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:

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

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:

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

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:

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

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:

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

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

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_i x_i y_i \tag{3}$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_i x_i y_i \tag{3}$$

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

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_i x_i y_i \tag{3}$$

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

$$A\mathbf{x} = ? \tag{5}$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_i x_i y_i \tag{3}$$

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

$$\boldsymbol{A}\mathbf{x} = ? \tag{5}$$

$$[\boldsymbol{A}\mathbf{x}]_j = \sum_j A_{ij}x_j \tag{6}$$

Index Notation of Vector Expressions

$$|\mathbf{x}|^2 = \sum_i x_i^2 \tag{2}$$

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_i x_i y_i \tag{3}$$

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

$$\boldsymbol{A}\mathbf{x} = ? \tag{5}$$

$$[\boldsymbol{A}\mathbf{x}]_j = \sum_j A_{ij} x_j \tag{6}$$

$$\boldsymbol{A}\mathbf{x} = \left[\sum_j A_{ij} x_j\right]_i \tag{7}$$

Index Notation of Vector Expressions

Back to our linear regression problem:

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

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

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:

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

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

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

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

Index Notation of Vector Expressions

Back to our linear regression problem:

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

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

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

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

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

Index Differentiation

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

Index Differentiation

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

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

Index Differentiation

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

- 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

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

- 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}.

Index Differentiation

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

- 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 $w_i = \sum_k \delta_{ik}$.

Index Differentiation

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

- 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 $w_i = \sum_k \delta_{ik}$.
- Think about what happens to $x_k = \sum_a v_a \delta_{ak}$.

Index Differentiation

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

- 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 $w_i = \sum_k \delta_{ik}$.
- Think about what happens to $x_k = \sum_a v_a \delta_{ak}$. Is δ_{ak} a matrix?

Index Differentiation

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

- 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 $w_i = \sum_k \delta_{ik}$.
- Think about what happens to $x_k = \sum_a v_a \delta_{ak}$. Is δ_{ak} a matrix?

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

Index Differentiation

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

- 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 $w_i = \sum_k \delta_{ik}$.
- Think about what happens to $x_k = \sum_a v_a \delta_{ak}$. Is δ_{ak} a matrix?

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

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

Vector Notation of Index Expressions

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

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

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\theta}} &= \\ &= \\ &= \\ &= \end{aligned} \tag{16}$$

Vector Notation of Index Expressions

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

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

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 2\sum_n (\boldsymbol{\phi}(x_n)^\mathsf{T}\boldsymbol{\theta} - y_n)\boldsymbol{\phi}(x_n)^\mathsf{T} \quad \text{Remember: } \boldsymbol{\phi}(x_n) \in \mathbb{R}^M$$

$$=$$

$$=$$

$$= \tag{16}$$

Vector Notation of Index Expressions

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

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

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

Vector Notation of Index Expressions

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

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

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

Vector Notation of Index Expressions

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

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

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

Overview

Index Notation

Differentiation of vector-valued functions

Multivariate Chain Rule

Chain Rule

Back to our linear regression problem:

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

Chain Rule

Back to our linear regression problem:

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

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

Chain Rule

Back to our linear regression problem:

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

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

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

Chain Rule

Back to our linear regression problem:

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

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

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

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

Chain Rule

Back to our linear regression problem:

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

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

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

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

Chain Rule

Back to our linear regression problem:

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

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

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

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

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t} \tag{18}$$

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:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{\mathbf{x}(t+\Delta t) - \mathbf{x}(t)}{\Delta t} \tag{18}$$

  But now, derivative is a vector!
- Subtraction is elementwise, so we can reduce to scalar case:

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

Example: Circle

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

$$\mathbf{x} = \begin{bmatrix} \cos 2\pi t & \sin 2\pi t \end{bmatrix}^\intercal \tag{20}$$

Draw on board

Example: Circle

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

$$\mathbf{x} = \begin{bmatrix} \cos 2\pi t & \sin 2\pi t \end{bmatrix}^\top \tag{20}$$

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:

$$\mathbf{x} = \begin{bmatrix} \cos 2\pi t & \sin 2\pi t \end{bmatrix}^\intercal \tag{20}$$

