Machine Learning lecture – recapitulation of gradients knowledge

Alexander Binder

ISTD Pillar, Singapore University of Technology and Design

February 3, 2019

Takeaway:

- directional derivatives and gradient
- directional derivatives $\leftrightarrow$ partial derivatives
- side topics in the deep learning lecture:
  - vector-valued function $\rightarrow$ Jacobi-matrix
  - chain rule with directional argument written as matrix multiplications
  - derivative of a linear function and matrix multiplications
  - derivative of a bilinear function

Limits

- a sequence $(s_n)_{n=1}^{\infty}$ converges to a value s if:
  for each value of $\delta > 0$ there exists an index K such that

  $$|s_n - e| < \delta$$

  for all indices $n > K$
- a limit $\lim_{\epsilon \to 0} g(\epsilon)$ exists if
  for each sequence of numbers $(s_n)_{n=1}^{\infty}$ which converges to 0
  (.i.e. $\lim_{n\to\infty} s_n = 0$)
  the limit of the sequence $\lim_{n\to\infty} g(s_n)$
  exists and has the same value for all such sequences $(s_n)_{n=1}^{\infty}$

application: functions with kinks do not have a derivative in the point of kink

Gradient in 1 dimension

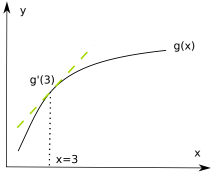
- $f : \mathbb{R}^1 \to \mathbb{R}^1$ is differentiable in input x if the limit exists:

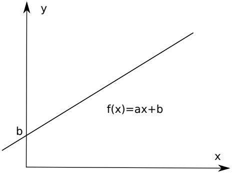
$$\lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \; (=: f'(x))$$

- intuition: slope of the function f at point x
- example: $f(x) = ax + b$ (affine with slope a),then

$$\begin{aligned} \frac{f(x+\epsilon) - f(x)}{\epsilon} &= \frac{a(x+\epsilon) + b - (ax+b)}{\epsilon} \\ &= \frac{ax + a\epsilon + b - ax - b}{\epsilon} \\ &= \frac{a\epsilon}{\epsilon} = a \end{aligned}$$

$$\Rightarrow \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} = a$$





Gradient in 1 dimension

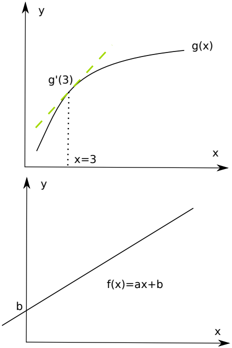
- $f : \mathbb{R}^1 \to \mathbb{R}^1$ is differentiable in input x if the limit exists:

$$\lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \ (=: f'(x))$$

- intuition: slope of the function f at point x
- example: $f(x) = ax + b$ (affine with slope a),then

$$\begin{aligned} \frac{f(x+\epsilon) - f(x)}{\epsilon} &= \frac{a(x+\epsilon) + b - (ax+b)}{\epsilon} \\ &= \frac{ax + a\epsilon + b - ax - b}{\epsilon} \\ &= \frac{a\epsilon}{\epsilon} = a \end{aligned}$$

$$\Rightarrow \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} = a$$



Gradient in 1 dimension

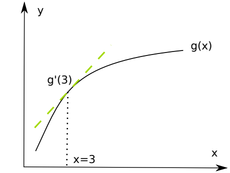
- $f : \mathbb{R}^1 \to \mathbb{R}^1$ is differentiable in input x if the limit exists:

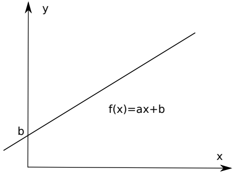
$$\lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \; (=: f'(x))$$

