Practice Problems for Midterm 1

1. Calculators are allowed
2. Computers are not allowed
3. Show your work

Python Basics

The midterm will be on paper, no computers will be allowed. Make sure you know what the python code output should be.

Python questions will be restriced to content covered in Python_1.ipynb and Python_2.ipynb

Q1. What will the following code print?

In [ ]:
```
hello = "'Hello'"
name = '"ECE"'
pi = 3.1419
print(f'{hello:s} {name}. pi is {pi:.03f}')  # string formatting
```

Q2. What will the following code print?

In [ ]:
```
xs = [1, 2, 3, 'hello', [4, 5, 6]]    # Create a list
print(xs[-1])
```

Q3. What will the following code print?

In [ ]:
```
nums = list(range(5))    # range is a built-in function that creates a list
print(nums[-2:])
```

Q4. Which code is faster? Option 1 or Option 2?

In [ ]:
```
# Code Option 1:
d = {'cat': 'cute', 'dog': 'furry'}  # Create a new dictionary with some dat
print(d['dog'])
# Code option 2:
keys = ['cat', 'dog'] # Create the dictionary with keys as lists
values = ['cute', 'furry'] # # Create the dictionary with values as lists
print(values[keys.index('dog')])
```

Q5. Which code is faster? Option 1 or Option 2?

In [ ]:
```
# Code Option 1:
d = {0: 'cute', 1: 'furry'}  # Create a new dictionary with some data
print(d[1])
# Code option 2:
```

```
values = ['cute', 'furry'] # # Create the dictionary with values as lists
print(values[1])
```

Q6. What is the output of the following code?

In [ ]:
```
class Value:
    def __init__(self, v):
        self.v = v

    def __add__(self, other):
        return self.v * other

print(Value(3) + 2)
```

Numpy basics

Python questions will be restriced to content covered in NumpyTutorial.ipynb

Q7: What is the output of the following code?

In [ ]:
```
import numpy as np
x = np.array([[1, 2], [3, 4]])
y = np.array([[5, 6]])
np.concatenate((x.T, y.T), axis=-1)
```

Q8. What is the output of the following code?

In [ ]:
```
x = np.array([[1, 2], [3, 4]])
y = np.array([[5, 6]])
x @ y.T
```

Q9. What is the output of the following code?

In [ ]:
```
x = np.array([[1, 2], [3, 4]])
y = np.array([[5, 6]])
(x * y).sum(axis=-1)
```

Linear algebra and it's geometry

Q10.

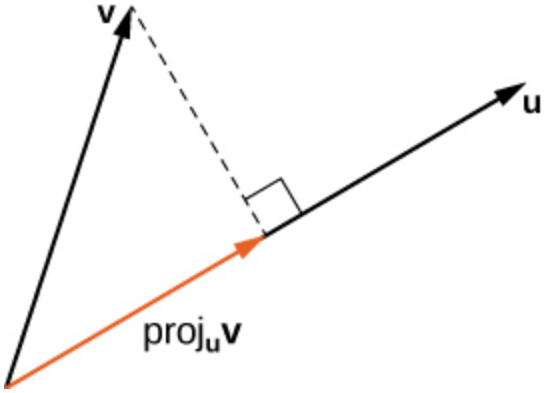
Show that for any vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]$, it's magnitude squared is same as dot product with itself i.e. $||\mathbf{a}||^2 = \mathbf{a}^\top \mathbf{a}$

A10. The mangitude of n-D vector is given by $||\mathbf{a}|| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$ and dot product the vector with itself is given by

$\mathbf{a}^\top \mathbf{a} = a_1 a_1 + a_2 a_2 + \cdots + a_n a_n = a_1^2 + a_2^2 + \cdots + a_n^2$. Squaring the magnitude gives us $\|\mathbf{a}\| = a_1^2 + a_2^2 + \cdots + a_n^2$, which is same as $\mathbf{a}^\top \mathbf{a}$.

Q11. For given vectors $\mathbf{v}$ and $\mathbf{u}$ find the projection of $\mathbf{v}$ on $\mathbf{u}$ $\text{proj}_{\mathbf{u}}\, \mathbf{v}$. Also find the equation of dotted line which is perpendicular to $\mathbf{u}$ and passes through $\mathbf{v}$. Convert the equation of line to the form $y = mx + c$.

$\mathbf{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$.



A11.

1. $\text{proj}_{\mathbf{u}}\, \mathbf{v} = \mathbf{v}^\top \frac{\mathbf{u}}{\|\mathbf{u}\|} = \frac{12}{\sqrt{13}}$
2. The dotted line is the set of all points $\mathbf{x} \in \mathbb{R}^2$ that satisfy $\mathbf{u}^\top \mathbf{x} = \mathbf{u}^\top \mathbf{v}$
3. Let $\mathbf{x} = [x, y]$. Then the above equation of line can be written as $[3, 2] \begin{bmatrix} x \\ y \end{bmatrix} = 12$ or $3x + 2y = 12$

Q12.

Convert the following scalar equation into vector form. Your end result should contain $\mathbf{m} = [m; c]$, $\mathbf{y} = [y_1; y_2; \ldots; y_n]$ and $\mathbf{x} = [x_1; x_2, \ldots, x_n]$. You can define other vectors and matrices as needed, included a vector of ones like 1_n.

$$e(m, c, (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) = (y_1 - (x_1 m + c))^2 + (y_2 - (x_2 m + c))^2 + (y_n - (x_n m + c))^2$$

A12. Recall that the magnitude of a vector $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^n}$ has a similar form to the error function. This suggests that we can define an error vector with the signed error for each data point as it's elements

$$\mathbf{e} = \begin{bmatrix} y_1 - (mx_1 + c) \\ y_2 - (mx_2 + c) \\ \vdots \\ y_n - (mx_n + c) \end{bmatrix}$$

The total error is same as minimizing the square of error vector magnitude which is further same as vector product with itself.

$$e(m, c, (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) = \|\mathbf{e}\|^2 = \mathbf{e}^\top \mathbf{e}$$

Let us define $\mathbf{x} = [x_1; \ldots; x_n]$ to denote the vector of all x coordinates of the dataset and $\mathbf{y} = [y_1; \ldots; y_n]$ to denote y coordinates. Then the error vector is:

$$\mathbf{e} = \mathbf{y} - (\mathbf{x}m + \mathbf{1}_n c)$$

where $\mathbf{1}_n$ is a n-D vector of all ones. Finally, we vectorize parameters of the line $\mathbf{m} = [m; c]$. We will also need to horizontally concatenate $\mathbf{x}$ and $\mathbf{1}_n$. Let's call the result $\mathbf{X} = [\mathbf{x}, \mathbf{1}_n] \in \mathbb{R}^{n \times 2}$. Now, the error vector looks like this:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{m}$$

Expanding the error magnitude:

$$\begin{aligned} \|\mathbf{e}\|^2 &= (\mathbf{y} - \mathbf{X}\mathbf{m})^\top (\mathbf{y} - \mathbf{X}\mathbf{m}) \\ &= \mathbf{y}^\top \mathbf{y} + \mathbf{m}^\top \mathbf{X}^\top \mathbf{X}\mathbf{m} - 2\mathbf{y}^\top \mathbf{X}\mathbf{m} \end{aligned}$$

Q13:

Convert the following scalar equation into vector form. Your end result should contain $\mathbf{m} = [a; b; c]$, $\mathbf{z} = [z_1; z_2; \ldots; z_n]$, $\mathbf{y} = [y_1; y_2; \ldots; y_n]$ and $\mathbf{x} = [x_1; x_2, \ldots, x_n]$. You can define other vectors and matrices as needed, included a vector of all ones like 1_n.

$$\begin{aligned} e(a, b, c, (x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)) &= (z_1 - (x_1 a + y_1 b + c))^2 \\ &+ (z_2 - (x_2 a + y_2 b + c))^2 + \cdots + (z_n - (x_n a + y_n b + c))^2 \end{aligned}$$

A13: A variation of A12

Q14

Convert the following vector equation into even more vectorized form.

$$\begin{aligned} e(m_0, \mathbf{m}, (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)) &= (y_1 - (\mathbf{x}_1^\top \mathbf{m} + m_0))^2 \\ &+ (y_2 - (\mathbf{x}_2^\top \mathbf{m} + m_0))^2 + \cdots + (y_n - (\mathbf{x}_n^\top \mathbf{m} + m_0))^2 \end{aligned}$$

where $\mathbf{m} = [m_1; m_2; \ldots; m_p] \in \mathbb{R}^p$ is a p-dimensional vector and
$\mathbf{x}_i = [x_{i1}; x_{i2}; \ldots; x_{ip}] \in \mathbb{R}^p$ are p-dimensional vectors for all $i = \{1, 2, \ldots n\}$

Your end result should contain $\mathbf{q} = [m_0, m_1, m_2, \ldots, m_p] \in \mathbb{R}^{p+1}$,
$\mathbf{y} = [y_1; y_2; \ldots; y_n] \in \mathbb{R}^n$ and

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}$$

.

