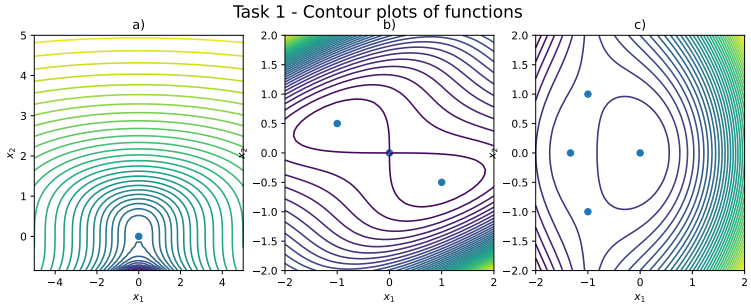
Numerical Optimization
Assignment 1

s.guda@student.tugraz.at
patrick.gaisberger@student.tugraz.at

October 2022

1 Characterization of Functions



- For each of the given functions :
  1. Compute the Gradient and the Hessian
  2. Determine the set of stationary points
  3. Characterize every stationary point whether its a saddle point, local/global minimum or maximum.

(a) $f(x) = \ln\left(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\right)$
Gradient :

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \ln\left(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\right) \\ \frac{\partial f}{\partial x_2} \ln\left(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\right) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{2x_1}{2+x_1^2+3x_2^2} \\ \frac{9x_2^2}{2+x_1^2+3x_2 2} \end{bmatrix}$$

Now we solve this system which will help us find stationary points.

$$\begin{cases} \frac{2x_1}{2+x_1^2+3x_2^2} = 0 \\ \frac{9x_2^2}{2+x_1^2+3x_2 2} = 0 \end{cases}$$
$$\{x_1 = 0 \quad x_2 = 0$$

Hessian :

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial f}{\partial x_1, x_2} \\ \frac{\partial f}{\partial x_1, x_2} & \frac{\partial^2 f}{\partial x_1^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} \ln\big(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\big) & \frac{\partial f}{\partial x_1, x_2} \ln\big(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\big) \\ \frac{\partial f}{\partial x_1, x_2} \ln\big(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\big) & \frac{\partial^2 f}{\partial x_1^2} \ln\big(1 + \frac{1}{2}(x_1^2 + 3x_2^3)\big) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\frac{1}{2}(x_1^2+3x_2^3)+1} - \frac{x_1^2}{(\frac{1}{2}(3x_2^3+x_1^2)+1)^2} & -\frac{9x_1x_2^2}{2(\frac{1}{2}(x_1^2+3x_2^3)+1)^2} \\ -\frac{9x_1x_2^2}{2(\frac{1}{2}(x_1^2+3x_2^3)+1)^2} & \frac{9x_2}{\frac{1}{2}(3x_2^3+x_1^2)+1} - \frac{81x_2^4}{4(\frac{1}{2}(3x_2^3+x_1^2)+1)^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2(-x_1^2+3x_2^3+2)}{(x_1^2+3x_2^3+2)^2} & -\frac{18x_1x_2^2}{(x_1^2+3x_2^3+2)^2} \\ -\frac{18x_1x_2^2}{(x_1^2+3x_2^3+2)^2} & \frac{9x_2(2x_1^2-3x_2^3+4)}{(x_1^2+3x_2^3+2)^3} \end{bmatrix}$$

Hessian matrix Determinant :

$$\det(H) = \frac{2(-x_1^2+3x_2^3+2)}{(x_1^2+3x_2^3+2)^2} \frac{9x_2(2x_1^2-3x_2^3+4)}{(x_1^2+3x_2^3+2)^3} - \frac{18x_1x_2^2}{(x_1^2+3x_2^3+2)^2} \frac{18x_1x_2^2}{(x_1^2+3x_2^3+2)^2}$$

Stationary point at $x_1 = 0$ and $x_2 = 0$
$\det(H)(0,0) = 0$
Stationary point cannot be classified!
No global/local maximas and minimas!