- intuition: slope of the function f at point x
- example: $f(x) = ax + b$ (affine with slope a),then

$$\begin{aligned} \frac{f(x+\epsilon) - f(x)}{\epsilon} &= \frac{a(x+\epsilon) + b - (ax+b)}{\epsilon} \\ &= \frac{ax + a\epsilon + b - ax - b}{\epsilon} \\ &= \frac{a\epsilon}{\epsilon} = a \end{aligned}$$

$$\Rightarrow \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} = a$$





Gradient in 1 dimension

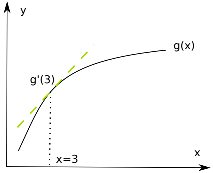
- $f : \mathbb{R}^1 \to \mathbb{R}^1$ is differentiable in input x if the limit exists:

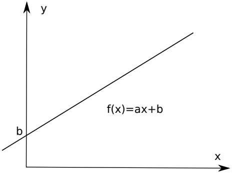
$$\lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \quad (=: f'(x))$$

- intuition: slope of the function f at point x
- example: $f(x) = ax + b$ (affine with slope a),then

$$\begin{aligned} \frac{f(x+\epsilon) - f(x)}{\epsilon} &= \frac{a(x+\epsilon) + b - (ax+b)}{\epsilon} \\ &= \frac{ax + a\epsilon + b - ax - b}{\epsilon} \\ &= \frac{a\epsilon}{\epsilon} = a \end{aligned}$$

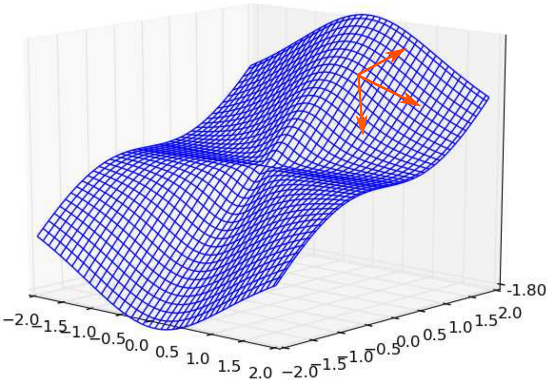
$$\Rightarrow \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} = a$$





2-dimensions: directional derivatives

Function of 2 input variables: $f(x_1, x_2) \in \mathbb{R}^1$



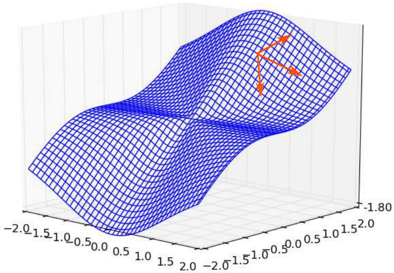
in every point (x_1, x_2): a two dimensional vector space of directions to move from it

in every direction there is a slope – the directional derivative

n-dimensions: directional derivatives

Function of n input variables:

$$f(x_1, x_2, \ldots, x_n) \in \mathbb{R}^1$$



in every point $(x_1, x_2, \ldots, x_n)$: a n-dimensional vector space of directions to move from it

in every direction there is a slope – the directional derivative – provides information about function value change in this direction

n-dimensions: directional derivatives and gradient

The directional derivative of function f in point x in direction v is defined as:

$$\delta_{\mathbf{v}} f(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{v}) - f(\mathbf{x})}{\epsilon}$$

Proposition: If the function is differentiable in x, then this is equivalent to the inner product of the gradient of x in v

$$\delta_{\mathbf{v}} f(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{v} = \sum_{d} \frac{\partial f}{\partial x_d}(\mathbf{x}) v_d$$

Takeaway

- directional derivatives tell you how the function grows from x in direction v when you take an infitinely small step
- the gradient contains infotmation about all directional derivatives, if differentiable in x
- need to define gradient, via partial derivatives

n-dimensions: the partial derivative

- consider $f : \mathbb{R}^n \to \mathbb{R}^1$, $f(\mathbf{x}) = f(x_1, \ldots, x_n) \in \mathbb{R}^1$
- we can define analogously a partial derivative for variable input x_i

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{e}_i) - f(\mathbf{x})}{\epsilon}$$

$$\mathbf{e}_i = (0, \ldots, \quad 0, \quad \underbrace{1}_{i} \;, \quad 0 \;, \ldots, 0)^\top$$

$$\mathbf{x} + \epsilon \mathbf{e}_i = (x_1, \ldots, x_{i-1}, \; \underbrace{x_i + \epsilon}_{\text{i-th comp.}} \;, x_{i+1}, \ldots, x_n)^\top$$

- why we did not simply say f is differentiable if all its partial derivatives exist? (later)

the gradient

- define the gradient $\nabla f(\mathbf{x})$ of function f in $\mathbf{x}$ as the vector of all partial derivatives in input point $\mathbf{x}$:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

- the gradient stores information about all slopes of a function at $\mathbf{x}$ for every direction $\mathbf{v}$ away from that point.

