


Vector Calculus

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

October 12, 2021

Overview

Index Notation

Differentiation of vector-valued functions

Multivariate Chain Rule

Vector Differentiation: Index Notation

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (1)$$

How to find $\frac{\partial L}{\partial \boldsymbol{\theta}} = [\frac{\partial L}{\partial \theta_i}]_i$?

Vector Differentiation: Index Notation

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (1)$$

How to find $\frac{\partial L}{\partial \boldsymbol{\theta}} = [\frac{\partial L}{\partial \theta_i}]_i$?

- ▶ Vector expressions can always be written in terms of operations on scalars.

Vector Differentiation: Index Notation

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (1)$$

How to find $\frac{\partial L}{\partial \boldsymbol{\theta}} = [\frac{\partial L}{\partial \theta_i}]_i$?

- ▶ Vector expressions can always be written in terms of operations on scalars.
- ▶ We know how to find partial derivatives of scalar expressions \implies rewrite into index notation.

Vector Differentiation: Index Notation

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (1)$$

How to find $\frac{\partial L}{\partial \boldsymbol{\theta}} = [\frac{\partial L}{\partial \theta_i}]_i$?

- ▶ Vector expressions can always be written in terms of operations on scalars.
- ▶ We know how to find partial derivatives of scalar expressions \implies rewrite into index notation.
- ▶ Take partial derivatives using tricks.

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i \quad (3)$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i \quad (3)$$

$$\mathbf{y} = f(\mathbf{x})\mathbf{x} \implies y_i = f(\mathbf{x})x_i \quad (4)$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i \quad (3)$$

$$\mathbf{y} = f(\mathbf{x})\mathbf{x} \implies y_i = f(\mathbf{x})x_i \quad (4)$$

$$A\mathbf{x} = ? \quad (5)$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i \quad (3)$$

$$\mathbf{y} = f(\mathbf{x})\mathbf{x} \implies y_i = f(\mathbf{x})x_i \quad (4)$$

$$A\mathbf{x} = ? \quad (5)$$

$$[A\mathbf{x}]_j = \sum_j A_{ij}x_j \quad (6)$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \quad (2)$$

$$\mathbf{x}^\top \mathbf{y} = \sum_i x_i y_i \quad (3)$$

$$\mathbf{y} = f(\mathbf{x})\mathbf{x} \implies y_i = f(\mathbf{x})x_i \quad (4)$$

$$A\mathbf{x} = ? \quad (5)$$

$$[A\mathbf{x}]_j = \sum_j A_{ij}x_j \quad (6)$$

$$A\mathbf{x} = \left[\sum_j A_{ij}x_j \right]_i \quad (7)$$

Index Notation of Vector Expressions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (8)$$

$$= \sum_{n=1}^N \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right)^2 \quad (9)$$

We want to find $\frac{\partial L}{\partial \boldsymbol{\theta}}$, i.e. $[\frac{\partial L}{\partial \theta_i}]_i$ for all i .

Index Notation of Vector Expressions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (8)$$

$$= \sum_{n=1}^N \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right)^2 \quad (9)$$

We want to find $\frac{\partial L}{\partial \boldsymbol{\theta}}$, i.e. $[\frac{\partial L}{\partial \theta_i}]_i$ for all i .

$$\frac{\partial}{\partial \theta_i} \sum_{n=1}^N \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right)^2 \quad (10)$$

Index Notation of Vector Expressions

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (8)$$

$$= \sum_{n=1}^N \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right)^2 \quad (9)$$

We want to find $\frac{\partial L}{\partial \boldsymbol{\theta}}$, i.e. $[\frac{\partial L}{\partial \theta_i}]_i$ for all i .

$$\frac{\partial}{\partial \theta_i} \sum_{n=1}^N \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right)^2 \quad (10)$$

$$= \sum_n 2 \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \frac{\partial}{\partial \theta_i} \left(y_n - \sum_{k=1}^M \phi_k(x_n) \theta_k \right) \quad (11)$$

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.
- ▶ Denoted as **kroncker delta** δ_{ik} .

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.
- ▶ Denoted as **kroncker delta** δ_{ik} .
- ▶ Think about what happens to $\sum_k \delta_{ik}$.

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.
- ▶ Denoted as **kronecker delta** δ_{ik} .
- ▶ Think about what happens to $\sum_k \delta_{ik}$.
- ▶ Think about what happens to $\sum_a v_a \delta_{ai}$.

