Vector calculus

K. Bouanane

12 mars 2023

Preliminaries

Function

We say that f is a function if it maps each element of $\mathbb{R}^D$ to a unique element in $\mathbb{R}$. We usually write

$$\begin{aligned} f: \quad & \mathbb{R}^D \mapsto \mathbb{R} \\ & \mathbf{x} \rightarrow f(\mathbf{x}) \end{aligned}$$

Differenciation of univariate function

Let

$$\begin{aligned} f: \quad & \mathbb{R} \mapsto \mathbb{R} \\ & x \to f(x) \end{aligned}$$

be an univariate function. Then the derivative of f at x is given by

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

It refers to the rate of change in the function's f value when moving from x to $x+h$.

Derivation rules

Consider $f, g : \quad \mathbb{R} \mapsto \mathbb{R}$. We have :

1. Product rule : $(f(x)g(x))' = f'(x)g(x) + g'(x)f(x)$
2. Quotient rule : $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2}$
3. Sum rule : $(f(x) + g(x))' = f'(x) + g'(x)$
4. Chain rule : $(g(f(x)))' = (g \circ f)'(x) = g'(f(x)).f'(x)$.

Example : Chain rule

We want to compute the derivative of the function $h(x) = (2x^2 + 1)^4$.

Partial Derivatives and Gradient Vector

Partial derivatives

Consider the multivariate function

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \mathbf{x} \rightarrow f(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.

Partial derivatives

Consider the multivariate function

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \mathbf{x} \to f(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.
A partial derivative of f with respect to $x_i, i = 1..n$ is defined by

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_i + h, x_{i+1}, \ldots, x_n) - f((x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_n)}{h}$$

Gradient vector

The vector containing first order partial derivatives

$$\nabla f = \frac{df}{d\mathbf{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right] \in \mathbb{R}^{1\times n}$$

is called the Gradient or the Jacobian of f.

We can get the partial derivatives using the chain rule.
Example : $f(x, y) = (x + 2y^3)^2$.

Basic rules of partial differentiation

Basic rules for differentiation still apply, but we must take into consideration that gradients are vectors :

Basic rules of partial differentiation

Basic rules for differentiation still apply, but we must take into consideration that gradients are vectors :

1. Product : $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}).g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}}.g(\mathbf{x}) + \frac{\partial g}{\partial \mathbf{x}}.f(\mathbf{x}) \in \mathbb{R}^{1\times n}$.
2. Sum : $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$.
3. Chain rule : $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f} \times \frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$.

Basic rules of partial differentiation

Basic rules for differentiation still apply, but we must take into consideration that gradients are vectors :

1. Product : $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}).g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}}.g(\mathbf{x}) + \frac{\partial g}{\partial \mathbf{x}}.f(\mathbf{x}) \in \mathbb{R}^{1\times n}$.
2. Sum : $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$.
3. Chain rule : $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f} \times \frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$.

In the case of the chain rule, we have to make sure that dimensions are matching.

Example

Consider the function $g(x, y) = (x^2 + y^2)^3$.

Always with the chain rule

consider the bivariate function $f(x_1, x_2) : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $x_1 = x_1(t)$ and $x_2 = x_2(t)$.

Always with the chain rule

consider the bivariate function $f(x_1, x_2) : \mathbb{R}^2 \mapsto \mathbb{R}$ such that $x_1 = x_1(t)$ and $x_2 = x_2(t)$.
To compute the derivative of f with respect to t, we have to apply the chain rule as follows :

$$\frac{df}{dt} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right] \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1} . \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} . \frac{\partial x_2}{\partial t}$$

Chain rule

In general,

Chain rule

In general,

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \mathbf{x} \to f(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, with

$$\begin{aligned} x_i: \quad & \mathbb{R} \mapsto \mathbb{R} \\ & t \to x_i(t) \end{aligned}$$

Chain rule

In general,

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \mathbf{x} \to f(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$, with

$$\begin{aligned} x_i: \quad & \mathbb{R} \mapsto \mathbb{R} \\ & t \to x_i(t) \end{aligned}$$

Then

$$\frac{df}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial t}$$

Example

$f(x, y) = x^2 + 2y$, with $x = \sin t$ and $y = \cos t$.