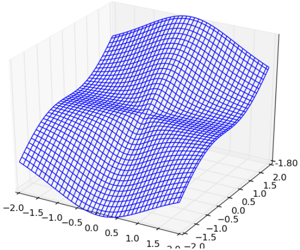
n-dimensions: differentiability

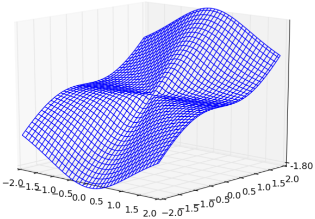
- $f : \mathbb{R}^n \rightarrow \mathbb{R}^1$, $f(\mathbf{x}) = f(x_1, \ldots, x_n) \in \mathbb{R}^1$
  is differentiable in input $\mathbf{x}$ if
  - all directional derivatives (for all vector $\mathbf{v}$) exist
  - the directional derivatives satisfy a linear relationship (with real numbers a_1, a_2)

$$\partial_{a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2} f(\mathbf{x}) = a_1 \partial_{\mathbf{v}_1} f(\mathbf{x}) + a_2 \partial_{\mathbf{v}_2} f(\mathbf{x})$$

n-dimensions: the partial derivative

- why we did not simply say f is differentiable if all its partial derivatives exist?
- $f((x,y)) =$ $$\begin{cases} \frac{y^3}{x^2+y^2} & \text{if } (x,y) \neq (0,0 \\ 0 & \text{if } (x,y) = (0,0 \end{cases}$$
- both partial derivatives exist, but function has kinks in its *directional* derivatives





n-dimensions: vector-valued functions

$$\begin{aligned} f(x) =& (f_1(x_1, \dots, x_n), f_2(x_1, \dots, x_n), \dots, f_s(x_1, \dots, x_n)) \\ f :& x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}^s \end{aligned}$$

Apply above for every component f_i.
Gradient becomes the Jabobi-matrix, which are the gradients of f_i concatenated.

$$\begin{aligned} \nabla f(\mathbf{x}) &= \left(\nabla f_1(x), \nabla f_2(x), \dots, \nabla f_s(x)\right) \\ &= \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x), \frac{\partial f_2}{\partial x_1}(x), \dots, \frac{\partial f_s}{\partial x_1}(x) \\ \frac{\partial f_1}{\partial x_2}(x), \frac{\partial f_2}{\partial x_2}(x), \dots, \frac{\partial f_s}{\partial x_2}(x) \\ \dots \\ \frac{\partial f_1}{\partial x_n}(x), \frac{\partial f_2}{\partial x_n}(x), \dots, \frac{\partial f_s}{\partial x_n}(x) \end{pmatrix} \end{aligned}$$

For every component f_i one looks at its slopes in all possible directions.

Its piece of cake!

n-dimensions: derivatives as linear mappings into the space of all directional derivatives

- above counter-example (with kinks) does not satisfy the linearity!
- linear relationship means: the directional derivatives define a linear mapping $Df(\mathbf{x})[\mathbf{v}]$ in point x:

$$\begin{aligned} \mathbf{v} &\mapsto \partial_{\mathbf{v}} f(\mathbf{x}), \\ Df(\mathbf{x})[\mathbf{v}] &:= \partial_{\mathbf{v}} f(\mathbf{x}) \end{aligned}$$

- $Df(\mathbf{x})[\mathbf{v}]$ has two arguments
  - the point $\mathbf{x}$ in which the derivative is computed
  - the vector $\mathbf{v}$ for the direction, in which one wants to know the slope
  - the mapping $Df(\mathbf{x})[\mathbf{v}]$ is linear only in its second argument $\mathbf{v}$ (the direction)

n-dimensions: derivatives as linear mappings into the space of all directional derivatives

What this is good for ?

