

Vector Calculus 1 of 2

Functions

The most common function we will consider is

$$f : \mathbb{R}^D \rightarrow \mathbb{R} \tag{1}$$

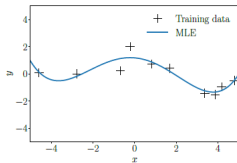
$$\mathbf{x} \mapsto f(\mathbf{x}) \tag{2}$$

A function f assigns every input \mathbf{x} exactly one function value $f(\mathbf{x})$.

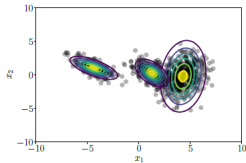
Functions and Models

Typically in machine learning we quantify the error as a function from $\mathbb{R}^D \rightarrow \mathbb{R}$

- Where the model parameters are usually \mathbb{R}^D and the how well we approximate target value is in \mathbb{R}
- Actually we will most often be in a situation where f has data pattern input as well as a set of modeling parameters.
- Knowing the gradient of f with respect to one or more of the model parameters allows us to optimize them to reduce the error in our fit/prediction.

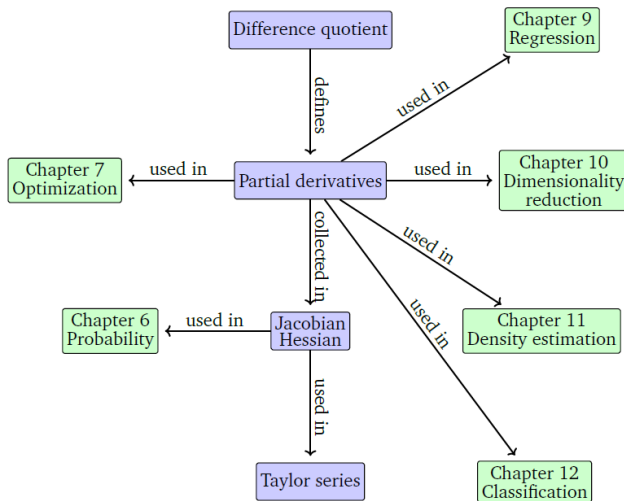


(a) Regression problem: Find parameters, such that the curve explains the observations (crosses) well.



(b) Density estimation with a Gaussian mixture model: Find means and covariances, such that the data (dots) can be explained well.

Mind Map



Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

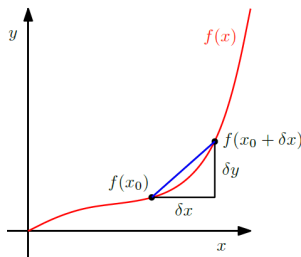
Differentiation of Univariate Functions

Difference Quotient

The *difference quotient*

$$\frac{\delta y}{\delta x} = \frac{f(x + \delta x) - f(x)}{\delta x} \quad (3)$$

computes the slope of the secant line through two points on the graph of f .



Source: M.P. Deisenroth et al, Mathematics for Machine Learning (First Edition)

Derivative of a Polynomial

$$(x+y)^2 = x^2 + 2xy + y^2$$

While we don't always derive a derivative from first principles it is informative to do so for the $f(x) = x^n$.

- We can actually be even more rigorous, but this is sufficient for our purposes

Let us use our definition of a derivative

$$\frac{df}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (4)$$

$$= \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} \quad (5)$$

$$= \lim_{h \rightarrow 0} \frac{\sum_{i=0}^n \binom{n}{i} x^{n-i} h^i - x^n}{h} \quad (6)$$

The last step is an application of the binomial formula, and

$$\binom{n}{i} = \frac{n(n-1)\dots(n-i+1)}{i!} = \frac{n!}{i!(n-i)!} \text{ and } n! = \prod_{i=1}^n i$$

