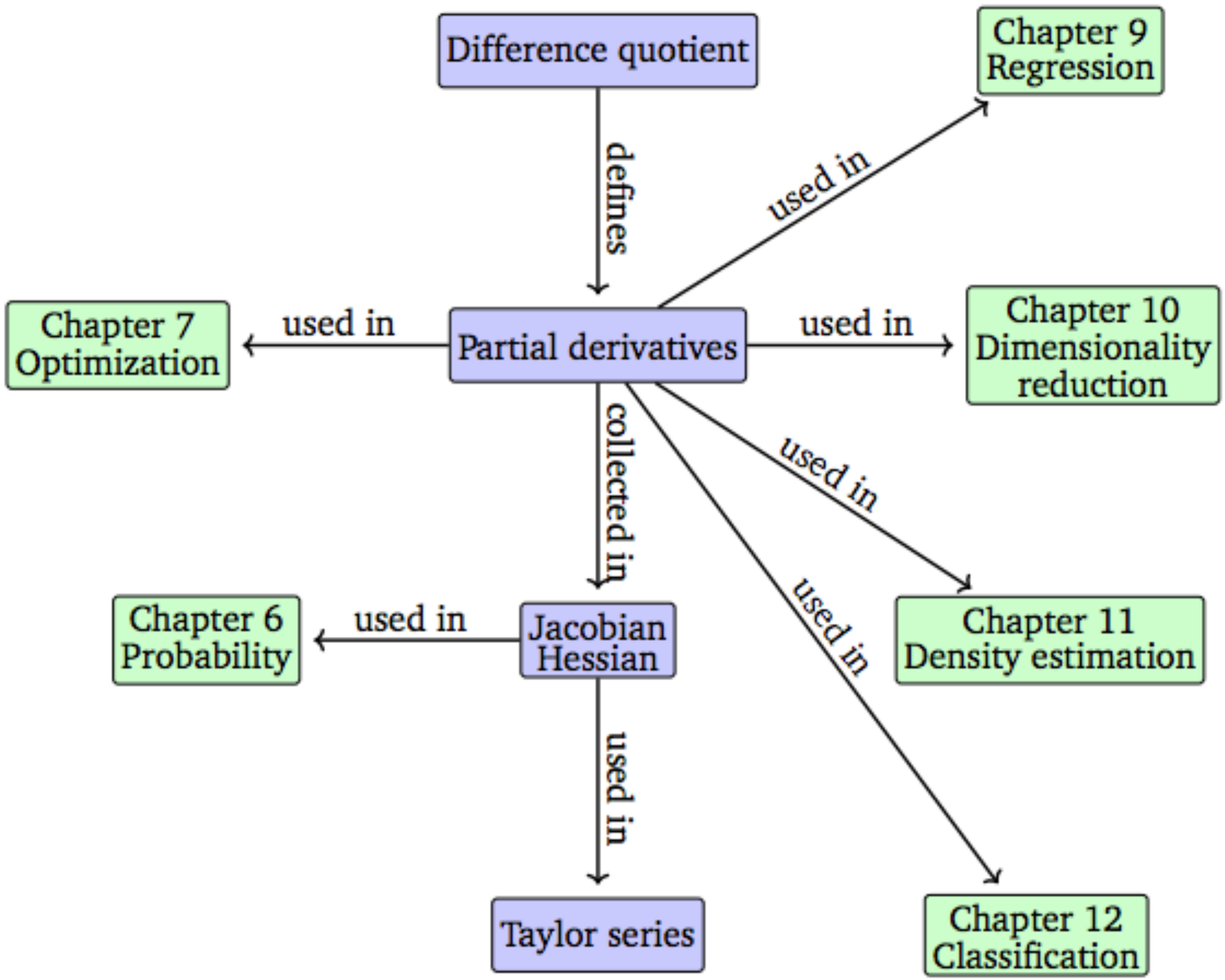
Vector Calculus

Liang Zheng
Australian National University
liang.zheng@anu.edu.au


Difference quotient
defines
Chapter 9
Regression
used in
Chapter 7
Optimization
used in
Partial derivatives
used in
Chapter 10
Dimensionality
reduction
collected in
used in
used in
Chapter 6
Probability
used in
Jacobian
Hessian
Chapter 11
Density estimation
used in
Taylor series
Chapter 12
Classification

5 Vector Calculus

- We discuss functions

$$f : \mathbb{R}^D \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto f(\boldsymbol{x})$$

where $\mathbb{R}^D$ is the domain of f, and the function values $f(\boldsymbol{x})$ are the image/codomain of f.