- Writing directional derivatives conveniently in matrix form
- partial derivatives $\frac{\partial f}{\partial x_i}$ are directional derivatives along canonical vectors e_i
- writing chain rule terms in matrix form

the derivative of a linear function

- let f be linear in $\mathbf{x}$:

$$f(\mathbf{x}) = \mathbf{u}^\top \mathbf{x} = \sum_i u_i x_i$$

- then $\frac{\partial f}{\partial x_i}(\mathbf{x}) = u_i$ and in general

$$Df(\mathbf{x})[\mathbf{v}] = f(\mathbf{v})$$

- the derivative of a linear function is a linear function
- practical for matrix algebra!

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{x}^\top \mathbf{z} \\ \Rightarrow Df(\mathbf{x})[\mathbf{v}] &= \\ f(\mathbf{x}) &= \mathbf{x}^\top \mathbf{A} \\ \Rightarrow Df(\mathbf{x})[\mathbf{v}] &= \end{aligned}$$

can get all directional derivatives as matrix-multiplications (GPU-implementations!)

example $\mathbf{A} \in \mathbb{R}^{k \times m}, \mathbf{X} \in \mathbb{R}^{m \times l}, \mathbf{C} \in \mathbb{R}^{l \times r}$

$$f(\mathbf{X}) = \mathbf{AXC}$$

$$\Rightarrow Df(\mathbf{X})[\mathbf{V}] =$$

example $\mathbf{A} \in \mathbb{R}^{k\times m}, \mathbf{X} \in \mathbb{R}^{m\times l}, \mathbf{C} \in \mathbb{R}^{l\times r}$

$$
\begin{aligned}
f(\mathbf{X}) &= \mathbf{AXC} \\
&\qquad \text{linear in } \mathbf{X} \text{ !!!} \\
\Rightarrow Df(\mathbf{X})[\mathbf{V}] &=
\end{aligned}
$$

example $\mathbf{A} \in \mathbb{R}^{k \times m}, \mathbf{X} \in \mathbb{R}^{m \times l}, \mathbf{C} \in \mathbb{R}^{l \times r}$

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{AXC} \\ &\quad \text{linear in } \mathbf{X} \text{ !!!} \\ \Rightarrow Df(\mathbf{X})[\mathbf{V}] &= \mathbf{AVC} \end{aligned}$$

sooo easy!!!

Chain rule

- 1-dim case:

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

- $L_{g'(h(x))}[v] = g'(h(x)) * v$ is a linear mapping in v
- $f'(x) = L_{g'(h(x))}[h'(x)]$
- N-dim case: the multiplication $g'(h(x)) * h'(x)$ will be replaced by a linear mapping in higher dimensions
- N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x}))$$
$$Df(\mathbf{x})[\mathbf{v}] = Dg(h(\mathbf{x}))[Dh(\mathbf{x})[\mathbf{v}]]$$

Chain rule

N-dim case: concatenation of linear mappings

$$\begin{aligned} f(\mathbf{x}) &= g(h(\mathbf{x})) \\ Df(\mathbf{x})[\mathbf{v}] &= Dg(h(\mathbf{x}))[Dh(\mathbf{x})[\mathbf{v}]] \\ &\quad Dg \text{ is derivated at point } h(\mathbf{x}) \text{ in direction } \mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] \end{aligned}$$

Why the linear view helps here?

