Mathematics for Machine Learning

Vazgen Mikayelyan

August 25, 2020

Partial Differentiation and Gradients

Partial Differentiation and Gradients

Definition

For a function $f : \mathbb{R}^n \to \mathbb{R}$ *(of* n *variables* $x_1, \ldots, x_n$*)*

Partial Differentiation and Gradients

Definition

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

$$f'_{x_1}(\mathbf{x}) = \frac{\partial f}{\partial x_1}(\mathbf{x}) = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

$$\vdots$$

$$f'_{x_n}(\mathbf{x}) = \frac{\partial f}{\partial x_n}(\mathbf{x}) = \lim_{h \to 0} \frac{f(x_1, x_2, \dots, x_n + h) - f(x_1, x_2, \dots, x_n)}{h}.$$

Partial Differentiation and Gradients

Definition

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

$$f'_{x_1}(\mathbf{x}) = \frac{\partial f}{\partial x_1}(\mathbf{x}) = \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$
$$\vdots$$
$$f'_{x_n}(\mathbf{x}) = \frac{\partial f}{\partial x_n}(\mathbf{x}) = \lim_{h \to 0} \frac{f(x_1, x_2, \ldots, x_n + h) - f(x_1, x_2, \ldots, x_n)}{h}.$$

The row vector

$$\nabla f = \frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n},$$

is called the **gradient** *of f or the Jacobian.*

Partial Differentiation and Gradients

Differentiation Rules

If the functions f and g have partial derivatives, then

Sum Rule: $$\frac{\partial}{\partial x_i}(f(\mathrm{x}) + g(\mathrm{x})) = \frac{\partial f(\mathrm{x})}{\partial x_i} + \frac{\partial g(\mathrm{x})}{\partial x_i}$$

Product Rule: $$\frac{\partial}{\partial x_i}(f(\mathrm{x})\; g(\mathrm{x})) = \frac{\partial f(\mathrm{x})}{\partial x_i} g(\mathrm{x}) + f(\mathrm{x}) \frac{\partial g(\mathrm{x})}{\partial x_i}$$

Partial Differentiation and Gradients

Differentiation Rules

If the functions f and g have partial derivatives, then

Sum Rule: $$\frac{\partial}{\partial x_i}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f(\mathbf{x})}{\partial x_i} + \frac{\partial g(\mathbf{x})}{\partial x_i}$$

Product Rule: $$\frac{\partial}{\partial x_i}(f(\mathbf{x}) \cdot g(\mathbf{x})) = \frac{\partial f(\mathbf{x})}{\partial x_i} g(\mathbf{x}) + f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial x_i}$$

Example

Find the gradient of the following functions:

a) $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by* $f(x_1, x_2) = (2x_1 + 3x_2)^3$

b) $f : \mathbb{R}^3 \to \mathbb{R}$ *defined by* $f(x, y, z) = e^{2x} + y^2 z^3$

Partial Differentiation and Gradients

Differentiation Rules

If the functions f and g have partial derivatives, then

Sum Rule: $$\frac{\partial}{\partial x_i}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f(\mathbf{x})}{\partial x_i} + \frac{\partial g(\mathbf{x})}{\partial x_i}$$

Product Rule: $$\frac{\partial}{\partial x_i}(f(\mathbf{x}) \cdot g(\mathbf{x})) = \frac{\partial f(\mathbf{x})}{\partial x_i} g(\mathbf{x}) + f(\mathbf{x}) \frac{\partial g(\mathbf{x})}{\partial x_i}$$

Example

Find the gradient of the following functions:

a) $f : \mathbb{R}^2 \to \mathbb{R}$ *defined by* $f(x_1, x_2) = (2x_1 + 3x_2)^3$

b) $f : \mathbb{R}^3 \to \mathbb{R}$ *defined by* $f(x, y, z) = e^{2x} + y^2 z^3$

Example

Given $z(x, y) = x^2 + y^2$ *where* $x(r, t) = r\cos(t)$ *and* $y(r, t) = r + t$, *determine the value of* $\frac{\partial z}{\partial t}$ *and* $\frac{\partial z}{\partial r}$.

