⑤ Differential

Derivative

$$\begin{cases} f: U \subset \mathbb{R} \to \mathbb{R} \\ U \text{ open set} \end{cases}$$

f admit a derivative at a point x of U

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = a$$

where a is finite

$a = f'(x)$ derivative of f at x

Equivalently,

$$f(x+h) = f(x) + f'(x)h + h\varepsilon(h)$$

where $\lim_{h \to 0} \varepsilon(h) = 0$

This means that the function f is approximated by the affine function $L(h) = f'(x)h$ in the neighborhood of x

$f(x+h) \approx f(x) + L(h)$ where h is small Geometrically

Let E, F be two vectorial spaces and

$$f: E \longrightarrow F$$

Goal: extend the notion of derivative for f

Def: Let U be an open set of E and $x \in E$.

We say that f is differentiable at x if there exists a continuous and linear mapping L from E into F such that

$$f(x+h) = f(x) + L(h) + \|h\|_E \varepsilon(h)$$

$\forall h \in E$ where $\lim_{\|h\| \to 0} |\varepsilon(h)|_F = 0$

When it is the case, L is called the differential of f at x
we will denote L by df_x

$$\begin{cases} df_x : E \to F \\ \text{linear, continuous} \end{cases}$$

$$df_x \in \underbrace{\mathcal{L}(E, F)}_{\text{space of linear and continuous function from } E \text{ to } F.}$$

if h is small, then

$$f(x+h) \simeq \underbrace{f(x) + df_x(h)}_{\text{affine approximation of } f \text{ on the neighborhood of } x.}$$

Link with the derivative

$$E = F = \mathbb{R}$$

$$f : \mathbb{R} \to \mathbb{R}$$

If f has a derivative $f'(x)$ at x then

$$f(x+h) = f(x) + f'(x)h + h\varepsilon(h)$$

let $L(h) = f'(x)h$

L is linear and continuous. from $\mathbb{R} \to \mathbb{R}$

Moreover $\lim_{h \to 0} \varepsilon(h) = 0$

So $L = df_x$

$$df_x : \mathbb{R} \to \mathbb{R}$$

$$h \mapsto df_x(h) = f'(x)h$$

Examples:

① $f : E \to F$

$\forall x \quad f(x) = c \quad$ where $c \in F$

$f(x+h) - f(x) = 0$

let $L(h) = 0 \quad \forall h$

$\varepsilon = 0$

$f(x+h) = f(x) + L(h) + \|h\| \varepsilon(h)$

$df_x : E \to F$

$h \mapsto 0$

$df_x = 0 \qquad df_x(h) = 0 \quad \forall h$

②) Let f be continuous and linear from E to F.

$f(x+h) - f(x) = f(x) + f(h) - f(x) = f(h)$

f is linear and continuous

so $df_x = f$

③ Let f be a bilinear continuous mapping from $E \times E$ into

II- Link with partial derivatives.

Let E be a finite normed vectorial space equipped with a basis $(e_1, \dots, e_n)$

Let $f : E \to \mathbb{R}$

Def: (partial derivative)

Let U be an open set of E and $x \in U$

For $i = 1, \dots, n$ let

$g : \mathbb{R} \to \mathbb{R}$, $h \mapsto f(x + h e_i)$

if g_i admits a derivative at 0 then we say that f has a partial derivative at x in the direction e_i. we denote $\partial_i f(x)$ this derivative.

$$\partial_i f(x) = \lim_{h \to 0} \frac{g_i(h) - g_i(0)}{h}$$
$$= \lim_{h \to 0} \frac{f(x + h e_i) - f(x)}{h}$$

Assume that f is differentiable at a point x of U

then, for any h

$$f(x+h) = f(x) + df_x(h) + \varepsilon(h)$$

where $\frac{\varepsilon(h)}{|h|} \longrightarrow 0, \quad |h| \to 0$

Let $h = t e_i$ with $t \in \mathbb{R}$

$$f(x + t e_i) = f(x) + df_x(t e_i) + \varepsilon(t e_i)$$

but by linearity

$$df_x(t e_i) = t\, df_x(e_i)$$

So that $$\frac{f(x + t e_i) - f(x)}{t} = df_x(e_i) + \frac{\varepsilon(t e_i)}{t}$$

