MMD Exercise 3

November 6th 2024

1. Exercise

a)

w1

function:

$$f = (w1 \cdot \text{vector}(1) + w2 \cdot \text{vector}(1) \odot x - y)^2$$

gradient with regards to w1:

$$\frac{\partial f}{\partial w1} = 2 \cdot (w1 \cdot \text{vector}(1) + w2 \cdot x - y)$$

where

- $w1$ is a scalar
- $w2$ is a scalar
- x is a vector
- y is a vector

w2

function:

$$f = (w1 \cdot \text{vector}(1) + w2 \cdot \text{vector}(1) \odot x - y)^2$$

gradient with regards to w2:

$$\frac{\partial f}{\partial w2} = 2 \cdot (w1 \cdot \text{vector}(1) + w2 \cdot x - y) \odot x$$

where

- $w1$ is a scalar
- $w2$ is a scalar
- x is a vector
- y is a vector

b)

q

function:

$$f = (r \cdot \text{vector}(1) - q \odot p)^2 + \|p\|_2 \cdot l1 \cdot \text{vector}(1) + \|q\|_2 \cdot l2 \cdot \text{vector}(1)$$

gradient with regards to q:

$$\frac{\partial f}{\partial q} = l2/\|q\|_2 \cdot \text{vector}(1) \cdot q^\top - 2 \cdot \text{diag}((r \cdot \text{vector}(1) - q \odot p) \odot p)$$

where

- $l1$ is a scalar
- $l2$ is a scalar
- p is a vector
- q is a vector
- r is a scalar

p

function:

$$f = (r \cdot \text{vector}(1) - q \odot p)^2 + \|p\|_2 \cdot l1 \cdot \text{vector}(1) + \|q\|_2 \cdot l2 \cdot \text{vector}(1)$$

gradient with regards to p:

$$\frac{\partial f}{\partial p} = l1/\|p\|_2 \cdot \text{vector}(1) \cdot p^\top - 2 \cdot \text{diag}((r \cdot \text{vector}(1) - q \odot p) \odot q)$$

where

- $l1$ is a scalar
- $l2$ is a scalar
- p is a vector
- q is a vector
- r is a scalar

Explanation

To be honest I am lost here

2

a)

Automatic Differentiation (AD) is used to compute gradients by breaking functions into simple operations and uses the cain rules to compute exact derivatives. In comparison numerical differentiation approximates derivatives which can leads to rounding errors. Symbolic differentiation computes the exact derivatives which can result in high computation time for complex functions. However, these issues are minimized by using AD.

b)

The forward mode computes derivatives alongside the function values by propagating the derivatives. The reverse mode propagates the derivates backwards starting from the output.

c)

The forward mode is more efficient for number of inputs being less than the number of outputs,as it propagates the inputs individually. In terms of the reverse mode it is reversed as it calculates gradients for each input in one backward pass.

d)

Checkpoints reduce memory usage as they store only certain data intermediate and recalculate necessary data. An application example would be deep neural networks as multiple layers can lead to an high increase of memory usage.