Chain Rule

Chain Rule

Let z be a function of two variables, $x; y$ and each of these variables $x; y$ be in turn functions of two variables, $t; s$.

Chain Rule

Let z be a function of two variables, x, y and each of these variables x, y be in turn functions of two variables, t, s. Then

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t} =$$

Chain Rule

Let z be a function of two variables, x, y and each of these variables x, y be in turn functions of two variables, t, s. Then

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t} = \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial t} \end{bmatrix} = \nabla z \frac{\partial \mathbf{x}}{\partial t}$$

Chain Rule

Let z be a function of two variables, x, y and each of these variables x, y be in turn functions of two variables, t, s. Then

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t} = \begin{pmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{pmatrix} \begin{bmatrix} \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial t} \end{bmatrix} = \nabla z \frac{\partial \mathbf{x}}{\partial t}$$

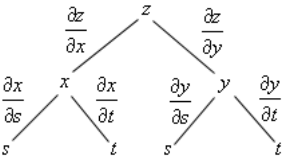
$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s} = \begin{pmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{pmatrix} \begin{bmatrix} \frac{\partial x}{\partial s} \\ \frac{\partial y}{\partial s} \end{bmatrix} = \nabla z \frac{\partial \mathbf{x}}{\partial s}.$$

Chain Rule

Let z be a function of two variables, x, y and each of these variables x, y be in turn functions of two variables, t, s. Then

$$\frac{\partial z}{\partial t} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial t} = \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial t} \end{bmatrix} = \nabla z \frac{\partial \mathbf{x}}{\partial t}$$

$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s} = \begin{bmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial s} \\ \frac{\partial y}{\partial s} \end{bmatrix} = \nabla z \frac{\partial \mathbf{x}}{\partial s}.$$



Example

Given $z(x, y) = x^2 + y^2$ *where* $x(r, t) = r\cos(t)$ *and* $y(r, t) = r\sin(t)$*, determine the value of* $\frac{\partial z}{\partial t}$ *and* $\frac{\partial z}{\partial r}$ *using the chain rule. Verify the results by expressing* z *as a function of* r, t *and computing the partial derivatives directly.*

Example

Given $z(x, y) = x^2 + y^2$ *where* $x(r, t) = r\cos(t)$ *and* $y(r, t) = r\sin(t)$, *determine the value of* $\frac{\partial z}{\partial t}$ *and* $\frac{\partial z}{\partial r}$ *using the chain rule. Verify the results by expressing* z *as a function of* r, t *and computing the partial derivatives directly.*

In general, assume z is a function of n variables, $x_1, \ldots, x_n$ and each of these variables are in turn functions of m variables, $t_1, t_2, \ldots, t_m$. Then for any variable $t_i, i = 1, 2, \ldots, m$ we have the following,

$$\frac{\partial z}{\partial t_i} = \frac{\partial z}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \frac{\partial z}{\partial x_2}\frac{\partial x_2}{\partial t_i} + \ldots + \frac{\partial z}{\partial x_n}\frac{\partial x_n}{\partial t_i}$$

Example

Given $z(x, y) = x^2 + y^2$ *where* $x(r, t) = r\cos(t)$ *and* $y(r, t) = r\sin(t)$*, determine the value of* $\frac{\partial z}{\partial t}$ *and* $\frac{\partial z}{\partial r}$ *using the chain rule. Verify the results by expressing* z *as a function of* r, t *and computing the partial derivatives directly.*

In general, assume z is a function of n variables, $x_1, \ldots, x_n$ and each of these variables are in turn functions of m variables, $t_1, t_2, \ldots, t_m$. Then for any variable $t_i, i = 1, 2, \ldots, m$ we have the following,

$$\frac{\partial z}{\partial t_i} = \frac{\partial z}{\partial x_1}\frac{\partial x_1}{\partial t_i} + \frac{\partial z}{\partial x_2}\frac{\partial x_2}{\partial t_i} + \ldots + \frac{\partial z}{\partial x_n}\frac{\partial x_n}{\partial t_i}$$

Example

Find the partial derivatives $\frac{\partial z}{\partial t_i}$, $i = 1, 2, 3$ *of the function* $z(x, y)$ *where* $x = t_1 + 2t_2 + 4t_3$ *and* $y = t_1 - 3t_2 + 5t_3$.

