Chapter 5. Vector Calculus

FPT UNIVERSITY

Chapter 5. Vector Calculus

FPT Education
FPT UNIVERSITY

FPT Education
FPT UNIVERSITY

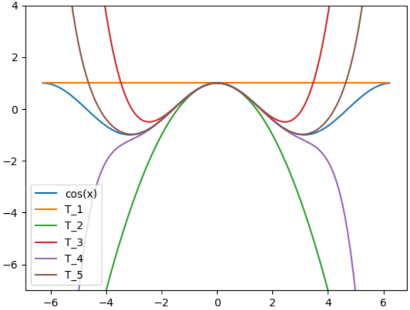
Definition

Let D be a domain of $\mathbb{R}$, and let $f : D \to \mathbb{R}$ be a function.

- The Taylor polynomial of degree n of a function f at $x_0 \in D$ is defined as
$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$
- For a smooth functions $f \in C^\infty(D)$, the Taylor series of f at x_0 is defined as
$$T_\infty(x) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$
- For $x_0 = 0$, we obtain the Maclaurin series.
- f is called analytic if $T_\infty(x) = f(x)$.

Note.

1. Taylor polynomial of degree n is an approximation of a function. For $n = 1$, we obtain the linear approximation.
2. The Taylor polynomial is similar to f in a neighborhood around x_0.
3. For $k \leq n$, Taylor polynomial of degree n is an exact representation of a polynomial f of degree k.



The function $f(x) = \cos(x)$ is approximated by Taylor polynomials around $x_0 = 1$.

Example

Consider the function $f(x) = x^3 + x - 3$ and find the Taylor polynomial T_4 at $x_0 = 1$.

Answer. We have

$$f'(x) = 3x^2 + 1, f''(x) = 6x, f'''(x) = 6, f^{(4)}(x) = 0.$$

Evaluate them at $x_0 = 1$, we get

$$f(1) = -1, f'(1) = 4, f''(1) = 6, f'''(1) = 6, f^{(4)}(1) = 0.$$

Thus, the Taylor polynomial T_4 of f at $x_0 = 1$ is

$$\begin{aligned} T_4(x) &= -1 + 4(x-1) + \frac{6}{2!}(x-1)^2 + \frac{6}{3!}(x-1)^3 \\ &= -1 + 4(x-1) + 3(x-1)^2 + (x-1)^3 \\ &= f(x). \end{aligned}$$

Maclaurin series of some basic functions

FPT Education
FPT UNIVERSITY

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^n}{n!}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$$

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots = \sum_{k=0}^{\infty} x^k.$$

5.2 Partial Differentiation and Gradients

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is a rule that assigns to each $x \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$ to a real value $f(x) = f(x_1, \cdots, x_n)$.

Definition

For a function $f : \mathbb{R}^n \to \mathbb{R}$, the partial derivatives of f are defined by

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

$$\vdots$$

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

Note. When finding the partial derivative $\frac{\partial f}{\partial x_i}$, we consider only x_i varies and keep the others constant.

Example

Find the partial derivatives of function $f(x, y) = \frac{1}{1+x^2+2y^4}$.

Answer.

$$\frac{\partial f(x,y)}{\partial x} = -\frac{1}{(1+x^2+3y^4)^2}\frac{\partial}{\partial x}(1+x^2+3y^4) = -\frac{2x}{(1+x^2+3y^4)^2}$$

$$\frac{\partial f(x,y)}{\partial y} = -\frac{1}{(1+x^2+3y^4)^2}\frac{\partial}{\partial y}(1+x^2+3y^4) = -\frac{12y^3}{(1+x^2+3y^4)^2}.$$

Definition

For a function $f : \mathbb{R}^n \to \mathbb{R}$ of n variables $x_1, \cdots, x_n$, the gradient of f is defined as:

$$\nabla_x f = \operatorname{grad} f = \frac{df}{dx} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1\times n}.$$

Example

Find the gradient of function $f(x_1, x_2, x_3) = x_1^2 x_2^3 - 4x_2^2 x_3$.