Derivative of a Polynomial

Consider in $\sum_{i=0}^n \binom{n}{i} x^{n-i} h^i - x^n$ that $\binom{n}{0} x^{n-0} h^0 = x^n$ it follows that

$$\frac{df}{dx} = \lim_{h \rightarrow 0} \frac{\sum_{i=1}^n \binom{n}{i} x^{n-i} h^i}{h} \quad (7)$$

$$= \lim_{h \rightarrow 0} \sum_{i=1}^n \binom{n}{i} x^{n-i} h^{i-1} \quad (8)$$

$$= \lim_{h \rightarrow 0} \binom{n}{1} x^{n-1} + \underbrace{\sum_{i=2}^n \binom{n}{i} x^{n-i} h^{i-1}}_{\rightarrow 0} \quad (9)$$

$$= \frac{n!}{1!(n-1)!} x^{n-1} \quad (10)$$

$$= nx^{n-1} \quad (11)$$

$$n! = n \cdot (n-1)!$$

Differentiation Rules: Foundational rules

- **Product** rule: $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
- **Quotient** rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$
- **Sum** rule: $(f(x) + g(x))' = f'(x) + g'(x)$
- **Chain** rule: $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$
 - ▶ $(g \circ f)$ denotes function composition $x \mapsto f(x) \mapsto g(f(x))$.

Chain rule example

Consider $h(x) = (2x + 1)^4$ and we want to find h' . We can think h as

$$h(x) = g(f(x)) \quad (14)$$

$$f(x) = 2x + 1 \quad (15)$$

$$g(f) = f^4 \quad (16)$$

we can get the derivatives separately and then use the chain rule. Specifically,

$$f'(x) = 2 \quad (17)$$

$$g'(f) = 4f^3 \quad (18)$$

$$h'(x) = g'(f)f'(x) = (4f^3)2 = 8(2x + 1)^3 \quad (19)$$

Partial Differentiation and Gradients

The generalization of the derivative to functions of several variables is the *gradient*. Which we will need partial differentiation to obtain

Partial Derivative and Gradient

For $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $\mathbf{x} \mapsto f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$. The component of \mathbf{x} are represented with x_1, \dots, x_n we define the *partial derivatives* as

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \rightarrow 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(\mathbf{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \rightarrow 0} \frac{f(x_1 + h, \dots, x_{n-1}, x_n + h) - f(\mathbf{x})}{h} \end{aligned} \tag{26}$$

and collect them in the row vector, called the *gradient* of f

$$\nabla_{\mathbf{x}} f = \text{grad} f = \frac{df}{d\mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \dots \frac{\partial f(\mathbf{x})}{\partial x_n} \right] \in \mathbb{R}^{1 \times n} \tag{27}$$

Chain Rule Example

Consider

$$f(x, y) = (x + 2y^3)^2 \quad (28)$$

then

$$\frac{\partial f(x, y)}{\partial x} = 2(x + 2y^3) \frac{\partial}{\partial x} (x + 2y^3) = 2(x + 2y^3) \quad (29)$$

and

$$\frac{\partial f(x, y)}{\partial y} = 2(x + 2y^3) \frac{\partial}{\partial y} (x + 2y^3) = 12(x + 2y^3)y^2 \quad (30)$$

Gradient Example

Consider

$$f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \quad (31)$$

and we want to find $\nabla_{\mathbf{x}} f$ then we need

$$\nabla_{\mathbf{x}} f = \frac{df}{d\mathbf{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right] \in \mathbb{R}^{1 \times n} \quad (32)$$

where

$$\frac{\partial f}{\partial x_1} = 2x_1 x_2 + x_2^3 \quad (33)$$

$$\frac{\partial f}{\partial x_2} = x_1^2 + 3x_1 x_2^2 \quad (34)$$

Partial Differentiation Rules

- **Product** rule:

$$\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}}g(\mathbf{x}) + f(\mathbf{x})\frac{\partial g}{\partial \mathbf{x}} \quad (35)$$

- **Sum** rule:

$$\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}) + g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}} \quad (36)$$

- **Chain** rule:

$$\frac{\partial}{\partial \mathbf{x}}(f \circ g)(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}f(g(\mathbf{x})) = \frac{\partial f}{\partial g} \frac{\partial g}{\partial \mathbf{x}}. \quad (37)$$