Usually linear operations are (when looking only at one argument!):

- inner products
- matrix-vector
- matrix-matrix multiplications

Can express chain rule in terms of matrix multiplications for gradient vectors and Jacobi-matrices

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \dots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$$\begin{aligned} Df(\mathbf{x})[\mathbf{v}] &= Dg(h(\mathbf{x}))[\mathbf{c}], \mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] \\ &= (\nabla g_{(h(\mathbf{x}))})^\top Dh_{(x)}^\top \mathbf{v} \text{ as matrix multiplications} \\ &= \mathbf{v}^\top Dh_{(x)} \nabla g_{(h(\mathbf{x}))} \text{ as matrix multiplications} \end{aligned}$$

$Dh_{(x)} = \left(\nabla h_1, \nabla h_2, \dots, \nabla h_n\right) \in \mathbb{R}^{k \times n}$ is Jacobi matrix of h in $\mathbf{x}$,

$\top$ is the transpose

$D(g \circ h)(\mathbf{x})[\mathbf{v}] = \mathbf{v}^\top \times$ (Jacobi of h in $\mathbf{x}$) $\times$ (Gradient of g in $h(\mathbf{x})$)
or its transpose, bcs its a real number

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$\begin{aligned}
f(\mathbf{x}) &= g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \ldots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k \\
Df(\mathbf{x})[\mathbf{v}] &= Dg(h(\mathbf{x}))[\mathbf{c}], \mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] \\
&= \nabla g(h(\mathbf{x})) \cdot \mathbf{c} \text{ as inner product} \\
&= (\nabla g_{(h(\mathbf{x}))})^\top \mathbf{c} \text{ as matrix-vector product} \\
&= \mathbf{c}^\top \nabla g_{(h(\mathbf{x}))} \text{ as matrix-vector product}
\end{aligned}$$

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$\begin{aligned}
f(\mathbf{x}) &= g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \dots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k \\
Df(\mathbf{x})[\mathbf{v}] &= Dg(h(\mathbf{x}))[\mathbf{c}], \mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] \\
&= \nabla g(h(\mathbf{x})) \cdot \mathbf{c} \text{ as inner product} \\
&= (\nabla g_{(h(\mathbf{x}))})^\top \mathbf{c} \text{ as matrix-vector product} \\
&= \mathbf{c}^\top \nabla g_{(h(\mathbf{x}))} \text{ as matrix-vector product}
\end{aligned}$$

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \dots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] =$??? as component-wise inner products for each h_i

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \nabla h_1(\mathbf{x}) \cdot \mathbf{v} \\ \nabla h_2(\mathbf{x}) \cdot \mathbf{v} \\ \dots \\ \nabla h_n(\mathbf{x}) \cdot \mathbf{v} \end{pmatrix} \text{ as comp-wise inner product}$$

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \mathbf{v}^\top \nabla h_{1,(\mathbf{x})} \\ \vdots \\ \mathbf{v}^\top \nabla h_{n,(\mathbf{x})} \end{pmatrix} \text{ as matrix-vector product}$$

$\mathbf{c}^\top = \left(\mathbf{v}^\top \nabla h_{1,(\mathbf{x})}, \dots, \mathbf{v}^\top \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top \left(\nabla h_{1,(\mathbf{x})}, \dots, \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top($ Jacobi-matrix of h in $x)$ as matrix-vector product

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \ldots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] =???$ as component-wise inner products for each h_i

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \nabla h_1(\mathbf{x}) \cdot \mathbf{v} \\ \nabla h_2(\mathbf{x}) \cdot \mathbf{v} \\ \ldots \\ \nabla h_n(\mathbf{x}) \cdot \mathbf{v} \end{pmatrix} \text{ as comp-wise inner product}$$

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \mathbf{v}^\top \nabla h_{1,(\mathbf{x})} \\ \vdots \\ \mathbf{v}^\top \nabla h_{n,(\mathbf{x})} \end{pmatrix} \text{ as matrix-vector product}$$

$\mathbf{c}^\top = \left(\mathbf{v}^\top \nabla h_{1,(\mathbf{x})}, \ldots, \mathbf{v}^\top \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top \left(\nabla h_{1,(\mathbf{x})}, \ldots, \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top($ Jacobi-matrix of h in $x)$ as matrix-vector product

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \ldots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] =$??? as component-wise inner products for each h_i

