



Lecture 6

Math Foundations Team



BITS Pilani

Pilani | Dubai | Goa | Hyderabad



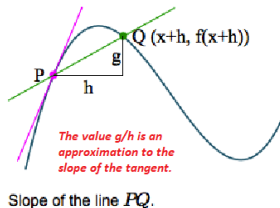
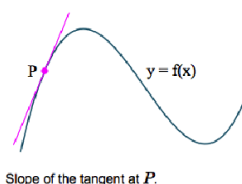
Many algorithms in machine learning optimize an objective function with respect to a set of desired model parameters that control how well a model explains the data: Finding good parameters can be phrased as an optimization problem.

Examples include: linear regression, where we look at curve-fitting problems and optimize linear weight parameters to maximize the likelihood; neural-network auto-encoders for dimensionality reduction and data compression.

For $h > 0$, the derivative of f at x is defined as the limit

$$\frac{df}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (1)$$

The derivative of f points in the direction of steepest ascent of f .



To compute the derivative of $f(x) = x^n$ $n \in \mathbb{N}$ using the definition

$$\begin{aligned}\frac{df}{dx} &= \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{\sum_{i=0}^n \binom{n}{i} x^{n-i} h^i - x^n}{h} \\ &= \lim_{h \rightarrow 0} \frac{\sum_{i=1}^n \binom{n}{i} x^{n-i} h^i}{h}\end{aligned}\tag{2}$$

Derivative of a Polynomial



$$\begin{aligned}\frac{df}{dx} &= \lim_{h \rightarrow 0} \sum_{i=1}^n \binom{n}{i} x^{n-i} h^{i-1} \\ &= \lim_{h \rightarrow 0} \binom{n}{1} x^{n-1} + \lim_{h \rightarrow 0} \sum_{i=2}^n \binom{n}{i} x^{n-i} h^{i-1} \\ &= nx^{n-1}\end{aligned}\tag{3}$$



The Taylor polynomial is a representation of a function f as an finite sum of terms. These terms are determined using derivatives of f evaluated at x_0 .

Definition: The Taylor polynomial of degree n of $f : \mathbb{R} \rightarrow \mathbb{R}$ at x_0 is defined as

$$T_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \quad (4)$$

where $f^{(k)}(x_0)$ is the k th derivative of f at x_0 which we assume exists.



Definition: The Taylor series of smooth (continuously differentiable infinite many times) function $f : \mathbb{R} \rightarrow \mathbb{R}$ at x_0 is defined as

$$T_{\infty}(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \quad (5)$$

For $x_0 = 0$, we obtain the Maclaurin series as a special instance of the Taylor series.

Remark: In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0 . However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leq n$ since all derivatives $f^{(i)} = 0$, for $i > k$.

Taylor Polynomial example



Consider the polynomial $f(x) = x^4$. Find the Taylor polynomial T_6 evaluated at $x_0 = 1$.

We compute $f^{(k)}(1)$ for $k = 0, 1, 2, \dots, 6$

$f(1) = 1$, $f'(1) = 4$, $f''(1) = 12$, $f^{(3)}(1) = 24$, $f^{(4)}(1) = 24$,
 $f^{(5)}(1) = 0$, $f^{(6)}(1) = 0$. The desired Taylor polynomial is

$$\begin{aligned} T_6(x) &= \sum_{k=0}^6 \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \\ &= 1 + 4(x - 1) + 12(x - 1)^2 + 24(x - 1)^3 + 24(x - 1)^4 \\ &= x^4 = f(x) \end{aligned} \tag{6}$$

we obtain an exact representation of the original function.

Taylor Series example



Consider the smooth function $f(x) = \sin(x) + \cos(x)$. We compute Taylor series expansion of f at $x_0 = 0$, which is the Maclaurin series expansion of f . We obtain the following derivatives:

$$f(0) = \sin(0) + \cos(0) = 1$$

$$f'(0) = \cos(0) - \sin(0) = 1$$

$$f''(0) = -\sin(0) - \cos(0) = -1$$

$$f^{(3)}(0) = -\cos(0) + \sin(0) = -1$$

$$f^{(4)}(0) = \sin(0) + \cos(0) = f(0) = 1$$

The coefficients in our Taylor series are only ± 1 (since $\sin(0) = 0$), each of which occurs twice before switching to the other one.