Moreover

$$\lim_{t \to 0} \frac{\varepsilon(t e_i)}{t} = \lim_{t \to 0} \frac{\varepsilon(t e_i)}{|t e_i|} = 0$$

then

$$\lim_{t \to 0} \frac{f(x + t e_i) - f(x)}{t} = df_x(e_i)$$

Prop: if f is differentiable at x then it admits partial derivative in any direction e_i and

$$\partial_i f(x) = df_x(e_i)$$

Moreover

$$df_x(h) = \sum_{i=1}^{n} \partial_i f(x) h_i$$

where h_i are the coordinates of h in basis $(e_1, \dots, e_n)$

Indeed

$$df_x(h) = df_x\left(\sum_{i=1}^{n} h_i e_i\right) = \sum_{i=1}^{n} h_i \, df_x(e_i)$$

$$= \sum_{i=1}^{n} h_i \, \partial_i f(x)$$

Now assume that

$$f : E \to \mathbb{R}^m$$

where E is finite dimensional with a basis $(e_1, \dots, e_n)$

We can describe f through its so called components:

$j = 1, \dots, m$ $\quad f_j : E \to \mathbb{R}$

$$x \mapsto f_j(x)$$

We can compute partial direction of f_i if they exist

Def: Let $x \in U$ open set of E and $f : U \to \mathbb{R}^m$.

if the components of f admit partial derivative in all direction

then we can define the Jacobian matrix, denoted

$Df(x)$

in the special case where $m = n$
we call jacobien the determinant of this matrix

Let us assume that f is differentiable at x

$$df_x \in \mathcal{L}(E, \mathbb{R}^m)$$

$$\begin{pmatrix} df_x^1(h) \\ \vdots \\ df_x^m(h) \end{pmatrix} = \begin{pmatrix} \sum \partial_i f_1(x)\, h_i \\ \vdots \\ \sum \partial_i f_n(x)\, h_i \end{pmatrix}$$

Prop:

$$df_x(h) = Df(x)\, h$$

III – Gradient

Let E be an Hilbert space equiped with an inner product $\langle \cdot, \cdot \rangle$ and $f : E \to \mathbb{R}$

if f is differentiable at a point x of an open set U of E then.

$$\begin{cases} df_x \in \mathcal{L}(E, \mathbb{R}) := \mathcal{L}(E) \\ df_x \text{ is a linear continuous form on } E. \end{cases}$$

In an Hilbert space a linear form L on E is uniquely determined an element of E:

$$L(h) = \langle a, h \rangle$$

(Riesz representation theorem)

Def: we call gradient of f at x the element of E, denoted $\nabla f(x)$, such that

$$\forall h \in E \quad df_x(h) = \langle \nabla f(x), h \rangle$$

Let E be finite dimensioned with a basis $(e_1, \dots, e_n)$

Assume that $f : \mathcal{O} \to \mathbb{R}$ is differentiable at x

then $df_x(h) = Df(x)\,h$

$Df(x)$ is a matrix of size $(1, n)$

$df_x(h) = \langle Df(x)^T, h \rangle$

By identification

$$\nabla f(x) = Df(x)^T$$

In other words

$$\nabla f(x)_i = \partial_i f(x).$$

Ex. 6. 1

Exercise 6.1: First and Second Differentials of a Bilinear Map

Let f be a continuous bilinear map from E^2 to F. We equip the product space E^2 with the norm

$$|(h,k)|_{E^2} = \sqrt{|h|_E^2 + |k|_E^2}, \quad \forall (h,k) \in E^2.$$

1. Show that f is differentiable on E^2 and compute its differential.
2. Show that f is twice differentiable and compute its second differential.
3. Suppose E is finite-dimensional and $F = \mathbb{R}$. Let $(e_1, \dots, e_p)$ be an orthonormal basis for E. We equip E with the inner product

$$\langle x, y \rangle = \sum_{i=1}^{p} x_i y_i,$$

where $(x_i)_{i=1}^p$ and $(y_i)_{i=1}^p$ are the coordinates of x and y in this basis, respectively.

