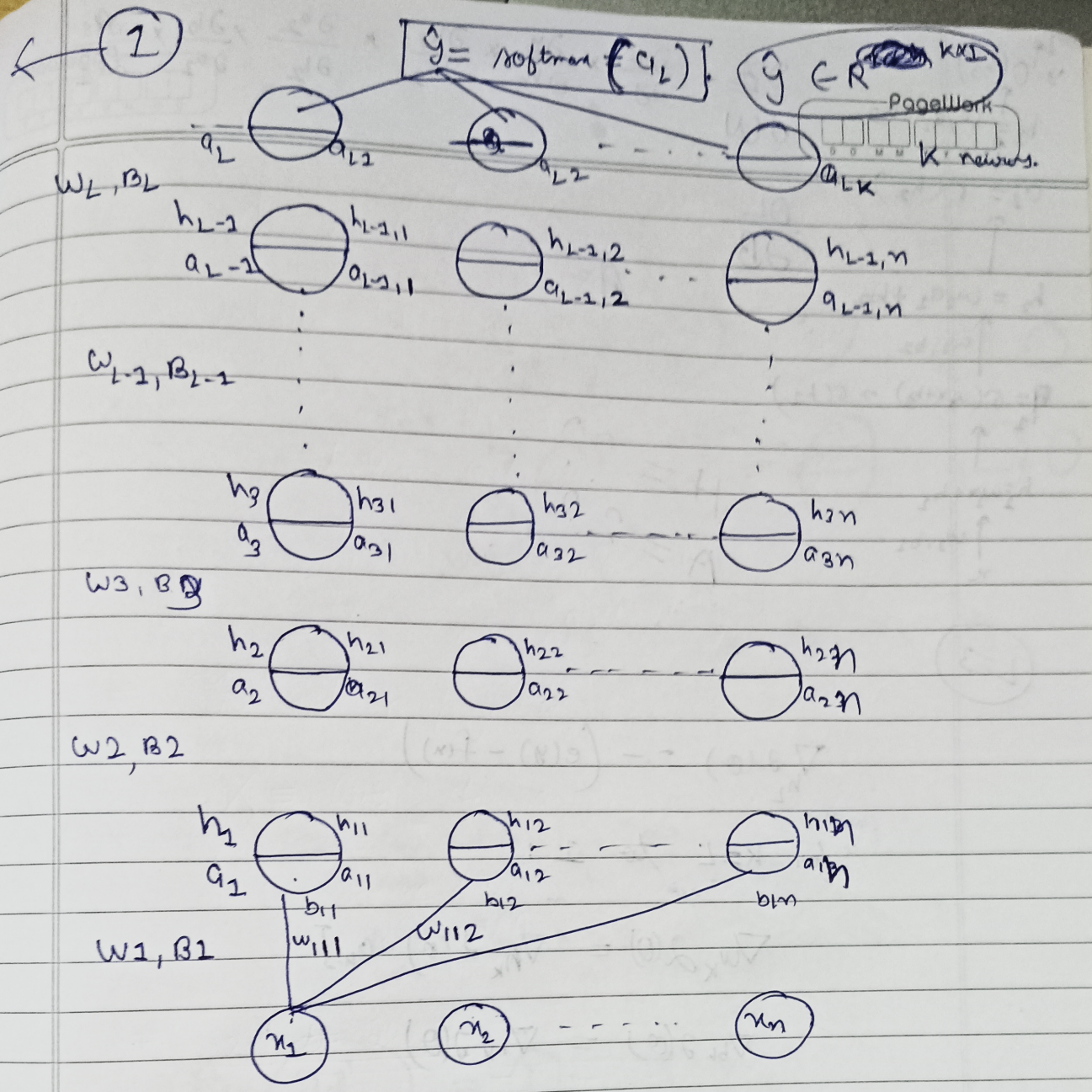

W_1 to $W_{L-1} \equiv$ $(n \times n)$ matrices having weights

$W_L = n \times K$ matrix

B_1 to $B_{L-1} \equiv (1 \times n)$ vector having biases.

$B_L \equiv 1 \times K$ vector.