Answer.

$$\begin{aligned} \nabla_x f &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \frac{\partial f}{\partial x_3} \end{bmatrix} \\ &= \begin{bmatrix} 2x_1 x_2^3 & 3x_1^2 x_2^2 - 8x_2 x_3 & -4x_2^2 \end{bmatrix} \in \mathbb{R}^{1\times 3}. \end{aligned}$$

Basic rules of Partial Differentiation

For $f, g : \mathbb{R}^n \to \mathbb{R}$ functions of variables $x \in \mathbb{R}^n$.

- Sum rule: $\nabla_x [f + g] = \nabla_x f + \nabla_x g$.
- Product rule: $\nabla_x [fg] = g(x) \nabla_x f + f(x) \nabla_x g$.

Consider a function $f : \mathbb{R}^n \to R$ of n variables $x = (x_1, \ldots, x_n)$. Moreover, $x_i(t)$ are themselves functions of t. Then we have

$$\text{Chain rule}: \quad f'(t) = \frac{df}{dt} = \frac{\partial f}{\partial x_1} \frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_n} \frac{dx_n}{dt}$$
$$= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}.$$

Example

Consider $f(x_1, x_2) = x_1^2 + x_1x_2$, where $x_1 = \sin t, x_2 = \cos t$. Find the derivative of f with respect to t.

Answer: We have

$$\frac{\partial f}{\partial x_1} = 2x_1 + x_2,\ \frac{\partial f}{\partial x_2} = x_1.$$
$$\frac{dx_1}{dt} = \cos t, \frac{dx_2}{dt} = -\sin t.$$

Hence

$$\begin{aligned}\frac{df}{dt} &= (2x_1 + x_2)\cos t + x_1(-\sin t)\\ &= (2\sin t + \cos t)\cos t - \sin t \sin t\\ &= 2\sin t \cos t + \cos^2 t - \sin^2 t\\ &= \sin(2t) + \cos(2t).\end{aligned}$$

5.3 Gradients of Vector-Valued Functions

Definition

A vector-valued function of n variables $x = (x_1, \ldots, x_n)$, $f : \mathbb{R}^n \to \mathbb{R}^m$ is given as

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{bmatrix},$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is a function of x.

Definition

The partial derivative of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i, i = 1, \ldots, n$, is given as the vector

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

Example

Consider a linear transformation (operator) $f : \mathbb{R}^n \to \mathbb{R}^m$

$$f(x) = Ax, \text{ where } A = \left[a_{ij}\right] \in \mathbb{R}^{m\times n}.$$

It is easy to see that the partial derivatives of f are

$$\frac{\partial f}{\partial x_j} = A_j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix} \in \mathbb{R}^m, \ j = 1, \ldots, n,$$

Example

Consider a vector-valued function

$$f(x, y) = \begin{bmatrix} xy^2 & y^3 & x^2 - y^2 \end{bmatrix}^T.$$

Then

$$\frac{\partial f}{\partial x} = \begin{bmatrix} y^2 \\ 0 \\ 2x \end{bmatrix},$$

$$\frac{\partial f}{\partial y} = \begin{bmatrix} 2xy \\ 3y^2 \\ -2y \end{bmatrix}.$$

Definition

The collection of all first-order partial derivatives of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called the Jacobian. The Jacobian J is an $m \times n$ matrix, which we define and arrange as follows:

$$
\begin{aligned}
J = \nabla_x f = \frac{df}{dx} &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \dots & \dots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.
\end{aligned}
$$

Example

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator given by

$$f(x) = Ax,$$

where A be a matrix in $\mathbb{R}^{m\times n}$. Then it is clear that the Jacobian of f:

$$\nabla_x f = A.$$

Example

Consider a vector-valued function $f : \mathbb{R}^3 \to \mathbb{R}^2$, $f(x_1, x_2, x_3) = \begin{bmatrix} e^{-x_1+x_2} \\ x_1 x_2^2 \\ \sin(x_1) \end{bmatrix}$.