You can define other vectors and matrices as needed, included a vector of all ones like 1_n.

A15. Recall that the magnitude of a vector $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^n}$ has a similar form to the error function. This suggests that we can define an error vector with the signed error for each data point as it's elements

$$\mathbf{e} = \begin{bmatrix} y_1 - (\mathbf{x}_1^\top \mathbf{m} + m_0) \\ y_2 - (\mathbf{x}_2^\top \mathbf{m} + m_0) \\ \vdots \\ y_n - (\mathbf{x}_2^\top \mathbf{m}_2 + m_0) \end{bmatrix}$$

The total error is same as minimizing the square of error vector magnitude which is further same as vector product with itself.

$$e(m_0, \mathbf{m}, (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)) = \|\mathbf{e}\|^2 = \mathbf{e}^\top \mathbf{e}$$

Let us define $\mathbf{X} = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_n^\top]$ to denote the vector of all x coordinates of the dataset and $\mathbf{y} = [y_1; \ldots; y_n]$ to denote y coordinates. Then the error vector is:

$$\mathbf{e} = \mathbf{y} - (\mathbf{1}_n m_0 + \mathbf{X}\mathbf{m})$$

where $\mathbf{1}_n$ is a n-D vector of all ones. Finally, we call parameters of the line $\mathbf{q} = [m_0; \mathbf{m}]$. We will also need to horizontally concatenate $\mathbf{X}$ and $\mathbf{1}_n$. Let's call the result $\bar{\mathbf{X}} = [\mathbf{1}_n, \mathbf{X}] \in \mathbb{R}^{n \times (p+1)}$. Now, the error vector looks like this:

$$\mathbf{e} = \mathbf{y} - \bar{\mathbf{X}}\mathbf{q}$$

Expanding the error magnitude:

$$\|\mathbf{e}\|^2 = (\mathbf{y} - \bar{\mathbf{X}}\mathbf{q})^\top (\mathbf{y} - \bar{\mathbf{X}}\mathbf{q})$$
$$= \mathbf{y}^\top \mathbf{y} + \mathbf{m}^\top \bar{\mathbf{X}}^\top \bar{\mathbf{X}}\mathbf{q} - 2\mathbf{y}^\top \bar{\mathbf{X}}\mathbf{q}$$

Q16:

Convert the following scalar equation into vector form. Your end result should contain $\mathbf{m} = [m; c]$, the matrix $\mathbf{W} = \text{Diag}([w_1; w_2; \ldots; w_n])$, $\mathbf{y} = [y_1; y_2; \ldots; y_n]$ and $\mathbf{x} = [x_1; x_2, \ldots, x_n]$. You can define other vectors and matrices as needed, included a vector of all ones like 1_n.

$$e(m, c, (x_1, y_1, w_1), (x_2, y_2, w_2), \ldots, (x_n, y_n, w_n)) = w_1^2(y_1 - (x_1 m + c))^2 + w_2^2(y_2 - (x_2 m + c))^2 + \cdots + w_n^2(y_n - (x_n m + c))^2$$

The matrix $\mathbf{W}$ is defined as $\text{Diag}([w_1; w_2; \ldots; w_n])$ which indicates that $\mathbf{W}$ is diagonal matrix of $[w_1; w_2; \ldots; w_n]$.

$$\mathbf{W} = \text{Diag}([w_1; w_2; \ldots; w_n]) = \begin{bmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{bmatrix}$$

A16:

Recall that the magnitude of a vector $\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$ has a similar form to the error function. This suggests that we can define an error vector with the signed error for each data point as it's elements

$$\mathbf{e} = \begin{bmatrix} y_1 - (mx_1 + c) \\ y_2 - (mx_2 + c) \\ \vdots \\ y_n - (mx_n + c) \end{bmatrix}$$

and let $\mathbf{W} = \text{Diag}([w_1; w_2; \ldots; w_n])$.

Note that

$$\mathbf{We} = \begin{bmatrix} w_1(y_1 - (mx_1 + c)) \\ w_2(y_2 - (mx_2 + c)) \\ \vdots \\ w_3(y_n - (mx_n + c)) \end{bmatrix}$$

The total error is same as the square of error vector magnitude

$$e(m, c, (x_1, y_1, w_1), (x_2, y_2, w_2), \ldots, (x_n, y_n, w_n)) = w_1^2(y_1 - (x_1 m + c))^2$$
$$+ w_2^2(y_2 - (x_2 m + c))^2 + \cdots + w_n^2(y_n - (x_n m + c))^2 = \|\mathbf{We}\|^2$$

The square of error vector magnitude is same as dot product with itself,

$$\|\mathbf{We}\|^2 = (\mathbf{We})^\top(\mathbf{We}) = \mathbf{e}^\top\mathbf{W}^\top\mathbf{We}$$

Let us define $\mathbf{x} = [x_1; \ldots; x_n]$ to denote the vector of all x coordinates of the dataset and $\mathbf{y} = [y_1; \ldots; y_n]$ to denote y coordinates. Then the error vector is:

$$\mathbf{e} = \mathbf{y} - (\mathbf{x}m + \mathbf{1}_n c)$$

where $\mathbf{1}_n$ is a n-D vector of all ones. Finally, we vectorize parameters of the line $\mathbf{m} = [m; c]$. We will also need to horizontally concatenate $\mathbf{x}$ and $\mathbf{1}_n$. Let's call the result $\mathbf{X} = [\mathbf{x}, \mathbf{1}_n] \in \mathbb{R}^{n \times 2}$. Now, the error vector looks like this:

$$\mathbf{e} = \mathbf{y} - \mathbf{Xm}$$

Expanding the error magnitude:

$$\|\mathbf{We}\|^2 = (\mathbf{y} - \mathbf{Xm})^\top\mathbf{W}^\top\mathbf{W}(\mathbf{y} - \mathbf{Xm})$$
$$= \mathbf{y}^\top\mathbf{W}^\top\mathbf{Wy} + \mathbf{m}^\top\mathbf{X}^\top\mathbf{W}^\top\mathbf{WXm} - 2\mathbf{y}^\top\mathbf{W}^\top\mathbf{WXm}$$

Q17:

Using vector derivatives find the minimum of the following vector quadratic function in $\mathbf{m}$:

$$\arg\min_{\mathbf{m}} e(\mathbf{m}) = \mathbf{y}^\top\mathbf{W}^\top\mathbf{Wy} + \mathbf{m}^\top\mathbf{X}^\top\mathbf{W}^\top\mathbf{WXm} - 2\mathbf{y}^\top\mathbf{W}^\top\mathbf{WXm}$$

The dimensions of the each of the variables are given $\mathbf{m} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{n \times n}$, $\mathbf{X} \in \mathbb{R}^{n \times p}$.

A17:

$$\mathbf{0}^\top = \frac{\partial}{\partial \mathbf{m}}(\mathbf{y}^\top\mathbf{W}^\top\mathbf{Wy} + \mathbf{m}^\top\mathbf{X}^\top\mathbf{W}^\top\mathbf{WXm} - 2\mathbf{y}^\top\mathbf{W}^\top\mathbf{WXm}) \quad (1)$$
$$= 2\mathbf{m}^{*\top}\mathbf{X}^\top\mathbf{W}^\top\mathbf{WX} - 2\mathbf{y}^\top\mathbf{W}^\top\mathbf{WX} \quad (2)$$

This gives us the solution

$$\mathbf{m}^* = (\mathbf{X}^\top\mathbf{W}^\top\mathbf{WX})^{-1}\mathbf{X}^\top\mathbf{W}^\top\mathbf{Wy}$$

Vector derivatives

Q18:

Find the derivative of $f(\mathbf{x}) = (\mathbf{x} - \mathbf{a}_1)^\top A(\mathbf{x} - \mathbf{a}_2)$ with respecto to $\mathbf{x}$.