Gradients of Vector-Valued Functions

Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector valued function.

Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector valued function. Then for a vector $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T \in \mathbb{R}^n$, the value of the function $\mathbf{f}$ at point $\mathbf{x}$ is a *vector* given as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m$$

Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be a vector valued function. Then for a vector $\mathbf{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T \in \mathbb{R}^n$, the value of the function $\mathbf{f}$ at point $\mathbf{x}$ is a *vector* given as

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^m$$

Here the functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are real-valued functions.

Gradients of Vector-Valued Functions

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

$$\frac{\partial \mathbf{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \frac{\partial f_2}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} \in \mathbb{R}^m.$$

Gradients of Vector-Valued Functions

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

$$\frac{\partial \mathbf{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \frac{\partial f_2}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} \in \mathbb{R}^m.$$

Hence the **gradient of the vector-valued function** $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is

$$\nabla \mathbf{f} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \dots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Gradients of Vector-Valued Functions

Note that gradient of the vector-valued function can also be represented as follows

$$\nabla \mathbf{f} = \begin{bmatrix} \nabla f_1 \\ \vdots \\ \nabla f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Note that gradient of the vector-valued function can also be represented as follows

$$\nabla \mathbf{f} = \begin{bmatrix} \nabla f_1 \\ \vdots \\ \nabla f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Definition

The **Jacobian matrix** *is the matrix of all first-order partial derivatives of a vector-valued function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$*. The Jacobian matrix* $\mathbf{J}$ *of* $\mathbf{f}$ *is an* $m \times n$ *matrix, usually defined and arranged as follows:*

$$\mathbf{J} = \nabla \mathbf{f} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}; \quad \mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j}.$$

Gradients of Vector-Valued Functions

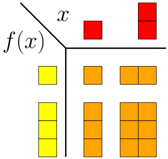


Figure: The dimension of the Jacobian $\mathbf{J_f}$

Gradients of Vector-Valued Functions

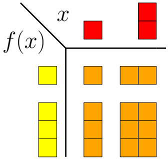


Figure: The dimension of the Jacobian $\mathbf{J_f}$

In particular, the Jacobian of a function $f : \mathbb{R}^n \to \mathbb{R}^1$, which maps a vector $\mathbf{x} \in \mathbb{R}^n$ onto a scalar, is a row vector (matrix of dimension $1 \times n$).

Gradients of Vector-Valued Functions

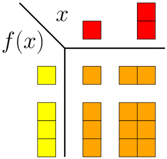


Figure: The dimension of the Jacobian $\mathbf{J_f}$

In particular, the Jacobian of a function $f : \mathbb{R}^n \to \mathbb{R}^1$, which maps a vector $\mathbf{x} \in \mathbb{R}^n$ onto a scalar, is a row vector (matrix of dimension $1 \times n$).

Example

Find the Jacobian of the following functions:

a) $f : \mathbb{R}^2 \to \mathbb{R}$; *given by* $f(\mathbf{x}) = x_1 + x_2^3$

b) $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^3$; *given by* $f(\mathbf{x}) = [2x_1;\ x_1x_2;\ x_1 + 3x_2]^T$

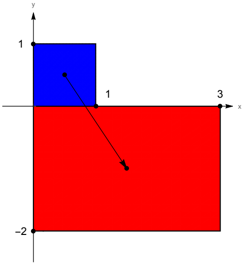
Gradients of Vector-Valued Functions

Consider the vector valued function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x, y) = \begin{pmatrix} 3x \\ 2y \end{pmatrix}.$$

Consider the vector valued function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

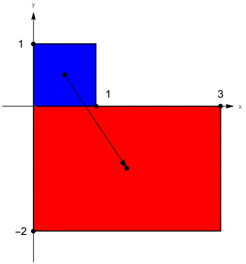
$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x, y) = \begin{pmatrix} 3x \\ -2y \end{pmatrix}.$$



Gradients of Vector-Valued Functions

Consider the vector valued function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(x, y) = \begin{pmatrix} 3x \\ -2y \end{pmatrix}.$$