(b) $f(x) = (x_1 - 2x_2)^4 + 64x_1x_2$

Gradient :

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_1} * (x_1 - 2x_2)^4 + 64x_1x_2 \\ \frac{\partial f}{\partial x_2} * (x_1 - 2x_2)^4 + 64x_1x_2 \end{bmatrix}$$
$$= \begin{bmatrix} 64x_2 + 4(x_1 - 2x_2)^3 \\ 64x_1 - 8(x_1 - 2x_2)^3 \end{bmatrix}$$

Now we solve this system which will help us find stationary points.

$$\begin{cases} 64x_2 + 4(x_1 - 3x_2)^3 \\ 64x_1 - 8(x_1 - 3x_2)^3 \end{cases}$$
$$\begin{cases} x_1 = 0, & x_2 = 0 \\ x_1 = -1, & x_2 = 1/2 \\ x_1 = -1/2, & x_2 = 1 \end{cases}$$

Hessian :

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial f}{\partial x_1, x_2} \\ \frac{\partial f}{\partial x_1, x_2} & \frac{\partial^2 f}{\partial x_1^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} * (x_1 - 2x_2)^4 + 64x_1x_2 & \frac{\partial f}{\partial x_1, x_2} * (x_1 - 2x_2)^4 + 64x_1x_2 \\ \frac{\partial f}{\partial x_1, x_2} * (x_1 - 2x_2)^4 + 64x_1x_2 & \frac{\partial^2 f}{\partial x_2^2} * (x_1 - 2x_2)^4 + 64x_1x_2 \end{bmatrix}$$
$$= \begin{bmatrix} 12(x_1 - 2x_2)^2 & 64 - 24(x_1 - 2x_2)^2 \\ 64 - 24(x_1 - 2x_2)^2 & 48(x_1 - 2x_2)^2 \end{bmatrix}$$

Hessian matrix Determinant :

$$\det(H) = 12(x_1 - 2x_2)^2 48(x_1 - 2x_2)^2 - ((64 - 24(x_1 - 2x_2)^2)(64 - 24(x_1 - 2x_2)^2)$$

Stationary point at $x_1 = 0$ and $x_2 = 0$
$\det(H)(0, 0) = -4096$
It is less than 0 so that makes P_1(0,0) is a saddle point.

Stationary point at $x_1 = -1$ and $x_2 = 1/2$
$\det(H)(-1, 1/2) = 8192$
It is more than 0 so we need P_2(-1,1/2) to check for minimum or maximum.

$$\frac{\partial^2 f}{\partial x_1^2}(x_1 - 2x_2)^4 + 64x_1x_2 \quad for \quad x_1 = -1, x_2 = 1/2$$
$$= 48$$

The result is greater than 0 so we can say point P_2 is a local minimum

$$(x_1 - 2x_2)^4 + 64x_1x_2 \quad for \quad x_1 = -1, x_2 = 1/2$$
$$= (-1 - 2\frac{1}{2})^4 + 64 - 1\frac{1}{2}$$
$$f(-1, \frac{1}{2}) = -16$$

Stationary point at $x_1 = 1$ and $x_2 = -1/2$
$\det(H)(-1, 1/2) = 8192$
It is more than 0 so we need P_3(1,-1/2) to check for minimum or maximum.

$$\frac{\partial^2 f}{\partial x_1^2}(x_1 - 2x_2)^4 + 64x_1x_2 \quad for \quad x_1 = -1, x_2 = 1/2$$
$$= 48$$

The result is greater than 0 so we can say point P_3 is a local minimum

$$(x_1 - 2x_2)^4 + 64x_1x_2 \quad for \quad x_1 = -1, x_2 = 1/2$$
$$= (1 - 2 - \frac{1}{2})^4 + 641 - \frac{1}{2}$$
$$f(-1, \frac{1}{2}) = -16$$

For Point $P_2(-1, \frac{1}{2})$ and $P_3(1, -\frac{1}{2})$ we need to check whether it is a global or local minima.

$$\nabla^2 f(-1, \frac{1}{2}) = \begin{bmatrix} 48 & -32 \\ -32 & 192 \end{bmatrix} \quad \nabla^2 f(1, -\frac{1}{2}) = \begin{bmatrix} 48 & -32 \\ -32 & 192 \end{bmatrix}$$

The Eigenvalues of $\nabla^2(f(-1, \frac{1}{2}))$ and $\nabla^2(f(1, -\frac{1}{2}))$ both come out as positive.