You can assume $A \in \mathbb{R}^{n \times n}$ to be symmetric. The size of vectors are $\mathbf{x}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{b} \in \mathbb{R}^n$

A18

$$\begin{aligned} f(\mathbf{x}) &= (\mathbf{x} - \mathbf{a}_1)^\top A(\mathbf{x} - \mathbf{a}_2) \\ &= \mathbf{x}^\top A\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2)^\top A\mathbf{x} + \mathbf{a}_1^\top \mathbf{a}_2 \\ \frac{\partial f}{\partial \mathbf{x}} &= 2\mathbf{x}^\top A - (\mathbf{a}_1 + \mathbf{a}_2)^\top A \\ &= (2\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2))^\top A \end{aligned}$$

Q19:

Find the quadratic approximation of the following function near the point $\mathbf{x}_0$:

$$f(\mathbf{x}) = \left((\mathbf{x} - \mathbf{a}_1)^\top A(\mathbf{x} - \mathbf{a}_2)\right)\left((\mathbf{x} - \mathbf{a}_3)^\top \mathbf{b}\right)$$

You can assume $A \in \mathbb{R}^{n \times n}$ to be symmetric. The size of vectors are $\mathbf{x}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{b} \in \mathbb{R}^n$

A19:

$$[\nabla_{\mathbf{x}} f(\mathbf{x})]^\top = \left((\mathbf{x} - \mathbf{a}_1)^\top A(\mathbf{x} - \mathbf{a}_2)\right)\mathbf{b}^\top + \left((\mathbf{x} - \mathbf{a}_3)^\top \mathbf{b}\right)\left((2\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2))^\top A\right)$$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left((\mathbf{x} - \mathbf{a}_1)^\top A(\mathbf{x} - \mathbf{a}_2)\right)\mathbf{b} + \left((\mathbf{x} - \mathbf{a}_3)^\top \mathbf{b}\right)\left(A(2\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2))\right)$$

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0) = \left((\mathbf{x}_0 - \mathbf{a}_1)^\top A(\mathbf{x}_0 - \mathbf{a}_2)\right)\mathbf{b} + \left((\mathbf{x}_0 - \mathbf{a}_3)^\top \mathbf{b}\right)\left(A(2\mathbf{x}_0 - (\mathbf{a}_1 + \mathbf{a}_2))\right)$$

$$\begin{aligned} Hf(\mathbf{x}) = \nabla^2_{\mathbf{x}} f(\mathbf{x}) = \mathbf{b}(2\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2))^\top A + \left(A(2\mathbf{x} - (\mathbf{a}_1 + \mathbf{a}_2))\right)\mathbf{b}^\top \\ + \left((\mathbf{x} - \mathbf{a}_3)^\top \mathbf{b}\right)(2A) \end{aligned}$$

$$\begin{aligned} Hf(\mathbf{x}_0) = \nabla^2_{\mathbf{x}} f(\mathbf{x}_0) = \mathbf{b}(2\mathbf{x}_0 - (\mathbf{a}_1 + \mathbf{a}_2))^\top A + \left(A(2\mathbf{x}_0 - (\mathbf{a}_1 + \mathbf{a}_2))\right)\mathbf{b}^\top \\ + \left((\mathbf{x}_0 - \mathbf{a}_3)^\top \mathbf{b}\right)(2A) \end{aligned}$$

The quadratic approximation by Taylor series is:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + [\nabla_{\mathbf{x}} f(\mathbf{x}_0)]^\top (\mathbf{x} - \mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^\top Hf(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

Q20

Show that for $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{c}^\top \mathbf{x} = \mathbf{c}^\top \tag{3}$$

A20: Let $\mathbf{c} = [c_1, c_2, \ldots, c_n]$ and $\mathbf{x} = [x_1, x_2, \ldots x_n]$

Let $f(\mathbf{x}) = \mathbf{c}^\top \mathbf{x} = c_1 x_1 + c_2 x_2 + \ldots c_n x_n$

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= c_1 \\ \frac{\partial f}{\partial x_2} &= c_2 \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= c_n \end{aligned}$$

By Jacobian convention, we arrange the partial derivatives in a row vector:

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{c}^\top \mathbf{x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \ldots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{4}$$

$$= \begin{bmatrix} c_1 & c_2 & \ldots & c_n \end{bmatrix} = \mathbf{c}^\top \tag{5}$$

Q21:

Show that for $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{x} \in \mathbb{R}^n$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A} \tag{6}$$

A21: Let $\mathbf{x} = [x_1; x_2; \ldots x_n]$

Let $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix}$, where $\mathbf{a}_i^\top \in \mathbb{R}^{1 \times n}$ are the row vectors of matrix $\mathbf{A}$.

Then

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_n^\top \mathbf{x} \end{bmatrix}$$

Let

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{bmatrix} = \mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_n^\top \mathbf{x} \end{bmatrix}$$

By Jacobian convention we arrange the partial derivatives of each function component column-wise

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial \mathbf{x}} \\ \frac{\partial f_2(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{a}_1^\top \mathbf{x}}{\partial \mathbf{x}} \\ \frac{\partial \mathbf{a}_2^\top \mathbf{x}}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial \mathbf{a}_2^\top \mathbf{x}}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} = \mathbf{A}$$

Q22:

Use vector-derivative chain rule:

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

,

for any function $\mathbf{g} : \mathbb{R}^n \mapsto \mathbb{R}^m$ and $\mathbf{f} : \mathbb{R}^m \mapsto \mathbb{R}^o$.

Show that for $\mathbf{x} \in \mathbb{R}^n$ amd $\mathbf{A} \in \mathbb{R}^{n \times n}$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{x}^\top (\mathbf{A}^\top + \mathbf{A}) \tag{7}$$

A22:

For product of any two vectors

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \tag{8}$$

If $\mathbf{y}$ is a function of $\mathbf{x}$, then

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top + \left(\frac{\partial}{\partial \mathbf{y}} \mathbf{x}^\top \mathbf{y} \right) \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right) \tag{9}$$

$$= \mathbf{y}^\top + \mathbf{x}^\top \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right) \tag{10}$$

If $\mathbf{y} = \mathbf{A}\mathbf{x}$, then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \mathbf{A}\mathbf{x} = \mathbf{A}$$

and

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{y}^\top + \mathbf{x}^\top \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right) = \mathbf{x}^\top \mathbf{A}^\top + \mathbf{x}^\top \mathbf{A} = \mathbf{x}^\top (\mathbf{A}^\top + \mathbf{A})$$

Perceptron

Q23:

You are given 2D points and corresponding labels as a training dataset $\{(x_1, y_1, l_1), (x_2, y_2, l_2), \ldots, (x_n, y_n, l_n)\}$, where $x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$ and the labels $l_i \in \{-1, 1\}$. Use the model $\hat{l}_i = \text{sign}(y_i - (mx_i + c))$ to construct a loss (or error) function. Find the gradient of the loss function with respect to the vector $\mathbf{m} = [m; c]$.

A23

$$e(y_i, x_i; m, c) = \begin{cases} 0 & \text{if } \text{sign}(y_i - mx_i + c) = l_i \\ |y_i - (mx_i + c)| & \text{if } \text{sign}(y_i - mx_i + c) \neq l_i \end{cases}$$

$$e(y_i, x_i; m, c) = \begin{cases} 0 & \text{if } \text{sign}(y_i - mx_i + c) = l_i \\ |y_i - (mx_i + c)| & \text{if } \text{sign}(y_i - mx_i + c) \neq l_i \end{cases}$$

$$\mathbf{m} = \begin{bmatrix} m \\ c \end{bmatrix}$$

$$e(y_i, x_i; \mathbf{m}) = \begin{cases} 0 & \text{if } \begin{bmatrix} x_i & 1 \end{bmatrix} \mathbf{m} = l_i \\ |y_i - \begin{bmatrix} x_i & 1 \end{bmatrix} \mathbf{m}| & \text{if } \begin{bmatrix} x_i & 1 \end{bmatrix} \mathbf{m} \neq l_i \end{cases}$$

If $l_i \in \{-1, 1\}$, then we can write

$$e(y_i, x_i; \mathbf{m}) = \max\{0, -l_i(y_i - \begin{bmatrix} x_i & 1 \end{bmatrix} \mathbf{m})\}$$

$$\nabla_{\mathbf{m}} e(y_i, x_i; \mathbf{m}) = \max\{0, l_i(\begin{bmatrix} x_i & 1 \end{bmatrix})\}$$

For the entire dataset, we have $\mathbf{y} = [y_1; \ldots; y_n]$ and $\mathbf{x} = [x_1; \ldots; x_n]$, $\mathbf{l} = [l_1; \ldots; l_n]$ the average error is:

$$e(\mathbf{x}, \mathbf{y}; \mathbf{m}) = \frac{1}{n} 1_n^\top \max\{0, -\mathbf{l} \odot (\mathbf{y} - \begin{bmatrix} \mathbf{x} & 1_n \end{bmatrix} \mathbf{m})\}$$

and the average gradient is:

$$\nabla_{\mathbf{m}}^{\top} e(\mathbf{x}, \mathbf{y}; \mathbf{m}) = \frac{1}{n} 1_n^{\top} \max\{0, \mathbf{l} \odot ([\, \mathbf{x} \quad 1_n \,])\}$$

Q24

You are given p-D points $\mathbf{x}_i \in \mathbb{R}^p$ and corresponding labels as a training dataset $\{(\mathbf{x}_1, l_1), (\mathbf{x}_2, l_2), \ldots, (\mathbf{x}_n, l_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^p$, and the labels $l_i \in \{-1, 1\}$. Use the model $\hat{l}_i = \text{sign}(\mathbf{x}_i^{\top} \mathbf{m} + m_0))$ to construct a loss (or error) function. Find the gradient of the loss function with respect to the vector $\mathbf{q} = [m_0; \mathbf{m}]$.