Chain rule

Now, suppose that we have $f(x, y)$ is such that $x = x(s, t)$ and $y = y(s, t)$. In this case we get

Chain rule

Now, suppose that we have $f(x, y)$ is such that $x = x(s, t)$ and $y = y(s, t)$. In this case we get

$$\frac{df}{dt} = \frac{\partial f}{\partial x}.\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}.\frac{\partial y}{\partial t}$$
$$\frac{df}{ds} = \frac{\partial f}{\partial x}.\frac{\partial x}{\partial s} + \frac{\partial f}{\partial y}.\frac{\partial y}{\partial s}$$

The gradient is obtained by the matrix multiplication :

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial (s,t)} = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right] \begin{bmatrix} \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \end{bmatrix}$$

Example

$f(x, y) = xy$
$x = t^2 + 2s$, $y = 3st$.

Gradients of Vector-Valued Functions

Consider the function

$$\begin{aligned} \mathbf{f}: \quad & \mathbb{R}^n \mapsto \mathbb{R}^m \\ & \mathbf{x} \to \mathbf{f}(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.

Consider the function

$$\begin{aligned} \mathbf{f}: \quad & \mathbb{R}^n \mapsto \mathbb{R}^m \\ & \mathbf{x} \to \mathbf{f}(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.
The corresponding vector of function values is given by

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \tag{1}$$

$\mathbf{f}$ is a vector of functions $[f_1, \ldots, f_m]^\top$ such that $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1..m$.

$\mathbf{f}$ is a vector of functions $[f_1, \ldots, f_m]^\top$ such that $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1..m$.
The differentiation rule for every function f_i is exactly the same as previously.

$\mathbf{f}$ is a vector of functions $[f_1, \ldots, f_m]^\top$ such that $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1..m$.

The differentiation rule for every function f_i is exactly the same as previously. Therefore, the partial derivative of a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}, i = 1..n$, is given as the vector

$\mathbf{f}$ is a vector of functions $[f_1, \ldots, f_m]^\top$ such that $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1..m$.
The differentiation rule for every function f_i is exactly the same as previously. Therefore, the partial derivative of a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}, i = 1..n$, is given as the vector

$$\frac{\partial \mathbf{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} \in \mathbb{R}^m$$

Since every partial derivative $\frac{\partial \mathbf{f}}{\partial x_i}$ is a vector, we obtain the gradient of $\mathbf{f}$ with respect to $\mathbf{x}$ by collecting these vectors and therefore we have :

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Since every partial derivative $\frac{\partial \mathbf{f}}{\partial x_i}$ is a vector, we obtain the gradient of $\mathbf{f}$ with respect to $\mathbf{x}$ by collecting these vectors and therefore we have :

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

This matrix is called the Jacobian of $\mathbf{f}$ and is denoted $\mathbf{J}$.

Example 1

Consider the function $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$, avec $\mathbf{f} = \begin{bmatrix} x^2 - 3y + z \\ x - y^3 + z^2 \end{bmatrix}$

Example 2

Consider the function $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x} \in \mathbb{R}^n$, with $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^n$.

Example 2

Consider the function $\mathbf{f}(\mathbf{x}) = \mathbf{A}\,\mathbf{x} \in \mathbb{R}^n$, with $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^n$.

Each function $f_i(\mathbf{x}) = \sum_{i=1}^{n} A_{ij}x_j$. Therefore, $\dfrac{\partial f_i}{\partial x_j} = A_{ij}$.

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \mathbf{A} = \begin{bmatrix} A_{11} & \dots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \dots & A_{mn} \end{bmatrix}$$

Gradient of matrices

Gradient of a matrix with respect to a matrix

We may need to compute the gradient of a matrix with respect to a vector or to another matrix.

The gradient of an $m \times n$ matrix of functions $\mathbf{A}$ with respect to an $p \times q$ matrix $\mathbf{B}$ is the Jacobian of size $m \times n \times p \times q$ (tensor), such that $\mathbf{J}_{ijkl} = \dfrac{\partial A_{ij}}{\partial B_{kl}}$.