   (a) Specify the gradient and Hessian of f in this basis of E.
   (b) Write the first and second differentials in terms of the gradient and Hessian, respectively.

1) $f(x+h) = f(x) + L(h) + R(h)$

with L : linear and continuous

$$\frac{R(h)}{\|h\|} \xrightarrow[\|h\| \to 0]{} 0$$

$x = (x_1, x_2)$ with $x_1 \in E$, $x_2 \in \mathcal{O}$

$f(x) = f(x_1, x_2)$

$x = (x_1, x_2) \quad h = (h_1, h_2)$

$f(x+h) - f(x) = f(x_1+h_1, x_2+h_2) - f(x_1, x_2)$

Since f is bilinear, we can write

$f(x+h) - f(x) = f(x_1, x_2+h_2) + f(h_1, x_2+h_2) - f(x_1, x_2)$

$= f(x_1, h_2) + f(h_1, x_2) + f(h_1, h_2)$

let $L(h) = f(x_1, h_2) + f(h_1, x_2)$

L is linear $\lambda, \mu \in \mathbb{R}$, $h, k \in \mathbb{R} \times \mathbb{R}$

$L(h+k)$

$= f(x_1, h_2 + k_2) + f(h_1 + k_1, x_2)$

$= f(x_1, h_2) + f(x_1, k_2) + f(h_1, x_2) + f(k_1, x_2)$

$= f(x_1, h_2) + f(h_1, x_2) + f(x_1, k_2) + f(k_1, x_2)$

$= L(h) + L(k)$

$L(\lambda h) = f(x_1, \lambda h_2) + f(\lambda h_1, x_2)$

$= \lambda f(x_1, h_2) + \lambda f(h_1, x_2)$

$= \lambda (f(x_1, h_2) + f(h_1, x_2))$

$= \lambda L(h)$

L is also continuous as it is a sum of continuous functionals

$h_1 \to f(h_1, x_2)$

and $h_2 \to f(x_1, h_2)$

Let $E(h) = f(h_1, h_2)$

we study the limit of

$$R(h) = \frac{\|\varepsilon(h_1, h_2)\|_F}{\|h\|_{E\times E}}$$

since f is linear continuous, there exists a positive constant k such that

$$\|\varepsilon(h_1, h_2)\|_F \leq k \|h_1\|_E \|h_2\|_E$$

So $$R(h) \leq k \frac{\|h_1\|_E \|h_2\|_E}{\|h\|_{E\times E}}$$

with $$\|h\|^2_{E\times E} = \|h_1\|^2_E + \|h_2\|^2_E$$

Noticing that $$(a-b)^2 = a^2 + b^2 - 2ab \geq 0$$
$$= ab \leq \frac{1}{2}(a^2 + b^2)$$

we obtain

$$R(h) \leq \frac{k}{2} \|h\|_{E\times E}$$

$$R(h) \xrightarrow[\|h\|_{E\times E} \to 0]{} 0$$

We conclued that f is differentiable at x and

$$df_x \quad E\times E \to F$$
$$(h_1, h_2) \mapsto f(x_1, h_2) + f(h_1, x_2)$$

Exercise 6.4: Linear Model

Consider the generalized linear model

$$Y = X\theta + \epsilon,$$

where the design matrix X is of size $n \times p$ and rank p, the parameters θ are in $\mathbb{R}^p$, and the errors ϵ are distributed according to $\mathcal{N}(0, \Sigma)$ with a positive definite covariance matrix Σ.
We define the generalized least squares criterion used to estimate

$$J(\theta) = (Y - X\theta)^T \Sigma^{-1} (Y - X\theta).$$

1. Justify that this criterion is of class $\mathcal{C}^\infty$.
2. Determine the first and second-order differentials of this criterion.
3. Equip $\mathbb{R}^p$ with the usual inner product. Specify the gradient and the Hessian matrix of J in the canonical basis of $\mathbb{R}^p$.
4. Verify that the Hessian matrix of J is positive definite.