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \nabla h_1(\mathbf{x}) \cdot \mathbf{v} \\ \nabla h_2(\mathbf{x}) \cdot \mathbf{v} \\ \ldots \\ \nabla h_n(\mathbf{x}) \cdot \mathbf{v} \end{pmatrix} \text{ as comp-wise inner product}$$

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \mathbf{v}^\top \nabla h_{1,(\mathbf{x})} \\ \vdots \\ \mathbf{v}^\top \nabla h_{n,(\mathbf{x})} \end{pmatrix} \text{ as matrix-vector product}$$

$\mathbf{c}^\top = \left(\mathbf{v}^\top \nabla h_{1,(\mathbf{x})}, \ldots, \mathbf{v}^\top \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top \left(\nabla h_{1,(\mathbf{x})}, \ldots, \nabla h_{n,(\mathbf{x})}\right)$ as matrix-vector product

$\mathbf{c}^\top = \mathbf{v}^\top($ Jacobi-matrix of h in $x)$ as matrix-vector product

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \dots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] =???$ as component-wise inner products for each h_i

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \nabla h_1(\mathbf{x}) \cdot \mathbf{v} \\ \nabla h_2(\mathbf{x}) \cdot \mathbf{v} \\ \dots \\ \nabla h_n(\mathbf{x}) \cdot \mathbf{v} \end{pmatrix} \text{ as comp-wise inner product}$$

$$\mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] = \begin{pmatrix} \mathbf{v}^\top \nabla h_{1,(\mathbf{x})} \\ \vdots \\ \mathbf{v}^\top \nabla h_{n,(\mathbf{x})} \end{pmatrix} \text{ as matrix-vector product}$$

$$\mathbf{c}^\top = \left(\mathbf{v}^\top \nabla h_{1,(\mathbf{x})}, \dots, \mathbf{v}^\top \nabla h_{n,(\mathbf{x})}\right) \text{ as matrix-vector product}$$

$$\mathbf{c}^\top = \mathbf{v}^\top \left(\nabla h_{1,(\mathbf{x})}, \dots, \nabla h_{n,(\mathbf{x})}\right) \text{ as matrix-vector product}$$

$$\mathbf{c}^\top = \mathbf{v}^\top (\text{ Jacobi-matrix of } h \text{ in } x) \text{ as matrix-vector product}$$

Chain rule with gradients

N-dim case: concatenation of linear mappings

$$f(\mathbf{x}) = g(h(\mathbf{x})) = g(h_1(\mathbf{x}), \dots, h_n(\mathbf{x})), \mathbf{x} \in \mathbb{R}^k$$

$$\begin{aligned}
Df(\mathbf{x})[\mathbf{v}] &= Dg(h(\mathbf{x}))[\mathbf{c}], \mathbf{c} = Dh(\mathbf{x})[\mathbf{v}] \\
&= (\nabla g_{(h(\mathbf{x}))})^\top Dh_{(x)}^\top \mathbf{v} \text{ as matrix multiplications} \\
&= \mathbf{v}^\top Dh_{(x)} \nabla g_{(h(\mathbf{x}))} \text{ as matrix multiplications} \\
Dh_{(x)} &= \left(\nabla h_1, \nabla h_2, \dots, \nabla h_n\right) \in \mathbb{R}^{k \times n} \text{ is Jacobi matrix of } h \text{ in } \mathbf{x}, \\
&\top \text{ is the transpose}
\end{aligned}$$

$$D(g \circ h)(\mathbf{x})[\mathbf{v}] = \mathbf{v}^\top \times (\text{Jacobi of } h \text{ in } \mathbf{x}) \times (\text{Gradient of } g \text{ in } h(\mathbf{x}))$$