A24:

$$e(m_0, \mathbf{m}; \mathbf{x}_i) = \begin{cases} 0 & \text{if } \text{sign}(\mathbf{x}_i^{\top} \mathbf{m} + m_0) = l_i \\ |\mathbf{x}_i^{\top} \mathbf{m} + m_0| & \text{if } \text{sign}(\mathbf{x}_i^{\top} \mathbf{m} + m_0) \neq l_i \end{cases}$$

$$e(y_i, x_i; m, c) = \begin{cases} 0 & \text{if } \text{sign}(\mathbf{x}_i^{\top} \mathbf{m} + m_0) = l_i \\ |\mathbf{x}_i^{\top} \mathbf{m} + m_0| & \text{if } \text{sign}(\mathbf{x}_i^{\top} \mathbf{m} + m_0) \neq l_i \end{cases}$$

$$\mathbf{q} = \begin{bmatrix} m_0 \\ \mathbf{m} \end{bmatrix}$$

$$e(m_0, \mathbf{m}; \mathbf{x}_i) = \begin{cases} 0 & \text{if } \begin{bmatrix} 1 & \mathbf{x}_i^{\top} \end{bmatrix} \mathbf{q} = l_i \\ |\begin{bmatrix} 1 & \mathbf{x}_i^{\top} \end{bmatrix} \mathbf{q}| & \text{if } \begin{bmatrix} 1 & \mathbf{x}_i^{\top} \end{bmatrix} \mathbf{q} \neq l_i \end{cases}$$

If $l_i \in \{-1, 1\}$, then we can write

$$e(m_0, \mathbf{m}; \mathbf{x}_i) = \max\{0, -l_i(\begin{bmatrix} 1 & \mathbf{x}_i^{\top} \end{bmatrix} \mathbf{q})\}$$

$$\nabla_{\mathbf{m}} e(m_0, \mathbf{m}; \mathbf{x}_i) = \max\{0, -l_i(\begin{bmatrix} 1 & \mathbf{x}_i^{\top} \end{bmatrix})\}$$

For the entire dataset, we have $\mathbf{X} = [\mathbf{x}_1^{\top}; \ldots; \mathbf{x}_n^{\top}]$, $\mathbf{l} = [l_1; \ldots; l_n]$ the average error is:

$$e(\mathbf{m}; \mathbf{X}, \mathbf{l}) = \frac{1}{n} 1_n^{\top} \max\{0, -\mathbf{l} \odot (\begin{bmatrix} 1_n & \mathbf{X} \end{bmatrix} \mathbf{q})\}$$

and the average gradient is:

$$\nabla_{\mathbf{m}}^{\top} e(\mathbf{m}; \mathbf{X}, \mathbf{l}) = \frac{1}{n} 1_n^{\top} \max\{0, \mathbf{l} \odot ([\, 1_n \quad \mathbf{X} \,])\}$$

Autograd

Q25:

Describe the Forward mode and reverse mode differentitation and their differences?

Consider the following functions which one of the two will you use for:

1. $\mathbf{f}(\mathbf{x}) : \mathbb{R}^2 \mapsto \mathbb{R}^{100}$
2. $\mathbf{f}(\mathbf{x}) : \mathbb{R}^{100} \mapsto \mathbb{R}^2$

A25:

1. Forward mode and reverse mode differentiation differ by the order in which the chain rule jacobians get multiplied. For example, if you are required to take the derivative of the the function by chain rule $\mathbf{f}(\mathbf{g}(\mathbf{h}(\mathbf{x})))$, where $\mathbf{h} : \mathbb{R}^n \mapsto \mathbb{R}^m$, $\mathbf{g} : \mathbb{R}^m \mapsto \mathbb{R}^o$, and $\mathbf{h} : \mathbb{R}^o \mapsto \mathbb{R}^p$ then by chain rule:

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{x}}$$

There are two options for multipliying the jacobians

a. Forward mode

$$\left(\frac{\partial \mathbf{f}}{\partial \mathbf{g}} \left(\frac{\partial \mathbf{g}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right) \right)$$

b. Reverse mode

$$\left(\left(\frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{h}} \right) \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right)$$

Q26:

How many operations (additions and multiplications) does it take to multiple two matrices of size $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$?

A26: $mp(2n - 1)$.

There exist matrix algorithms that are faster than $O(n^3)$. They speed up matrix multiplication to $O(n^{2.7})$.

Q27:

Write the reverse mode vector-Jacobian product(s) for the following operations:

1. $f(x) = exp(x)$ where $x \in \mathbb{R}$
2. $\mathbf{f}(\alpha, \mathbf{v}) = \alpha \mathbf{v}$ where $\alpha \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$

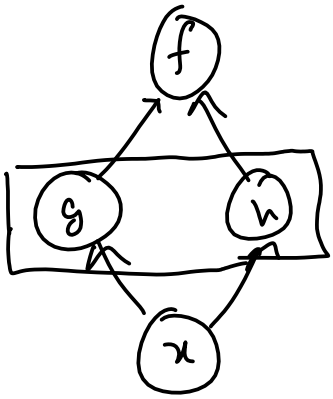
Auto differentiation / Autograd

① Forward mode $\frac{\partial}{\partial x} f(g(h(x))) = \left(\frac{\partial f}{\partial g} \cdot \left(\frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial x}\right)\right)$

② Reverse mode.

$$= \left(\left(\frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial h}\right) \cdot \frac{\partial h}{\partial x}\right)$$

Chain rule for vector derivatives



$\frac{\partial}{\partial x} f(g(x), h(x)) = ?$

$\frac{\partial}{\partial x} f\left(\begin{bmatrix} g(x) \\ h(x) \end{bmatrix}\right) = ?$

$\frac{\partial}{\partial x} f\left(\underline{g}(x)\right) = ?$

$\varepsilon \to 0 \quad \varepsilon \approx 10^{-6} \quad \varepsilon^2 \approx 10^{-12}$

$f(g(x+\varepsilon), h(x+\varepsilon)) = ?$

$g(x+\varepsilon) = g(x) + \varepsilon \frac{\partial g}{\partial x}$

$h(x+\varepsilon) = h(x) + \varepsilon \frac{\partial h}{\partial x}$

$\Leftarrow \quad \frac{\partial g}{\partial x} = \lim_{\varepsilon \to 0} \frac{g(x+\varepsilon) - g(x)}{\varepsilon}$

$f(g(x+\varepsilon), h(x+\varepsilon)) = f\left(g(x) + \varepsilon \frac{\partial g}{\partial x}, h(x) + \varepsilon \frac{\partial h}{\partial x}\right)$

$$f\left(g(x)+\varepsilon\frac{\partial g}{\partial x},\ h(x)+\varepsilon\frac{\partial h}{\partial x}\right) \quad ①$$

$$= f\left(g(x), h(x)+\varepsilon\frac{\partial h}{\partial x}\right) + \varepsilon\left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial x}\right)$$

$$f(g+\varepsilon_g) = f(g) + \varepsilon_g \frac{\partial f}{\partial g}$$

$$= f(g(x), h(x)) \quad ② + \varepsilon \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} \quad 3a + \varepsilon \frac{\partial f}{\partial h}\frac{\partial h}{\partial x} \quad 3b + \varepsilon^2 \to 0$$

$$\frac{\partial}{\partial x} f(g(x), h(x)) = \frac{\partial f}{\partial g}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial h}\frac{\partial h}{\partial x} \qquad \lim_{\varepsilon\to 0} \frac{① - ②}{\varepsilon}$$

$$= \begin{bmatrix} \frac{\partial f}{\partial g} & \frac{\partial f}{\partial h} \end{bmatrix} \begin{Bmatrix} \frac{\partial g}{\partial x} \\ \frac{\partial h}{\partial x} \end{Bmatrix}$$

$$\frac{\partial}{\partial x} f\left(\underbrace{\begin{bmatrix} g(x) \\ h(x) \end{bmatrix}}_{\underline{g}(x)}\right) = \frac{\partial f}{\partial \underline{g}}\frac{\partial \underline{g}}{\partial x}$$

$$\frac{\partial}{\partial \underline{x}} f(\underline{g}(\underline{x})) = \frac{\partial f}{\partial \underline{g}}\underbrace{\frac{\partial \underline{g}}{\partial \underline{x}}}_{\text{Jacobian matrix}}$$

$$\frac{\partial}{\partial \underline{x}} \left\{ \underline{f}\left(\underline{g}\left(\underline{h}(\underline{x})\right)\right) \right\}$$

$$= \frac{\partial \underline{f}}{\partial \underline{g}} \frac{\partial \underline{g}}{\partial \underline{h}} \frac{\partial \underline{h}}{\partial \underline{x}}$$