Remark

We can rewrite matrices as vectors of dimensions mn and pq. In this case, the Jacobian is of size $mn \times pq$.

Consider the matrix $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ With $A_{11} = 2x + 4y$, $A_{12} = xy^2$, $A_{21} = xy$, $A_{22} = x^2 + y$.

Gradient of a vector with respect to a matrix

Consider $\mathbf{f} = \mathbf{A}\,\mathbf{x}$, $\mathbf{f} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$.

Gradient of a vector with respect to a matrix

Consider $\mathbf{f} = \mathbf{A}\,\mathbf{x}$, $\mathbf{f} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$.

We want to compute the gradient $\dfrac{d\,\mathbf{f}}{d\mathbf{A}} \in \mathbb{R}^{m \times m \times n}$.

Gradient of a vector with respect to a matrix

Consider $\mathbf{f} = \mathbf{A}\,\mathbf{x}$, $\mathbf{f} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$.

We want to compute the gradient $\dfrac{d\,\mathbf{f}}{d\mathbf{A}} \in \mathbb{R}^{m \times m \times n}$.

From the definition, we have $\dfrac{d\,\mathbf{f}}{d\mathbf{A}} = \begin{bmatrix} \dfrac{\partial f_1}{\partial \mathbf{A}} \\ \dfrac{\partial f_2}{\partial \mathbf{A}} \\ \vdots \\ \dfrac{\partial f_m}{\partial \mathbf{A}} \end{bmatrix}$, such that

$\dfrac{\partial f_i}{\partial \mathbf{A}} \in \mathbb{R}^{1 \times m \times n}$.

We have $f_i = \sum_{j=1}^{n} A_{ij} xj$.

We have $f_i = \sum_{j=1}^{n} A_{ij}xj$.

Thus $\frac{\partial f_i}{\partial A_{iq}} = x_q$.

We have $f_i = \sum_{j=1}^{n} A_{ij}xj$.

Thus $\dfrac{\partial f_i}{\partial A_{iq}} = x_q$. For a row in $\mathbf{A}$, we get

We have $f_i = \sum_{j=1}^{n} A_{ij} xj$.

Thus $\frac{\partial f_i}{\partial A_{iq}} = x_q$. For a row in $\mathbf{A}$, we get

$$\frac{\partial f_i}{\partial A_{i,:}} = \mathbf{x}^\top$$
$$\frac{\partial f_i}{\partial A_{k \neq i,:}} = \mathbf{0}^\top$$

Therefore, we have

Therefore, we have

$$\frac{\partial f_i}{\mathbf{A}} = \begin{bmatrix} \mathbf{0}^\top \\ \mathbf{0}^\top \\ \vdots \\ \mathbf{0}^\top \\ \mathbf{x}^\top \\ \vdots \\ \mathbf{0}^\top \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ x_1 & \dots & x_n \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}$$

$$\frac{\partial \mathbf{f}}{\partial \mathbf{A}} = \begin{bmatrix} \begin{pmatrix} x_1 & \dots & x_n \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & \dots & 0 \\ x_1 & \dots & x_n \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix} \\ \vdots \\ \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \end{pmatrix} \end{bmatrix}$$

Some useful Formulas

- $\dfrac{\partial}{\partial \mathbf{X}} \mathbf{f}(\mathbf{X})^\top = \left(\dfrac{\partial\, \mathbf{f}(\mathbf{X})}{\partial \mathbf{X}}\right)^\top$
- $\dfrac{\partial\, \mathrm{tr}(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}} = \mathrm{tr}\left(\dfrac{\partial\, \mathbf{f}(\mathbf{X})}{\partial \mathbf{X}}\right)$
- $\dfrac{\partial \det(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}} = \det(\mathbf{f}(\mathbf{X})).\, \mathrm{tr}\left(\mathbf{f}(\mathbf{X})^{-1} \dfrac{\partial\, \mathbf{f}(\mathbf{X})}{\partial \mathbf{X}}\right)$
- $\dfrac{\partial}{\partial \mathbf{X}} \mathbf{f}(\mathbf{X})^{-1} = -\,\mathbf{f}(\mathbf{X})^{-1} \dfrac{\partial\, \mathbf{f}(\mathbf{X})}{\partial \mathbf{X}}\, \mathbf{f}(\mathbf{X})^{-1}.$
- $\dfrac{\partial a^\top \mathbf{X} b}{\partial \mathbf{X}} = ab^\top.$
- $\dfrac{\partial a^\top \mathbf{X}^{-1} b}{\partial \mathbf{X}} = -\mathbf{X}^{-1} ab^\top (\mathbf{X}^{-1})^\top.$

