive semi-definite** for all $\mathbf{x} \in \mathbb{R}^n$, i.e., $\nabla^2 f(\mathbf{x}) \succeq 0, \forall \mathbf{x} \in \mathbb{R}^n$, where $\nabla^2 f(\mathbf{x}) \succeq 0$ indicates that for any vector $\mathbf{z} \in \mathbb{R}^n$,

$$\mathbf{z}^\top \nabla^2 f(\mathbf{x})\, \mathbf{z} \geq 0.$$

- This condition indicates that that all eigenvalues of $\nabla^2 f(\mathbf{x})$ are non-negative.

Interpretation :

- The positive semi-definiteness of $\nabla^2 f(\mathbf{x})$ means that the Hessian matrix does not have any negative eigenvalues, ensuring non-negative curvature of the function f in all directions

Theorem : A is positive definite if and only if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

Proof : Assume, for contradiction, that there exists $\mathbf{x} \neq 0$ such that $\mathbf{x}^\top A\mathbf{x} \leq 0$ and A is positive definite.
Since A is symmetric and positive definite, there exists an orthogonal matrix Q (where $Q^\top Q = I$) such that $A = Q^\top \Lambda Q$, where Λ is a diagonal matrix with entries $\Lambda_{ii} = \lambda_i > 0$ (the eigenvalues of A).
Now, let $\mathbf{y} \neq 0$ be such that $\mathbf{x} = Q^\top \mathbf{y}$.
Then,

$$0 \geq \mathbf{x}^\top A\mathbf{x} = \mathbf{y}^\top QAQ^\top \mathbf{y} = \mathbf{y}^\top QQ^\top \Lambda QQ^\top \mathbf{y} = \mathbf{y}^\top \Lambda \mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 > 0.$$

This is a contradiction, as the right side is strictly positive (since $\lambda_i > 0$ and $\mathbf{y} \neq 0$). Therefore, if A is positive definite, it must hold that $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$.
Conversely, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$, this implies that all eigenvalues of A are positive, making A positive definite.

Theorem : A is positive definite if and only if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

Proof : Assume, for contradiction, that there exists $\mathbf{x} \neq 0$ such that $\mathbf{x}^\top A\mathbf{x} \leq 0$ and A is positive definite.
Since A is symmetric and positive definite, there exists an orthogonal matrix Q (where $Q^\top Q = I$) such that $A = Q^\top \Lambda Q$, where Λ is a diagonal matrix with entries $\Lambda_{ii} = \lambda_i > 0$ (the eigenvalues of A).
Now, let $\mathbf{y} \neq 0$ be such that $\mathbf{x} = Q^\top \mathbf{y}$.
Then,

$$0 \geq \mathbf{x}^\top A\mathbf{x} = \mathbf{y}^\top QAQ^\top \mathbf{y} = \mathbf{y}^\top QQ^\top \Lambda QQ^\top \mathbf{y} = \mathbf{y}^\top \Lambda \mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 > 0.$$

This is a contradiction, as the right side is strictly positive (since $\lambda_i > 0$ and $\mathbf{y} \neq 0$). Therefore, if A is positive definite, it must hold that $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$.
Conversely, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$, this implies that all eigenvalues of A are positive, making A positive definite.

Ex : Least Squares Convexity (Exercise)

université PARIS-SACLAY

- Show that the least squares objective function in linear regression is convex iff $X^\top X$ is positive semi-definite.
- The objective function is : $f(w) = \frac{1}{2}\|y - Xw\|^2 = \frac{1}{2}(y - Xw)^\top (y - Xw)$.
- The gradient is : $\nabla f(w) = -X^\top (y - Xw)$.
- The Hessian is : $\nabla^2 f(w) = X^\top X$, which is :
  - Positive semi-definite because for any vector $v \in \mathbb{R}^n$:

$$v^\top (X^\top X) v = (Xv)^\top (Xv) = \|Xv\|^2 \geq 0.$$

  - Positive definite (and the least squares function strictly convex) if $X^\top X$ has full column rank, ensuring $\|Xv\|^2 > 0$ for all $v \neq 0$.
- Conclusion : $f(w)$ is convex because $X^\top X$ is positive semi-definite. Moreover, $f(w)$ is strictly convex iff $X^\top X$ is positive definite (has full column rank).

Full column rank implies $Xv = 0$ only when $v = 0$, ensuring $\|Xv\|^2 > 0$ for $v \neq 0$. A matrix $X \in \mathbb{R}^{m \times n}$ is said to have **full column rank** if its columns are linearly independent. Equivalently, $Xv = 0$ implies $v = 0$ for all $v \in \mathbb{R}^n$.

Ex : Least Squares Convexity (Exercise)

université PARIS-SACLAY

- Show that the least squares objective function in linear regression is convex iff $X^\top X$ is positive semi-definite.
- The objective function is : $f(w) = \frac{1}{2}\|y - Xw\|^2 = \frac{1}{2}(y - Xw)^\top (y - Xw)$.
- The gradient is : $\nabla f(w) = -X^\top (y - Xw)$.
- The Hessian is : $\nabla^2 f(w) = X^\top X$, which is :
  - Positive semi-definite because for any vector $v \in \mathbb{R}^n$:

$$v^\top (X^\top X) v = (Xv)^\top (Xv) = \|Xv\|^2 \geq 0.$$

  - Positive definite (and the least squares function strictly convex) if $X^\top X$ has full column rank, ensuring $\|Xv\|^2 > 0$ for all $v \neq 0$.
- Conclusion : $f(w)$ is convex because $X^\top X$ is positive semi-definite. Moreover, $f(w)$ is strictly convex iff $X^\top X$ is positive definite (has full column rank).

Full column rank implies $Xv = 0$ only when $v = 0$, ensuring $\|Xv\|^2 > 0$ for $v \neq 0$. A matrix $X \in \mathbb{R}^{m \times n}$ is said to have **full column rank** if its columns are linearly independent. Equivalently, $Xv = 0$ implies $v = 0$ for all $v \in \mathbb{R}^n$.

Ex : Least Squares Convexity (Exercise)

université PARIS-SACLAY

- Show that the least squares objective function in linear regression is convex iff $X^\top X$ is positive semi-definite.
- The objective function is : $f(w) = \frac{1}{2}\|y - Xw\|^2 = \frac{1}{2}(y - Xw)^\top (y - Xw)$.
- The gradient is : $\nabla f(w) = -X^\top (y - Xw)$.
- The Hessian is : $\nabla^2 f(w) = X^\top X$, which is :
  - Positive semi-definite because for any vector $v \in \mathbb{R}^n$:

$$v^\top (X^\top X)v = (Xv)^\top (Xv) = \|Xv\|^2 \geq 0.$$

  - Positive definite (and the least squares function strictly convex) if $X^\top X$ has full column rank, ensuring $\|Xv\|^2 > 0$ for all $v \neq 0$.
- Conclusion : $f(w)$ is convex because $X^\top X$ is positive semi-definite. Moreover, $f(w)$ is strictly convex iff $X^\top X$ is positive definite (has full column rank).

Full column rank implies $Xv = 0$ only when $v = 0$, ensuring $\|Xv\|^2 > 0$ for $v \neq 0$. A matrix $X \in \mathbb{R}^{m \times n}$ is said to have **full column rank** if its columns are linearly independent. Equivalently, $Xv = 0$ implies $v = 0$ for all $v \in \mathbb{R}^n$.