Note that the image of the $[0, 1]^2$ (the blue square) is the rectangle $[0, 3] \times [-2, 0]$ (depicted in red). The quotient of the areas of the rectangle and the square is 6 and it is equal to $|\det \mathbf{J_f}| = \left|\det \begin{pmatrix} 3 & 0 \\ 0 & -2 \end{pmatrix}\right|$

Gradients of Vector-Valued Functions

Gradients of Vector-Valued Functions

A nonlinear map $f : \mathbb{R}^2 \to \mathbb{R}^2$ sends a small square (left, in red) to a distorted parallelogram (right, in red). The Jacobian at a point gives the best linear approximation of the distorted parallelogram near that point (right, in translucent white), and the Jacobian determinant gives the ratio of the area of the approximating parallelogram to that of the original square.

Gradients of Vector-Valued Functions

Example

Find the gradient of the vector-valued function $f : \mathbb{R}^2 \to \mathbb{R}^2$, given by

$$f(x) = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Gradients of Vector-Valued Functions

Example

Find the gradient of the vector-valued function $f : \mathbb{R}^2 \to \mathbb{R}^2$; given by

$$f(x) = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Example

Prove that the Jacobian of the vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$; given by $f(x) = Ax$; where $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$ is

Gradients of Vector-Valued Functions

Example

Find the gradient of the vector-valued function $f : \mathbb{R}^2 \to \mathbb{R}^2$, given by $f(x) = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

Example

Prove that the Jacobian of the vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$, given by $f(x) = Ax$, where $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$ is the matrix A, i.e.

$$J_f = A.$$

Chain Rule (Matrix form)

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable functions, then

$$J_{f \circ g}(a) = J_f(g(a)) J_g(a); \quad a \in \mathbb{R}^n.$$

Chain Rule (Matrix form)

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable functions, then

$$J_{f \circ g}(a) = J_f(g(a)) J_g(a); \quad \forall a \in \mathbb{R}^n.$$

Note that $J_{f \circ g}(a) \in \mathbb{R}^{k \times n}$; $J_f(g(a)) \in \mathbb{R}^{k \times m}$; and $J_g(a) \in \mathbb{R}^{m \times n}$.

Chain Rule (Matrix form)

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable functions, then

$$J_{f \circ g}(a) = J_f(g(a)) J_g(a), \quad a \in \mathbb{R}^n.$$

Note that $J_{f \circ g}(a) \in \mathbb{R}^{k \times n}$, $J_f(g(a)) \in \mathbb{R}^{k \times m}$, and $J_g(a) \in \mathbb{R}^{m \times n}$.
Equivalently, if $z = f(y)$ and $y = g(x)$ then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}.$$

Chain Rule (Matrix form)

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable functions, then

$$J_{f \circ g}(a) = J_f(g(a)) J_g(a), \quad a \in \mathbb{R}^n.$$

Note that $J_{f \circ g}(a) \in \mathbb{R}^{k \times n}$, $J_f(g(a)) \in \mathbb{R}^{k \times m}$, and $J_g(a) \in \mathbb{R}^{m \times n}$.
Equivalently, if $z = f(y)$ and $y = g(x)$ then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x}.$$

Example

Find the Jacobians of the functions $f \circ g$ and $g \circ f$, where $f : \mathbb{R}^2 \to \mathbb{R}$, given by $f(x) = x_1 + x_2^2$ and $g : \mathbb{R} \to \mathbb{R}^2$ given by $g(t) = \begin{pmatrix} e^t \\ t^2 \end{pmatrix}$.

Gradients of Vector-Valued Functions

Example

Using the formula $\frac{\partial x^T B x}{\partial x} = x^T(B + B^T)$ deduce that $\frac{\partial x^T x}{\partial x} = 2x^T$.