$$\nabla^2 f(-1, \frac{1}{2} - \lambda I) = \begin{bmatrix} 48 - \lambda & -32 \\ -32 & 192 - \lambda \end{bmatrix}$$

$$\begin{aligned} det(\nabla^2(f - \lambda I)) &= (48 - \lambda)(192 - \lambda) - (-32)^2 \\ &= \lambda^2 - 240\lambda + 8192 \\ \lambda_1 &= 8(15 + \sqrt{97}) \\ \lambda_2 &= -8(\sqrt{97} - 15) \end{aligned}$$

Point P_3 will have the same Eigenvalue.
Next we check the span of the function.

$$\lim_{x_1, \to +\infty} f \textbf{ then } f \to +\infty$$

$$\lim_{x_1, \to -\infty} f \textbf{ then } f \to +\infty$$

This means the function will have a global minimum. In our case bot P_2 and P_3 when plugged into the function give the same value so they will be at the same height in the function. We can say that our function has a global maximum that occurs at $f(x_1, x_2 = -16)$ by 2 different points P_2 and P_3.

(c) $f(x) = x_1^2 + x_1\|x\|^2 + \|x\|^2$

$f(x) = x_1^2 + x_1(\sqrt{x_1^2 + x_2^2})^2 + (\sqrt{x_1^2 + x_2^2})^2$
$f(x) = 2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2$
Gradient :

$$\begin{aligned} \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} &= \begin{bmatrix} \frac{\partial f}{\partial x_1} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \\ \frac{\partial f}{\partial x_2} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \end{bmatrix} \\ &= \begin{bmatrix} 3x_1^2 + 4x_1 + x_2 2 \\ 2x_1x_2 + 2x_2 \end{bmatrix} \end{aligned}$$

Now we solve this system which will help us find stationary points.

$$\begin{cases} 3x_1^2 + 4x_1 + x_2^2 \\ 2x_1x_2 + 2x_2 \end{cases}$$

$$\begin{cases} x_1 = 0, \quad x_2 = 0 \\ x_1 = -\dfrac{4}{3}, \quad x_2 = 0 \\ x_1 = -1, \quad x_2 = 1 \\ x_1 = -1, \quad x_2 = -1 \end{cases}$$

Hessian :

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial f}{\partial x_1, x_2} \\ \frac{\partial f}{\partial x_1, x_2} & \frac{\partial^2 f}{\partial x_1^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) & \frac{\partial f}{\partial x_1, x_2} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \\ \frac{\partial f}{\partial x_1, x_2} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) & \frac{\partial^2 f}{\partial x_1^2} * (2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \end{bmatrix}$$

$$= \begin{bmatrix} 4 + 6x_1 & 2x_2 \\ 2x_2 & 2x_1 + 2 \end{bmatrix}$$

Hessian matrix Determinant :

$$\det(H) = (4 + 6x_1)(2x_1 + 2) - (2x_2 2x_2)$$

Stationary point at $x_1 = 0$ and $x_2 = 0$
$\det(H)(0, 0) = 8$
It is more than 0 so we need P_1(0,0) to check for minimum or maximum.

$$\frac{\partial^2 f}{\partial x_1^2}(2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \quad for \quad x_1 = 0, x_2 = 0$$
$$= 4$$

The result is greater than 0 so we can say point P_1 is a local minimum

$$(2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \quad for \quad x_1 = 0, x_2 = 0$$
$$= 0$$
$$f(0, 0) = 0$$

Stationary point at $x_1 = -\frac{4}{3}$ and $x_2 = 0$
$\det(H)(-\frac{4}{3}, 0) = \frac{8}{3}$
It is more than 0 so we need P_2(-1,1/2) to check for minimum or maximum.

$$\frac{\partial^2 f}{\partial x_1^2}(2x_1^2 + x_2^2 + x_1^3 + x_1x_2^2) \quad for \quad x_1 = -\frac{4}{3}, x_2 = 0$$
$$= -4$$

The result is lesser than 0 so we can say point P_2 is a local maximum

Stationary point at $x_1 = -1$ and $x_2 = -1$
$\det(H)(-1,-1) = -4$
It is less than 0 so we can say P_3(-1,-1) is a saddle point.

Stationary point at $x_1 = -1$ and $x_2 = 1$
$\det(H)(-1,1) = -4$
It is less than 0 so we can say P_4(-1,1) is a saddle point.