Furthermore, $f^{(k+4)}(0) = f^k(0)$

Taylor Series example



Therefore, the full Taylor series expansion of f at $x_0 = 0$ is given by

$$\begin{aligned} T_{\infty}(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \\ &= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \dots \\ &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \dots x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \dots \quad (7) \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k} + \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1} \\ &= \cos(x) + \sin(x) \end{aligned}$$



We denote the derivative of f by f'

- ▶ Product Rule: $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
- ▶ Sum Rule: $(f(x) + g(x))' = f'(x) + g'(x)$
- ▶ Quotient Rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
- ▶ Chain Rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$

Example: Chain Rule



Compute the derivative of function $h(x) = (2x + 1)^4$

$$h(x) = (2x + 1)^4 = g(f(x))$$

$$f(x) = 2x + 1,$$

$$g(f) = f^4$$

Derivatives of f and g are

$$f'(x) = 2$$

$$g'(f) = 4f^3$$

$$h'(x) = g'(f)f'(x) = (4f^3).2 = 8(2x + 1)^3$$



Differentiation applies to functions f of a scalar variable $x \in R$. In the following, we consider the general case where the function f depends on one or more variables $x \in R^n$, e.g., $f(x) = f(x_1, x_2)$.

The generalization of the derivative to functions of several variables is the gradient. We find the gradient of the function f with respect to x by varying one variable at a time and keeping the others constant. The gradient is then the collection of these partial derivatives.



Definition: For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $x \mapsto f(x)$, $x \in \mathbb{R}^n$ of n variables x_1, \dots, x_n we define the *partial derivatives* as

$$\frac{\partial f}{\partial x_1} = \lim_{h \rightarrow 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

$$\frac{\partial f}{\partial x_2} = \lim_{h \rightarrow 0} \frac{f(x_1, x_2 + h, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

\vdots

$$\frac{\partial f}{\partial x_n} = \lim_{h \rightarrow 0} \frac{f(x_1, x_2, \dots, x_n + h) - f(x_1, x_2, \dots, x_n)}{h}$$

We collect them in the row vector called the gradient of f or Jacobian

$$\Delta_x f = \text{grad} f = \frac{df}{dx} = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \dots, \frac{\partial f(x)}{\partial x_n} \right] \quad (8)$$

Example 1: Find the partial derivatives of $f(x, y) = (x + 2y^3)^2$

$$\frac{\partial f(x, y)}{\partial x} = 2(x + 2y^3) \frac{\partial (x + 2y^3)}{\partial x} = 2(x + 2y^3) \quad (9)$$

$$\frac{\partial f(x, y)}{\partial y} = 2(x + 2y^3) \frac{\partial (x + 2y^3)}{\partial y} = 12y^2(x + 2y^3) \quad (10)$$

here we used the chain rule to compute the partial derivatives.

Example 2



Find the partial derivatives of $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3 \quad (11)$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2 \quad (12)$$

So the gradient is then

$$\frac{df}{dx} = \left[\frac{\partial f(x_1, x_2)}{\partial x_1}, \frac{\partial f(x_1, x_2)}{\partial x_2} \right] = [2x_1 x_2 + x_2^3, x_1^2 + 3x_1 x_2^2] \in \mathbb{R}^{1 \times 2} \quad (13)$$



When we compute derivatives with respect to vectors $x \in \mathbb{R}^n$ we need to pay attention: Our gradients now involve vectors and matrices, and matrix multiplication is not commutative i.e., the order matters.

$$\text{Product rule: } \frac{\partial}{\partial x}(f(x)g(x)) = \frac{\partial f}{\partial x}g(x) + f(x)\frac{\partial g}{\partial x} \quad (14)$$

$$\text{Sum rule: } \frac{\partial}{\partial x}(f(x) + g(x)) = \frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} \quad (15)$$

$$\text{chain rule: } \frac{\partial}{\partial x}(g \circ f)(x) = \frac{\partial}{\partial x}(g(f(x))) = \frac{\partial g}{\partial f} \frac{\partial f}{\partial x} \quad (16)$$



Consider a function $f : \mathbb{R} \rightarrow \mathbb{R}$ of two variables x_1, x_2 .
Furthermore, $x_1(t)$ and $x_2(t)$ are themselves functions of t .