Gradients of Matrices

	Scalar y		Vector $\mathbf{y}$ (size m)	
	Notation	Type	Notation	Type
Scalar x	$\frac{\partial y}{\partial x}$	scalar	$\frac{\partial \mathbf{y}}{\partial x}$	size-m col. vector
Vector $\mathbf{x}$ (size n)	$\frac{\partial y}{\partial \mathbf{x}}$	size-n row vector	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	$m \times n$ matrix
Matrix X (size $p \times q$)	$\frac{\partial y}{\partial X}$	$p \times q$ matrix	$\frac{\partial \mathbf{y}}{\partial X}$	$m \times (p \times q)$ tensor

	Scalar y Notation Type	Vector y (size m) Notation Type
Scalar x	$\frac{\partial y}{\partial x}$ scalar	$\frac{\partial y}{\partial x}$ size-m col. vector
Vector x (size n)	$\frac{\partial y}{\partial x}$ size-n row vector	$\frac{\partial y}{\partial x}$ $m \times n$ matrix
Matrix X (size $p \times q$)	$\frac{\partial y}{\partial X}$ $p \times q$ matrix	$\frac{\partial y}{\partial X}$ $m \times (p \times q)$ tensor

Example

(Gradient of Scalars with respect to Matrices) Let

$$y = y(X) = \operatorname{tr}(X); \text{ where } X \in \mathbb{R}^{p \times p}.$$

Find the gradient $\frac{\partial y}{\partial X}$.

Gradients of Matrices

Example

(Gradient of Vectors with respect to Matrices)

Let $v \in \mathbb{R}^q$ be a fixed vector and $f : \mathbb{R}^{p \times q} \to \mathbb{R}^p$ be a function given by

$$f(X) = Xv; \quad \text{where } X \in \mathbb{R}^{p \times q}.$$

Find the gradient $\frac{\partial y}{\partial X}$ of the function $y = f(X)$.

Gradients of Matrices

	Matrix (size $m \times k$) Notation Type
Scalar x	$\frac{\partial Y}{\partial x}$ $m \times k$ matrix
Vector x (size n)	$\frac{\partial Y}{\partial x}$ $(m \times k) \times n$ tensor
Matrix X (size $p \times q$)	$\frac{\partial Y}{\partial X}$ $(m \times k) \times (p \times q)$ tensor

Gradients of Matrices

	Matrix Y (size $m \times k$) Notation Type
Scalar x	$\frac{\partial Y}{\partial x}$ $m \times k$ matrix
Vector x (size n)	$\frac{\partial Y}{\partial x}$ $(m \times k) \times n$ tensor
Matrix X (size $p \times q$)	$\frac{\partial Y}{\partial X}$ $(m \times k) \times (p \times q)$ tensor

Example

(Gradient of Matrices with respect to Matrices)

Let $f : \mathbb{R}^{p \times q} \to \mathbb{R}^{q \times q}$ be a function given by

$$f(X) = X^T X, \text{ where } X \in \mathbb{R}^{p \times q}.$$

Find the gradient $\frac{\partial}{\partial X}$ of the function $= f(X)$.

Useful Identities for Computing Gradients

Let $a \in \mathbb{R}^n$ and $\varepsilon > 0$. Denote $B(a;\varepsilon) = \{x \in \mathbb{R}^n : \|x - a\| < \varepsilon\}$.

Limits of Multivariable Functions

Let $a \in \mathbb{R}^n$ and $\varepsilon > 0$. Denote $B(a;\varepsilon) = \{x \in \mathbb{R}^n : \|x - a\| < \varepsilon\}$.

Definition

Let $f : X \to \mathbb{R}^m$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^n$ and $A \in \mathbb{R}^m$. We will say that $\lim_{x \to a} f(x) = A$ if for all $\varepsilon > 0$ there exists $\delta > 0$, such that from $0 < \|x - a\|_n < \delta; x \in X$, follows that $\|f(x) - A\|_m < \varepsilon$.

Limits of Multivariable Functions

Let $a \in \mathbb{R}^n$ and $\varepsilon > 0$. Denote $B(a;\varepsilon) = \{x \in \mathbb{R}^n : \|x - a\| < \varepsilon\}$.

Definition

Let $f : X \to \mathbb{R}^m$, $X \subset \mathbb{R}^n$, $a \in \mathbb{R}^n$ and $A \in \mathbb{R}^m$. We will say that $\lim_{x \to a} f(x) = A$ if for all $\varepsilon > 0$ there exists $\delta > 0$, such that from $0 < \|x - a\|_n < \delta; x \in X$, follows that $\|f(x) - A\|_m < \varepsilon$.