This function does not have a **global** minimum or maximum. The range of this function spans from $-\infty$ to ∞. We can prove this by having a look at the limits of the function when x_1 and x_2 aproaches ∞.

$$\lim_{x_1,\to+\infty} f \textbf{ then } f \to +\infty$$

$$\lim_{x_1,\to-\infty} f \textbf{ then } f \to -\infty$$

The reason why the functions spans to $-\infty$ is because the power of 3 will be dominant and as x_1 aproaches $-\infty$ the function will be strictly negative.

2 Matrix Calculus

(a) Compute the Gradient and the Hessian for

$$\begin{aligned} f(x) &= \frac{1}{2}||c \odot (x-k)||^2 \\ &= \frac{1}{2}\left(\sqrt{(c_1(x_1-k_1))^2 + (c_2(x_2-k_2))^2 + (c_n(x_n-k_n))^2}\right)^2 \\ &= \frac{1}{2}(\sum_{i=1}^{n}(c_i(x_i-k_i))^2) \end{aligned}$$

Gradient calculation. First derivative of the function.

$$\begin{aligned} \nabla f &= \frac{\partial f}{\partial x_k}\frac{1}{2}(\sum_{i=1}^{n}(c_i(x_i-k_i))^2) \\ &= \frac{1}{2}(\sum_{i=1}^{n}\frac{\partial f}{\partial x_k} * (c_i(x_i-k_i))^2) \end{aligned}$$

for $x_k = x_i$ result of the derivation will be $2c_k^2(x_k - k_k)$ otherwise 0
Gradient will equal :

$$\nabla f = \frac{1}{2}(2c^2(x-k)) = c^2(x-k)$$

For the hessian we calculate the second derivative of the function or derivative of the gradient

$$\nabla^2 f = \frac{\partial f}{\partial x} c^2(x-k) = c^2$$

(b) Compute the Gradient and the Hessian for

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} h((Ax)_i) \quad for \quad h(t) = \frac{1}{2}t^2 + 2t,\ t \in R, A \in R^{nxn} \\ &= \sum_{i=1}^{n} h(\sum_{j=1}^{n} A_{ij}x_j) \\ &= \sum_{i=1}^{n} \frac{1}{2}(\sum_{j=1}^{n} A_{ij}x_j)^2 + 2\sum_{j=1}^{n} A_{ij}x_j \end{aligned}$$

Gradient calculation. First derivative of the function.

$$\begin{aligned} \nabla f &= \frac{\partial f}{\partial x_k} * (\sum_{i=1}^{n} \frac{1}{2}(\sum_{j=1}^{n} A_{ij}x_j)^2 + 2\sum_{j=1}^{n} A_{ij}x_j) \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_k} * (\frac{1}{2}(\sum_{j=1}^{n} A_{ij}x_j)^2 + 2\sum_{j=1}^{n} A_{ij}x_j) \\ &= \sum_{i=1}^{n} (\frac{1}{2}(2\sum_{j=1}^{n} A_{ij}x_j \frac{\partial f}{\partial x_k} * (\sum_{j=1}^{n} A_{ij}x_j)) + 2\frac{\partial f}{\partial x_k} * \sum_{j=1}^{n} A_{ij}x_j) \\ &= \sum_{i=1}^{n} (\sum_{j=1}^{n} A_{ij}x_j)A_{ik} + 2A_{ik}) \\ &= \sum_{i=1}^{n} A_{ik}(\sum_{j=1}^{n} A_{ij}x_j) + 2) \\ &= A^T(Ax + 2) \end{aligned}$$

for $x_k = x_i$ result of the derivation will be A_{ij} otherwise 0
Gradient will equal :

$$\nabla f = A^T(Ax + 2)$$

For the hessian we calculate the second derivative of the function or derivative of the gradient

$$\begin{aligned}\nabla^2 f = \frac{\partial f}{\partial x} * A^T(Ax+2) &= \frac{\partial f}{\partial x} * \sum_{i=1}^{n} A_{ik}(\sum_{j=1}^{n} A_{ij}x_j) + 2) \\ &= \sum_{i=1}^{n} A_{ik}(\sum_{j=1}^{n} \frac{\partial f}{\partial x_l} * (A_{ij}x_j) + \frac{\partial f}{\partial x_l}2) \\ &= \sum_{i=1}^{n} A_{ik}A_{il} = A^T \cdot A\end{aligned}$$

