6.(a) $$\frac{\partial y_L}{\partial b_L} = \frac{\partial}{\partial b_L}(A_L y_{L-1} + b_L) = \boxed{1}$$

$$\frac{\partial y_L}{\partial y_{L-1}} = \frac{\partial}{\partial y_{L-1}}(A_L y_{L-1} + b_L) = \boxed{A_L}$$

For $\ell = 1, \dots, L-1$,

$$\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell} = \begin{pmatrix} \frac{\partial y_{\ell 1}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell 1}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell 1}}{\partial b_{\ell n_\ell}} \\ \frac{\partial y_{\ell 2}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell 2}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell 2}}{\partial b_{\ell n_\ell}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell 1}} & \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell 2}} & \cdots & \frac{\partial y_{\ell n_\ell}}{\partial b_{\ell n_\ell}} \end{pmatrix}$$

Define $\underline{\tilde{y}}_\ell = A_\ell \underline{y}_{\ell-1} + \underline{b}_\ell$ (i.e. $\underline{y}_\ell$ pre-activation)

$$\frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}} = \frac{\partial}{\partial b_{\ell j}}\left([A_\ell \underline{y}_{\ell-1} + \underline{b}_\ell]_i\right) = \frac{\partial}{\partial b_{\ell j}}\left([A_\ell \underline{y}_{\ell-1}]_i + b_{\ell i}\right) = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

by chain rule and elementwise application

$$\frac{\partial y_{\ell i}}{\partial b_{\ell j}} = \frac{\partial}{\partial b_{\ell j}}\left([\underline{\sigma}(\underline{\tilde{y}}_\ell)]_i\right) = \frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}}[\underline{\sigma}'(\underline{\tilde{y}}_\ell)]_i = \frac{\partial \tilde{y}_{\ell i}}{\partial b_{\ell j}} \sigma'(\tilde{y}_{\ell i}) = \begin{cases} 0 & \text{if } i \neq j \\ \sigma'(\tilde{y}_{\ell i}) & \text{if } i = j \end{cases} \text{ by above}$$

Hence, only diagonal entries of $\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell}$ remain are non-zero $(i=j)$

with $\frac{\partial y_{\ell i}}{\partial b_{\ell i}} = \sigma'(\tilde{y}_{\ell i}) = \sigma'([A_\ell y_{\ell-1} + b_\ell]_i)$

That is, $\boxed{\frac{\partial \underline{y}_\ell}{\partial \underline{b}_\ell} = \mathrm{diag}(\underline{\sigma}'(A_\ell \underline{y}_{\ell-1} + \underline{b}_\ell))}$ ∎

For $\ell = 2, \dots, L-1$,

For $\frac{\partial \underline{y}_\ell}{\partial \underline{y}_{\ell-1}}$, consider entries once more:

A_ℓ $\quad y_{\ell-1}$

i

$$\frac{\partial \tilde{y}_{\ell i}}{\partial (y_{\ell-1})_j} = \frac{\partial}{\partial (y_{\ell-1})_j}\left([A_\ell \underline{y}_{\ell-1}]_i + b_{\ell i}\right) = \frac{\partial}{\partial (y_{\ell-1})_j}\left(\sum_{k=1}^{n_{\ell-1}} (A_\ell)_{ik}(y_{\ell-1})_k + b_{\ell i}\right) = (A_\ell)_{ij}$$

So
$$\frac{\partial (y_\ell)_i}{\partial (y_{\ell-1})_j} = \frac{\partial}{\partial (y_{\ell-1})_j}\left(\sigma((\tilde{y}_\ell)_i)\right) = \frac{\partial (\tilde{y}_\ell)_i}{\partial (y_{\ell-1})_j}\sigma'(\tilde{y}_{\ell i}) = (A_\ell)_{ij}\sigma'((\tilde{y}_\ell)_i) = \sigma'((\tilde{y}_\ell)_i)(A_\ell)_{ij} = [\underline{\sigma}'(A_\ell \underline{y}_{\ell-1} + \underline{b}_\ell)]_i (A_\ell)_{ij}$$

$\updownarrow$ elements equivalent

$$[\mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + b_\ell))A_\ell]_{ij} = \sum_{k=1}^{n_\ell} \underbrace{[\mathrm{diag}(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell))]_{ik}}_{\neq 0 \text{ iff } k=i} (A_\ell)_{kj} = [\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell)]_i (A_\ell)_{ij}$$

$i \quad j$

diag $\in \mathbb{R}^{n_\ell \times n_\ell}$ $\quad A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$

$$\Rightarrow \boxed{\frac{\partial \underline{y}_\ell}{\partial \underline{y}_{\ell-1}} = \mathrm{diag}(\underline{\sigma}'(A_\ell \underline{y}_{\ell-1} + \underline{b}_\ell))A_\ell}$$

as given ∎

6.(b) Consider again, $y_\ell = \underline{\sigma}(A_\ell y_{\ell-1} + \underline{b}_\ell)$

$$= \sum_{i=1}^{n_\ell}\Big(\sigma\big([A_\ell y_{\ell-1} + \underline{b}_\ell]_i\big)\Big)\cdot \underline{e}_i$$

$$= \sum_{i=1}^{n_\ell}\Big(\sigma\Big(\sum_{j=1}^{n_{\ell-1}} (A_\ell)_{ij}(y_{\ell-1})_j + (\underline{b}_\ell)_i\Big)\Big)\cdot \underline{e}_i$$

$$\Rightarrow \frac{\partial y_\ell}{\partial (A_\ell)_{ij}} = \sigma'\big([A_\ell y_{\ell-1} + \underline{b}_\ell]_i\big)\cdot (y_{\ell-1})_j \cdot \underline{e}_i$$

$$\left(\frac{\partial y_L}{\partial A_\ell}\right)_{ij} = \frac{\partial y_L}{\partial (A_\ell)_{ij}} = \frac{\partial y_L}{\partial y_\ell}\frac{\partial y_\ell}{\partial (A_\ell)_{ij}} \quad \text{by Chain Rule}$$

$$= \frac{\partial y_L}{\partial y_\ell}\,\sigma'\big([A_\ell y_{\ell-1} + \underline{b}_\ell]_i\big)(y_{\ell-1})_j\, \underline{e}_i$$

$$= \sigma'\big([A_\ell y_{\ell-1} + \underline{b}_\ell]_i\big)\left(\frac{\partial y_L}{\partial y_\ell}\right)^T_i (y_{\ell-1})_j$$

$$\Rightarrow \frac{\partial y_L}{\partial A_\ell} = \underbrace{\mathrm{diag}\big(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell)\big)}_{\in \mathbb{R}^{n_\ell \times n_\ell}}\underbrace{\left(\frac{\partial y_L}{\partial y_\ell}\right)^T}_{\in \mathbb{R}^{n_\ell \times 1}}\underbrace{(y_{\ell-1})^T}_{\in \mathbb{R}^{1 \times n_\ell}}$$

$$\left[\left(\frac{\partial y_L}{\partial A_\ell}\right)_{ij} = \sum_{k=1}^{n_\ell}\left(\mathrm{diag}\big(\underline{\sigma}'(A_\ell y_{\ell-1} + \underline{b}_\ell)\big)_{ik}\left(\frac{\partial y_L}{\partial y_\ell}\right)^T_k\right)\cdot (y_{\ell-1})^T_j\right.$$

$$\left. = [\sigma'(A_\ell y_{\ell-1} + \underline{b}_\ell)]_i\left(\frac{\partial y_L}{\partial y_\ell}\right)^T_i (y_{\ell-1})^T_j \quad \text{so matches above} \checkmark \right]$$