Lecture 6

Math Foundations Team

BITS Pilani

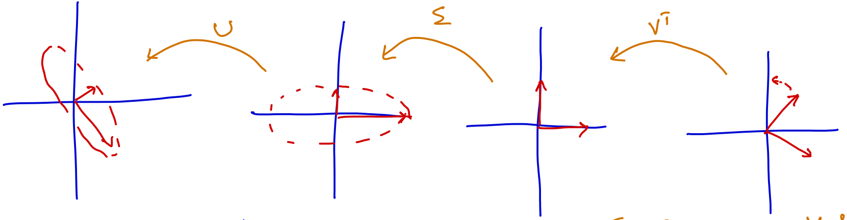
Pilani | Dubai | Goa | Hyderabad

SVD:

Geometric interpretation

$$Ax = \underbrace{U}_{\text{Rotation}}\ \underbrace{\Sigma}_{\text{scale}}\ \underbrace{V^T x}_{\text{Rotation}}$$

V & U are square orthogonal
$\Rightarrow$ preserve length &
preserve relative angle
$\Rightarrow$ Rotation in R^n



* if $A \in R^{n \times n}$ $\quad A = Q \Lambda Q^T$ $\quad$ then: Q^T & Q were inverses
(Q^T "undoes" effect of Q)

NOT so in SVD

if A is square & cols are orthogonal

$$A^T A = I \Rightarrow \underbrace{A^{-1} A = I}_{\text{by defn of inv.}} \Rightarrow \underbrace{A A^{-1} = I}_{\text{inverse should undo from both sides}} \Rightarrow A A^T = I \Rightarrow \text{Rows are ortho}$$

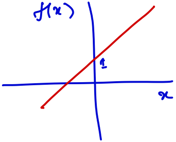
Introduction

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.

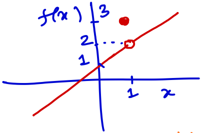
Examples include: linear regression, where we look at curve-fitting problems and optimize linear weight parameters to maximize the likelihood; neural-network auto-encoders for dimensionality reduction and data compression.

limits:

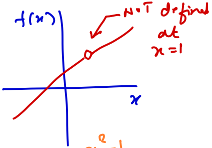
$$\lim_{x \to c} f(x) = L$$



$f(x) = x+1$



$f(x) = \begin{cases} x+1 & x \neq 1 \\ 2 & x = 1 \end{cases}$



$f(x) : \dfrac{x^2-1}{x-1} \quad x \neq 1$

$\to \dfrac{(x+1)(x-1)}{(x-1)}$

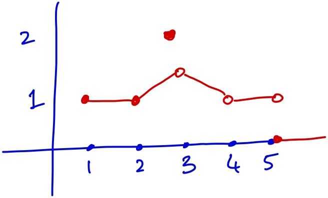
$$\lim_{x \to 1} f(x) = 2$$

even though $f(x) \,!= 2$ at $x=1$

Formally For every $\varepsilon > 0$, if there exists $\delta > 0$ s.T.

$|f(x) - L| < \varepsilon$ for $0 < |x - c| < \delta$ $\Rightarrow$ limit exists at δ

ie, arbitrarily small range in f(x) is defined using a δ range around c



$x \to$	1	2	3	4	5
$\lim_{x \to c^-}$	NA	1	1.5	1	1
$\lim_{x \to c^+}$	1	1	1.5	1	NA
$f(c)$	1	1	2	NA	0

f(x) is continuous at c if:

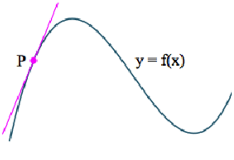
$$\lim_{x \to c^-} f(x) = \lim_{x \to c^+} f(x) = f(c)$$

Differentiation of Univariate Functions

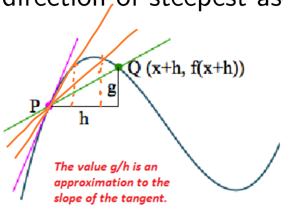
For $h > 0$, the derivative of f at x is defined as the limit

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

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



Slope of the tangent at P.