Bilinear functions

- consider $B(\cdot,\cdot)$ to be any bilinear function of both variables. Let $\mathbf{a}$ and $\mathbf{c}$ be two vector-valued functions which take a vector $\mathbf{x}$ as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

$$\begin{aligned} f(\mathbf{x}) &= B(\mathbf{a}(\mathbf{x}), \mathbf{c}(\mathbf{x})) \\ Df(\mathbf{x})[\mathbf{v}] &= B(D\mathbf{a}(\mathbf{x})[\mathbf{v}], \mathbf{c}(\mathbf{x})) + B(\mathbf{a}(\mathbf{x}), D\mathbf{c}(\mathbf{x})[\mathbf{v}]) \end{aligned}$$

is sum of two terms ... each time plug in derivative in one component

Product rule – more general

All kind of matrix-matrix $A * B$, matrix-vector products $A\mathbf{v}$ and the like are bilinear as a function of both arguments. So apply

$$\begin{aligned} f(\mathbf{x}) &= B(\mathbf{a}(\mathbf{x}), \mathbf{c}(\mathbf{x})) \\ Df(\mathbf{x})[\mathbf{v}] &= B(D\mathbf{a}(\mathbf{x})[\mathbf{v}], \mathbf{c}(\mathbf{x})) + B(\mathbf{a}(\mathbf{x}), D\mathbf{c}(\mathbf{x})[\mathbf{v}]) \end{aligned}$$

Product rule

- 1-dim case application of chain rule to the product – a bilinear mapping

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

- N-dim case: consider matrix multiplication of two functions **a** and **c** which take a vector **x** as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

$$f(\mathbf{x}) = \mathbf{a}(\mathbf{x})^\top \mathbf{c}(\mathbf{x})$$
$$g(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} \text{ bilinear in } (\mathbf{u}, \mathbf{v})$$
$$Df(\mathbf{x})[\mathbf{v}] = (D\mathbf{a}(\mathbf{x})[\mathbf{v}])^\top \mathbf{c}(\mathbf{x}) + \mathbf{a}(\mathbf{x})^\top (D\mathbf{c}(\mathbf{x})[\mathbf{v}])$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions **a** and **c** which take a vector **x** as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

$$\begin{aligned}\text{example: } f(\mathbf{x}) &= \mathbf{x}^\top \mathbf{A}\mathbf{x} \\ &= Matmul(\mathbf{x}^\top, \mathbf{A}\mathbf{x}), g(\mathbf{x}) = \mathbf{A}\mathbf{x} \\ Df(\mathbf{x})[\mathbf{v}] &= Matmul(DId^\top(\mathbf{x})[\mathbf{v}], \mathbf{A}\mathbf{x}) + Matmul(\mathbf{x}^\top, Dg(\mathbf{x})[\mathbf{v}])\end{aligned}$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions $\mathbf{a}$ and $\mathbf{c}$ which take a vector $\mathbf{x}$ as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

  example: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$

  by above rule

  $$Df(\mathbf{x})[\mathbf{v}] = D(\mathbf{x}^\top)[\mathbf{v}](\mathbf{A}\mathbf{x}) + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{x}^\top$ is linear, so $D(\mathbf{x}^\top)[\mathbf{v}] = \mathbf{v}^\top$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is linear, so $D(\mathbf{A}\mathbf{x})[\mathbf{v}] = \mathbf{A}\mathbf{v}$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathbf{v}$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions $\mathbf{a}$ and $\mathbf{c}$ which take a vector $\mathbf{x}$ as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

  example: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$

  by above rule

  $$Df(\mathbf{x})[\mathbf{v}] = D(\mathbf{x}^\top)[\mathbf{v}](\mathbf{A}\mathbf{x}) + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{x}^\top$ is linear, so $D(\mathbf{x}^\top)[\mathbf{v}] = \mathbf{v}^\top$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is linear, so $D(\mathbf{A}\mathbf{x})[\mathbf{v}] = \mathbf{A}\mathbf{v}$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathbf{v}$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions $\mathbf{a}$ and $\mathbf{c}$ which take a vector $\mathbf{x}$ as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

  example: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$

  by above rule

  $$Df(\mathbf{x})[\mathbf{v}] = D(\mathbf{x}^\top)[\mathbf{v}](\mathbf{A}\mathbf{x}) + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{x}^\top$ is linear, so $D(\mathbf{x}^\top)[\mathbf{v}] = \mathbf{v}^\top$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is linear, so $D(\mathbf{A}\mathbf{x})[\mathbf{v}] = \mathbf{A}\mathbf{v}$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathbf{v}$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions **a** and **c** which take a vector **x** as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