The Jacobian of f is

$$J = \nabla_x f = \begin{bmatrix} -e^{-x_1+x_2} & e^{-x_1+x_2} \\ x_2^2 & 2x_1x_2 \\ \cos x_1 & 0 \end{bmatrix}.$$

Chain rule

Consider a valued-vector function $f : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f(x) = f(x_1, \ldots, x_n) = \begin{bmatrix} f_1(x) & \cdots & f_m(x) \end{bmatrix}^T$$

and $x_i = x_i(t_1, \ldots, t_l)$ are themselves function of l-variables $t = (t_1, \ldots, t_l)$. It means $x : \mathbb{R}^l \to \mathbb{R}^m$

$$x(t) = x(t_1, \ldots, t_l) = \begin{bmatrix} x_1(t) & \cdots & x_n(t) \end{bmatrix}^T.$$

Then $f \circ x : \mathbb{R}^l \to \mathbb{R}^m$ given by $f(t) = f(x(t))$ and

$$\frac{\partial f_j}{\partial t_k} = \sum_{i=1}^{n} \frac{\partial f_j}{\partial x_i} \frac{\partial x_i}{\partial t_k}, \text{ for all } j = 1, \ldots, m \text{ and } k = 1, \ldots, l$$

$$\nabla_t f = \nabla_x f \nabla_t x.$$

Example

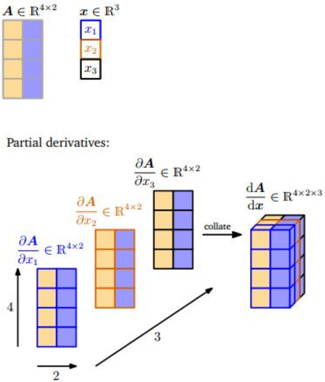
Consider a function $f(x_1, x_2) = x_1^2 + 2x_1x_2$ and $x_1(s,t) = s.\cos t$, $x_2 = s.\sin t$. Find the partial derivative of f with respect to s and t.

Answer.

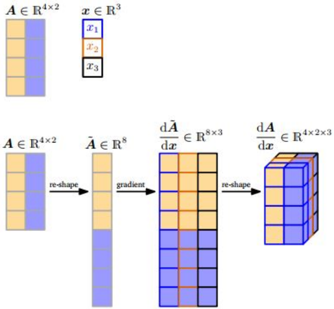
$$\begin{aligned}
\frac{\partial f}{\partial s} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s} \\
&= (2x_1 + 2x_2)\cos t + 2x_1 \sin t \\
&= s.\cos^2 t + 2s.\sin t.\cos t = s.\cos^2 t + 2s.\sin(2t) \\
\frac{\partial f}{\partial t} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \\
&= (2x_1 + 2x_2)(-s.\sin t) + 2x_1.(s.\cos t) \\
&= -2s^2.\sin^2 t
\end{aligned}$$

FPT Education FPT UNIVERSITY

Let $A \in \mathbb{R}^{4\times 2}$ and let $x \in \mathbb{R}^3$. How to define $\frac{dA}{dx}$?



Compute the partial derivatives $\frac{\partial A}{\partial x_1}, \frac{\partial A}{\partial x_2}, \frac{\partial A}{\partial x_3}$, each is a 4×2 matrix, and collate them in a $4 \times 2 \times 3$ tensor.



Re-shape A into a vector $\tilde{A} \in \mathbb{R}^8$. Then, compute the gradient $\frac{d\tilde{A}}{dx} \in \mathbb{R}^{8\times 3}$. Re-shaping this gradient to obtain the gradient tensor.

The gradient of an $m \times n$ matrix A with respect to a $p \times q$ matrix B is a four-dimensional tensor J, whose entries are given as

$$J_{ijkl} = \frac{\partial A_{ij}}{\partial B_{kl}}.$$

We can also identity the space $\mathbb{R}^{m \times n}$ of $m \times n$ matrices and the space $\mathbb{R}^{mn}$. Hence, we reshape A and B into vectors of lengths mn and pq, respectively. The obtained Jacobian is in size $mn \times pq$.