Slope of the line PQ.

lim $h \to 0$ moves the secant PQ towards tangent at P

$\therefore \frac{df}{dx}$ defines slope

Differentiable

Continuous	Yes	No
Y		ReLU
N	NA diff ⇒ cont	

Derivative of a Polynomial

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

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

$(x+h)^n = \sum_{i=0}^{n} {}^nC_i \, x^{n-i} h^i$ $\rightarrow$

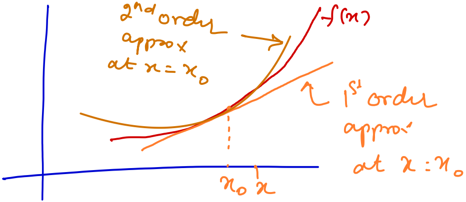
$x^n = x^{n-0} h^0$ $\rightarrow$

Derivative of a Polynomial

innovate achieve lead

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

$^nC_1 = \frac{n!}{(n-1)!} = n$



approx. of function at some point near x_0

1st order: $\underbrace{f(x)}_{\text{some point near } x_0} = \underbrace{f(x_0)}_{\text{value of fn at } x_0} + \underbrace{f'(x)\Big|_{x_0}}_{\text{slope of approximating line (1st order)}} \underbrace{(x-x_0)}_{\text{How far is } x \text{ from } x_0}$

2nd order: $f(x) = f(x_0) + f'(x)\Big|_{x_0}(x-x_0) + \frac{1}{2} f''(x)\Big|_{x_0}(x-x_0)^2$

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

3rd order: ...

:

$$f(x) = \sum_{K=0}^{n} \frac{f^{(K)}(x)}{K!}\Bigg|_{x_0} (x-x_0)^K$$ Taylor polynomial

$$f(x) = \sum_{K=0}^{\infty} \frac{f^{(K)}(x)\Big|_{x_0}}{K!} (x-x_0)^K$$ Taylor Series

Taylor polynomial

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 $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

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

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

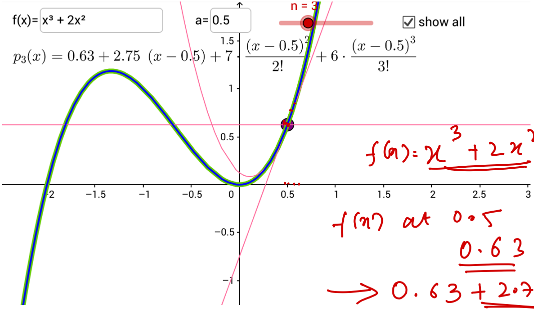
Taylor series

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

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

For $x_0 = 0$, we obtain the Maclaurin series as a 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 $i > k$.



f(x) is 3rd order poly.

Exactly represented using 4 terms

$f(x) = \underline{x^3 + 2x^2}$

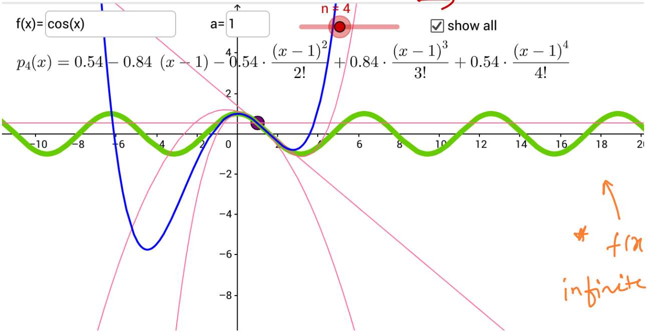
f(x) at 0.5

$\underline{0.63}$ → 0.5

→ $0.63 + \underline{2.75}\,(x - 0.5)$ → $0.48 - 0.53$

$+ 7 \cdot \frac{(x-0.5)^2}{2!}$

→



f(x) Cannot be exactly represented

* f(x) is differentiable infinite number of times

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 $k = 0, 1, 2..., 6$
$f(1) = 1,\ f'(1) = 4,\ f''(1) = 12,\ f^{(3)}(1) = 24,\ f^{(4)}(1) = 24,$
$f^{(5)}(1) = 0,\ f^{(6)}(1) = 0.$ 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

$$
\begin{aligned}
T_\infty(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k \\
&= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \dots \\
&= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \dots x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \dots \qquad (7) \\
&= \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k)!}x^{2k} + \sum_{k=0}^{\infty}(-1)^k \frac{1}{(2k+1)!}x^{2k+1} \\
&= \underbrace{cos(x) + sin(x)}_{\text{exact with infinite terms}}
\end{aligned}
$$

HW 1^{st} order:

if $f(x)$ is approximated using 1^{st} order poly:

$$f(x) = a_0 + a_1(x - x_0)$$

at $x = x_0$; $f(x_0) = a_0$

$$f'(x)\Big|_{x_0} = 0 + a_1\Big|_{x_0} = a_1$$

$$\left.\begin{aligned} a_0 &= f(x_0) \\ a_1 &= f'(x)\big|_{x_0} \end{aligned}\right.$$

$$\Rightarrow f(x) = f(x_0) + f'(x)\Big|_{x_0}(x - x_0)$$

2^{nd} order

$$f(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2$$

at $x = x_0$; $f(x_0) = a_0$

$$\begin{cases} f'(x)\Big|_{x_0} = 0 + a_1 + a_2\, 2\cancel{(x - x_0)}\Big|_{x_0} = a_1 \\ f''(x)\Big|_{x_0} = 0 + 0 + 2a_2\Big|_{x_0} = 2a_2 \end{cases}$$