(c) Compute the Gradient and the Hessian for

$$\begin{aligned}f(\alpha) &= \frac{1}{2}\|A(x+\alpha y) - b\|^2 \text{ for } \alpha \in R,\ x, y \in R^n,\ b \in R^m, A \in R^{n \times m} \\ &= \frac{1}{2}(\sum_{j=1}^{m}(\sum_{i=1}^{n} A_{ji}(x_i + \alpha y_i) - b_j)^2)\end{aligned}$$

Gradient calculation. First derivative of the function.

$$\begin{aligned}\nabla f &= \frac{\partial f}{\partial x_k} * \frac{1}{2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} A_{ji}(x_i + \alpha y_i) - b_j\right)^2 \\ &= \frac{1}{2}2(\sum_{j=1}^{m}(\sum_{i=1}^{n} A_{ji}(x_i + \alpha y_i) - b_j)\frac{\partial f}{\partial a} * (\sum_{i=1}^{n} A_{ji}(x_i + \alpha y_i) - b_j)) \\ &= \sum_{j=1}^{m}(\sum_{i=1}^{n}(A_{ji}(x_i + \alpha y_i) - b_j)A_{ji}y_i)) \\ &= (A(x+\alpha y) - b)(Ay)^T\end{aligned}$$

Gradient will equal :

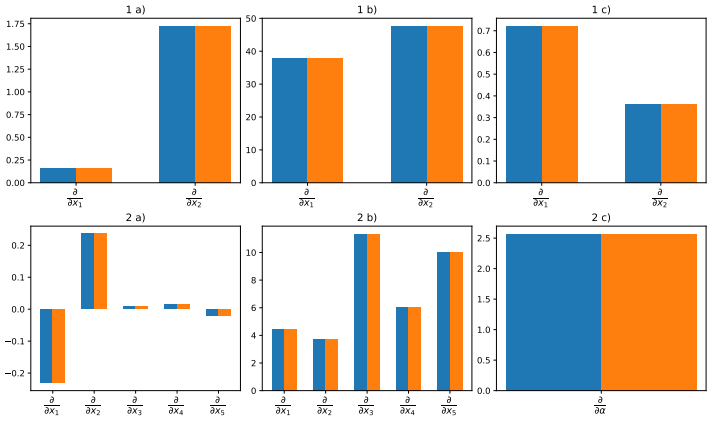
$$\nabla f = (A(x+\alpha y) - b)(Ay)^T$$

For the hessian we calculate the second derivative of the function or derivative of the gradient

$$\begin{aligned}
\nabla^2 f &= \frac{\partial f}{\partial x} * A(x + \alpha y) - b)(Ay)^T \\
&= \frac{\partial f}{\partial x} * \sum_{j=1}^{m}(\sum_{i=1}^{n} A_{ji}(x_i + \alpha y_i) - b_j)A_{ji}y_i \\
&= \sum_{j=1}^{m}(\sum_{i=1}^{n} \frac{\partial f}{\partial x} * ((A_{ji}(x_i + \alpha y_i) - b_j)A_{ji}y_i)) \\
&= \sum_{j=1}^{m}(\sum_{i=1}^{n}(A_{ji}y_i)(A_{ji}y_i)) \\
&= (Ay)(Ay)^T
\end{aligned}$$