Gradients of Vector-Valued Functions

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

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \quad (38)$$

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

- So we can use our rules of partial differentiation from earlier for each f_i

Gradients of Vector-Valued Functions

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

$$\frac{\partial \mathbf{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \rightarrow 0} \frac{f_1(x_1, \dots, x_i + h, \dots, x_n) - f_1(\mathbf{x})}{h} \\ \vdots \\ \lim_{h \rightarrow 0} \frac{f_m(x_1, \dots, x_i + h, \dots, x_n) - f_m(\mathbf{x})}{h} \end{bmatrix} \quad (39)$$

Gradients of Vector-Valued Functions

We are now in a good position to define the gradient of a vector valued function. Specifically,

$$\frac{d\mathbf{f}}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} \quad (40)$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad (41)$$

$\frac{d\mathbf{f}}{d\mathbf{x}}$ is also referred to as the **Jacobian**, $J = \nabla_{\mathbf{x}}\mathbf{f}$.

Chain rule with the gradient

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function of x_1 and x_2 where both are functions of t . Then what is the gradient with respect to t ?

$$\frac{df}{dt} = \frac{df}{d\mathbf{x}} \frac{d\mathbf{x}}{dt} \quad (42)$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \quad (43)$$

$$= \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial t} \quad (44)$$

Chain rule with the gradient

Let $f(x_1, x_2)$ be a function of x_1, x_2 where $x_1(t_1, t_2)$ and $x_2(t_1, t_2)$ are functions of two variables. Can we find $\frac{df}{dt}$?

$$\frac{df}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{t}} \quad (45)$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t_1} & \frac{\partial x_1}{\partial t_2} \\ \frac{\partial x_2}{\partial t_1} & \frac{\partial x_2}{\partial t_2} \end{bmatrix} \quad (46)$$

Gradient of a Vector-Valued Function)

Let $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A} \in \mathbb{R}^{M \times N}$ and $\mathbf{x} \in \mathbb{R}^n$ then

$$\frac{d\mathbf{f}}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} \quad (47)$$

$$= \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \mathbf{A} \quad (48)$$

since

$$\frac{\partial f_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{j=1}^N A_{ij} x_j \quad (49)$$

$$= A_{ij} \quad (50)$$

Chain Rule Example with a Vector-Valued Function

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

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad \mathbf{g} : \mathbb{R} \rightarrow \mathbb{R}^2 \quad (51)$$

$$f(\mathbf{x}) = e^{x_1 x_2^2} \quad (52)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{g}(t) = \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix} \quad (53)$$

Let us find $\frac{dh}{dt}$.

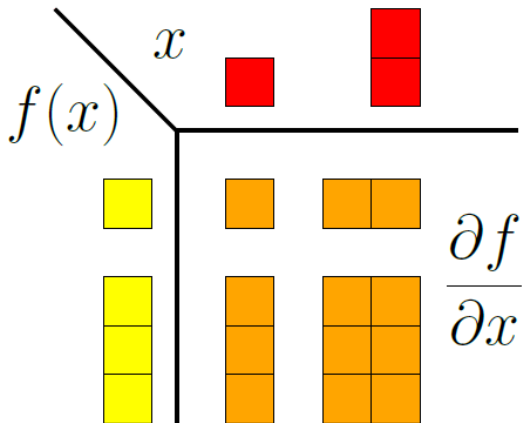
$$\frac{dh}{dt} = \frac{\partial f}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial t} \quad (54)$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} \quad (55)$$

$$= \begin{bmatrix} e^{x_1 x_2^2} x_2^2 & 2e^{x_1 x_2^2} x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t \sin t \\ \sin t + t \cos t \end{bmatrix} \quad (56)$$

$$= e^{x_1 x_2^2} (x_2^2 (\cos t - t \sin t) + 2x_1 x_2 (\sin t - t \cos t)) \quad (57)$$

Dimensionality of (partial) derivatives



Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Jacobian and the Area/Volume Change under Variable Transformations

Can we determine how much scaling a change of variables causes?