  example: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$

  by above rule

  $$Df(\mathbf{x})[\mathbf{v}] = D(\mathbf{x}^\top)[\mathbf{v}](\mathbf{A}\mathbf{x}) + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{x}^\top$ is linear, so $D(\mathbf{x}^\top)[\mathbf{v}] = \mathbf{v}^\top$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}]$$

  the mapping $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is linear, so $D(\mathbf{A}\mathbf{x})[\mathbf{v}] = \mathbf{A}\mathbf{v}$

  $$\Rightarrow \quad = \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathbf{v}$$

Product rule – Application

- N-dim case: consider matrix multiplication of two functions $\mathbf{a}$ and $\mathbf{c}$ which take a vector $\mathbf{x}$ as input and compute a vector as output, i.e. $\mathbf{a}(\mathbf{x})$ and $\mathbf{c}(\mathbf{x})$ are vectors

example: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$

$$
\begin{aligned}
& && \text{by above rule} \\
Df(\mathbf{x})[\mathbf{v}] &= D(\mathbf{x}^\top)[\mathbf{v}](\mathbf{A}\mathbf{x}) + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}] && \\
& && \text{the mapping } \mathbf{x} \mapsto \mathbf{x}^\top \text{ is linear, so } D(\mathbf{x}^\top)[\mathbf{v}] = \mathbf{v}^\top \\
\Rightarrow \quad &= \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top D(\mathbf{A}\mathbf{x})[\mathbf{v}] && \\
& && \text{the mapping } \mathbf{x} \mapsto \mathbf{A}\mathbf{x} \text{ is linear, so } D(\mathbf{A}\mathbf{x})[\mathbf{v}] = \mathbf{A}\mathbf{v} \\
\Rightarrow \quad &= \mathbf{v}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}\mathbf{v} &&
\end{aligned}
$$

Product rule – more general

- N-dim case: consider matrix multiplication of two functions $\mathbf{A}$ and $\mathbf{C}$ which take a vector $\mathbf{x}$ as input and compute a matrix as output, i.e. $\mathbf{A}(\mathbf{x})$ and $\mathbf{C}(\mathbf{x})$ are matrices

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{A}(\mathbf{x}) * \mathbf{C}(\mathbf{x}) \\ Df(\mathbf{x})[\mathbf{v}] &= D\mathbf{A}(\mathbf{x})[\mathbf{v}] * \mathbf{C}(\mathbf{x}) + \mathbf{A}(\mathbf{x}) * D\mathbf{C}(\mathbf{x})[\mathbf{v}] \end{aligned}$$

Recap: Linear mapping to partial derivatives

- if $f = f(u)$, and u is a vector, then $Df(u)[1_k] = \frac{\partial f}{\partial u_k}$
- if $f = f(u)$, and u is a $k \times m$ matrix, then $Df(u)[1_{i,j}] = \frac{\partial f}{\partial u_{i,j}}$
- the general rule: plugging in into the linear part $[\cdot]$ an object which is 1 in one entry and zero otherwise, returns the partial derivative for the entry in which one has a 1
- practical example: $f(X) = u^T X v = \sum_{r,s} u_r X_{r,s} v_s$.
  $Df(X)[H] = u^T H v$,
  $\frac{\partial f}{\partial X_{i,j}} = Df(X)[1_{i,j}] = \sum_{r,s} u_r \text{``}(H = 1_{i,j})''_{r,s} v_s = u_i v_j$

I hope this messing around with gradients helps you!!

gradients are not that dangerous

Takeaway:

- directional derivatives and gradient
- directional derivatives ↔ partial derivatives
- side topics in the deep learning lecture:
  - vector-valued function → Jacobi-matrix
  - chain rule with directional argument written as matrix multiplications
  - derivative of a linear function and matrix multiplications
  - derivative of a bilinear function