3 Numerical Gradient Approximation

Task	Analytical	Approximated
1a	$\nabla f(x) = \begin{bmatrix} 0.15876359484002991 \\ 1.72411072038896 \end{bmatrix}$	$\nabla f(x) = \begin{bmatrix} 0.1587635948818722 \\ 1.724110720413563 \end{bmatrix}$
1b	$\nabla f(x) = \begin{bmatrix} 37.91112574914982 \\ 47.65364675485097 \end{bmatrix}$	$\nabla f(x) = \begin{bmatrix} 37.911125749801045 \\ 47.65364675485273 \end{bmatrix}$
1c	$\nabla f(x) = \begin{bmatrix} 0.7214343006411869 \\ 0.360665146631054 \end{bmatrix}$	$\nabla f(x) = \begin{bmatrix} 0.7214343006461954 \\ 0.3606651466231403 \end{bmatrix}$
2a	$\nabla f(x) = \begin{bmatrix} -0.23156760781536873 \\ 0.2365904612531706 3 \\ 0.007627519380685171 \\ 0.013780976894936542 \\ -0.0198886218552852 \end{bmatrix}$	$\nabla f(x) = \begin{bmatrix} -0.2315671374650734 \\ 0.236590807729944 \\ 0.007627541917301573 \\ 0.013780993429349865 \\ -0.01988860504063226 \end{bmatrix}$
2b	$\nabla f(x) = \begin{bmatrix} 4.487123665446781 \\ 3.7596287728711992 \\ 11.339896254936967 \\ 6.021859148912858 \\ 10.042643652593142 \end{bmatrix}$	$\nabla f(x) = \begin{bmatrix} 4.48712391509338 \\ 3.7596290472958804 \\ 11.339897677179781 \\ 6.021859763000317 \\ 10.042644861196646 \end{bmatrix}$
2c	$\nabla f(\alpha) = (2.568274108756583)$	$\nabla f(a) = 2.5682746302252197$



4 Vectors, Norms and Matrices

1. Show that $\|\cdot\|_{\frac{1}{2}}$ for $x \in \mathbb{R}^n$ is not a norm.
   We can show that it violates the property of nonnegativity $\|\cdot\| \geq 0$ and the triangle inequality, but proving one wrong is sufficient.

$$n = 1; x = -1$$
$$\|x\|_{\frac{1}{2}} = (\sqrt{-1})^2 = i^2 = -1 \leq 0$$

   Although the prove goes over the imaginary number i the method still maps to $\mathbb{R}$

2. Show that for any $x, y, z \in \mathbb{R}^n$ the following inequality holds

$$\begin{aligned} \|x - z\| &\leq \|x - y\| + \|y - z\| \\ a &= x - y \\ b &= y - z \\ \|x - z\| &= \|x - y + y - z\| = \|a + b\| \leq \\ &\leq \|a\| + \|b\| = \|x - y\| + \|y - z\| \\ \|x - z\| &\leq \|x - y\| + \|y - z\| \end{aligned}$$

3. Let $x, y \in \mathbb{R}^n$ be two orthogonal vectors, prove that

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2$$

$$\|x+y\|^2 = \sum_i^n (x_i + y_i)^2 = \sum_i^n x_i^2 + 2x_i y_i + y_i^2 =$$

$$= \sum_i^n x_i^2 + 2\sum_i^n x_i y_i + \sum_i^n y_i^2 =$$

$$= \|x\|^2 + 2 \cdot 0 + \|y\|^2 = \|x\|^2 + \|y\|^2$$

dot product between two orthogonal vectors is 0

4. Let $Q \in \mathbb{R}^{n \times n}$ be a positive definite matrix, Show that the following is a norm.

$$\|x\|_Q = \sqrt{x^T Q x}$$

Prove $x^T Qy$ to be a proper inner product $\langle x, y \rangle$

Positive definite matrix are symmetric

$$\langle x, y \rangle = x^T Qy = (x^T Qy)^T = y^T Q^T x = y^T Qy = \langle y, x \rangle$$

$\Rightarrow$**symmetry✓**

$$\langle x, y+z \rangle = x^T Q(y+z) = x^T Qy + x^T Qz = \langle x, y \rangle + \langle x, z \rangle$$

$\Rightarrow$**additivity✓**

$$\langle \lambda x, y \rangle = \lambda x^T Qy = \lambda \langle x, y \rangle$$

$\Rightarrow$**homogeneity✓**

From the definition of a positive definite matrix we take

$$x^T Qx \geq 0 \;\; \& \;\; x^T Qx = 0 \textbf{ iff } x = 0$$

$\Rightarrow$**pos. definiteness✓**

$\Rightarrow$**$x^T Qy$ is a proper inner product $\langle x, y \rangle$**

Hence $\sqrt{\langle x, x \rangle} = \sqrt{x^T Qx} = \|x\|_Q$ **is a norm!**

From the lecture notes (math preliminaries) we know that any square root of an inner product $\sqrt{\langle x, x \rangle}$ is a norm. The proof for that is also more or less simple. Nonnegativity and positive homogeneity emerge from the pos. definiteness, symmetry and the homogenity. For the triangle inequality we can proof Cauchy Schwarz Ineq and then use it together with the additivity to proof the the triangle inequality correct.