- Example (dot product)
- Previously, we write dot product as

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}, \qquad \boldsymbol{x} \in \mathbb{R}^2$$

- In this chapter, we write it as

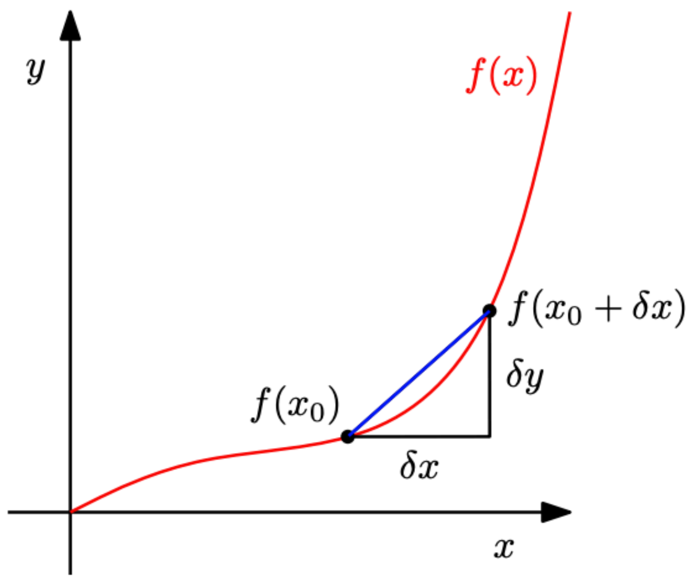
$$f : \mathbb{R}^2 \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto x_1^2 + x_2^2$$

5.1 Differentiation of Univariate Functions

- Given $y = f(x)$, the difference quotient is defined as

$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x}$$

- It computes the slope of the secant line through two points on the graph of f. In this figure, these are the points with x–coordinates x_0 and $x_0 + \delta x_0$.
- In the limit for $\delta x \to 0$, we obtain the tangent of f at x (if f is differentiable). The tangent is then the derivative of f at x.



5.1 Differentiation of Univariate Functions

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

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

- The derivative of f points in the direction of steepest ascent of f.
- Example - Derivative of a Polynomial
- Compute the derivative of $f(x) = x^n,\ n \in \mathbb{N}$. (From our high school knowledge, the derivative is nx^{n-1}.)

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &:= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{(x+h)^n - x^n}{h} \\ &= \lim_{h \to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h}\end{aligned}$$

we see that $x^n = \binom{n}{0} x^{n-0} h^0$. By starting the sum at 1, the x^n cancels.

5.1 Differentiation of Univariate Functions

$$
\begin{aligned}
\frac{\mathrm{d}f}{\mathrm{d}x} &= \lim_{h\to 0} \frac{\sum_{i=0}^{n}\binom{n}{i}x^{n-i}h^{i} - x^{n}}{h} \\
&= \lim_{h\to 0} \frac{\sum_{i=1}^{n}\binom{n}{i}x^{n-i}h^{i}}{h} \\
&= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i}h^{i-1} \\
&= \lim_{h\to 0} \left\{ \binom{n}{1}x^{n-1} + \underbrace{\sum_{i=2}^{n}\binom{n}{i} x^{n-i}h^{i-1}}_{\to\, 0 \text{ as } h \to 0} \right\} \\
&= nx^{n-1}
\end{aligned}
$$

5.1.2 Differentiation Rules

- Product rule

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

- Quotient rule:

$$(\frac{f(x)}{g(x)})' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$$

- Sum rule:

$$(f(x) + g(x))' = f'(x) + g'(x)$$

- Chain rule:

$$(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$$

Here, $g \circ f$ denotes function composition $g\big(f(x)\big)$

Example -- Chain rule

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

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

- We obtain the derivatives of f and g as,

$$f'(x) = 2$$
$$g'(f) = 4f^3$$

- The derivative of h is given as

$$h'^{(x)} = g'^{(f)} f'^{(x)} = (4f^3) \cdot 2 = 4(2x + 1)^3 \cdot 2 = 8(2x + 1)^3$$