$S \times 1$

$S \times 1$

$S \times 1$

$n \times K \cdot n \times 1$

$K \times n \cdot n \times 1 \; K \times 1 +$

$$W_1 = \begin{bmatrix} w_{111} & w_{112} & w_{113} & \cdots & w_{11n} \\ w_{121} & w_{122} & w_{123} & \cdots & w_{12n} \\ w_{131} & w_{132} & w_{133} & \cdots & w_{13n} \\ \vdots & \vdots & \vdots & & \vdots \\ w_{1n1} & w_{1n2} & w_{1n3} & \cdots & w_{1nn} \end{bmatrix}_{n \times n \equiv R^{n \times n}}$$

$$B_1 = \begin{bmatrix} b_{11} & b_{12} & b_{13} & \cdots & b_{1n} \end{bmatrix}_{1 \times n \equiv R^n}$$

$$x \in R^n \equiv [x_1, x_2 \cdots x_n]_{1 \times n}$$

$$\boxed{a_1 = W_1 \cdot x^T + B_1^T}$$

We need to have,

a_1 same size as x

$$\underset{R}{n \times n} \cdot \underset{R}{n \times 1} \quad + \underset{R}{n \times 1} \quad = \underset{R}{n \times 1}$$

$$\Rightarrow a_1 \in R^{n \times 1}$$

$$h_1 = \text{sigmoid}(a_1) \text{ (or) } \text{act\_f}^n(a_1)$$

$$h_1 \in R^{n \times 1} \equiv \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \end{bmatrix}^T$$

$$a_2 = W_2 \cdot h_1 + B_2^T$$

$$h_2 = \text{act\_f}^n(a_2).$$

Forward

(#) Weights & Bias Initialisation

(*) Initialise $W_1, W_2 \dots W_{L-1} \in R^{n \times n}$

$W_L \in R^{n \times k}$

$B_1, B_2, \dots B_{L-1} \in R^{1 \times n}$

$B_L \in R^{1 \times k}$

$x \in R^n$.

(#) Forward propagation

$X \equiv$ list of m data points.

for x in X:

$x \in R^{1 \times n}$

(#.) Input layer

$$a_1 = W_1 \cdot x^T + B_1^T$$

$$h_1 = act\text{-}f^n(a_1)$$

(#) Middle layers

$k = 2$ to $k = L-1$

$$a_k = W_k \cdot h_{k-1} + B_k^T$$

$$h_k = act\text{-}f^n(a_k)$$

(#) Output layer

$$a_L = W_L^T \cdot h_{L-1} + B_L^T$$

$$\hat{y} = Output\text{-}f^n(a_L)$$

ROUGH.

W → mine → 1 2 3 ... L-1 L
list comprehension → 0 1 2 ... L-2 L-1

X → m points.

Y → m outputs for m points

A1 → m A1 values for m points.
⋮
AL-1

$A_L = k\ A_1$ ....

$H_1 \rightarrow H_{L-1} \rightarrow$ m pnA.

W → mine → 1 2 3 ... L-1 L
list comprehension → 0 1 2 ... L-2 L-1

A → mine → A1 A2 A3 ... AL-1 AL
list comp. → A[0] A[1] A[2] ... A[L-2] A[L-1]

H → mine → H1 H2 H3 ... HL-1
list comprehension → H[0] H[1] H[2] ... H[L-2]

(#) Backprop Theory (# Mathematically Derived).

$\nabla_{a_L} \mathcal{L}(\theta) = -(e(y) - f(x)) = \hat{y} - y$ (for softmax-fn)

For $k = L$ to $k = 2$:

$\nabla_{W_k} \mathcal{L}(\theta) = \nabla_{a_k} \mathcal{L}(\theta) \cdot h_{k-1}^T$

$\nabla_{b_k} \mathcal{L}(\theta) = \nabla_{a_k} \mathcal{L}(\theta)$

$\nabla_{h_{k-1}} \mathcal{L}(\theta) = W_k^T (\nabla_{a_k} \mathcal{L}(\theta))$

$\nabla_{a_{k-1}} \mathcal{L}(\theta) = \nabla_{h_{k-1}} \mathcal{L}(\theta) \odot (\dots, g'(a_{k-1,j}), \dots)$

↳ element wise product of vectors.

$\nabla_{W_1} \mathcal{L}(\theta) = x^T$

$\nabla_{b_1} \mathcal{L}(\theta) = 1$

Rough

$a_1 = W_1 \cdot x^T + B_1^T$

$\frac{\partial a_1}{\partial W_1} = x^T$

$\frac{\partial a_1}{\partial B_1} = 1$

(#) $g = act\_f^n$

$h_i = act\_f^n(a_i)$

(#) if g = sigmoid

$g' = g(a_{ij})(1 - g(a_{ij}))$

$h = g(a_{ij})$

(#) Pseudo Code

```
for i in range (max_iter):
    dW = [[0.0]]
    dB = [[0.0]]
    (# forward-propagate (x).
    for x, y, y_hat, a, h in zip(X, Y, Y_hat, A, H):
        (#) back propagate.
    (#) update weights
    (#) update biases
```

S,4 4,1 S,1

4,S 4,1 S,1

S,1 H,1 dA.T 2,4