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.
- ▶ Denoted as **kronecker delta** δ_{ik} .
- ▶ Think about what happens to $\sum_k \delta_{ik}$.
- ▶ Think about what happens to $\sum_a v_a \delta_{ai}$.

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \delta_{ki} \quad (13)$$

Index Differentiation

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \frac{\partial \theta_k}{\partial \theta_i} \quad (12)$$

- ▶ Think about $\frac{\partial \theta_k}{\partial \theta_i}$. What value does it take?
- ▶ If $i \neq k$, then 0; if $i = k$, then 1.
- ▶ Denoted as **kronecker delta** δ_{ik} .
- ▶ Think about what happens to $\sum_k \delta_{ik}$.
- ▶ Think about what happens to $\sum_a v_a \delta_{ai}$.

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \sum_k \phi_k(x_n) \delta_{ki} \quad (13)$$

$$= -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (14)$$

Vector Notation of Index Expressions

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (15)$$

Sometimes we prefer a compact vector expression again. Index notation is ambiguous about row/col vector. If we remember convention, we can go back:

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

=

=

=

(16)

Vector Notation of Index Expressions

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (15)$$

Sometimes we prefer a compact vector expression again. Index notation is ambiguous about row/col vector. If we remember convention, we can go back:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}} &= 2 \sum_n (\boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta} - y_n) \boldsymbol{\phi}(x_n)^\top \quad \text{Remember: } \boldsymbol{\phi}(x_n) \in \mathbb{R}^M \\ &= \\ &= \\ &= \end{aligned} \quad (16)$$

Vector Notation of Index Expressions

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (15)$$

Sometimes we prefer a compact vector expression again. Index notation is ambiguous about row/col vector. If we remember convention, we can go back:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}} &= 2 \sum_n (\boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta} - y_n) \boldsymbol{\phi}(x_n)^\top && \text{Remember: } \boldsymbol{\phi}(x_n) \in \mathbb{R}^M \\ &= 2 \sum_n [\Phi(X) \boldsymbol{\theta} - \mathbf{y}]_n \boldsymbol{\phi}(x_n)^\top && \text{Define: } \Phi(X) \in \mathbb{R}^{N \times M}, \Phi_{n:}(X) = \boldsymbol{\phi}(x_n)^\top \\ &= \\ &= \end{aligned} \quad (16)$$

Vector Notation of Index Expressions

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (15)$$

Sometimes we prefer a compact vector expression again. Index notation is ambiguous about row/col vector. If we remember convention, we can go back:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}} &= 2 \sum_n (\boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta} - y_n) \boldsymbol{\phi}(x_n)^\top && \text{Remember: } \boldsymbol{\phi}(x_n) \in \mathbb{R}^M \\ &= 2 \sum_n [\Phi(X) \boldsymbol{\theta} - \mathbf{y}]_n \boldsymbol{\phi}(x_n)^\top && \text{Define: } \Phi(X) \in \mathbb{R}^{N \times M}, \Phi_{n:}(X) = \boldsymbol{\phi}(x_n)^\top \\ &= 2(\Phi(X) \boldsymbol{\theta} - \mathbf{y})^\top \Phi(X) \\ &= \end{aligned} \quad (16)$$

Vector Notation of Index Expressions

$$\frac{\partial L}{\partial \theta_i} = -2 \sum_n \left(y_n - \sum_{m=1}^M \phi_m(x_n) \theta_m \right) \phi_i(x_n) \quad (15)$$

Sometimes we prefer a compact vector expression again. Index notation is ambiguous about row/col vector. If we remember convention, we can go back:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}} &= 2 \sum_n (\boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta} - y_n) \boldsymbol{\phi}(x_n)^\top && \text{Remember: } \boldsymbol{\phi}(x_n) \in \mathbb{R}^M \\ &= 2 \sum_n [\Phi(X) \boldsymbol{\theta} - \mathbf{y}]_n \boldsymbol{\phi}(x_n)^\top && \text{Define: } \Phi(X) \in \mathbb{R}^{N \times M}, \Phi_n(X) = \boldsymbol{\phi}(x_n)^\top \\ &= 2(\Phi(X) \boldsymbol{\theta} - \mathbf{y})^\top \Phi(X) \\ &= 2(\boldsymbol{\theta}^\top \Phi(X)^\top \Phi(X) - \mathbf{y}^\top \Phi(X)) \end{aligned} \quad (16)$$