5.2 Partial Differentiation and Gradients

- Instead of considering $x \in \mathbb{R}$, we consider $\boldsymbol{x} \in \mathbb{R}^n$, e.g., $f(\boldsymbol{x}) = f(x_1, x_2)$
- The generalization of the derivative to functions of several variables is the gradient.
- We find the gradient of the function f with respect to $\boldsymbol{x}$ by
  - varying one variable at a time and keeping the others constant.
  - The gradient is the collection of these partial derivatives.
- For a function $f: \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \dots, x_n$, we define the partial derivatives as

$$\frac{\partial f}{\partial x_1} := \lim_{h \to 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(\boldsymbol{x})}{h}$$
$$\vdots$$
$$\frac{\partial f}{\partial x_n} := \lim_{h \to 0} \frac{f(x_1, \dots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h}$$

and collect them in the row vector

$$\nabla_x f = \mathrm{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1} \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n} \quad \dots \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \mathbb{R}^{1 \times n}$$

5.2 Partial Differentiation and Gradients

- $\nabla_x f = \mathrm{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1} \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n} \quad \dots \quad \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right] \in \mathbb{R}^{1\times n}$
- n is the number of variables and 1 is the dimension of the image/range/codomain of f
- The row vector $\nabla_x f \in \mathbb{R}^{1\times n}$ is called the gradient of f or the Jacobian.

- Example - Partial Derivatives Using the Chain Rule
- For $f(x,y) = (x+2y^3)^2$, we obtain the partial derivatives

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

5.2 Partial Differentiation and Gradients

- For $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$, the partial derivatives (i.e., the derivatives of f with respect to x_1 and x_2 are

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

- and the gradient is then

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(x_1, x_2)}{\partial x_1} & \frac{\partial f(x_1, x_2)}{\partial x_2} \end{bmatrix} = [2x_1 x_2 + x_2^3 \quad x_1^2 + 3x_1 x_2^2] \in \mathbb{R}^{1\times 2}$$

5.2.1 Basic Rules of Partial Differentiation

- Product rule:

$$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x})g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} g(\boldsymbol{x}) + f(\boldsymbol{x})\frac{\partial g}{\partial \boldsymbol{x}}$$

- Sum rule:

$$\frac{\partial}{\partial \boldsymbol{x}}\big(f(\boldsymbol{x}) + g(\boldsymbol{x})\big) = \frac{\partial f}{\partial \boldsymbol{x}} + \frac{\partial g}{\partial \boldsymbol{x}}$$

- Chain rule:

$$\frac{\partial}{\partial \boldsymbol{x}}(g \circ f)(x) = \frac{\partial}{\partial \boldsymbol{x}}\Big(g\big(f(\boldsymbol{x})\big)\Big) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \boldsymbol{x}}$$

5.2.2 Chain Rule

- Consider a function $f: \mathbb{R}^2 \to \mathbb{R}$ of two variables x_1 and x_2.
- $x_1(t)$ and $x_2(t)$ are themselves functions of t.
- To compute the gradient of f with respect to t, we apply the chain rule:

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

Where d denotes the gradient and ∂ partial derivates.

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

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t}$$

$$= 2 \sin t \frac{\partial \sin t}{\partial t} + 2 \frac{\partial \cos t}{\partial t}$$

$$= 2 \sin t \cos t - 2 \sin t = 2 \sin t (\cos t - 1)$$

- The above is the corresponding derivative of f with respect to t.

5.2.2 Chain Rule

- If $f(x_1, x_2)$ is a function of x_1 and x_2, where $f: \mathbb{R}^2 \to \mathbb{R}$, $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

- The gradient can be obtained by the matrix multiplication

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

5.3 Gradients of Vector-Valued Functions

- We discussed partial derivatives and gradients of function $f: \mathbb{R}^n \to \mathbb{R}$
- We will generalize the concept of the gradient to vector-valued functions (vector fields) $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.
- For a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\mathrm{T} \in \mathbb{R}^n$, the corresponding vector of function values is given as

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

- Writing the vector-valued function in this way allows us to view a vector valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ as a vector of functions $[f_1, \ldots, f_m]^\mathrm{T}$, $f_i: \mathbb{R}^n \to \mathbb{R}$ that map onto $\mathbb{R}$.
- The differentiation rules for every f_i are exactly the ones we discussed before.