Linear model:

$$Y = X\theta + \varepsilon$$

- Y : n v in $\mathbb{R}$
- $X \in M_{n \times p}$
- $\theta \in \mathbb{R}^p$
- ε : n v in $\mathbb{R}^n$

$\varepsilon \sim N(0, \sigma^2 I)$

MLE of θ minimizes

$J(\theta) = \frac{1}{2}\|Y - X\theta\|^2$

least square criterion

① If we assume that $\varepsilon \sim N(0, \Sigma)$
where Σ is a positive definite matrix
then the MLE of θ minimizes the generalized least squares criterion

$$J(\theta) = (Y - X\theta)^T \Sigma^{-1} (Y - X\theta)$$

Let us show that J is differentiable $h \in \mathbb{R}^n$

$$\begin{aligned} J(\theta + h) &= \frac{1}{2}\|Y - X(\theta + h)\|^2 \\ &= \frac{1}{2}\|Y - X\theta - Xh\|^2 \\ &= \frac{1}{2}\langle Y - X\theta - Xh, Y - X\theta - Xh \rangle \\ &= \frac{1}{2}\left(\|Y - X\theta\|^2 + \|Xh\|^2 - 2\langle Y - X\theta, Xh \rangle\right) \end{aligned}$$

then $J(\theta + h) - J(\theta) = \langle Y - X\theta, Xh \rangle + \frac{1}{2}\|Xh\|^2$

Let $L(h) = \langle X\theta - Y, Xh \rangle = (X\theta - Y)^T Xh$

$$= (X^T(X\theta - Y))^T h$$
$$= \langle X^T(X\theta - Y), h \rangle$$
$$= \langle X^T X\theta - X^T Y, h \rangle$$

L is linear and since it is defined on a finite dimensional space it is continuous

let us study the limit

$$R(h) = \frac{1}{2} \frac{\|Xh\|^2}{\|h\|}$$

quite easy!

as $h \to 0$

$$\|Xh\|^2 = \langle Xh, Xh \rangle$$
$$= h^T X^T X h$$

II- Higher order differentials

E, F

$f : U \to F$

Def

Let U be an open set of E. if f is differentiable on every point x of U, we say that f is differentiable on U.

if f is differentiable on U, we can define on U the mapping

$$df : U \subset E \to \mathcal{L}(E, F)$$
$$x \mapsto df_x$$

Def! let df be defined on $U \subset E$, and $x \in U$.

if df is differentiable at x, then we can define

$d(df)_x$

In this case, we say that f is twice differentiable at x, and we call $d(df)_x$ the second order differential of f at x.

We denote

$$d^2 f_x = d(df)_x$$

Remark:

$$d^2 f_x \in \mathcal{L}(E; \mathcal{L}(E;F))$$

More specifically

Let $h \in E$

then $d^2 f_x(h) \in \mathcal{L}(E;F)$

Let $k \in E$

then $d^2 f_x(h)(k) \in F$

$$g : (E;E) \longrightarrow F$$
$$(h,k) \longrightarrow d^2 f_x(h)(k)$$

g is linear with respect to each argument.

it is also continuous on $E \times E$

We can identify $d^2 f$ with a bilinear continuous mapping

$$\mathcal{L}(E; \mathcal{L}(E;F)) = \mathcal{L}^2(E;F)$$
$$\simeq \mathcal{L}_2(E;F) \text{ space of}$$