Overview

Index Notation

Differentiation of vector-valued functions

Multivariate Chain Rule

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

- We could manually take partial derivatives of $L(\boldsymbol{\theta})$

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

- ▶ We could manually take partial derivatives of $L(\boldsymbol{\theta})$
- ▶ Or, we could see our scalar function $L(\boldsymbol{\theta})$ as a function composition:

$$f(g(\boldsymbol{\theta})) \qquad f : \mathbb{R}^D \rightarrow \mathbb{R}, \qquad g : \mathbb{R}^E \rightarrow \mathbb{R}^D$$

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

- ▶ We could manually take partial derivatives of $L(\boldsymbol{\theta})$
- ▶ Or, we could see our scalar function $L(\boldsymbol{\theta})$ as a function composition:

$$f(g(\boldsymbol{\theta})) \qquad f : \mathbb{R}^D \rightarrow \mathbb{R}, \qquad g : \mathbb{R}^E \rightarrow \mathbb{R}^D$$

- ▶ Wouldn't it be nice to have a chain rule?

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

- ▶ We could manually take partial derivatives of $L(\boldsymbol{\theta})$
- ▶ Or, we could see our scalar function $L(\boldsymbol{\theta})$ as a function composition:

$$f(g(\boldsymbol{\theta})) \qquad f : \mathbb{R}^D \rightarrow \mathbb{R}, \qquad g : \mathbb{R}^E \rightarrow \mathbb{R}^D$$

- ▶ Wouldn't it be nice to have a chain rule? $\frac{df(g(\boldsymbol{\theta}))}{d\boldsymbol{\theta}} = \frac{df}{dg} \frac{dg}{d\boldsymbol{\theta}}$

Chain Rule

Back to our linear regression problem:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^N (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2 = \|\mathbf{y} - \Phi(X)\boldsymbol{\theta}\|^2 \quad (17)$$

- ▶ We could manually take partial derivatives of $L(\boldsymbol{\theta})$
- ▶ Or, we could see our scalar function $L(\boldsymbol{\theta})$ as a function composition:

$$f(g(\boldsymbol{\theta})) \qquad f : \mathbb{R}^D \rightarrow \mathbb{R}, \qquad g : \mathbb{R}^E \rightarrow \mathbb{R}^D$$

- ▶ Wouldn't it be nice to have a chain rule? $\frac{df(g(\boldsymbol{\theta}))}{d\boldsymbol{\theta}} = \frac{df}{dg} \frac{dg}{d\boldsymbol{\theta}}$
- ▶ But how does a derivative of a vector w.r.t. a vector work?

Differentiation of a vector w.r.t. a scalar

- Differentiation quantifies how an output changes,
in response to an input change.

Differentiation of a vector w.r.t. a scalar

- ▶ Differentiation quantifies how an output changes,
in response to an input change.
- ▶ Change is measured by subtraction. Draw vector subtraction.

Differentiation of a vector w.r.t. a scalar

- ▶ Differentiation quantifies how an output changes,
in response to an input change.
- ▶ Change is measured by subtraction. Draw vector subtraction.
- ▶ Can use the same limiting argument as with scalars:

Differentiation of a vector w.r.t. a scalar

- ▶ Differentiation quantifies how an output changes,
in response to an input change.
- ▶ Change is measured by subtraction. Draw vector subtraction.
- ▶ Can use the same limiting argument as with scalars:

$$\frac{d\mathbf{x}}{dt} = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t} \quad (18)$$

But now, derivative is a vector!

Differentiation of a vector w.r.t. a scalar

- ▶ Differentiation quantifies how an output changes,
in response to an input change.
- ▶ Change is measured by subtraction. Draw vector subtraction.
- ▶ Can use the same limiting argument as with scalars:

$$\frac{d\mathbf{x}}{dt} = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t} \quad (18)$$

But now, derivative is a vector!

- ▶ Subtraction is elementwise, so we can reduce to scalar case:

$$\left[\frac{d\mathbf{x}}{dt} \right]_i = \lim_{\Delta t \rightarrow 0} \frac{x_i(t + \Delta t) - x_i(t)}{\Delta t} \quad (19)$$

Example: Circle

Function that describes a point going round a circle, with period 1s:

$$\mathbf{x} = [\cos 2\pi t \quad \sin 2\pi t]^T \quad (20)$$

Draw on board

Example: Circle

Function that describes a point going round a circle, with period 1s:

$$\mathbf{x} = [\cos 2\pi t \quad \sin 2\pi t]^T \quad (20)$$

Draw on board

- Based on intuition, what should the speed be?