$\frac{\partial \underline{f}}{\partial \underline{g}} \in \mathbb{R}^{p \times o}$, $\frac{\partial \underline{g}}{\partial \underline{h}} \in \mathbb{R}^{o \times m}$, $\frac{\partial \underline{h}}{\partial \underline{x}} \in \mathbb{R}^{m \times n}$

$\frac{\partial \underline{f}}{\partial \underline{g}} \frac{\partial \underline{g}}{\partial \underline{h}} \in \mathbb{R}^{p \times m}$

$\frac{\partial \underline{f}}{\partial \underline{x}} \in \mathbb{R}^{p \times n}$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_n} \end{bmatrix}$$

$\underline{x} \in \mathbb{R}^n$

$\underline{h}(\underline{x}) : \mathbb{R}^n \longmapsto \mathbb{R}^m$

$\frac{\partial \underline{h}(\underline{x})}{\partial \underline{x}} = ?$

$\frac{\partial \underline{h}}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m}{\partial x_1} & \cdots & \frac{\partial h_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$

$\underline{g}(\underline{h}) : \mathbb{R}^m \longmapsto \mathbb{R}^o$

$\underline{f}(\underline{g}) : \mathbb{R}^o \longmapsto \mathbb{R}^p$

Computation cost of multiplying matrices

$A \in \mathbb{R}^{m \times n}$

$B \in \mathbb{R}^{n \times p}$

$C = AB \quad C \in \mathbb{R}^{m \times p}$

$$= \begin{bmatrix} & & \\ & c_{ij} & \\ & & \end{bmatrix} = \begin{bmatrix} \underline{a}_1^T \\ \underline{a}_2^T \\ \vdots \\ \underline{a}_m^T \end{bmatrix} \begin{bmatrix} \underline{b}_1, \underline{b}_2, \ldots, \underline{b}_p \end{bmatrix}$$

$c_{ij} = \underline{a}_i^T \underline{b}_j = a_{i1} b_{j1} + a_{i2} b_{j2} + \cdots + a_{in} b_{jn}$

multiplications $= nmp$

additions $= (n-1)mp$

cost $= (2n-1)mp \simeq O(nmp)$

when $n = m = p$, $= O(n^3)$

there are algo $\to O(n^{2.7})$

$$\frac{\partial}{\partial x}\left\{f\left(g\left(h(x)\right)\right)\right\}$$

$$= \left(\left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial h}\right)\right)\frac{\partial h}{\partial x}$$

$\frac{\partial f}{\partial g} \in \mathbb{R}^{p \times o}$, $\frac{\partial g}{\partial h} \in \mathbb{R}^{o \times m}$, $\left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial h}\right) \in \mathbb{R}^{p \times m}$, $\frac{\partial h}{\partial x} \in \mathbb{R}^{m \times n}$, $\left(\left(\frac{\partial f}{\partial g}\right)\left(\frac{\partial g}{\partial h}\right)\right)\frac{\partial h}{\partial x} \in \mathbb{R}^{p \times n}$

$f(g(h(k(x))))$

$q \quad p \quad o \quad m \quad n$

$q \times o$

$O(qpo) + O(qom) + O(qmn)$

① Reverse mode : what is the cost? $= O(pom) + O(pmn)$ $(n > p)$

② Forward mode $(n < p)$

$$\frac{\partial}{\partial x}\left\{f\left(g\left(h(x)\right)\right)\right\}$$

$$= \left(\frac{\partial f}{\partial g}\right)\left(\left(\frac{\partial g}{\partial h}\right)\frac{\partial h}{\partial x}\right)$$

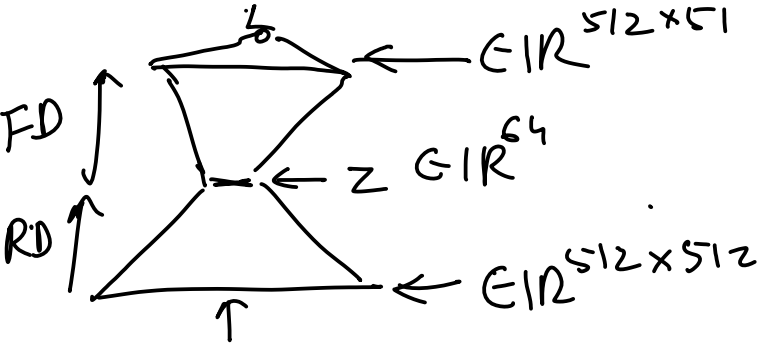
$\frac{\partial f}{\partial g} \in \mathbb{R}^{p \times o}$, $\left(\frac{\partial g}{\partial h}\right)\frac{\partial h}{\partial x} \in \mathbb{R}^{o \times n}$, $\frac{\partial h}{\partial x} \in \mathbb{R}^{m \times n}$

$= O(omn) + O(pon)$

In most optimization problems } Reverse mode

$\min L(x) \in \mathbb{R}$

$x \in \mathbb{R}^{512 \times 512}$ $\quad n = 512 \times 512$



$p = 1$

Autodiff Libraries

① Reverse mode

$$f(g(h(k(z))))$$

$\in \mathbb{R}^{1\times p}$ $\in \mathbb{R}^{1\times o}$ $\in \mathbb{R}^{1\times n}$ $f: \mathbb{R}^m \longmapsto \mathbb{R}$

vector – Jacobian product

$$\frac{\partial f}{\partial g}\,\frac{\partial g}{\partial h}\,\frac{\partial h}{\partial k}\,\frac{\partial k}{\partial z}$$

② Forward mode

$$f(g(h(k(t))))$$

$t \in \mathbb{R}$

$$\frac{\partial f}{\partial g}\,\frac{\partial g}{\partial h}\,\frac{\partial h}{\partial k}\,\frac{\partial k}{\partial t}$$

$\frac{\partial k}{\partial t} \in \mathbb{R}^{m\times 1}$

$\frac{\partial h}{\partial k}\frac{\partial k}{\partial t} \in \mathbb{R}^{n\times 1}$

$\frac{\partial g}{\partial h}\left(\frac{\partial h}{\partial k}\frac{\partial k}{\partial t}\right) \in \mathbb{R}^{o\times 1}$

$\frac{\partial f}{\partial g}\left(\frac{\partial g}{\partial h}\left(\frac{\partial h}{\partial k}\frac{\partial k}{\partial t}\right)\right) \in \mathbb{R}^{p\times 1}$

Jacobian – vector product (JVP)

---

① Product rule → Differentiating product

Chain rule → i=D, RD

(a) $f(\alpha,\beta) = \alpha\beta$

↳ FD → JVP →

$$\frac{df}{d\alpha}(\alpha,\beta) = \beta \qquad \frac{\partial f}{\partial \beta} = \alpha$$

Jacobian

$$\frac{df}{dt}(\alpha,\beta) = \beta\frac{d\alpha}{dt} + \alpha\frac{d\beta}{dt}$$

↖ vector $\frac{\partial}{\partial t}$

↳ RD → VJP →

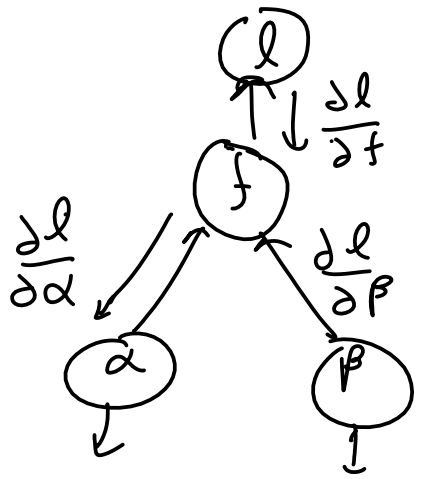
$l(f(\alpha, \beta)) \in \mathbb{R}$

$\frac{\partial l}{\partial \alpha} = ?, \quad \frac{\partial l}{\partial \beta} = ?$

vector ↙ Jacobian ↙

$\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial f} \frac{\partial f}{\partial \alpha} = \frac{\partial l}{\partial f} B$