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:

$$\mathbf{x} = \begin{bmatrix} \cos 2\pi t & \sin 2\pi t \end{bmatrix}^{\mathsf{T}} \tag{20}$$

Draw on board

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

Example: Circle

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

$$\mathbf{x} = \begin{bmatrix} \cos 2\pi t & \sin 2\pi t \end{bmatrix}^\intercal \tag{20}$$

Draw on board

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

$$\frac{\mathrm{d}f(a(t), b(t))}{\mathrm{d}t} = \frac{\partial f}{\partial a}\frac{\mathrm{d}a}{\mathrm{d}t} + \frac{\partial f}{\partial b}\frac{\mathrm{d}b}{\mathrm{d}t}$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i} \frac{\mathrm{d}g_i}{\mathrm{d}t} \qquad\qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i} \frac{\mathrm{d}g_i}{\mathrm{d}t} \qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

Inner product!

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i} \frac{\mathrm{d}g_i}{\mathrm{d}t} \qquad\qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

Inner product! **Given our convention**, we can write in vector form:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}}}_{\text{row}} \cdot \underbrace{\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}}_{\text{column}}$$

Multivariate Chain Rule w.r.t. scalar

It turns out, there is a multivariate chain rule:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i} \frac{\mathrm{d}g_i}{\mathrm{d}t} \qquad\qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

Inner product! **Given our convention**, we can write in vector form:

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}}}_{\text{row}} \cdot \underbrace{\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}}_{\text{column}}$$

- $\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}$ 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 $f : \mathbb{R}^2 \to \mathbb{R}, \quad \boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$

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

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

Example: Chain Rule

- Consider $f : \mathbb{R}^2 \to \mathbb{R}, \quad \boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$

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

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

- What are the dimensions of $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}$ and $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$?

Work it out with your neighbors

Example: Chain Rule

- Consider $f : \mathbb{R}^2 \to \mathbb{R}, \quad \boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$

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

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

- What are the dimensions of $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}$ and $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$?

1×2 and 2×1

- Compute the gradient $\frac{\mathrm{d}f}{\mathrm{d}t}$ using the chain rule:

Example: Chain Rule

- Consider $f : \mathbb{R}^2 \to \mathbb{R}, \quad \boldsymbol{x} : \mathbb{R} \to \mathbb{R}^2$

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

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

- What are the dimensions of $\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}}$ and $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$?

1×2 and 2×1

- Compute the gradient $\frac{\mathrm{d}f}{\mathrm{d}t}$ using the chain rule:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix} = \begin{bmatrix} 2\sin t & 2 \end{bmatrix} \begin{bmatrix} \cos t \\ -\sin t \end{bmatrix}$$

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

Derivative of vector w.r.t. vector

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

$$\frac{\mathrm{d}f(a(t), b(t))}{\mathrm{d}t} = \frac{\partial f}{\partial a}\frac{\mathrm{d}a}{\mathrm{d}t} + \frac{\partial f}{\partial b}\frac{\mathrm{d}b}{\mathrm{d}t}$$

Derivative of vector w.r.t. vector

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

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i}\frac{\mathrm{d}g_i}{\mathrm{d}t} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}}}_{\text{row}} \cdot \underbrace{\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}}_{\text{column}} \qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

Derivative of vector w.r.t. vector

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

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i}\frac{\mathrm{d}g_i}{\mathrm{d}t} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}}}_{\text{row}} \cdot \underbrace{\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}}_{\text{column}} \qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

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

Derivative of vector w.r.t. vector

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

$$\frac{\mathrm{d}f(\boldsymbol{g}(t))}{\mathrm{d}t} = \sum_{i=1}^{D} \frac{\partial f}{\partial g_i} \frac{\mathrm{d}g_i}{\mathrm{d}t} = \underbrace{\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}}}_{\text{row}} \cdot \underbrace{\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t}}_{\text{column}} \qquad \boldsymbol{g}(t) \in \mathbb{R}^D$$

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.