Example: Circle

Function that describes a point going round a circle, with period 1s:

$$\mathbf{x} = [\cos 2\pi t \quad \sin 2\pi t]^T \quad (20)$$

Draw on board

- Based on intuition, what should the speed be? 2π m/s.

Example: Circle

Function that describes a point going round a circle, with period 1s:

$$\mathbf{x} = [\cos 2\pi t \quad \sin 2\pi t]^T \quad (20)$$

Draw on board

- ▶ Based on intuition, what should the speed be? 2π m/s.
- ▶ Find velocity vector $d\mathbf{x}/dt$.

Example: Circle

Function that describes a point going round a circle, with period 1s:

$$\mathbf{x} = [\cos 2\pi t \quad \sin 2\pi t]^T \quad (20)$$

Draw on board

- ▶ Based on intuition, what should the speed be? 2π m/s.
- ▶ Find velocity vector $d\mathbf{x}/dt$.
- ▶ Find speed from the norm.

Summary: Differentiation of a vector w.r.t. a scalar

Derivative of a vector

The derivative of a vector-valued function is given by the derivative of each of its elements.

Overview

Index Notation

Differentiation of vector-valued functions

Multivariate Chain Rule

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{df(a(t), b(t))}{dt} = \frac{\partial f}{\partial a} \frac{da}{dt} + \frac{\partial f}{\partial b} \frac{db}{dt}$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{df(\mathbf{g}(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} \quad \mathbf{g}(t) \in \mathbb{R}^D$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{df(\mathbf{g}(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} \quad \mathbf{g}(t) \in \mathbb{R}^D$$

This is an inner product. Can write in vector form:

$$\frac{df(\mathbf{g}(t))}{dt} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{dt}}_{\text{column}}$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{df(g(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} \quad g(t) \in \mathbb{R}^D$$

This is an inner product. Can write in vector form:

$$\frac{df(g(t))}{dt} = \underbrace{\frac{df}{dg}}_{\text{row}} \cdot \underbrace{\frac{dg}{dt}}_{\text{column}}$$

- ▶ $\frac{dg}{dt}$ is the derivative of a column vector. We keep this to be a column vector.
- ▶ Can also be derived from a limit argument, like the scalar derivative (**board**: Circle Example).

Example: Chain Rule

- ▶ Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^2$

$$f(\mathbf{x}) = f(x_1, x_2) = x_1^2 + 2x_2,$$

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

Example: Chain Rule

- ▶ Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^2$

$$f(\mathbf{x}) = f(x_1, x_2) = x_1^2 + 2x_2,$$

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

- ▶ What are the dimensions of $\frac{df}{d\mathbf{x}}$ and $\frac{d\mathbf{x}}{dt}$?

Work it out with your neighbors

Example: Chain Rule

- ▶ Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^2$

$$f(\mathbf{x}) = f(x_1, x_2) = x_1^2 + 2x_2,$$

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

- ▶ What are the dimensions of $\frac{df}{d\mathbf{x}}$ and $\frac{d\mathbf{x}}{dt}$?

$$1 \times 2 \text{ and } 2 \times 1$$

- ▶ Compute the gradient $\frac{df}{dt}$ using the chain rule:

Example: Chain Rule

- Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\mathbf{x} : \mathbb{R} \rightarrow \mathbb{R}^2$

$$f(\mathbf{x}) = f(x_1, x_2) = x_1^2 + 2x_2,$$

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

- What are the dimensions of $\frac{df}{d\mathbf{x}}$ and $\frac{d\mathbf{x}}{dt}$?

$$1 \times 2 \text{ and } 2 \times 1$$

- Compute the gradient $\frac{df}{dt}$ using the chain rule:

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

Multivariate Chain Rule

We saw the chain rule if we were differentiating w.r.t. a scalar:

$$\frac{df(a(t), b(t))}{dt} = \frac{\partial f}{\partial a} \frac{da}{dt} + \frac{\partial f}{\partial b} \frac{db}{dt}$$

Multivariate Chain Rule

We saw the chain rule if we were differentiating w.r.t. a scalar:

$$\frac{df(\mathbf{g}(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{dt}}_{\text{column}} \quad \mathbf{g}(t) \in \mathbb{R}^D$$

Multivariate Chain Rule

We saw the chain rule if we were differentiating w.r.t. a scalar:

$$\frac{df(\mathbf{g}(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{dt}}_{\text{column}} \quad \mathbf{g}(t) \in \mathbb{R}^D$$

What happens if we differentiate w.r.t. a vector?