$a_2 = \frac{1}{2} f''(x)\big|_{x_0}$ $\quad a_1 = f'(x)\big|_{x_0}$ $\quad a_0 = f(x_0)$

$$\Rightarrow f(x) = f(x_0) + f'(x)\Big|_{x_0}(x - x_0) + \frac{1}{2} f''(x)\Big|_{x_0}(x - x_0)^2$$

.... K^{th} order $\rightarrow$

Differentiation Rules

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: $(\frac{f(x)}{g(x)})' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$
- Chain Rule: $(g(f(x))' = (g \circ f)'(x) = g'(f(x))f'(x)$

Example: Chain Rule

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

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

$g(\underbrace{f(x)}_{2x+1})$

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

$g(f) = f^4$

Derivatives of f and g are

$f'(x) = 2$

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

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

$(2x+1)^2$

Sum Rule $\rightarrow$ $h(x) = 4x^2 + 4x + 1 \Rightarrow \frac{df}{dx} = 8x + 4$

Chain Rule $\rightarrow$ $f: 2x+1$, $g: f^2$

$\frac{dh}{dx} = 2(f) \cdot f' = 2(2x+1) \cdot 2 = 8x+4$

Partial Differentiation and Gradients

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.

Partial derivatives and Gradients

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

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

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

$$\vdots$$

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

to find partial diff w.r.t x_i $\left(\frac{\partial f}{\partial x_i}\right)$, change only x_i to $x_i + h$

$\circledast \quad f : \underset{\text{i/p}}{R^n} \to \underset{\text{o/p}}{R} \qquad \nabla_x f \in R^{\overset{\text{o/p}}{1} \times \overset{\text{i/p}}{n}}$

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

How does f change w.r.t x_1

How does f change w.r.t x_2

$$\Delta_x f = gradf = \frac{df}{dx} = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}\right] \tag{8}$$

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

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

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

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

Example 2

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

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 \underbrace{x_2}_{\text{const}} + \overbrace{x_2^3}^{\text{const}} \tag{11}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = \underbrace{x_1^2}_{\text{const}} + 3\underbrace{x_1}_{\text{const}} x_2^2 \tag{12}$$

So the gradient is then

$$\frac{df}{dx} = [\frac{\partial f(x_1, x_2)}{\partial x_1}, \frac{\partial f(x_1, x_2)}{\partial x_2}] = [2x_1x_2 + x_2^3, x_1^2 + 3x_1x_2^2] \in \mathbb{R}^{1\times 2} \tag{13}$$

Basic rules of partial differentiation

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.

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

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

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

Chain Rule

Consider a function $f : \mathbb{R} \to \mathbb{R}$ 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

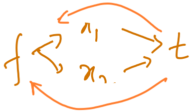
$$\frac{df}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial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

where d denotes the gradient and ∂ partial derivatives.

$$\frac{\partial f}{\partial t_1} = \frac{\partial f}{\partial x_1} \cdot \frac{\partial x_1}{\partial t_1} + \frac{\partial f}{\partial x_2} \cdot \frac{\partial x_2}{\partial t_1} + \cdots$$

Example

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:

$$\frac{\partial f}{\partial s} = \overset{a}{\frac{\partial f}{\partial x_1}} \overset{b}{\frac{\partial x_1}{\partial s}} + \overset{p}{\frac{\partial f}{\partial x_2}} \overset{q}{\frac{\partial x_2}{\partial s}} \tag{18}$$

$$\frac{\partial f}{\partial t} = \overset{a}{\frac{\partial f}{\partial x_1}} \overset{c}{\frac{\partial x_1}{\partial t}} + \overset{p}{\frac{\partial f}{\partial x_2}} \overset{r}{\frac{\partial x_2}{\partial t}} \tag{19}$$

and the gradient is obtained by the matrix multiplication

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial (s,t)}$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix} \qquad \begin{bmatrix} a & p \end{bmatrix} \begin{pmatrix} b & c \\ q & r \end{pmatrix}$$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial s} & \frac{\partial f}{\partial t} \end{bmatrix}$$

vector valued function.

$$f: R^n \to R^m$$

$$f(\bar{x}) = f\left(\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\in R^n}\right) = \underbrace{\begin{bmatrix} f_1(\bar{x}) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}}_{\in R^m}$$