Example

Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{n\times n}$ be given $f(A) = A^T A := K \in \mathbb{R}^{n\times n}$. Find the gradient dK/dA.

Answer. The gradient $dK/dA \in \mathbb{R}^{(n\times n)\times(m\times n)}$ is a tensor. Moreover,

$$\frac{dK_{pq}}{dA} \in \mathbb{R}^{1\times m\times n}.$$

Denote by A_i the i^{th} column of A and by K_{pq} by the (p,q)-entry of K, $p, q = 1, \ldots, n$. Since

$$K_{pq} = A_p^T A_q = \sum_{k=1}^{m} A_{km} A_{mk},$$

$$\Rightarrow \frac{\partial K_{pq}}{\partial A_{ij}} = \sum_{k=1}^{m} \frac{\partial}{\partial A_{ij}}(A_{km}A_{mk}) = \begin{cases} A_{iq} & \text{if } j = p, p \neq q \\ A_{ip} & \text{if } j = q = p \neq q \\ 2A_{iq} & \text{if } j = p = q \\ 0 & \text{otherwise} \end{cases}$$

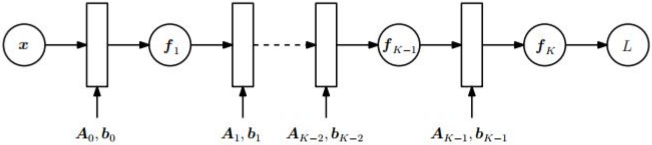
5.6.1 Gradients in a Deep Network

In deep learning, the function value y is computed as a many-level function composition

$$y = (f_K \circ f_{K-1} \circ \cdots \circ f_1)(x) = f_K(f_{K-1}(\cdots(f_1(x))\cdots))$$

where x are the inputs (e.g., images), y are the observations (e.g., class labels), and every function $f_i, i = 1, \cdots, K$, possesses its own parameters.

Given a neural network with multiple layers:



In the i^{th} layer:

$$f_i(x_{i-1}) = \sigma(A_{i-1}x_{i-1} + b_{i-1})$$

where x_{i-1} is the output of the layer $i-1$, σ is an activation function (sigmoid or ReLU or tanh, ... functions).

Training these model requires us to compute the gradient of a loss function L w.r.t all model parameters $\theta_j = \{A_j, b_j\}, j = 0, \dots, K-1$.

Suppose we have inputs x and observations y and a network structure

$$
\begin{aligned}
f_0 &:= x \\
f_i &:= \sigma_i(A_i f_{i-1} + b_{i-1}), \; i = 1, \dots, K.
\end{aligned}
$$

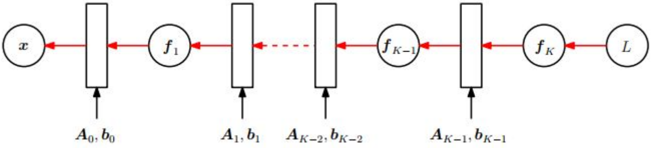
We need find $\theta = \{\theta_j\} = \{A_j, b_j\}, j = 0, \dots, K-1$ which minimize the loss function

$$L(\theta) = \left\| y - f_K(\theta, x) \right\|^2 .$$

The gradients of L w.r.t θ:

$$\frac{\partial L}{\partial \theta_{K-1}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial \theta_{K-1}},$$
$$\frac{\partial L}{\partial \theta_{K-2}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}}\frac{\partial f_{K-1}}{\partial \theta_{K-2}}, \cdots$$
$$\frac{\partial L}{\partial \theta_{i}} = \frac{\partial L}{\partial f_K}\frac{\partial f_K}{\partial f_{K-1}} \cdots \frac{\partial f_{i+2}}{\partial f_{i+1}}\frac{\partial f_{i+1}}{\partial \theta_{i}}$$

Most of the computation of $\frac{\partial L}{\partial \theta_{i+1}}$ can be reused to compute $\frac{\partial L}{\partial \theta_i}$.