- Consider we have two sets of vectors

$$\{\mathbf{b}_1, \mathbf{b}_2\} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \quad (67)$$

$$\{\mathbf{c}_1, \mathbf{c}_2\} = \left\{ \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\} \quad (68)$$

- If there is a mapping that will convert \mathbf{b}_1 to \mathbf{c}_1 and \mathbf{b}_2 to \mathbf{c}_2 how much does it scale the space by?

Jacobian and the Area/Volume Change under Variable Transformations

Approach 1:

- Note that both $(\mathbf{b}_1, \mathbf{b}_2)$ and $(\mathbf{c}_1, \mathbf{c}_2)$ form basis for \mathbb{R}^2 (Show)
- There means that there exists a linear transformation, \mathbf{J} that will ensure that $\mathbf{J}\mathbf{b}_1 = \mathbf{c}_1$ and $\mathbf{J}\mathbf{b}_2 = \mathbf{c}_2$.
 - ▶ Since the first basis is just the canonical one we can find \mathbf{J} directly as

$$\mathbf{J} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \quad (69)$$

- The amount of scaling is just

$$|\det(\mathbf{J})| = 3 \quad (70)$$

Jacobian and the Area/Volume Change under Variable Transformations

Approach 2:

- We can write this change of variable as the following

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} -2x_1 + x_2 \\ x_1 + x_2 \end{bmatrix} \quad (71)$$

which represents the functional relationship between \mathbf{x} and \mathbf{f} , which allows us to get the partial derivatives

$$\frac{\partial y_1}{\partial x_1} = -2, \quad \frac{\partial y_1}{\partial x_2} = 1, \quad \frac{\partial y_2}{\partial x_1} = 1, \quad \frac{\partial y_2}{\partial x_2} = 1 \quad (72)$$

which is just the Jacobian

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \quad (73)$$

- The amount of scaling is just

$$|\det(\mathbf{J})| = 3 \quad (74)$$

Jacobian and the Area/Volume Change under Variable Transformations

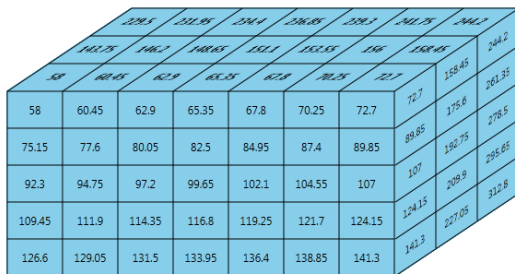
Approach 2:

- The Jacobian represents the coordinate transformation we are looking for.
 - ▶ It is exact if the coordinate transformation is linear
- If the coordinate transformation is nonlinear, the Jacobian approximates this nonlinear transformation locally with a linear one.

Gradients of Matrices

In machine learning we encounter situations where we need to take gradients of matrices with respect to vectors (or even other matrices),


- which results in a multidimensional tensor.
 - ▶ We can think of a tensor as a multidimensional array.



Gradients of Matrix valued Function with Respect to a Matrix

$$\mathbf{F}(\mathbf{B}) = \begin{bmatrix} F_{11}(\mathbf{B}) & \cdots & F_{1n}(\mathbf{B}) \\ \vdots & & \vdots \\ F_{m1}(\mathbf{B}) & \cdots & F_{mn}(\mathbf{B}) \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} B_{11} & \cdots & B_{1q} \\ \vdots & & \vdots \\ \underline{B_{p1}} & \cdots & B_{pq} \end{bmatrix} \quad (1)$$

For a bit more context

- If we need to compute the gradient of a $m \times n$ matrix valued function \mathbf{F} with respect to a $p \times q$ matrix \mathbf{B}
 - ▶ The Jacobian will be a $(m \times n) \times (p \times q)$
 - ▶ A fourth order tensor \mathbf{J} , whose entries are given as
- 

