NNDL - 2025

Exercise 1

Prob. 1

c) Let $x \in \mathbb{R}$. Then, if $x \geq 0$, $\psi(x) = x = \max(\alpha x, x)$, because $\alpha x \leq x$ for $\alpha \in [0,1]$. Similarly, if $x < 0$, then $\alpha x \geq x$ and $\psi(x) = \alpha x = \max(\alpha x, x)$. Thus,

$$\psi(x) = \begin{cases} x & , x \geq 0 \\ \alpha x & , x < 0 \end{cases} = \max(\alpha x, x).$$

$\blacksquare$

a) Choose $\alpha = 0$, then

$$\psi(x) = \max(x, \alpha x) = \max(x, 0),$$

which shows that ReLu can be approximated by leaky ReLu, making it a special case.

b) Choose $\alpha := 1$. Then

$$\psi(x) = \max\{x, \alpha x\}$$

$$= \max\{x, x\} = x,$$

for every $x \in \mathbb{R}$

$\blacksquare$

Prob. 2

a) If $W_1 \in \mathbb{R}^{m \times n}$, i.e. first layer has n neurons, then $W_2 \in \mathbb{R}^{n \times k}$ for some $k \in \mathbb{N}$. Thus, W_2 has n rows.

b) If NN is invertible (i.e. the product is injective), then $m \geq n$ meaning that $\dim R(A_{n+1}) \geq \dim R(A_n)$.

Assume $m < n$. Then there exists $x \neq 0$ s.t. $W_1 x = 0$, which implies

$$W_k \dots W_1 x = 0 \quad \text{by linearity.}$$

Which would be a contradiction with injectivity since $W_k \dots W_1 0 = 0$.

$\square$

c) If all matrices are square then, the NN is invertible iff all matrices are invertible.

Optimization:

Prob 1:

Partial derivatives are given as

$$\partial_i f_1(w) = \partial_i \|w\|^2 = \partial_i \sum w_k^2$$
$$= \sum \partial_i w_k^2 = 2 w_i .$$

Thus, the gradient is

$$\nabla f_1(w) = \begin{bmatrix} \partial_1 \\ \vdots \\ \partial_n \end{bmatrix} = 2 \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

Prob 2

Hessian is given as

$$H_f = (D \nabla f)^T = \begin{bmatrix} D \nabla f_1 \\ \vdots \\ D \nabla f_2 \end{bmatrix}^T ,$$

where D is the differentiation operator.

Since $(D \nabla f_i)_j = \partial_j \partial_i f = \partial_j 2 w_i = \begin{cases} 0, & j \neq i \\ 2, & j = i \end{cases}$

we set

$$H_f = 2 I_{n \times n} \quad \text{(identity matrix)}.$$

$\square$

Prob 3.

The Newton's method is given by the recursion

$$w_{k+1} = w_k - H_f(w_k)^{-1} \nabla f(w_k)$$

which in our case is

$$w_{k+1} = w_k - w_k$$

always converging in one step.

Prob 4

The gradient is given as

$$\nabla f_0(w) = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = z$$

Prob 5

Using chain rule we obtain gradient at point w as

$$\nabla f_2(w) = \left(Dg(w^T z)\, D(w^T z) \right)^T$$

$$= g'(w^T z)\, \nabla f_0(w)$$

$$= g'(w^T z)\, z$$

☒

Prob 6

Stochastic gradient is given as

$$\nabla f_3(w) = \nabla \mathbb{E}(g(w^T z))$$
$$= \mathbb{E}(\nabla g(w^T z)) \quad \text{// By definition}$$
$$= \mathbb{E}(g'(w^T z)\, z)$$

$\boxed{3}$

Prob 7

We note partial derivative of the quadratic form at point w is given as

$$\partial_i = \frac{1}{h}\left((w+he_i)^T A (w+he_i) - w^T A w\right)$$
$$= \frac{1}{h}\left(w^T A h e_i + (he_i)^T A w\right)$$
$$= \frac{1}{h}\left(w^T (A + A^T) e_i h\right)$$
$$= 2 w^T A e_i = 2 w^T A_{\cdot i}$$

And by definition the gradient is

$$\nabla \tfrac{1}{2} w^T A w = w^T A$$

$\boxed{2}$

Prob 7

b) First note that $\|w\|^4 = (\|w\|^2)^2$. Then, by chain rule and solution 1., we get

$$D\|w\|^4 = D(\|w\|^2)^2 D\|w\|^2$$

$$= 2\|w\|^2 (2\nabla\|w\|^2)^T$$

Transposing to get the gradient we have

$$\nabla \tfrac{1}{4}\|w\|^4 = \|w\|^2 w$$

∎