Some useful Formulas

- $\dfrac{\partial a^\top \mathbf{x}}{\mathbf{x}} = a^\top$.
- $\dfrac{\partial \mathbf{x}^\top a}{\mathbf{x}} = a^\top$.
- $\dfrac{\partial \mathbf{x}^\top \mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{B} + \mathbf{B}^\top)$
- $\dfrac{\partial}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{A}\mathbf{s})\, \mathbf{W}\, (\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\, (\mathbf{x} - \mathbf{A}\mathbf{s})^\top \mathbf{W}\mathbf{A}$, for symmetric $\mathbf{W}$.

Automatic Differentiation and Backpropagation

Automatic Differentiation

Automatic differentiation refers to a technique that evaluates the exact gradient of a function by using intermediate variables and applying the chain rule.

Automatic Differentiation

Automatic differentiation refers to a technique that evaluates the exact gradient of a function by using intermediate variables and applying the chain rule.

It applies a series of elementary arithmetic operations (addition, multiplication) and elementary functions (sin, cos, exp, log) to compute the gradient of quite complicated functions automatically.

Automatic differentiation

given a function value $\mathbf{y}$ that is computed as a many level function composition :

$$\mathbf{y} = (f_k \circ f_{k-1} \circ \ldots \circ f_1)(\mathbf{x}) = f_k(f_{k-1}(\ldots(f_1(\mathbf{x}))))$$

Automatic differentiation

given a function value $\mathbf{y}$ that is computed as a many level function composition :

$$\mathbf{y} = (f_k \circ f_{k-1} \circ \ldots \circ f_1)(\mathbf{x}) = f_k(f_{k-1}(\ldots(f_1(\mathbf{x}))))$$

$$\mathbf{x} \longrightarrow f_1 \longrightarrow f_2 \dashrightarrow f_{k-1} \longrightarrow f_k$$

Automatic differentiation

Using the chain rule, we have

Automatic differentiation

Using the chain rule, we have

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \dots \cdot \frac{\partial f_1}{\partial \mathbf{x}}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \dots \cdot \left(\frac{\partial f_2}{\partial f_1} \frac{\partial f_1}{\partial \mathbf{x}} \right) \qquad \text{forward mode}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \right) \cdot \dots \cdot \frac{\partial f_1}{\partial \mathbf{x}} \qquad \text{Reverse mode}$$

Automatic differentiation

Using the chain rule, we have

$$
\begin{aligned}
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} &= \frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \dots \cdot \frac{\partial f_1}{\partial \mathbf{x}} \\
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} &= \frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \dots \cdot \left(\frac{\partial f_2}{\partial f_1} \frac{\partial f_1}{\partial \mathbf{x}} \right) \qquad \text{forward mode} \\
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} &= \left(\frac{\partial y}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \right) \cdot \dots \cdot \frac{\partial f_1}{\partial \mathbf{x}} \qquad \text{Reverse mode}
\end{aligned}
$$

In Deep Learning, it is the reverse mode that is used to compute the partial derivatives. This is called Backpropagation

Example

Consider the function

$$f(x) = \sqrt{x^2 + \exp\{(x^2)\}} + \cos\left(x^2 + \exp\{(x^2)\}\right).$$

Example

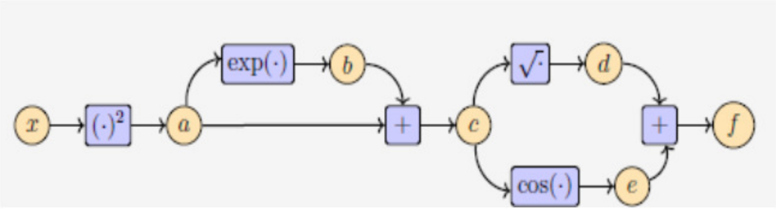
Consider the function

$f(x) = \sqrt{x^2 + \exp\{(x^2)\}} + \cos\left(x^2 + \exp\{(x^2)\}\right)$.