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

Multivariate Chain Rule w.r.t. vector

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

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

Multivariate Chain Rule w.r.t. vector

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

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

- $\frac{\mathrm{d}g}{\mathrm{d}\mathbf{x}}$ is the derivative of a column vector w.r.t. the input vector $\mathbf{x}$. We put the elements of $\boldsymbol{g}$ (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:

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

- $\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}\mathbf{x}}$ is the derivative of a column vector w.r.t. the input vector $\mathbf{x}$. We put the elements of $\boldsymbol{g}$ (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 $f : \mathbb{R}^N \to \mathbb{R}^M$

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{x} \in \mathbb{R}^N$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \ldots, x_N) \\ \vdots \\ f_M(x_1, \ldots, x_N) \end{bmatrix}$$

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

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{x} \in \mathbb{R}^N$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \ldots, x_N) \\ \vdots \\ f_M(x_1, \ldots, x_N) \end{bmatrix}$$

- Jacobian matrix (collection of all partial derivatives)

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

Dimensionality of the Gradient

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

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M \times N}, \qquad \mathrm{d}\boldsymbol{f}[m, n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions $\times$ # input dimensions

Dimensionality of the Gradient

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

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M\times N}, \qquad \mathrm{d}\boldsymbol{f}[m,n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions × # input dimensions

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

Dimensionality of the Gradient

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

$$\frac{\mathrm{d}f}{\mathrm{d}x} \in \mathbb{R}^{M \times N}, \qquad \mathrm{d}f[m,n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions $\times$ # input dimensions

A function composition $\mathbf{f}(\mathbf{x}) = \mathbf{f}_\mathbf{g}(\mathbf{g}(\mathbf{x}))$ (subscript to distinguish from overall function) has the constraint that the **output dimension** of $\mathbf{g}(\cdot)$ has to equal the **input dimension** of $\mathbf{f}_\mathbf{g}(\cdot)$, so that we can compute $\mathbf{f}(\mathbf{g}(\mathbf{x}))$.
This ensures that the shapes of the chain rule work out:

$$\mathbf{f} : \mathbb{R}^N \to \mathbb{R}^M \qquad \mathbf{f}_\mathbf{g} : \mathbb{R}^L \to \mathbb{R}^M \qquad \mathbf{g} : \mathbb{R}^N \to \mathbb{R}^L \tag{21}$$

$$\underbrace{\frac{\mathrm{d}\mathbf{f}}{\mathrm{d}\mathbf{x}}}_{M \times N} = \underbrace{\frac{\mathrm{d}\mathbf{f}_\mathbf{g}}{\mathrm{d}\mathbf{g}}}_{M \times L} \underbrace{\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{x}}}_{L \times N} \tag{22}$$

Example: Vector Field Differentiation

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

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

- Compute the gradient $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}}$

Example: Vector Field Differentiation

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

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

- Compute the gradient $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}}$
  - Gradient:

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

Example: Vector Field Differentiation

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

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

- Compute the gradient $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}}$
  - Gradient:

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

Example: Vector Field Differentiation

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

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

- Compute the gradient $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}}$
  - Gradient:

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

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

Example: Multivariate Chain Rule

- Consider the function

$$L(\boldsymbol{e}) = \tfrac{1}{2}\|\boldsymbol{e}\|^2 = \tfrac{1}{2}\boldsymbol{e}^\top \boldsymbol{e}$$

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

- Compute the gradient $\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}}$. What is the dimension/size of $\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}}$?

Work it out with your neighbours

Example: Multivariate Chain Rule

- Consider the function

$$L(\boldsymbol{e}) = \tfrac{1}{2}\|\boldsymbol{e}\|^2 = \tfrac{1}{2}\boldsymbol{e}^\top \boldsymbol{e}$$

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

- Compute the gradient $\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}}$. What is the dimension/size of $\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}}$?

$$\frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}} = \frac{\partial L}{\partial \boldsymbol{e}}\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{x}}$$

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

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

$$\implies \frac{\mathrm{d}L}{\mathrm{d}\boldsymbol{x}} = \boldsymbol{e}^\top(-\boldsymbol{A}) = -(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})^\top \boldsymbol{A} \in \mathbb{R}^{1\times N}$$

Summary

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