$$\mathbf{J}_{ijkl} = \frac{\partial F_{ij}}{\partial B_{kl}} \quad (2)$$

Gradients of Matrix valued Function with Respect to a Matrix

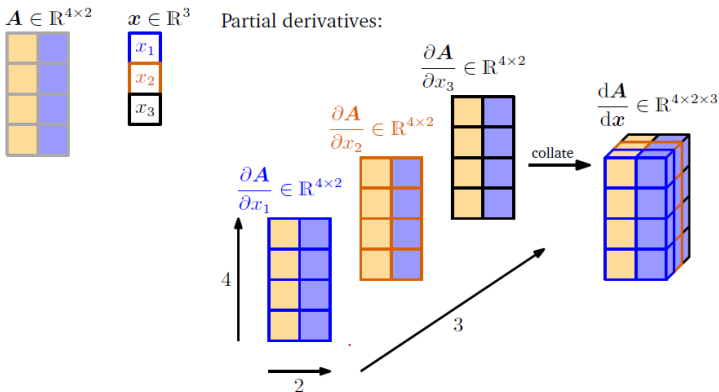
Since matrices represent linear mappings, we can exploit the fact that there is a vector-space isomorphism (linear, invertible mapping) between

- the space $\mathbb{R}^{m \times n}$ matrices and the space \mathbb{R}^{mn} of mn vectors

If we consider our earlier example as a $m \times n$ matrix valued function with a $p \times q$ matrix input we can reshape to rather work with:

- A mn vector valued function with a pq vector input.

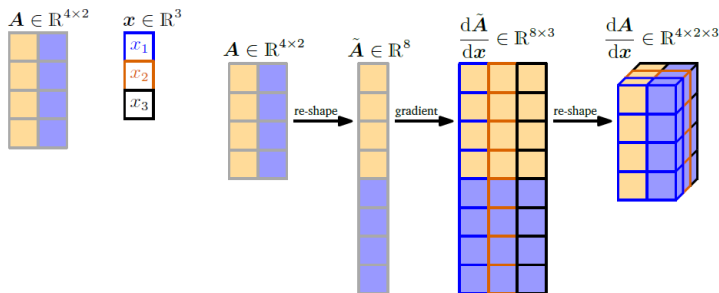
Approach 1: Visualization



(a) Approach 1: We compute the partial derivative $\frac{\partial A}{\partial x_1}$, $\frac{\partial A}{\partial x_2}$, $\frac{\partial A}{\partial x_3}$, each of which is a 4×2 matrix, and collate them in a $4 \times 2 \times 3$ tensor.

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Approach 2: Visualization



(b) Approach 2: We re-shape (flatten) $A \in \mathbb{R}^{4 \times 2}$ into a vector $\tilde{A} \in \mathbb{R}^8$. Then, we compute the gradient $\frac{d\tilde{A}}{dx} \in \mathbb{R}^{8 \times 3}$. We obtain the gradient tensor by re-shaping this gradient as illustrated above.

Source: M.P. Deisenroth *et al*, Mathematics for Machine Learning (First Edition)

Gradient of Vectors with Respect to Matrices

Consider the following

$$\mathbf{f} = \mathbf{A}\mathbf{x}, \quad \mathbf{f} \in \mathbb{R}^M, \quad \mathbf{A} \in \mathbb{R}^{M \times N}, \quad \mathbf{x} \in \mathbb{R}^N \quad (3)$$

We want to find $\frac{d\mathbf{f}}{d\mathbf{A}}$. Firstly it is informative to identify the expected size, namely:

$$\frac{d\mathbf{f}}{d\mathbf{A}} \in \mathbb{R}^{M \times (M \times N)} \quad (4)$$

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

$$\frac{d\mathbf{f}}{d\mathbf{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \mathbf{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \mathbf{A}} \end{bmatrix} \quad (5)$$

where

$$\frac{\partial f_i}{\partial \mathbf{A}} \in \mathbb{R}^{1 \times (M \times N)} \quad (6)$$

Gradient of Vectors with Respect to Matrices

To compute the partial derivatives, it will be helpful to explicitly write out the matrix vector multiplication:

$$f_i = \sum_{j=1}^N A_{ij} x_j, \quad i = 1, \dots, M \quad (7)$$

and the partial derivatives are then given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q \quad (8)$$