So we may use a notation above

By a recursion we can define differential of an arbitrary order n

Let $d^n f_x$ denotes the differential of order n at x

and $d^n f_x \in \mathcal{L}^n(E, F)$

Suite exercice

$$J_n(\theta) = \frac{1}{2}\|Y - X\theta\|^2$$

$$J_n(\theta) = \frac{1}{2}(Y - X\theta)^T \Sigma^{-1}(Y - X\theta)$$

$$J_n(\theta + h) = \frac{1}{2}\|Y - X(\theta + h)\|^2$$

$$= \frac{1}{2}\|Y - X\theta + Xh\|^2$$

$$= \frac{1}{2}\langle Y - X\theta + Xh, Y - X\theta + Xh\rangle \quad \text{fait avant}$$

$$= \langle X^T X\theta - X^T Y, h\rangle + \frac{1}{2}\|Xh\|^2$$

$$\lim \frac{R(h)}{\|h\|^2} = 0 \ ?$$

Let $A = X^T X$

A is symmetric

A is define positive since $h^T A h = h^T X^T X h = \|Xh\|^2 \geq 0$

H is diagonalisable in a orthonormal basis $(u_1, \dots, u_n)$

and it's eigenvalues $\lambda_1, \dots, \lambda_n$ are non negative

Soit $h = \sum_{i=1}^{p} h_i u_i$

$$Ah = A\left(\sum_{i=1}^{p} h_i u_i\right) = \sum_{i=1}^{p} h_i A u_i = \sum_{i=1}^{n} h_i \lambda_i u_i$$

$$h^T A h = \left(\sum_{j=1}^{n} h_j u_j\right)^T \left(\sum_{i=1}^{n} h_i \lambda_i u_i\right)$$

$$= \left(\sum_{j=1}^{n} h_j u_j\right)^T \left(\sum_{i=1}^{n} h_i \lambda_i u_i\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} h_i h_j \lambda_i u_j^T u_i$$

but $u_j^t u_i = \langle u_j, u_i \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$

then $h^T A h = \sum_{i=1}^{p} h_i^2 \lambda_i$

But $\lambda_i \leq \lambda_p \quad \forall i = 1, \dots, p$

so $h^T A h \leq \lambda_p \sum_{i=1}^{p} h_i^2$

But $\sum_{i=1}^{p} h_i^2 = \|h\|^2$

$$h^t A h \leq \lambda_p \|h\|^2$$

$$R(h) = \frac{1}{2} h^T A h \quad \text{with} \quad A = X^T X$$

$$\frac{R(h)}{\|h\|} \leq \frac{\lambda_p}{2} \|h\| \quad \text{where } \lambda_p = \text{largest eigenvalue of } A$$

$$\lim \frac{R(h)}{\|h\|} = 0$$

Conclusion J is differentiable at θ and $dJ_\theta(h) = \langle X^T X\theta - X^T y, h \rangle$ if $h \in \mathbb{R}^p$.

Remark: the gradient of J is:

$$\nabla J(\theta) = X^T X\theta - X^T y \in \mathbb{R}$$

$$dJ_\theta(h) = \langle X^T X\theta - X^T y, h \rangle$$

$$dJ_{\theta+h}(h) - dJ_\theta(h) = \langle X^T X(\theta+h) - X^T y, h \rangle - \langle X^T X\theta - X^T y, h \rangle$$

$$= \langle X^T X\theta + X^T X h - X^T y, h \rangle - \langle X^T X\theta - X^T y, h \rangle$$

$$= \langle X^T X\theta - X^T y, h \rangle + \langle X^T X h, h \rangle - \langle X^T X\theta - X^T y, h \rangle$$

$$= \langle X^T X h, h \rangle$$

$\ell : h \longrightarrow (h \to \langle X^T X h, a \rangle)$ is linear, continuous with values

in $\mathcal{L}(\mathbb{R}^p)$

So L is the differential of dJ

Hence $d^2 J_\theta : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$

$$(h,k) \mapsto \langle X^T X h, k \rangle$$

Condition for $d^2 J(\theta)$ to be definite positive

1) $X^T X$ has only positive eigenvalues
2) $\det(X^T X) > 0$
3) $\ker(X^T X) = \{0\}$
4) $\operatorname{rank}(X) = p$

Exercise 6.5: Logistic Regression

Consider a sample consisting of n independent random variables Y_i following a Bernoulli distribution with parameter p_i. We assume that for each i, the parameter p_i depends on explanatory variables $x_i \in \mathbb{R}^p$ and unknown parameters $\theta \in \mathbb{R}^p$ through the relation $p_i = \sigma(x_i^T \theta)$, where $\sigma(t) = \frac{1}{1+e^{-t}}$.