5.3 Gradients of Vector-Valued Functions

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

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

- In above, every partial derivative $\frac{\partial \boldsymbol{f}}{\partial x_i}$ is a column vector
- Recall that the gradient of f with respect to a vector is the row vector of the partial derivatives
- Therefore, we obtain the gradient of $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$, by collecting these partial derivatives:

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

5.3 Gradients of Vector-Valued Functions

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

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

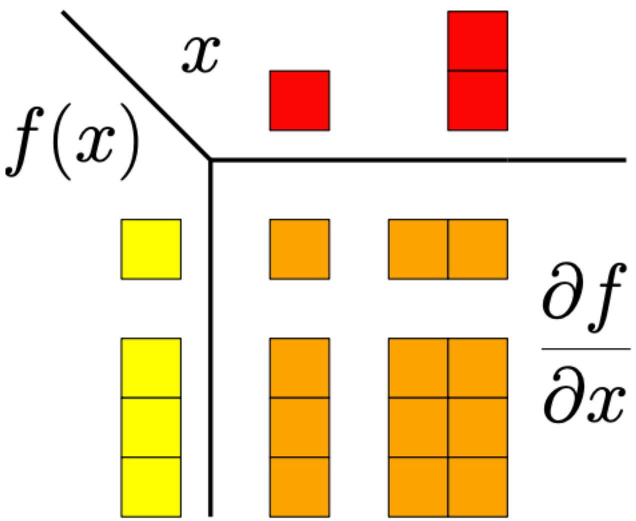
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \boldsymbol{J}(i,j) = \frac{\partial f_i}{\partial x_j}$$

- The elements of $\boldsymbol{f}$ define the rows and the elements of $\boldsymbol{x}$ define the columns of the corresponding Jacobian

- Special case: for a function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^1$ which maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a scalar, i.e., $m = 1$, the Jacobian is a row vector of dimension $1 \times n$.

5.3 Gradients of Vector-Valued Functions

- If $f: \mathbb{R} \to \mathbb{R}$, the gradient is a scalar
- If $f: \mathbb{R}^D \to \mathbb{R}$, the gradient is a $1 \times D$ row vector
- If $\boldsymbol{f}: \mathbb{R} \to \mathbb{R}^E$, the gradient is a $E \times 1$ column vector
- If $\boldsymbol{f}: \mathbb{R}^D \to \mathbb{R}^E$, the gradient is an $E \times D$ matrix



Example - Gradient of a Vector-Valued Function

- We are given $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \ \boldsymbol{A} \in \mathbb{R}^{M\times N}, \ \boldsymbol{x} \in \mathbb{R}^N$.
- To compute the gradient $d\boldsymbol{f}/d\boldsymbol{x}$ we first determine the dimension of $d\boldsymbol{f}/d\boldsymbol{x}$: Since $\boldsymbol{f}: \mathbb{R}^N \to \mathbb{R}^M$, it follows that $d\boldsymbol{f}/d\boldsymbol{x} \in \mathbb{R}^{M\times N}$.
- Then, we determine the partial derivatives of $\boldsymbol{f}$ with respect to every x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \Rightarrow \frac{\partial f_i}{\partial x_j} = A_{ij}$$

- We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \ \in \ \mathbb{R}^{M\times N}$$

Example - Chain Rule

- Consider the function $h: \mathbb{R} \to \mathbb{R},\ h(t) = (f \circ g)(t)$ with

$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$g: \mathbb{R} \to \mathbb{R}^2$$
$$f(\boldsymbol{x}) = \exp(x_1 x_2^2)$$
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$$

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

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}, \qquad \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1}$$

- The desired gradient is computed by applying the chain rule:

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

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

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

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

Example - Gradient of a Least-Squares Loss in a Linear Model

- Let us consider the linear model

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$ are input features and $\boldsymbol{y} \in \mathbb{R}^N$ are the corresponding observations. We define the functions

$$L(\boldsymbol{e}) := \| \boldsymbol{e} \|^2,$$
$$\boldsymbol{e}(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}$$

- We seek $\frac{\partial L}{\partial \boldsymbol{\theta}}$, and we will use the chain rule for this purpose. L is called a least-squares loss function.
- First, we determine the dimensionality of the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D}$$

- The chain rule allows us to compute the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$$