$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial f} \frac{\partial f}{\partial \beta} = \frac{\partial l}{\partial f} \alpha$



$\underline{f}(\alpha, \underline{v}) = \alpha \underline{v}$

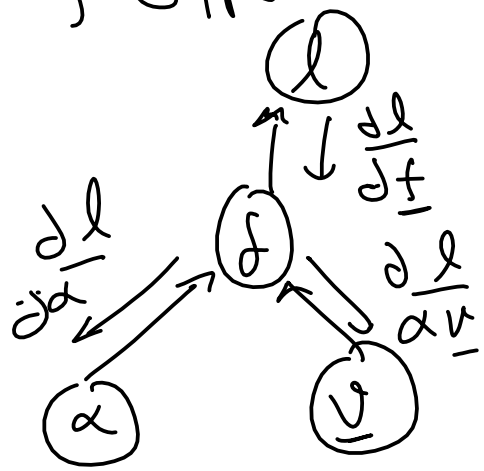
$\alpha \in \mathbb{R}$

$\underline{v} \in \mathbb{R}^n$

$\underline{f} \in \mathbb{R}^n$

FD: JVP

$\frac{\partial \underline{f}}{\partial t} = \frac{\partial \alpha}{\partial t} + ? \frac{\partial \underline{v}}{\partial t}$



RD: VJP

$l(\underline{f}(\alpha, \underline{v})) \in \mathbb{R}$

$\frac{\partial l}{\partial \alpha} = ? \left(\frac{\partial l}{\partial \underline{f}}\right)$

$\frac{\partial l}{\partial \underline{v}} = ?$ $\frac{\partial l}{\partial \underline{f}}$

$$\underline{f}(\alpha, \underline{v}) = \alpha \underline{v}$$

$\alpha \in \mathbb{R}$
$\underline{v} \in \mathbb{R}^n$
$\underline{f} \in \mathbb{R}^n$

FD on JVP

$$\boxed{\frac{\partial \underline{f}}{\partial t} = \left(\frac{\partial \alpha}{\partial t}\right)\underline{v} + \alpha \frac{\partial \underline{v}}{\partial t}} \quad \text{— ①}$$

vector

$$\boxed{\frac{\partial}{\partial \underline{v}}(\alpha \underline{v}) = \alpha \underline{I}_{n\times n}}$$

Jacobian

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1} & & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial(\alpha v_1)}{\partial v_1} & \cdots & \frac{\partial \alpha v_1}{\partial v_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \alpha v_n}{\partial v_1} & \cdots & \frac{\partial(\alpha v_n)}{\partial v_n} \end{bmatrix} = \begin{bmatrix} \alpha & & 0 \\ & \ddots & \\ 0 & & \alpha \end{bmatrix}_{n\times n}$$

$$= \alpha \underline{I}_{n\times n}$$

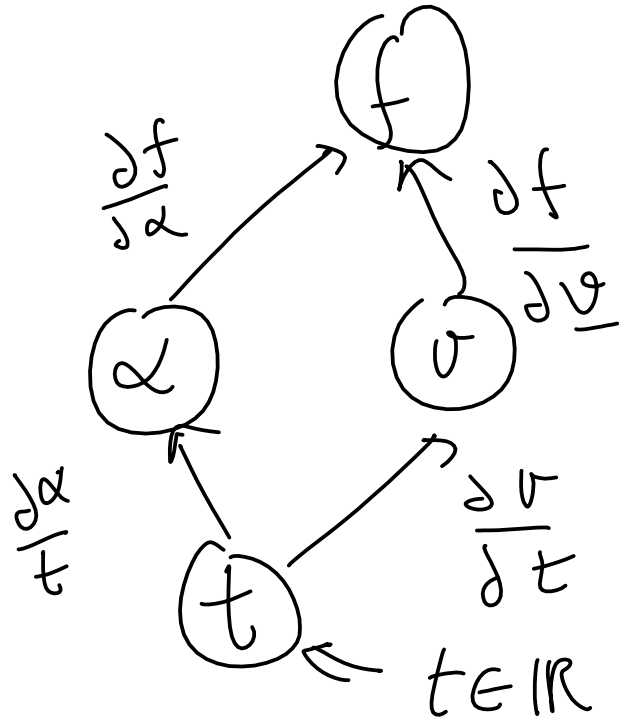
$$f(\alpha, \underline{v}) = \alpha \underline{v}$$

$$\frac{\partial}{\partial \underline{v}}(\alpha \underline{v}) = J_{\underline{v}} f(\alpha, \underline{v})$$

$$J_{\underline{v}} f(\alpha, \underline{v}) \frac{\partial \underline{v}}{\partial t}$$

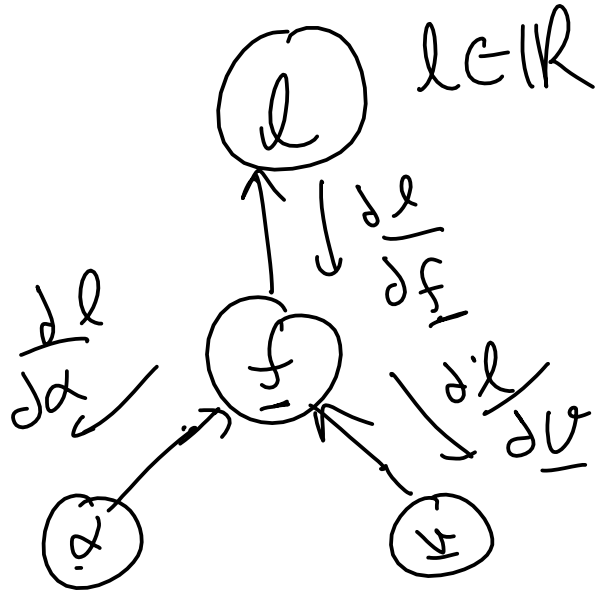
$$\alpha I_{n \times n} \frac{\partial \underline{v}}{\partial t}$$

Forward differentiation



Reverse differentiation

$$\ell(f(\alpha,v)) = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial \alpha} + \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial v}$$



$\ell(f(g(h(x))))$

$$\underbrace{\underbrace{\left(\frac{\partial \ell}{\partial f}\frac{\partial f}{\partial g}\right)\frac{\partial g}{\partial h}}_{\frac{\partial \ell}{\partial g}}\frac{\partial h}{\partial x}}_{\frac{\partial \ell}{\partial h}}$$

$$\frac{\partial \ell}{\partial x}$$

$$\underbrace{\underbrace{\underbrace{\left(\frac{\partial \ell}{\partial f}\left(\frac{\partial f}{\partial g}\left(\frac{\partial g}{\partial h}\frac{\partial h}{\partial x}\right)\right)\right)}_{\frac{\partial \ell}{\partial g}}\frac{\partial g}{\partial x}}_{\frac{\partial \ell}{\partial f}}\frac{\partial f}{\partial x}}_{\frac{\partial \ell}{\partial x}}$$

$\underline{f}(\alpha, \underline{v}) = \alpha \underline{v}$

$l \in \mathbb{R}$ $\alpha \in \mathbb{R}$

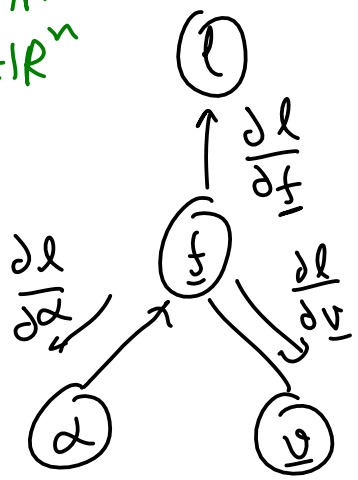
$\underline{f} \in \mathbb{R}^n$ $\underline{v} \in \mathbb{R}^n$