Gradients are passed backward through the network.

5.6.2 Automatic Differentiation

- Backpropagation is a special case of a general technique in numerical analysis called automatic differentiation.
- Automatic differentiation refers to a set of techniques to numerically evaluate the exact gradient of a function by working with intermediate variables and applying the chain rule.

Given a simple graph representing the data flow from inputs x to outputs y:

$$\text{x} \longrightarrow \text{a} \longrightarrow \text{b} \longrightarrow \text{y}.$$

To compute the derivative dy/dx:

$$\frac{dy}{dx} = \frac{dy}{db}\left(\frac{db}{da}\frac{da}{dx}\right)$$

Forward mode: the gradients flow with forward mode the data from left to right through the graph.

$$\frac{dy}{dx} = \left(\frac{dy}{db}\frac{db}{da}\right)\frac{da}{dx}$$

Reverse mode: gradients are prop- reverse mode agated backward through the graph, i.e., reverse to the data flow.

Definition

Reverse mode automatic differentiation is called backpropagation.

Example

Consider the function

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

Use intermediate variables:

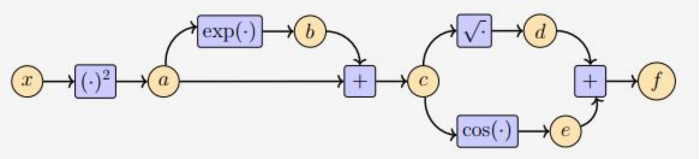
$$a = x^2, b = \exp(a), c = a + b, d = \sqrt{c}, e = \cos c, f = d + e.$$

We get

$$\frac{\partial a}{\partial x} = 2x \qquad \frac{\partial b}{\partial a} = \exp(a)$$

$$\frac{\partial c}{\partial a} = \frac{\partial c}{\partial b} = 1 \qquad \frac{\partial d}{\partial c} = \frac{1}{2\sqrt{c}}$$

$$\frac{\partial e}{\partial c} = -\sin c \qquad \frac{\partial f}{\partial d} = \frac{\partial f}{\partial e} = 1.$$



We can compute $\partial f/\partial x$ using backpropagation method

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial c} + \frac{\partial f}{\partial e}\frac{\partial e}{\partial c} = 1.\frac{1}{2\sqrt{c}} + 1.(-\sin c)$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial b} = \frac{\partial f}{\partial c}$$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial a} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial a} = \frac{\partial f}{\partial b}\exp(a) + \frac{\partial f}{\partial c}$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x} = \frac{\partial f}{\partial a}.2x$$

Let $x_1, \ldots, x_d$ be the input variables to the function, $x_{d+1}, \ldots, x_{D-1}$ be the intermediate variables, and x_D the output variable.

$$\text{For } i = d+1, \ldots, D : x_i = g_i(x_{\mathrm{Pa}(x_i)}),$$

where the $g_i(\cdot)$ are elementary functions and $x_{\mathrm{Pa}(x_i)}$ are the parent nodes of the variable x_i in the graph. Let $f = x_D$. By the chain rule,

$$\frac{\partial f}{\partial x_i} = \sum_{x_j :\, x_i \in \mathrm{Pa}(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial x_j}{\partial x_i} = \sum_{x_j :\, x_i \in \mathrm{Pa}(x_j)} \frac{\partial f}{\partial x_j} \frac{\partial g_j}{\partial x_i}$$

is the backpropagation of the gradient through the computation graph, where $\mathrm{Pa}(x_j)$ is the set of parent nodes of x_j.

5.7 Higher-Order Derivatives

Consider a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ of two variables x, y. We use the notation for higher-order partial derivatives:

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right), \quad \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right)$$

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \quad \frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right), \cdots$$

If $f(x, y)$ is a twice (continuously) differentiable function, then

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

Definition

Hessian matrix of of f is

$$H = \nabla^2_{x,y} f(x, y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}.$$

Example

Let $f(x,y) = e^{-x^2+2y}$. Find the Hessian matrix of f at $(0,0)$.