5 Automatic Differentiation

1. Compute the derivative of $\sigma(t)$ for $t \in \mathbb{R}$, which you will need in the subsequent tasks.

$$\begin{aligned}\sigma(t) &= \frac{1}{1+\exp(-t)} \\ \frac{\mathrm{d}\sigma(t)}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{1+\exp(-t)} = -\frac{\frac{\mathrm{d}}{\mathrm{d}t}[1+\exp(-t)]}{(1+\exp(-t))^2} \\ &= -\frac{0-\exp(-t)}{(1+\exp(-t))^2} = \frac{\exp(-t)}{(1+\exp(-t))^2}\end{aligned}$$

2. Compute all partial derivatives $\frac{\partial y}{\partial x_i}$ using automatic differentitaion in forward mode. Do this manually, showing how each partial derivative can be evaluated in parallel to one forward sweep.

$$W^{(0)} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \quad W^{(1)} = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

$$\begin{aligned} z_1 &= W_{11}^{(0)} x_1 + W_{12}^{(0)} x_2 \\ z_2 &= W_{21}^{(0)} x_1 + W_{22}^{(0)} x_2 \\ \alpha_1 &= \sigma(z_1) = \frac{1}{1+\exp(-z_1)} \\ \alpha_2 &= \sigma(z_2) = \frac{1}{1+\exp(-z_2)} \\ y &= W_1^{(1)} \alpha_1 + W_2^{(1)} \alpha_2 \end{aligned}$$

Computing partial derivatives $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$ in forward mode.
Seed for $\frac{\partial f}{\partial x_1}$:

$$\dot{x_1} = 1 \qquad \dot{x_2} = 0$$
$$\dot{z_1} = \frac{\partial z_1}{\partial x_1}\dot{x_1} + \frac{\partial z_1}{\partial x_2}\dot{x_2} = W_{11}^{(0)}\dot{x_1}$$
$$\dot{z_2} = \frac{\partial z_2}{\partial x_1}\dot{x_1} + \frac{\partial z_2}{\partial x_2}\dot{x_2} = W_{21}^{(0)}\dot{x_1}$$
$$\dot{\alpha_1} = \frac{\partial \alpha_1}{\partial z_1}\dot{z_1} + \frac{\partial \alpha_1}{\partial z_2}\dot{z_2} = \frac{\exp(-z_1)}{(1+\exp(-z_1))^2}\dot{z_1}$$
$$\dot{\alpha_2} = \frac{\partial \alpha_2}{\partial z_1}\dot{z_1} + \frac{\partial \alpha_2}{\partial z_2}\dot{z_2} = \frac{\exp(-z_2)}{(1+\exp(-z_2))^2}\dot{z_2}$$
$$\dot{y} = \frac{\partial y}{\partial \alpha_1}\dot{\alpha_1} + \frac{\partial y}{\partial \alpha_2}\dot{\alpha_2} = W_1^{(1)}\dot{\alpha_1} + W_2^{(1)}\dot{\alpha_2}$$
$$\frac{\partial f}{\partial x_1} = W_1^{(1)}\frac{\exp(-z_1)}{(1+\exp(-z_1))^2}W_{11}^{(0)} + W_2^{(1)}\frac{\exp(-z_2)}{(1+\exp(-z_2))^2}W_{21}^{(0)}$$

Seed for $\frac{\partial f}{\partial x_2}$:

$$\dot{x_2} = 0 \qquad \dot{x_2} = 1$$
$$\dot{z_1} = \frac{\partial z_1}{\partial x_1}\dot{x_1} + \frac{\partial z_1}{\partial x_2}\dot{x_2} = W_{12}^{(0)}\dot{x_2}$$
$$\dot{z_2} = \frac{\partial z_2}{\partial x_1}\dot{x_1} + \frac{\partial z_2}{\partial x_2}\dot{x_2} = W_{22}^{(0)}\dot{x_2}$$
$$\dot{\alpha_1} = \frac{\partial \alpha_1}{\partial z_1}\dot{z_1} + \frac{\partial \alpha_1}{\partial z_2}\dot{z_2} = \frac{\exp(-z_1)}{(1+\exp(-z_1))^2}\dot{z_1}$$
$$\dot{\alpha_2} = \frac{\partial \alpha_2}{\partial z_1}\dot{z_1} + \frac{\partial \alpha_2}{\partial z_2}\dot{z_2} + \frac{\partial \alpha_2}{\partial \alpha_1}\dot{\alpha_1} = \frac{\exp(-z_2)}{(1+\exp(-z_2))^2}\dot{z_2}$$
$$\dot{y} = \frac{\partial y}{\partial \alpha_1}\dot{\alpha_1} + \frac{\partial y}{\partial \alpha_2}\dot{\alpha_2} = W_1^{(1)}\dot{\alpha_1} + W_2^{(1)}\dot{\alpha_2}$$
$$\frac{\partial f}{\partial x_2} = W_1^{(1)}\frac{\exp(-z_1)}{(1+\exp(-z_1))^2}W_{12}^{(0)} + W_2^{(1)}\frac{\exp(-z_2)}{(1+\exp(-z_2))^2}W_{22}^{(0)}$$