Which means that

$$\frac{\partial f_i}{\partial A_{i,:}} = \mathbf{x}^T \in \mathbb{R}^{1 \times 1 \times N} \quad (9)$$

$$\frac{\partial f_i}{\partial A_{k \neq i,:}} = \mathbf{0}^T \in \mathbb{R}^{1 \times 1 \times N} \quad (10)$$

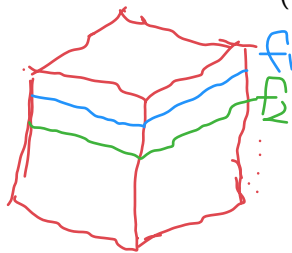
We typically don't write a row vector this way, but it makes it clear how it fits into the tensor form.

Gradient of Vectors with Respect to Matrices

All we need to do is stack the partial derivatives and get the desired gradients

$$\frac{\partial f_i}{\partial \mathbf{A}} = \begin{bmatrix} \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \\ \mathbf{x}^T \\ \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \end{bmatrix} \in \mathbb{R}^{1 \times (M \times N)} \quad (11)$$

How would you write $\frac{df}{d\mathbf{A}}$ though?



Gradient of Matrices with Respect to Matrices

Consider $\mathbf{R} \in \mathbb{R}^{M \times N}$ and $\mathbf{F} : \mathbb{R}^{M \times N} \rightarrow \mathbb{R}^{N \times N}$ with

$$\mathbf{F}(\mathbf{R}) = \mathbf{R}^T \mathbf{R} =: \mathbf{K} \in \mathbb{R}^{N \times N} \quad (12)$$

Can we find $\frac{d\mathbf{K}}{d\mathbf{R}}$?

- Once again it is informative to identify the expected size, namely:

$$\frac{d\mathbf{K}}{d\mathbf{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)} \quad (13)$$

a 4D tensor.

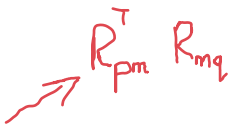
- Where the components of the “first matrix” are

$$\frac{dK_{pq}}{d\mathbf{R}} \quad (14)$$

for $p, q = 1, \dots, N$ where K_{pq} is the (p, q) th entry \mathbf{k}

Gradient of Matrices with Respect to Matrices

Let us first consider what K_{pq} is

$$K_{pq} = \mathbf{r}_p^T \mathbf{r}_q = \sum_{m=1}^M R_{mp} R_{mq} \quad (15)$$


where \mathbf{r}_i is the i th column of \mathbf{R} . We can now compute all the elements of $\frac{d\mathbf{K}}{d\mathbf{R}}$ namely

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^M \frac{\partial}{\partial R_{ij}} R_{mp} R_{mq} = \begin{cases} R_{iq} & \text{if } j = p, p \neq q \\ R_{ip} & \text{if } j = q, q \neq p \\ 2R_{iq} & \text{if } j = p, p = q \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

Useful Identities for Computing Gradients

There are a number useful gradient identities that are frequently required in a machine learning context listed in 5.5.

- These can be seen as given (unless stated otherwise)
- These will be provided in a test like setting if needed.

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^\top = \left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)^\top$$

$$\frac{\partial}{\partial \mathbf{X}} \text{tr}(f(\mathbf{X})) = \text{tr} \left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)$$

$$\frac{\partial}{\partial \mathbf{X}} \det(f(\mathbf{X})) = \det(f(\mathbf{X})) \text{tr} \left(f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \right)$$

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X})^{-1} = -f(\mathbf{X})^{-1} \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} f(\mathbf{X})^{-1}$$

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

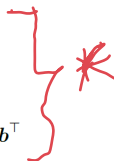
$$\frac{\partial \mathbf{a}^\top \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} = -(\mathbf{X}^{-1})^\top \mathbf{a} \mathbf{b}^\top (\mathbf{X}^{-1})^\top$$

$$\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}^\top$$

$$\frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}^\top$$

$$\frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^\top$$

$$\frac{\partial \mathbf{x}^\top \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{x}^\top (\mathbf{B} + \mathbf{B}^\top)$$



Source: M.P. Deisenroth et al, Mathematics for Machine Learning (First Edition)