RD or VJP (vector Jacobian product

$$\underbrace{\frac{\partial}{\partial \alpha}}_{1\times 1} l(\underline{f}(\alpha, \underline{v})) = \frac{\partial l}{\partial \underline{f}} \frac{\partial \underline{f}}{\partial \alpha} = \underbrace{\frac{\partial l}{\partial \underline{f}}}_{\in \mathbb{R}^{1\times n}} \underbrace{\underline{v}}_{n\times 1}$$

$$\underbrace{\frac{\partial}{\partial \underline{v}}}_{1\times n} l(\underline{f}(\alpha, \underline{v})) = \underbrace{\frac{\partial l}{\partial \underline{f}}}_{1\times n} \underbrace{\frac{\partial \underline{f}}{\partial \underline{v}}}_{n\times n}$$

$$\underbrace{\frac{\partial \underline{f}}{\partial \underline{v}}}_{n\times n} = \alpha I_{n\times n}$$



$$\frac{\partial l}{\partial \underline{f}} = \begin{bmatrix} \frac{\partial l_1}{\partial f_1} & \cdots & \frac{\partial l_1}{\partial f_n} \\ \vdots & \ddots & \\ \frac{\partial l_m}{\partial f_1} & & \frac{\partial l_m}{\partial f_n} \end{bmatrix}$$

$$\frac{\partial l}{\partial \underline{f}} = \begin{bmatrix} \frac{\partial l}{\partial f_1} & \cdots & \frac{\partial l}{\partial f_n} \end{bmatrix}$$

$$f(\underline{a}, \underline{b}) = \underline{a}^T \underline{b}$$

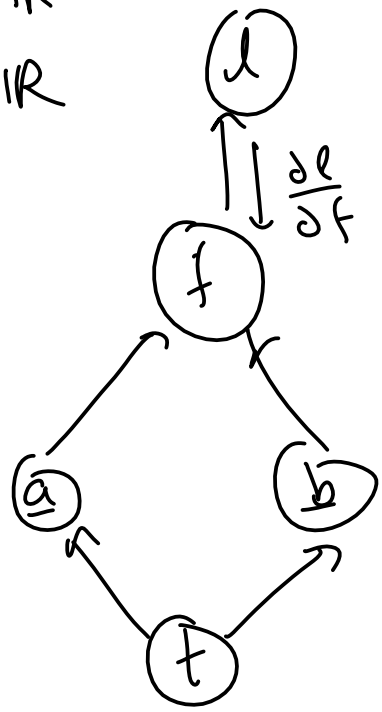
$\underline{a} \in \mathbb{R}^n$

$\underline{b} \in \mathbb{R}^n$

$f \in \mathbb{R}$

FD on JVP

$$\frac{\partial}{\partial t} f(\underline{a}, \underline{b}) = \frac{\underline{a}^T \partial \underline{b}}{\partial t} + \left(\frac{\partial \underline{a}}{\partial t}\right)^T \underline{b}$$



RD on VJP

$$\underbrace{\frac{\partial}{\partial \underline{a}} l(f(\underline{a}, \underline{b}))}_{1 \times n} = \frac{\partial l}{\partial f}\left(\frac{\partial f}{\partial \underline{a}}\right) = \underbrace{\frac{\partial l}{\partial f}}_{1 \times 1} \underbrace{\underline{b}^T}_{1 \times n}$$

$$\frac{\partial}{\partial \underline{b}} l(f(\underline{a}, \underline{b})) = \frac{\partial l}{\partial f}\left(\frac{\partial f}{\partial \underline{b}}\right) = \frac{\partial l}{\partial f} \underline{a}^T$$

$$\frac{\partial}{\partial t} \underline{a}^T \underline{x} = \left(\frac{\partial}{\partial \underline{x}} \underline{a}^T \underline{x}\right)\left(\frac{\partial \underline{x}}{\partial t}\right) + \left(\frac{\partial}{\partial \underline{a}} (\underline{a}^T \underline{x})\right)\frac{\partial \underline{a}}{\partial t}$$

$$\frac{\partial}{\partial \underline{x}} \underline{b}^T \underline{x} = \underline{b}^T$$

$$\frac{\partial}{\partial \underline{x}} \underline{x}^T \underline{b} = \underline{b}^T$$

$$\frac{\partial}{\partial t} \underline{a}^T \underline{x} = \frac{\partial}{\partial t}\left[a_1 x_1 + a_2 x_2 + \cdots a_n x_n\right]$$

$$= \left(\frac{\partial a_1}{\partial t}\right) x_1 + a_1 \frac{\partial x_1}{\partial t} + \frac{\partial a_2}{\partial t} x_2 + a_2 \frac{\partial x_2}{\partial t} + \cdots + a_n \frac{\partial x_n}{\partial t} + \frac{\partial a_n x_n}{\partial t}$$

If f is linear

$$f(\alpha \underline{x} + \beta \underline{y}) = \alpha f(\underline{x}) + \beta f(\underline{y})$$

$$\frac{\partial f(\underline{x})}{\partial t} = f\left(\frac{\partial \underline{x}}{\partial t}\right)$$

$$\frac{\partial}{\partial t} f(\underline{x}, \underline{y}) = f\left(\frac{\partial \underline{x}}{\partial t}, \underline{y}\right) + f\left(\underline{x}, \frac{\partial \underline{y}}{\partial t}\right)$$

$$\underline{f}(A, \underline{b}) = A\underline{b} \qquad A \in \mathbb{R}^{m \times n} \quad \underline{b} \in \mathbb{R}^{n} \quad \underline{f} \in \mathbb{R}^{m}$$

$$A = \begin{bmatrix} \underline{a}_1^T \\ \underline{a}_2^T \\ \vdots \\ \underline{a}_m^T \end{bmatrix}$$

FD or JVP

$$\frac{\partial}{\partial t}\underline{f}(A, \underline{b}) = \frac{\partial}{\partial t}\underbrace{\begin{bmatrix} \underline{a}_1^T \\ \underline{a}_2^T \\ \vdots \\ \underline{a}_m^T \end{bmatrix}}_{m\times n}\underbrace{\underline{b}}_{n\times 1} = \frac{\partial}{\partial t}\begin{bmatrix} \underline{a}_1^T\underline{b} \\ \underline{a}_2^T\underline{b} \\ \vdots \\ \underline{a}_m^T\underline{b} \end{bmatrix}_{1\times 1} \frac{\partial A}{\partial t} = \begin{bmatrix} \frac{\partial a_{11}}{\partial t} & \cdots & \frac{\partial a_{1n}}{\partial t} \\ \vdots & & \\ \frac{\partial a_{m1}}{\partial t} & \cdots & \frac{\partial a_{mn}}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial}{\partial \underline{a}_1}(\underline{a}_1^T\underline{b})\frac{\partial \underline{a}_1}{\partial t} + \frac{\partial}{\partial \underline{b}}(\underline{a}_1^T\underline{b})\frac{\partial \underline{b}}{\partial t} \\ \frac{\partial}{\partial \underline{a}_2}(\underline{a}_2^T\underline{b})\frac{\partial \underline{a}_2}{\partial t} + \frac{\partial}{\partial \underline{b}}(\underline{a}_2^T\underline{b})\frac{\partial \underline{b}}{\partial t} \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{\partial}{\partial t}\underline{a}_1^T\right)\underline{b} + \underline{a}_1^T\frac{\partial \underline{b}}{\partial t} \\ \left(\frac{\partial}{\partial t}\underline{a}_2^T\right)\underline{b} + \underline{a}_2^T\frac{\partial \underline{b}}{\partial t} \\ \vdots \end{bmatrix}$$

$$\frac{\partial}{\partial t}\underline{a}^T\underline{b} = \frac{\partial \underline{a}^T}{\partial t}\underline{b} + \underline{a}^T\frac{\partial \underline{b}}{\partial t}$$

$$= \begin{bmatrix} \frac{\partial \underline{a}_1^T}{\partial t} \\ \frac{\partial \underline{a}_2^T}{\partial t} \\ \vdots \\ \frac{\partial \underline{a}_m^T}{\partial t} \end{bmatrix}\underline{b} + \begin{bmatrix} \underline{a}_1^T \\ \underline{a}_2^T \\ \vdots \\ \underline{a}_m^T \end{bmatrix}\frac{\partial \underline{b}}{\partial t}$$

$$\frac{\partial (A\underline{b})}{\partial t} = \frac{\partial A}{\partial t}\underline{b} + A\frac{\partial \underline{b}}{\partial t}$$

$$\frac{\partial (AB)}{\partial t} = \frac{\partial A}{\partial t} B + A \frac{\partial B}{\partial t}$$

$f(AB) = AB$

$A \in \mathbb{R}^{m \times n}$

$B \in \mathbb{R}^{n \times p}$

$\underline{f} \in \mathbb{R}^m, \underline{x} \in \mathbb{R}^n$

$$\frac{\partial \underline{f}}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial A}{\partial \underline{x}} = ?$$

