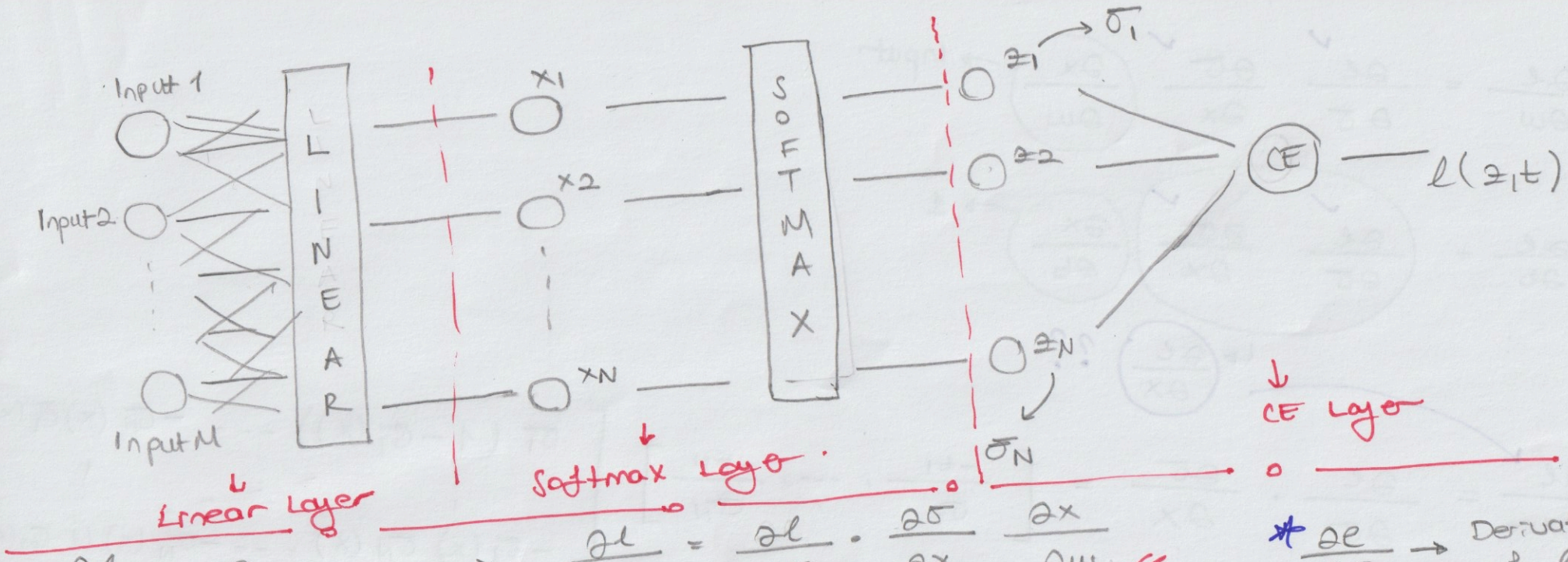

$$\frac{\partial l}{\partial w} = ? \longrightarrow \frac{\partial l}{\partial w} = \frac{\partial l}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial x} \frac{\partial x}{\partial w}$$

$$\frac{\partial l}{\partial b} = ? \longrightarrow \frac{\partial l}{\partial b} = \frac{\partial l}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial x} \frac{\partial x}{\partial b}$$

* $\frac{\partial l}{\partial \sigma}$ → Derivative of l w.r.t softmax output

* $\frac{\partial \sigma}{\partial x}$ → Derivative of softmax w.r.t its inputs (x_i).

* Let's stat with computing $\frac{\partial l}{\partial \sigma}$

$$\frac{\partial l}{\partial z} = \left[\frac{\partial l}{\partial z_1}, \dots, \frac{\partial l}{\partial z_N} \right]$$

$$= \left[\frac{-t_1}{\partial z_1}, \dots, \frac{-t_N}{\partial z_N} \right]$$

* Then compute $\frac{\partial \sigma}{\partial x} =$

$$\begin{bmatrix} \sigma_1(1-\sigma_1(x)) & \cdots & -\sigma_N(x)\,\sigma_1(x) \\ \vdots & & \vdots \\ -\sigma_1(x)\,\sigma_N(x) & \cdots & -\sigma_N(x)(1-\sigma_N(x)) \end{bmatrix}$$

* $\frac{\partial \ell}{\partial w} = \frac{\partial \ell}{\partial \sigma} \checkmark \quad \frac{\partial \sigma}{\partial x} \checkmark \quad \frac{\partial x}{\partial w}$ → Input

* $\frac{\partial \ell}{\partial b} = \frac{\partial \ell}{\partial \sigma} \checkmark \quad \frac{\partial \sigma}{\partial x} \checkmark \quad \frac{\partial x}{\partial b}$ → 1

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

* $\frac{\partial \ell}{\partial x} = \frac{\partial \ell}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial x} = \left[\frac{-t_1}{\sigma_1}, \dots, \frac{-t_N}{\sigma_N} \right] \begin{bmatrix} \sigma_1 (1 - \sigma_1(x)) & \cdots & -\sigma_N(x)\sigma_1(x) \\ \vdots & & \vdots \\ -\sigma_1(x)\sigma_N(x) & \cdots & -\sigma_N(x)(1-\sigma_N(x)) \end{bmatrix}$

$1 \times N$ $\quad N \times N$

$$= \sigma_i(x)\left(v_{i^\circ} - v\sigma(x)^T\right)$$

* Let's assume that we have two Linear layers and (W_1, b_1) is parameters of Linear 1 and (W_2, b_2) is parameters of Linear 2. (So Input → Linear 1 → Linear 2 → Softmax → Loss)

$$\frac{\partial \ell}{\partial w_2} = \frac{\partial \ell}{\partial \sigma} \frac{\partial \sigma}{\partial x} \frac{\partial x}{\partial w_2}$$

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial \sigma} \frac{\partial \sigma}{\partial x} \frac{\partial x}{\partial \text{Input2}} \frac{\partial \text{Input2}}{\partial w_1}$$