3. Compute the gradient using backward mode.
One time forward sweep. Store each result in memory and build calculation tree for the derivatives:

$$z_1 = w_{11}^{(0)}x1 + w_{12}^{(0)}x2$$
$$z_2 = w_{21}^{(0)}x1 + w_{22}^{(0)}x2$$
$$a_1 = \sigma(z_1)$$
$$a_2 = \sigma(z_2)$$
$$y = w_1^{(1)}a_1 + w_2^{(1)}a_2$$

Backward pass, using stored values, the nn and the respective derivatives of the intermediate steps:

$$
\begin{aligned}
\bar{y} &= 1 \\
\bar{a_1} &= \frac{\mathrm{d}y}{\mathrm{d}a_1}\bar{y} = W_1^{(1)} \\
\bar{a_2} &= \frac{\mathrm{d}y}{\mathrm{d}a_2}\bar{y} = W_2^{(1)} \\
\bar{z_1} &= \frac{\mathrm{d}a_1}{\mathrm{d}z_1}\bar{a_1} = \frac{\exp(-z_1)}{(1+\exp(-z_1))^2}W_1^{(1)} \\
\bar{z_2} &= \frac{\mathrm{d}a_2}{\mathrm{d}z_2}\bar{a_2} = \frac{\exp(-z_2)}{(1+\exp(-z_2))^2}W_2^{(1)} \\
\frac{\mathrm{d}f}{\mathrm{d}x_1} &= \frac{\mathrm{d}z_1}{\mathrm{d}x_1}\bar{z_1} + \frac{\mathrm{d}z_2}{\mathrm{d}x_1}\bar{z_2} = \\
&= W_{11}^{(0)}\frac{\exp(-z_1)}{(1+\exp(-z_1))^2}W_1^{(1)} + W_{21}^{(0)}\frac{\exp(-z_2)}{(1+\exp(-z_2))^2}W_2^{(1)} \\
\frac{\mathrm{d}f}{\mathrm{d}x_2} &= \frac{\mathrm{d}z_1}{\mathrm{d}x_2}\bar{z_1} + \frac{\mathrm{d}z_2}{\mathrm{d}x_2}\bar{z_2} = \\
&= W_{12}^{(0)}\frac{\exp(-z_1)}{(1+\exp(-z_1))^2}W_1^{(1)} + W_{22}^{(0)}\frac{\exp(-z_2)}{(1+\exp(-z_2))^2}W_2^{(1)}
\end{aligned}
$$

4. Consider a general function $f : \mathbb{R}^n \to \mathbb{R}^m$
   - $m << n \to$ we'd use backward mode since we only need one forward sweep and m backward sweeps to calculate the derivatives
   - $m >> n \to$ we'd use forward mode since we only need n forward sweeps to calculate the derivatives.
5. In python script.
6. We've implemented both **automatic** differentiation forward and backward mode, the formula we've calculated and used the approximation function. Results:
   - Forward mode gradient:
     $\nabla_x y = (0.2498698795475206, -0.08752602409049587)^T$
   - Backward mode gradient:
     $\nabla_x y = (0.2498698795475206, -0.08752602409049587)^T$
   - Calculated gradient:
     $\nabla_x y = (0.2498698795475206, -0.08752602409049587)^T$
   - Approximated gradient:
     $\nabla_x y = (0.2498696598276934, -0.08752598893396271)^T$

   As we can see there are numerical issues with the approximated gradient.