$A \in \mathbb{R}^{m \times n}$

$\underline{x} \in \mathbb{R}^p$

$$\begin{bmatrix} \frac{\partial a_{ij}}{\partial x_k} \end{bmatrix}$$

$i \in [1, m]$

$j \in [1, n]$

$k \in [1, p]$

3D Tensor

RD on VJP

$A \in \mathbb{R}^{m \times n}$

$\underline{f}(A, \underline{b}) = A\underline{b}$

$\underline{b} \in \mathbb{R}^n$

$\underline{f} \in \mathbb{R}^m$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$$\frac{\partial l}{\partial A} = \begin{bmatrix} \frac{\partial l}{\partial a_{11}} & \cdots & \frac{\partial l}{\partial a_{1n}} \\ \vdots & & \\ \frac{\partial l}{\partial a_{m1}} & \cdots & \frac{\partial l}{\partial a_{mn}} \end{bmatrix}_{m \times n} \qquad \frac{\partial l}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial l}{\partial x_1} & \cdots & \frac{\partial l}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial}{\partial A} l(\underline{f}(A, \underline{b})) = \frac{\partial l}{\partial \underline{f}} \frac{\partial}{\partial A}(A\underline{b}) = \frac{\partial l}{\partial \underline{f}} \frac{\partial}{\partial A} \begin{bmatrix} \underline{a}_1^T \underline{b} \\ \underline{a}_2^T \underline{b} \\ \vdots \\ \underline{a}_m^T \underline{b} \end{bmatrix}$$

$$\frac{\partial}{\partial \underline{b}} l(f(A, \underline{b})) = ?$$

$$\frac{\partial}{\partial f}\frac{\partial f}{\partial A}\begin{bmatrix} a_1^T b \\ a_2^T b \\ \vdots \\ a_m^T b \end{bmatrix} = \begin{bmatrix} \frac{\partial l}{\partial f}\frac{\partial}{\partial A} a_1^T b \\ \frac{\partial l}{\partial f}\frac{\partial}{\partial A} a_2^T b \\ \vdots \\ \frac{\partial l}{\partial f}\frac{\partial}{\partial A} a_m^T b \end{bmatrix}$$

$$\frac{\partial l}{\partial A} = \begin{bmatrix} \frac{\partial l}{\partial a_1^T} \\ \frac{\partial l}{\partial a_2^T} \\ \vdots \\ \frac{\partial l}{\partial a_m^T} \end{bmatrix}$$

$$\frac{\partial (a_1^T b)}{\partial A} = \begin{bmatrix} \frac{\partial (a_1^T b)}{\partial a_1^T} \\ \frac{\partial (a_1^T b)}{\partial a_2^T} \\ \vdots \end{bmatrix} = \begin{bmatrix} b^T \\ 0^T \\ \vdots \end{bmatrix}$$

$$\begin{bmatrix} \vdots \\ \frac{\partial (a_1^T b)}{\partial a_m^T} \end{bmatrix} \quad \begin{bmatrix} \vdots \\ 0^T \end{bmatrix}_{m \times n}$$

$$\frac{\partial l}{\partial f} \in 1 \times m \cdot \frac{\partial l}{\partial f} = \left[\frac{\partial l}{\partial f_1}, \ldots, \frac{\partial l}{\partial f_m} \right]$$

$$\frac{\partial l}{\partial f} \frac{\partial}{\partial A} (a_1^T b) = \left[\frac{\partial l}{\partial f_1} \cdots \frac{\partial l}{\partial f_m} \right] \begin{bmatrix} b^T \\ 0^T \\ \vdots \\ 0^T \end{bmatrix}$$

$$= \frac{\partial l}{\partial f_1} b^T$$

$$\frac{\partial l}{\partial f} \frac{\partial}{\partial A} (A b) = \begin{bmatrix} \frac{\partial l}{\partial f_1} b^T \\ \frac{\partial l}{\partial f_2} b^T \\ \vdots \\ \frac{\partial l}{\partial f_m} b^T \end{bmatrix}_{m \times n} = \underbrace{\left(\frac{\partial l}{\partial f} \right)^T}_{1 \times m} \underbrace{b^T}_{1 \times n}$$

$m \times 1$

3. $f(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}$ where $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$
4. $\mathbf{f}(\mathbf{A}, \mathbf{b}) = \mathbf{A}\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^n$
5. $F(\mathbf{A}, \mathbf{B}) = \mathbf{A}\mathbf{B}$ where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$

A27:

Let the vector be $\frac{\partial l}{\partial f}$

1. $\frac{\partial l}{\partial x} = \frac{\partial l}{\partial f} exp(x)$

2. Let the vector be $\frac{\partial l}{\partial \mathbf{f}}$. Then,

$\frac{\partial l}{\partial \alpha} = \frac{\partial l}{\partial \mathbf{f}} \mathbf{v}$ and $\frac{\partial l}{\partial \mathbf{v}} = \frac{\partial l}{\partial \mathbf{f}} \alpha \mathbf{I}_{n \times n}$

3. Let the vector be $\frac{\partial l}{\partial f}$. Then

$\frac{\partial l}{\partial \mathbf{a}} = \frac{\partial l}{\partial f} \mathbf{b}^\top$ and $\frac{\partial l}{\partial \mathbf{b}} = \frac{\partial l}{\partial f} \mathbf{a}^\top$

Define matrix multiplication to be: Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

then,

$$\frac{\partial l}{\partial \mathbf{A}} = \begin{bmatrix} \frac{\partial l}{\partial a_{11}} & \frac{\partial l}{\partial a_{12}} & \dots & \frac{\partial l}{\partial a_{1n}} \\ \frac{\partial l}{\partial a_{21}} & \frac{\partial l}{\partial a_{22}} & \dots & \frac{\partial l}{\partial a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial l}{\partial a_{m1}} & \frac{\partial l}{\partial a_{m2}} & \dots & \frac{\partial l}{\partial a_{mn}} \end{bmatrix}$$

4. Let the vector be $\frac{\partial l}{\partial \mathbf{f}}$. Then

$\frac{\partial l}{\partial \mathbf{b}} = \frac{\partial l}{\partial \mathbf{f}} \mathbf{A}$ and $\frac{\partial l}{\partial \mathbf{A}} = \frac{\partial l}{\partial \mathbf{f}}^\top \mathbf{b}^\top$.

5. Let the vector be $\frac{\partial l}{\partial F} \in \mathbb{R}^{p \times m}$. Then

$\frac{\partial l}{\partial \mathbf{A}} = \frac{\partial l}{\partial F} \mathbf{B}^\top$ and $\frac{\partial l}{\partial \mathbf{B}} = \mathbf{A}^\top \frac{\partial l}{\partial F}$

Q28:

Write the forward-mode Jacobian-vector product(s) for the following operations

1. $f(x) = exp(x)$ where $x \in \mathbb{R}$
2. $\mathbf{f}(\alpha, \mathbf{v}) = \alpha\mathbf{v}$ where $\alpha \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$
3. $f(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top\mathbf{b}$ where $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$
4. $\mathbf{f}(\mathbf{A}, \mathbf{b}) = \mathbf{A}\mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^n$
5. $F(\mathbf{A}, \mathbf{B}) = \mathbf{A}\mathbf{B}$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$

A28:

Let the vector be $\frac{\partial x}{\partial t}$

1. $\frac{\partial f}{\partial t} = exp(x)\frac{\partial x}{\partial t}$
2. Let the vectors be $\frac{\partial \alpha}{\partial t}$ and $\frac{\partial \mathbf{v}}{\partial t}$. Then,

$$\frac{\partial \mathbf{f}}{\partial t} = \mathbf{v}\frac{\partial \alpha}{\partial t} + \alpha\mathbf{I}_{n\times n}\frac{\partial \mathbf{v}}{\partial t}$$

3. Let the vectors be $\frac{\partial \mathbf{a}}{\partial t}$ and $\frac{\partial \mathbf{b}}{\partial t}$. Then

$$\frac{\partial f}{\partial t} = \mathbf{b}^\top\frac{\partial \mathbf{a}}{\partial t} + \mathbf{a}^\top\frac{\partial \mathbf{b}}{\partial t}$$

4. Let the vectors be $\frac{\partial \mathbf{A}}{\partial t}$ and $\frac{\partial \mathbf{b}}{\partial t}$. Then

$$\frac{\partial \mathbf{f}}{\partial t} = \frac{\partial \mathbf{A}}{\partial t}\mathbf{b} + \mathbf{A}\frac{\partial \mathbf{b}}{\partial t}$$

5. Let the vectors be $\frac{\partial \mathbf{A}}{\partial t}$ and $\frac{\partial \mathbf{B}}{\partial t}$. Then

$$\frac{\partial F}{\partial t} = \frac{\partial \mathbf{A}}{\partial t}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial t}$$