R^n

$f: R \to R$

$f: \underline{\underline{R^n}} \to R$

∇

$f: R^n \to R^m$

$$\frac{df(x)}{dx} = \nabla_x f = \begin{bmatrix} \frac{df_1(x)}{dx} \\ \frac{df_2(x)}{dx} \\ \vdots \end{bmatrix}$$

$$\frac{df_1(x)}{dx} = \begin{bmatrix} \frac{df_1(x_1)}{dx_1} & \frac{df_1(x_2)}{dx_2} & \cdots \end{bmatrix}$$

$$= \left.\underbrace{\begin{bmatrix} \frac{df_1(x_1)}{dx} & \frac{df_1(x_2)}{dx} & & \\ \frac{df_2(x_1)}{dx} & \cdots & & \\ \vdots & & & \\ \frac{df_m(x_1)}{dx} & \frac{df_m(x_2)}{dx} & & \frac{df_m(x_n)}{dx} \end{bmatrix}}_{n}\right\} m \quad \in R^{m \times n}$$

o/p

i/p

i/p

$$\nabla_x f = \#^{o/p}\begin{bmatrix} \quad \end{bmatrix}$$

Numerator layout

∝ Convention used in this course

Gradients of Vector-Valued Functions

We have discussed partial derivatives and gradients of functions $f : \mathbb{R}^n \to \mathbb{R}$ mapping to the real numbers. Now we will generalize the concept of the gradient to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.

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

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

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

Gradients of Vector-Valued Functions

innovate achieve lead

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

$$\frac{\partial f}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix}$$

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

Gradients of Vector-Valued Functions

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 $f : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\frac{df(x)}{dx} = \left[\frac{\partial f(x)}{\partial x_1} \cdots \frac{\partial f(x)}{\partial x_n}\right]$$

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

$f(x) = \underset{\hookrightarrow \mathbb{R}^{M\times N}}{A}\underset{\hookrightarrow \in \mathbb{R}^N}{x}$ $\qquad f: \mathbb{R}^N \to \mathbb{R}^M$

$\therefore \nabla f = M\Big\{\overbrace{\begin{bmatrix} & & \end{bmatrix}}^{N}$

e.g. 2x2:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix} \begin{matrix} \to f_1 \\ \to f_2 \end{matrix}$$

$$\frac{d f(x)}{dx} = \begin{bmatrix} \frac{d f_1(x)}{dx} \\ \frac{d f_2(x)}{dx} \end{bmatrix} = \begin{bmatrix} \frac{d f_1(x_1)}{d x_1} & \frac{d f_1(x_2)}{d x_2} \\ \frac{d f_2(x_1)}{d x_1} & \frac{d f_1(x_2)}{d x_2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d}{dx_1}(a_{11}x_1 + a_{12}x_2) & \frac{d}{dx_2}(a_{11}x_1 + a_{12}x_2) \\ \cdots & \cdots \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\Rightarrow \nabla_x f = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & \cdots & \\ \vdots & & & \end{bmatrix} = A \in \mathbb{R}^{M\times N}$$

Example 1: Gradients of Vector-Valued Functions

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

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

We obtain the gradient using Jacobian

$$\frac{df}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \dots \frac{\partial f_1}{\partial x_N} \\ \vdots \\ \frac{\partial f_M}{\partial x_1} \dots \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} \dots A_{1N} \\ \vdots \\ A_{M1} \dots A_{MN} \end{bmatrix} = \underline{A \in \mathbb{R}^{M \times N}} \tag{22}$$

Example 2: Gradients of Vector-Valued Functions

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

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

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

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

$f : R^2 \to R$

$\therefore \nabla f = {}_1[\quad]^2$

$g : R \to R^2$

$\nabla g = {}_2[\quad]$

The desired gradient is computed by applying the chain rule:

$\frac{\partial h}{\partial f} \to 1$

$$\frac{dh}{dt} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} exp(x_1 x_2^2) x_2^2 & 2exp(x_1 x_2^2) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t \sin t \\ \sin t + t \cos t \end{bmatrix}$$

$$= exp(x_1 x_2^2)(x_2^2(\cos t - t \sin t) + 2 x_1 x_2 (\sin t + t \cos t))$$

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

$$\frac{\partial f}{\partial x_1} = \frac{\partial}{\partial x_1}\left(exp(x_1 x_2^2)\right) = exp(x_1 x_2^2) \frac{d}{dx_1}(x_1 x_2^2) = exp(x_1 x_2^2) \cdot x_2^2$$

$$\frac{\partial f}{\partial x_2} = \frac{\partial}{\partial x_2}\left(exp(x_1 x_2^2)\right) = exp(x_1 x_2^2) \cdot \frac{\partial}{\partial x_2}(x_1 x_2^2) = exp(x_1 x_2^2) \cdot 2 x_1 x_2$$

$\vdots$