We define the variables :

$$
\begin{aligned}
a &= x^2, \\
b &= \exp(a), \\
c &= a + b, \\
d &= \sqrt{c}, \\
e &= \cos(c), \\
f &= d + e,
\end{aligned}
$$

Example of a computation graph



Partial Derivatives

We compute the partial derivatives :

Partial Derivatives

We compute the partial derivatives :

$$\frac{\partial a}{\partial x} = 2x,$$
$$\frac{\partial b}{\partial a} = \exp(a),$$
$$\frac{\partial c}{\partial a} = \frac{\partial c}{\partial b} = 1,$$
$$\frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}},$$
$$\frac{\partial e}{\partial c} = \sin(c),$$
$$\frac{\partial f}{\partial d} = \frac{\partial f}{\partial e} = 1.$$

Going back to the computation graph, we can deduce that

$$
\begin{aligned}
\frac{\partial f}{\partial c} &= \frac{\partial f}{\partial d} \cdot \frac{\partial d}{\partial c} + \frac{\partial f}{\partial e} \cdot \frac{\partial e}{\partial c} \\
\frac{\partial f}{\partial b} &= \frac{\partial f}{\partial c} \cdot \frac{\partial c}{\partial b} = \frac{\partial f}{\partial c} \\
\frac{\partial f}{\partial a} &= \frac{\partial f}{\partial c} \cdot \frac{\partial c}{\partial a} = \frac{\partial f}{\partial c} \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial a} \cdot \frac{\partial a}{\partial x}.
\end{aligned}
$$

Example 2

Consider the bivariate function

$$y = f(x_1, x_2) = \frac{x_1^2 - 2x_2}{\sqrt{x_1 x_2 - \sin(x_2)}}$$

Automatic differentiation

Let $x_1, \dots, x_d$ be the input variables to the function and $x_{d+1}, \dots, x_{D-1}$ be the intermediate variables and x_D the output variable. Then the computation graph can be expressed as follows :

Automatic differentiation

Let $x_1, \ldots, x_d$ be the input variables to the function and $x_{d+1}, \ldots, x_{D-1}$ be the intermediate variables and x_D the output variable. Then the computation graph can be expressed as follows : For $i = d+1, \ldots, D: \quad x_i = g_i(x_{P(x_i)})$

Automatic differentiation

Let $x_1, \ldots, x_d$ be the input variables to the function and $x_{d+1}, \ldots, x_{D-1}$ be the intermediate variables and x_D the output variable. Then the computation graph can be expressed as follows : For $i = d+1, \ldots, D : \quad x_i = g_i(x_{P(x_i)})$
where g_i is an elementary function and $x_{P(x_i)}$ are the parent nodes (predecessors) of the variable x_i in the graph.

Automatic differentiation

Let $x_1, \dots, x_d$ be the input variables to the function and $x_{d+1}, \dots, x_{D-1}$ be the intermediate variables and x_D the output variable. Then the computation graph can be expressed as follows : For $i = d+1, \dots, D : \quad x_i = g_i(x_{P(x_i)})$
where g_i is an elementary function and $x_{P(x_i)}$ are the parent nodes (predecessors) of the variable x_i in the graph. Therefore, we have

Automatic differentiation

Let $x_1, \dots, x_d$ be the input variables to the function and $x_{d+1}, \dots, x_{D-1}$ be the intermediate variables and x_D the output variable. Then the computation graph can be expressed as follows : For $i = d+1, \dots, D: \quad x_i = g_i(x_{P(x_i)})$
where g_i is an elementary function and $x_{P(x_i)}$ are the parent nodes (predecessors) of the variable x_i in the graph. Therefore, we have

$$\frac{\partial f}{\partial x_D} = 1$$
$$\frac{\partial f}{\partial x_i} = \sum_{x_j : x_i \in P(x_j)} \frac{\partial f}{\partial x_j} \cdot \frac{\partial x_j}{\partial x_i}$$