Multivariate Chain Rule

We saw the chain rule if we were differentiating w.r.t. a scalar:

$$\frac{df(\mathbf{g}(t))}{dt} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{dg_i}{dt} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{dt}}_{\text{column}} \quad \mathbf{g}(t) \in \mathbb{R}^D$$

What happens if we differentiate w.r.t. a vector?

\implies As before, we just stack the derivatives w.r.t. each of the inputs.

$$\frac{\partial f(\mathbf{g}(\mathbf{x}))}{\partial x_j} = \sum_{i=1}^D \frac{\partial f}{\partial g_i} \frac{\partial g_i}{\partial x_j} \quad \mathbf{g}(\mathbf{x}) \in \mathbb{R}^D$$

Multivariate Chain Rule w.r.t. vector

This is a matrix multiplication! Can write in vector form:

$$\frac{df(\mathbf{g}(\mathbf{x}))}{d\mathbf{x}} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{d\mathbf{x}}}_{\text{matrix}}$$

Multivariate Chain Rule w.r.t. vector

This is a matrix multiplication! Can write in vector form:

$$\frac{df(\mathbf{g}(\mathbf{x}))}{d\mathbf{x}} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{d\mathbf{x}}}_{\text{matrix}}$$

- ▶ $\frac{d\mathbf{g}}{d\mathbf{x}}$ is the derivative of a column vector w.r.t. the input vector \mathbf{x} . We put the elements of \mathbf{g} (i.e. i) along the column, and the dimensions of the derivative (i.e. j) along the row.

Multivariate Chain Rule w.r.t. vector

This is a matrix multiplication! Can write in vector form:

$$\frac{df(\mathbf{g}(\mathbf{x}))}{d\mathbf{x}} = \underbrace{\frac{df}{d\mathbf{g}}}_{\text{row}} \cdot \underbrace{\frac{d\mathbf{g}}{d\mathbf{x}}}_{\text{matrix}}$$

- ▶ $\frac{d\mathbf{g}}{d\mathbf{x}}$ is the derivative of a column vector w.r.t. the input vector \mathbf{x} . We put the elements of \mathbf{g} (i.e. i) along the column, and the dimensions of the derivative (i.e. j) along the row.
- ▶ Can also be derived from a directional derivative argument, but for a vector.

Vector Field Differentiation $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \in \mathbb{R}^M, \quad \mathbf{x} \in \mathbb{R}^N$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \dots, x_N) \\ \vdots \\ f_M(x_1, \dots, x_N) \end{bmatrix}$$

Vector Field Differentiation $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$

$$\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^M, \quad \mathbf{x} \in \mathbb{R}^N$$
$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \dots, x_N) \\ \vdots \\ f_M(x_1, \dots, x_N) \end{bmatrix}$$

- **Jacobian** matrix (collection of all partial derivatives)

$$\begin{bmatrix} \frac{dy_1}{dx} \\ \vdots \\ \frac{dy_M}{dx} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} \in \mathbb{R}^{M \times N}$$

Dimensionality of the Gradient

In general: A function $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ has a gradient that is an $M \times N$ -matrix with

$$\frac{df}{dx} \in \mathbb{R}^{M \times N}, \quad df[m, n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions \times # input dimensions

Dimensionality of the Gradient

In general: A function $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ has a gradient that is an $M \times N$ -matrix with

$$\frac{df}{dx} \in \mathbb{R}^{M \times N}, \quad df[m, n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions \times # input dimensions

A function composition $\mathbf{f}(\mathbf{x}) = (\mathbf{g} \circ \mathbf{h})(\mathbf{x})$ has the constraint that the **output dimension** of $\mathbf{h}(\cdot)$ has to equal the **input dimension** of $\mathbf{g}(\cdot)$, so that we can compute $\mathbf{g}(\mathbf{h}(\mathbf{x}))$.

Dimensionality of the Gradient

In general: A function $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ has a gradient that is an $M \times N$ -matrix with

$$\frac{df}{dx} \in \mathbb{R}^{M \times N}, \quad df[m, n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions \times # input dimensions

A function composition $\mathbf{f}(\mathbf{x}) = (\mathbf{g} \circ \mathbf{h})(\mathbf{x})$ has the constraint that the **output dimension** of $\mathbf{h}(\cdot)$ has to equal the **input dimension** of $\mathbf{g}(\cdot)$, so that we can compute $\mathbf{g}(\mathbf{h}(\mathbf{x}))$.