Answer: We have

$$\frac{\partial f}{\partial x} = -2xe^{-x^2+2y}, \ \frac{\partial f}{\partial y} = 2e^{-x^2+2y}$$
$$\frac{\partial^2 f}{\partial x^2} = (-2+4x^2)e^{-x^2+2y}, \ \frac{\partial^2 f}{\partial y^2} = 4e^{-x^2+2y}$$
$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = -4xe^{-x^2+y^2}.$$

Hence

$$\nabla^2_{(x,y)} f(0,0) = \begin{bmatrix} -2 & 0 \\ 0 & 4 \end{bmatrix}.$$

FPT Education FPT UNIVERSITY

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that has continuous partial derivatives up to order 2. Then

$$f(x) = f(a) + \nabla_x f(a) \cdot (x - a) + R_1(x, a),$$

where the error term $R_1(x, a)$ going to zero faster than a constant times $\|x - a\|^2$ as $x \to a$.

Definition

The first order Taylor polynomial of f at a is:

$$\begin{aligned} T_1(x) &= f(a) + \nabla_x f(a) \cdot (x - a) \\ &= f(a) + \frac{\partial f}{\partial x_1}(a)(x_1 - a_1) + \cdots + \frac{\partial f}{\partial x_n}(a)(x - a_n). \end{aligned}$$

Example

Find the first order Taylor polynomial of $f(x, y) = x^2 + 2xy^3$ at $(1, 2)$.

Answer: We have $f(1, 2) = 17$ and

$$\frac{\partial f}{\partial x} = 2x + 2y^3,\ \frac{\partial f}{\partial y} = 6xy^2 \Rightarrow \qquad \frac{\partial f}{\partial x}(1, 2) = 18,\ \frac{\partial f}{\partial y}(1, 2) = 24.$$

Hence the first order Taylor polynomial of f at $(1, 2)$ is

$$T_1(x, y) = 17 + 18(x - 1) + 24(y - 2).$$

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that has continuous partial derivatives up to order 3. Then we can write

$$f(x) = f(a) + \nabla_x f(a) \cdot (x-a) + \frac{1}{2}(x-a)^T \nabla_x^2 f(a)(x-a) + R_2$$

$$= f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a)(x_i - a_i) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a)(x_i - a_i)(x_j - a_j) + R_2,$$

where the error term $R_2 = R_2(x, a)$ going to zero faster than a constant times $\|x-a\|^3$ as $x \to a$.

Definition

The second order Taylor polynomial of f at a is

$$f(a) + \nabla_x f(a) \cdot (x-a) + \frac{1}{2}(x-a)^T \nabla_x^2 f(a)(x-a).$$

Find the second order Taylor polynomial of $f(x, y) = e^{x+y^2}$ about $(x, y) = (0, 0)$.

Answer: We can compute

$$\frac{\partial f}{\partial x} = e^{x+y^2},\ \frac{\partial f}{\partial y} = 2ye^{x+y^2},\ \frac{\partial^2 f}{\partial x^2} = e^{x+y^2},$$
$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = e^{x+y^2} = 2ye^{x+y^2},\ \frac{\partial^2 f}{\partial y^2} = (2 + 4y^2)e^{x+y^2}.$$

Hence

$$\nabla_{(x,y)}(0, 0) = \begin{bmatrix} 1 & 0 \end{bmatrix},\ \nabla^2_{(x,y)}(0, 0) = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

The second order Taylor polynomial is

$$1 + x + \frac{1}{2}x^2 + y^2.$$

We have studied:

- How to differentiate an univariate function?
- gradients of multivariable functions and vector-valued functions;
- backpropagation;
- higher-order derivatives;
- linearization and Taylor series.

Exercises for practice: 5.1-5.7 (pages 170, 171).