1. Show that
$$\sigma(t) + \sigma(-t) = 1$$
and that
$$(\log_e \circ \sigma)'(t) = \sigma(-t).$$
2. Determine the likelihood $\mathcal{L}(\theta; y_1, \ldots, y_n)$ of the model.
3. For fixed y_i, set $\ell(\theta) = -\log_e(\mathcal{L}(\theta; y_1, \ldots, y_n))$, the logistic loss function. Show that ℓ is of class $\mathcal{C}^\infty$.
4. Show that the gradient $\nabla\ell(\theta)$ and the Hessian matrix $\nabla^2\ell(\theta)$ of ℓ are given by:
$$\nabla\ell(\theta) = -\sum_{i=1}^{n} x_i(y_i - \sigma(x_i^T\theta)) \quad \text{and} \quad \nabla^2\ell(\theta) = \sum_{i=1}^{n} \sigma(x_i^T\theta)(1 - \sigma(x_i^T\theta)) x_i x_i^T.$$
5. Under what condition(s) is $\nabla^2\ell(\theta)$ a positive definite matrix?

Logistic regression

$$Y_i \in \{0,1\}$$
$$x_i \in \mathbb{R}^p$$

$$\begin{cases} \mathbb{P}(Y_i = 1) = \exp(\sigma(\langle X_i, \theta \rangle)) \\ \sigma(t) = \frac{1}{1+e^{-t}} \text{ is the logit function} \end{cases}$$

Assume Y_i are independant

Likelyhood of $(Y_1, \ldots, Y_n)$: $\mathcal{L}(y_1, \ldots, y_n, \theta) = \prod_{i=1}^{n} \mathbb{P}(Y_i=1)^{y_i} \mathbb{P}(Y_i=0)^{1-y_i}$

$$= \prod_{i=1}^{n} \exp(y_i \sigma(\langle X_i, \theta \rangle)) \cdot \exp((1-y_i)(1-\sigma(\langle X_i, \theta \rangle)))$$

$$= \exp\left(y_i \sigma(\langle x_i, \theta \rangle) + (1-y_i)(1-\sigma(\langle x_i, \theta \rangle)) \right).$$

The negative log-likelihood is

$$\begin{cases} \ell(\theta) = -\log(\mathcal{L}(\theta, y_1, \dots, y_n)) = -\sum_{i=1}^{n} \left[y_i \sigma(\langle x_i, \theta \rangle) + (1-y_i) \right. \\ \qquad \left. \times (1-\sigma(\langle x_i, \theta \rangle)) \right] \end{cases}$$

this function is known as the binary cross-entropy

c'est pas une bonne idée de travailler sur $\ell(\theta+h) - \ell(\theta)$

vaut mieux calculer les dérivées partielles

$$\frac{\partial g(\theta)}{\partial \theta} = \partial \frac{\sigma(\langle x, \theta \rangle)}{\partial \theta}$$

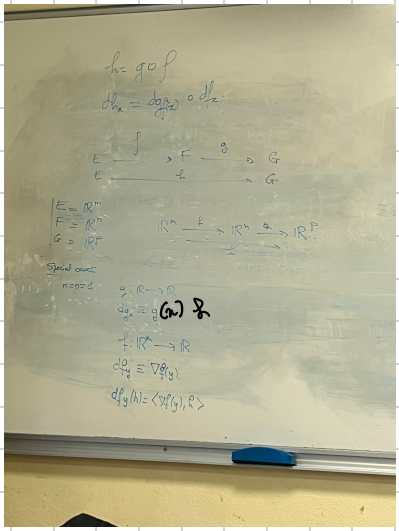
$$\frac{1}{1+e^{-}}$$

deux fonctions différentiables

$g = h \circ f$

$dg_x = dh_{f(x)} \circ df_x$

$g'_x = h'(f(x)) f'(x)$.



$$dh_x = dg_{f(x)} \circ df$$