0.1 Higher dimensional Derivative

a) $y_i = \sum_j x_j W_{j,i}$ implies

$$\frac{\partial y_i}{\partial W_{l,k}} = \begin{cases} 0, \text{when } k \neq i \\ x_l, \text{when } k = i \end{cases} = x_l$$

Since index of y coincides with the column index, the gradient is a row vector, while the partial derivative is a column vector. In other words, $\nabla_W(y) = x = \frac{\partial y}{\partial W}^\intercal$. Given $f(y) \in \mathbb{R}$, $\frac{\partial f}{\partial W} = \frac{\partial f}{\partial y} x^\intercal$.
b)$y_i = \sum_j x_j W^\intercal_{j,i} = \sum_j x_j W_{i,j}$ implies

$$\frac{\partial y_i}{\partial W_{l,k}} = \begin{cases} 0, \text{when } l \neq i \\ x_k, \text{when } l = i \end{cases} = x_k$$

Since index of y coincides with the row index, the gradient is a column vector, while the partial derivative is a column vector. In other words, $\nabla_W(y) = x^\intercal = \frac{\partial y}{\partial W}^\intercal$. Given $f(y) \in \mathbb{R}$, $\frac{\partial f}{\partial W} = \frac{\partial f}{\partial W^\intercal}^\intercal = x \frac{\partial f}{\partial y}^\intercal$.
c)$y_i = \sum_j x_j W^{-1}_{j,i} = \sum_j x_j det(W) adj(W)_{j,i}$ implies

$$\frac{\partial W^{-1}_{i,j}}{\partial W_{k,l}} \{$$

d)Let $W_1, W_2 \in \mathbb{R}^{n \times n}$. Then $y_i = \sum_j o_j (W_2^{-1})_{j,i} = \sum_j tanh(xW_1W_2)_j (W_2^{-1})_{j,i}$ implies

$$\begin{aligned} \frac{\partial y_i}{\partial x_k} &= \sum_j \frac{\partial y_i}{\partial o_j} \frac{\partial o_j}{\partial x_k} \\ &= \sum_j (W_2^{-1})_{j,i} (1 - tanh(xW_1W_2)_j^2)(W_2)_{k,j} \end{aligned}$$

In other words, $\frac{\partial y}{\partial x} = W_2^{-1} D W_1 W_2, D = diag[I - tanh(xW_1)^2]$.