Backpropagation in a Neural Network

Consider a function value $\mathbf{y}$ that is computed as a many level function composition :

$$\mathbf{y} = (f_k \circ f_{k-1} \circ \ldots \circ f_1)(\mathbf{x}) = f_k(f_{k-1}(\ldots(f_1(\mathbf{x}))))$$

where $\mathbf{x}$ are the input, $\mathbf{y}$ represents the observations, and every function $f_i, i = 1..k$ has its own parameters.

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

x_{i-1} denotes the output of the layer $i - 1$ ($x_0 = \mathbf{x}$) ,

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

x_{i-1} denotes the output of the layer $i-1$ ($x_0 = \mathbf{x}$) ,

$\mathbf{A}_{i-1}$ is the matrix of weights between layers $i-1$ and i,

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

x_{i-1} denotes the output of the layer $i-1$ ($x_0 = \mathbf{x}$) ,

$\mathbf{A}_{i-1}$ is the matrix of weights between layers $i-1$ and i,

b_{i-1} is a bias corresponding to the layer $i-1$,

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

x_{i-1} denotes the output of the layer $i-1$ ($x_0 = \mathbf{x}$) ,

$\mathbf{A}_{i-1}$ is the matrix of weights between layers $i-1$ and i,

b_{i-1} is a bias corresponding to the layer $i-1$,

σ is an activation function.

Backpropagation in a Neural Network

In a neural network, we have $f_i(x_{i-1}) = \sigma(\mathbf{A}_{i-1}x_{i-1} + b_{i-1})$, where

x_{i-1} denotes the output of the layer $i-1$ ($x_0 = \mathbf{x}$) ,

$\mathbf{A}_{i-1}$ is the matrix of weights between layers $i-1$ and i,

b_{i-1} is a bias corresponding to the layer $i-1$,

σ is an activation function.

Our goal is then to compute the gradient of the function

$$\mathcal{L}(\Theta) = \|\, \mathbf{y} - f_k(\Theta, \mathbf{x})\|^2$$

with $\Theta = \{\mathbf{A}_0, \mathbf{b}_0, \ldots, \mathbf{A}_{k-1}, \mathbf{b}_{k-1}\}$.

Chain rule allows us to compute the partial derivatives of $\mathcal{L}$ with respect to parameters $\theta_i = \{\mathbf{A}_i, \mathbf{b}_i\}$:

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-1}} = \frac{\partial \mathcal{L}}{\partial f_k} . \frac{\partial f_k}{\partial \theta_{k-1}}$$

Chain rule allows us to compute the partial derivatives of $\mathcal{L}$ with respect to parameters $\theta_i = \{\mathbf{A}_i, \mathbf{b}_i\}$:

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-1}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial \theta_{k-1}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-2}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \boxed{\frac{\partial f_k}{\partial f_{k-1}} \cdot \frac{\partial f_{k-1}}{\partial \theta_{k-2}}}$$

Chain rule allows us to compute the partial derivatives of $\mathcal{L}$ with respect to parameters $\theta_i = \{\mathbf{A}_i, \mathbf{b}_i\}$:

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-1}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial \theta_{k-1}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-2}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \boxed{\frac{\partial f_k}{\partial f_{k-1}} \cdot \frac{\partial f_{k-1}}{\partial \theta_{k-2}}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-3}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \boxed{\frac{\partial f_{k-1}}{\partial f_{k-2}} \cdot \frac{\partial f_{k-2}}{\partial \theta_{k-3}}}$$

Chain rule allows us to compute the partial derivatives of $\mathcal{L}$ with respect to parameters $\theta_i = \{\mathbf{A}_i, \mathbf{b}_i\}$:

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-1}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial \theta_{k-1}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-2}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \boxed{\frac{\partial f_k}{\partial f_{k-1}} \cdot \frac{\partial f_{k-1}}{\partial \theta_{k-2}}}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{k-3}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \boxed{\frac{\partial f_{k-1}}{\partial f_{k-2}} \cdot \frac{\partial f_{k-2}}{\partial \theta_{k-3}}}$$

$$\vdots$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{i}} = \frac{\partial \mathcal{L}}{\partial f_k} \cdot \frac{\partial f_k}{\partial f_{k-1}} \cdot \frac{\partial f_{k-1}}{\partial f_{k-2}} \cdots \boxed{\frac{\partial f_{i+2}}{\partial f_{i+1}} \cdot \frac{\partial f_{i+1}}{\partial \theta_{i}}}$$

Second Order Derivatives

$$
\begin{aligned}
f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\
& \mathbf{x} \to f(x)
\end{aligned}
$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.

$$f: \quad \mathbb{R}^n \mapsto \mathbb{R}$$
$$\mathbf{x} \rightarrow f(\ x)$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.
The second order partial derivatives are obtained by deriving the first order partial derivatives function with respect to all variables.

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \mathbf{x} \to f(\,x) \end{aligned}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top \in \mathbb{R}^n$.
The second order partial derivatives are obtained by deriving the first order partial derivatives function with respect to all variables.
Thus we have, for $i, j = 1..n$:

$$\frac{\partial(\frac{\partial f}{\partial x_i})}{\partial x_j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

The Hessian Matrix

The *Hessian* matrix of f is a matrix of second-order partial derivatives :

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$\text{i.e. } [(\mathbf{H})_{ij}] = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

The Hessian Matrix

If the partial derivatives are continuous, the order of differentiation can be interchanged :

The Hessian Matrix

If the partial derivatives are continuous, the order of differentiation can be interchanged :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

The Hessian Matrix

If the partial derivatives are continuous, the order of differentiation can be interchanged :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

This implies that the Hessian matrix is symmetric.

The Hessian Matrix

If the partial derivatives are continuous, the order of differentiation can be interchanged :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

This implies that the Hessian matrix is symmetric.

The Hessian is used in some optimization algorithms such as Newton's method.

The Hessian Matrix

If the partial derivatives are continuous, the order of differentiation can be interchanged :

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

This implies that the Hessian matrix is symmetric.

The Hessian is used in some optimization algorithms such as Newton's method.

It is expensive to calculate but can drastically reduce the number of iterations needed to converge to a local minimum by providing information about the curvature of f.

Example

Consider the function $f = xy^2 - 3z^2$.

Remark

If $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$, then the Hessian is a $m \times n \times n$ tensor.

For instance, $\mathbf{f} = \begin{pmatrix} x^2y + \exp(z) \\ xy^2z - \ln(x) \end{pmatrix}$

Taylor's Theorem

Taylor's theorem has natural generalizations to multivariate and vector functions :

Theorem ((Taylor's theorem))

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable, and let* $\mathbf{h} \in \mathbb{R}^d$*. Then there exists* $t \in [0, 1]$ *such that*

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

Furthermore, if f *is twice continuously differentiable, then there exists* $t \in [0, 1]$ *such that*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})\mathbf{h} + \frac{1}{2}\boldsymbol{h}^T\mathbf{H}(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Example

Consider the function $f(x, y) = x^2 + y^2$. We want to compute $f(\mathbf{x})$ with $\mathbf{x} = (1, 1; 1, 1)^\top$.

Taylor's theorem is used to give approximation of any differentiable function by a polynomial function and particularly linear and quadratic approximation in the neighborhood of a point $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$. We can thus write

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + o(\|\mathbf{x} - \mathbf{x_0}\|)$$

And

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^\top \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + o(\|\mathbf{x} - \mathbf{x_0}\|^2)$$

Taylor's theorem is used to give approximation of any differentiable function by a polynomial function and particularly linear and quadratic approximation in the neighborhood of a point $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$. We can thus write

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + o(\|\mathbf{x} - \mathbf{x_0}\|)$$

And

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + \frac{1}{2}(\mathbf{x} - \mathbf{x_0})^\top \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x_0}) + o(\|\mathbf{x} - \mathbf{x_0}\|^2)$$

It is also used in proofs about conditions for local minima of unconstrained optimization problems.