This ensures that the shapes of the chain rule work out:

$$\mathbf{f} : \mathbb{R}^N \rightarrow \mathbb{R}^M \qquad \mathbf{g} : \mathbb{R}^L \rightarrow \mathbb{R}^M \qquad \mathbf{h} : \mathbb{R}^N \rightarrow \mathbb{R}^L \quad (21)$$

$$\underbrace{\frac{df}{dx}}_{M \times N} = \underbrace{\frac{dg}{dh}}_{M \times L} \underbrace{\frac{dh}{dx}}_{L \times N} \quad (22)$$

Example: Vector Field Differentiation

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N \\ \vdots \\ A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N \end{bmatrix}$$

- Compute the gradient $\frac{df}{dx}$

Example: Vector Field Differentiation

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N \\ \vdots \\ A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N \end{bmatrix}$$

- Compute the gradient $\frac{df}{dx}$
 - Gradient:

$$f_i(x) = \sum_{k=1}^N A_{ik}x_k \quad \implies \quad \frac{\partial f_i}{\partial x_j} = \sum_k A_{ik} \frac{\partial x_k}{\partial x_j} = \sum_k A_{ik} \delta_{kj} = A_{ij}$$

Example: Vector Field Differentiation

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N \\ \vdots \\ A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N \end{bmatrix}$$

- Compute the gradient $\frac{df}{dx}$
 - Gradient:

$$f_i(x) = \sum_{k=1}^N A_{ik}x_k \quad \implies \quad \frac{\partial f_i}{\partial x_j} = \sum_k A_{ik} \frac{\partial x_k}{\partial x_j} = \sum_k A_{ik} \delta_{kj} = A_{ij}$$

Example: Vector Field Differentiation

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1N}x_N \\ \vdots \\ A_{M1}x_1 + A_{M2}x_2 + \cdots + A_{MN}x_N \end{bmatrix}$$

- Compute the gradient $\frac{df}{d\mathbf{x}}$
 - Gradient:

$$\begin{aligned} f_i(\mathbf{x}) &= \sum_{k=1}^N A_{ik}x_k \quad \implies \quad \frac{\partial f_i}{\partial x_j} = \sum_k A_{ik} \frac{\partial x_k}{\partial x_j} = \sum_k A_{ik} \delta_{kj} = A_{ij} \\ \implies \frac{d\mathbf{f}}{d\mathbf{x}} &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \mathbf{A} \in \mathbb{R}^{M \times N} \end{aligned}$$

Example: Multivariate Chain Rule

- ▶ Consider the function

$$L(\mathbf{e}) = \frac{1}{2} \|\mathbf{e}\|^2 = \frac{1}{2} \mathbf{e}^\top \mathbf{e}$$

$$\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{e}, \mathbf{y} \in \mathbb{R}^M$$

- ▶ Compute the gradient $\frac{dL}{dx}$. What is the dimension/size of $\frac{dL}{dx}$?

Work it out with your neighbours

Example: Multivariate Chain Rule

- Consider the function

$$L(\mathbf{e}) = \frac{1}{2} \|\mathbf{e}\|^2 = \frac{1}{2} \mathbf{e}^\top \mathbf{e}$$

$$\mathbf{e} = \mathbf{y} - \mathbf{A}\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{e}, \mathbf{y} \in \mathbb{R}^M$$

- Compute the gradient $\frac{dL}{d\mathbf{x}}$. What is the dimension/size of $\frac{dL}{d\mathbf{x}}$?

$$\frac{dL}{d\mathbf{x}} = \frac{\partial L}{\partial \mathbf{e}} \frac{\partial \mathbf{e}}{\partial \mathbf{x}}$$

$$\frac{\partial L}{\partial \mathbf{e}} = \mathbf{e}^\top \in \mathbb{R}^{1 \times M}, \quad \frac{\partial L}{\partial e_i} = \frac{\partial}{\partial e_i} \sum_j \frac{1}{2} e_j^2 = \sum_j \frac{1}{2} 2e_j \frac{\partial e_j}{\partial e_i} = e_i$$

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}} = -\mathbf{A} \in \mathbb{R}^{M \times N}$$

$$\Rightarrow \frac{dL}{d\mathbf{x}} = \mathbf{e}^\top (-\mathbf{A}) = -(\mathbf{y} - \mathbf{A}\mathbf{x})^\top \mathbf{A} \in \mathbb{R}^{1 \times N}$$

Summary

- ▶ Chain rule for multivariate functions
- ▶ Derivatives of vectors w.r.t. scalars.
- ▶ Derivatives of vectors w.r.t. vectors (and shapes).