Example - Gradient of a Least-Squares Loss in a Linear Model

- We know that $||\boldsymbol{e}||^2 = \boldsymbol{e}^\mathrm{T}\boldsymbol{e}$ and determine

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^\mathrm{T} \in \mathbb{R}^{1\times N}$$

- Further, we obtain

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$$

- Our desired derivative is

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\boldsymbol{e}^\mathrm{T}\boldsymbol{\Phi} = -\underbrace{2\left(\boldsymbol{y}^\mathrm{T} - \boldsymbol{\theta}^\mathrm{T}\boldsymbol{\Phi}^\mathrm{T}\right)}_{1\times \mathrm{N}} \underbrace{\boldsymbol{\Phi}}_{\mathrm{N}\times \mathrm{D}} \in \mathbb{R}^{1\times D}$$

5.4 Gradients of Matrices

- Consider the following example

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{f} \in \mathbb{R}^{M}, \qquad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \qquad \boldsymbol{x} \in \mathbb{R}^{N}$$

- We seek the gradient $\frac{d\boldsymbol{f}}{dA}$
- First, we determine the dimension of the gradient

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M\times(M\times N)}$$

- By definition, the gradient is the collection of the partial derivatives:

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}$$

- To compute the partial derivatives, we explicitly write out the matrix vector multiplication

$$f_i = \sum_{j=1}^{N} A_{ij}x_j, \qquad i = 1, \cdots, M,$$

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \qquad i = 1, \cdots, M,$$

- The partial derivatives are then given as

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

- Partial derivatives of f_i with respect to a row of $\boldsymbol{A}$ are given as

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^{\mathrm{T}} \in \mathbb{R}^{1\times 1\times N}, \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = \boldsymbol{0}^{\mathrm{T}} \in \mathbb{R}^{1\times 1\times N}$$

- Since f_i maps onto $\mathbb{R}$ and each row of $\boldsymbol{A}$ is of size $1\times N$, we obtain a $1\times 1\times N$ sized tensor as the partial derivative of f_i with respect to a row of $\boldsymbol{A}$.
- We stack the partial derivatives and get the desired gradient

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \boldsymbol{0}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{x}^{\mathrm{T}} \\ \boldsymbol{0}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{0}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{1\times(M\times N)}$$

Example - Gradient of Matrices with Respect to Matrices

- Consider a matrix $\boldsymbol{R} \in \mathbb{R}^{M \times N}$ and $\boldsymbol{f}: \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{N \times N}$ with

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}^{\mathrm{T}} \boldsymbol{R} =: \boldsymbol{K} \in \mathbb{R}^{N \times N}$$

- We seek the gradient $\frac{d\boldsymbol{K}}{d\boldsymbol{R}}$
- First, the dimension of the gradient is given as

$$\frac{d\boldsymbol{K}}{d\boldsymbol{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)}$$

$$\frac{dK_{pq}}{d\boldsymbol{R}} \in \mathbb{R}^{1 \times M \times N}$$

for $p, q = 1, \ldots, N$, where K_{pq} is the pqth entry of $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$.

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^{\mathrm{T}} \boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp} R_{mq}$$

Example - Gradient of Matrices with Respect to Matrices

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^{\mathrm{T}} \boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp} R_{mq}$$

- We now compute the partial derivative $\frac{\partial K_{pq}}{\partial R_{ij}}$, we obtain

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial R_{ij}} R_{mp} R_{mq} = \partial_{pqij}$$

$$\partial_{pqij} = \begin{cases} R_{iq} & if \;\; j = p, p \neq q \\ R_{ip} & if \;\; j = q, p \neq q \\ 2R_{iq} & if \;\; j = p, p = q \\ 0 & \text{otherwise} \end{cases}$$

- The desired gradient has the dimension $(N \times N) \times (M \times N)$, and every single entry of this tensor is given by ∂_{pqij}, where $p, q, j = 1, \ldots, N$ and $i = 1, \ldots, M$

5.5 Useful Identities for Computing Gradients

- Some useful gradients that are frequently required in machine learning
- $\mathrm{tr}(\cdot)$: trace $\det(\cdot)$: determinant $\boldsymbol{f}(\boldsymbol{X})^{-1}$: the inverse of $\boldsymbol{f}(\boldsymbol{X})$

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}})$$

$$\frac{\partial}{\partial \boldsymbol{s}}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W}$$

You should be able to calculate these gradients