Definition

Let $f : X \times Y \to \mathbb{R}$, $X; Y \subset \mathbb{R}$, $(x_0; y_0) \in \mathbb{R}^2$ and $A \in \mathbb{R}$. We will say that $\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x) = A$ if for all $\varepsilon > 0$ there exists $\delta > 0$, such that from $|x - x_0| < \delta$; $|y - y_0| < \delta$; $(x_0; y_0) \neq (0;0)$; $x \in X; y \in Y$, follows that $|f(x) - A| < \varepsilon$.

Limits of Multivariable Functions

Theorem

If $\lim\limits_{\substack{x \to x_0 \\ y \to y_0}} f(x; y) = A$ and $\lim\limits_{x \to x_0} f(x; y) = \varphi(y)$ for all $y \in Y$, $y \neq y_0$, then

$$\lim_{y \to y_0} \varphi(y) = \lim_{y \to y_0} \lim_{x \to x_0} f(x; y) = A.$$

Theorem

If $\lim\limits_{\substack{x \to x_0 \\ y \to y_0}} f(x;y) = A$ and $\lim\limits_{x \to x_0} f(x;y) = \varphi(y)$ for all $y \in Y$, $y \neq y_0$, then

$$\lim_{y \to y_0} \varphi(y) = \lim_{y \to y_0} \lim_{x \to x_0} f(x;y) = A.$$

Example

1. $f(x;y) = x \sin \frac{1}{y}$, $(x_0;y_0) = (0;0)$,

Theorem

If $\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x; y) = A$ and $\lim_{x \to x_0} f(x; y) = \varphi(y)$ for all $y \in Y$, $y \neq y_0$, then

$$\lim_{y \to y_0} \varphi(y) = \lim_{y \to y_0} \lim_{x \to x_0} f(x; y) = A.$$

Example

1. $f(x; y) = x \sin \frac{1}{y}$, $(x_0; y_0) = (0; 0)$,
2. $f(x; y) = \begin{cases} 0; & \text{if } x \neq y; \\ 1; & \text{if } x = y \end{cases}$; $(x_0; y_0) = (0; 0)$.

Di erential

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and $(x_0; y_0)$ is an interior point of X.

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and $(x_0; y_0)$ is an interior point of X.

Definition

f is called differentiable at the point $(x_0; y_0)$ if there exists $A; B \in \mathbb{R}$ such that

$$f(x_0 + \Delta x; y_0 + \Delta y) = f(x_0; y_0) + A\Delta x + B\Delta y + o(\rho); \quad \rho \to 0;$$

where $\rho = \sqrt{\Delta x^2 + \Delta y^2}$.

Differential

Let $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^2$ and $(x_0; y_0)$ is an interior point of X.

Definition

f is called differentiable at the point $(x_0; y_0)$ if there exists $A; B \in \mathbb{R}$ such that

$$f(x_0 + \Delta x; y_0 + \Delta y) = f(x_0; y_0) + A\Delta x + B\Delta y + o(\rho); \quad \rho \to 0;$$

where $\rho = \sqrt{\Delta x^2 + \Delta y^2}$.

Theorem

If partial derivatives of the first degree of f are continuous at $(x_0; y_0)$ then it is differentiable at $(x_0; y_0)$. The inverse is not true.

Differential

Example

$$f(x;y) = \begin{cases} x^2 \sin \dfrac{1}{x}; & if \ (x;y) \neq (0;0); \\ 0; & if \ (x;y) = (0;0) \end{cases}$$

Differential

Example

$$f(x;y) = \begin{cases} x^2 \sin\dfrac{1}{x}; & \text{if } (x;y) \neq (0;0); \\ 0; & \text{if } (x;y) = (0;0) \end{cases}$$

Definition

$df(x_0;y_0) = \dfrac{\partial f}{\partial x}(x_0;y_0) \quad x + \dfrac{\partial f}{\partial y}(x_0;y_0) \quad y$ *is called differential of* f.