To compute the gradient of f with respect to t , we need to apply the chain rule for multivariate functions as

$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t} \quad (17)$$

where d denotes the gradient and ∂ partial derivatives.

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$ then

$$\begin{aligned}\frac{df}{dt} &= \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t} \\ &= 2 \sin t \frac{\partial \sin t}{\partial t} + 2 \frac{\partial \cos t}{\partial t} \\ &= 2 \sin t \cos t - 2 \sin t = 2 \sin t (\cos t - 1)\end{aligned}$$

is the corresponding derivative of f with respect to t .



If $f(x_1, x_2)$ is a function of x_1 and x_2 , where $x_1(s, t)$ and $x_2(s, t)$ are themselves functions of two variables s and t , the chain rule yields the partial derivatives:

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial s} \quad (18)$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t} \quad (19)$$

and the gradient is obtained by the matrix multiplication

$$\begin{aligned} \frac{df}{d(s, t)} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial (s, t)} \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix} \end{aligned}$$



We have discussed partial derivatives and gradients of functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ mapping to the real numbers. Now we will generalize the concept of the gradient to vector-valued functions

$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.

For a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a vector $x = [x_1, \dots, x_n]^T$ corresponding vector of function values is given as

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \in \mathbb{R}^m \quad (20)$$

where each $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$



Therefore, the partial derivative of a vector-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ w.r.t. $x_i \in R$, $i = 1, \dots, n$ is given as the vector

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} \\ &= \begin{bmatrix} \lim_{h \rightarrow 0} \frac{f_1(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_1(x)}{h} \\ \vdots \\ \lim_{h \rightarrow 0} \frac{f_m(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots, x_n) - f_m(x)}{h} \end{bmatrix} \in \mathbb{R}^m \end{aligned}$$



We know that the gradient of f with respect to a vector is the row vector of the partial derivatives. Every partial derivative $\frac{\partial f}{\partial x_i}$ is itself a column vector. Therefore, we obtain the gradient of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with respect to $x \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\begin{aligned}\frac{df(x)}{dx} &= \left[\frac{\partial f(x)}{\partial x_1} \cdots \frac{\partial f(x)}{\partial x_n} \right] \\ &= \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}\end{aligned}$$

Example 1: Gradients of Vector-Valued Functions



Given $f(x) = Ax$, $f(x) \in \mathbb{R}^M$, $A \in \mathbb{R}^{M \times N}$, $x \in \mathbb{R}^N$

Since $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$, it follows that $df/dx \in \mathbb{R}^{M \times N}$. To compute the gradient we determine the partial derivatives of f w.r.t x_j :

$$f_i(x) = \sum_{j=1}^N A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \quad (21)$$

We obtain the gradient using Jacobian

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = A \in \mathbb{R}^{M \times N} \quad (22)$$

Example 2: Gradients of Vector-Valued Functions



Consider the function $h : \mathbb{R} \rightarrow \mathbb{R}$, $h(t) = (f \circ g)(t)$ with $f(x) = \exp(x_1 x_2^2)$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix} \quad (23)$$

and compute the gradient of h w.r.t. t . Since $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow \mathbb{R}^2$ we note that

$$\frac{\partial f}{\partial x} \in \mathbb{R}^{1 \times 2} \quad \text{and} \quad \frac{\partial g}{\partial t} \in \mathbb{R}^{2 \times 1} \quad (24)$$

The desired gradient is computed by applying the chain rule:

$$\begin{aligned}\frac{dh}{dt} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \\ &= \begin{bmatrix} \exp(x_1 x_2^2) x_2^2 & 2 \exp(x_1 x_2^2) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t \sin t \\ \sin t + t \cos t \end{bmatrix} \\ &= \exp(x_1 x_2^2) (x_2^2 (\cos t - t \sin t) + 2 x_1 x_2 (\sin t + t \cos t))\end{aligned}$$

where $x_1 = t \cos t$ and $x_2 = t \sin t$;