Backpropagation in Neural Networks

Nicholas Richardson

October 2021

1 Notation Preliminaries

For clarity when discussing scalars, vectors, and matricies, scalar functions and elements will be kept unbolded where as vectors and matricies will be bolded. Vectors will use lower-cased symbols and matricies will use upper-cased symbols. If we wish to identify the i^{th} element of a (column or row) vector $\boldsymbol{x}$, we will subscript the unbolded letter x_i or subscript parenthesis $(\boldsymbol{x})_i$. Similarly, we will use double subscripts X_{ij}, x_{ij} or subscripted parenthesis $(\boldsymbol{X})_{ij}$ to denote the element in the i^{th} row and j^{th} column of a matrix $\boldsymbol{X}$. Finally, we will use non-italics bolded symbols for higher dimentional tensors like so $\mathbf{X}$.

There are two common conventions for writing derivatives involving vectors and matrices. For consistency, we will use the "numerator style" or "Jacobian formulation". This involves keeping the dimension of the derived function, and transposing the dimensions of the element the derivative is being taken with respect to. Specifically,

1. If $\boldsymbol{y} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector valued function with a vector input, the partial derivative with respect to the vector input is the $m \times n$ Jacobian matrix

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}.$$

This has the convenient shorthand notation $\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)_{ij} = \frac{\partial y_i}{\partial x_j}$.

In particular,

(a) If $m = 1$, $y : \mathbb{R}^n \to \mathbb{R}$. The derivative with respect to the vector input $\boldsymbol{x}$ would be the transpose of the (column vector) gradient of y:

$$\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \dots & \frac{\partial y}{\partial x_n} \end{bmatrix} = (\nabla y)^\top .$$

(b) If $n = 1$, $\boldsymbol{y} : \mathbb{R} \to \mathbb{R}^m$. The derivative of the function with respect to its scalar input x will remain a column vector:

$$\frac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix} .$$

(c) If $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, we have $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{A}$.

2. We denote the derivative of a matrix valued function $\boldsymbol{Y} : \mathbb{R} \to \mathbb{R}^{m \times n}$ with respect to its scalar input x by the $m \times n$ matrix

$$\frac{\partial \boldsymbol{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \dots & \frac{\partial y_{1n}}{\partial x} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \dots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}, \quad \left(\frac{\partial \boldsymbol{Y}}{\partial x}\right)_{ij} = \frac{\partial y_{ij}}{\partial x}.$$

Note that if $n = 1$, $\boldsymbol{Y} : \mathbb{R} \to \mathbb{R}^m$, this notation agrees with the notation in (1b).

3. We denote the derivative of a scalar function $y : \mathbb{R}^{n \times m} \to \mathbb{R}$ with respect to its $n \times m$ matrix input $\boldsymbol{X}$ by the $m \times n$ matrix

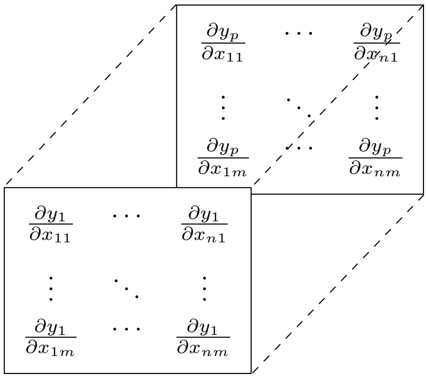
$$\frac{\partial y}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \dots & \frac{\partial y}{\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1m}} & \dots & \frac{\partial y}{\partial x_{nm}} \end{bmatrix}, \quad \left(\frac{\partial y}{\partial \boldsymbol{X}}\right)_{ij} = \frac{\partial y}{\partial x_{ji}}.$$

Note that if $m = 1$, $y : \mathbb{R}^n \to \mathbb{R}$, this notation agrees with the notation in (1a).

4. Let $\boldsymbol{y} : \mathbb{R}^{n \times m} \to \mathbb{R}^p$ be a vector valued function with a matrix input $\boldsymbol{X}$. The derivative with respect

to its $n \times m$ matrix input is the $p \times m \times n$ tensor

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}} = \begin{bmatrix} \left[\frac{\partial y_1}{\partial \boldsymbol{X}}\right] \\ \vdots \\ \left[\frac{\partial y_p}{\partial \boldsymbol{X}}\right] \end{bmatrix} =$$



$$\begin{matrix} \frac{\partial y_1}{\partial x_{11}} & \cdots & \frac{\partial y_1}{\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_{1m}} & \cdots & \frac{\partial y_1}{\partial x_{nm}} \end{matrix} \qquad \begin{matrix} \frac{\partial y_p}{\partial x_{11}} & \cdots & \frac{\partial y_p}{\partial x_{n1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_p}{\partial x_{1m}} & \cdots & \frac{\partial y_p}{\partial x_{nm}} \end{matrix}$$

In our shorthand component notation, we may write this as $\left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}}\right)_{ijk} = \frac{\partial y_i}{\partial x_{kj}}$. In particular, if $p = 1$, $y : \mathbb{R}^{n\times m} \to \mathbb{R}$, and the partial derivative is of dimension $1 \times m \times n = ((1)(m)) \times n = m \times n$, which is identical with the one we denote in item (3) of the list. If $m = 1$, $\boldsymbol{y} : \mathbb{R}^n \to \mathbb{R}^p$, the partial derivative is of dimension $p \times 1 \times n = ((p)(1)) \times n = p \times n$. Here we use the following convention when we go from 3D to 2D when the first or the second entry is one:

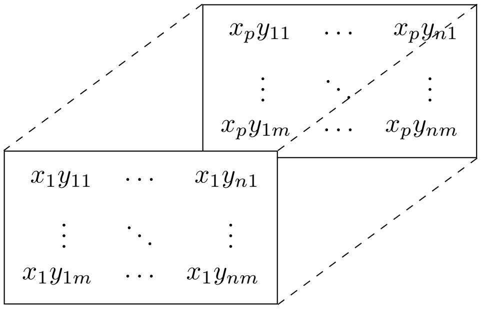
$$1 \times m \times n = (1m) \times n = m \times n, \quad m \times 1 \times n = (m1) \times n = m \times n.$$

5. Recall matrix product: If $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{x}\boldsymbol{y}^\top \in \mathbb{R}^{n\times m}$:

$$\boldsymbol{x}\boldsymbol{y}^\top = \begin{bmatrix} x_1y_1 & \cdots & x_1y_m \\ \vdots & \ddots & \vdots \\ x_ny_1 & \cdots & x_ny_m \end{bmatrix}, \quad \left(\boldsymbol{x}\boldsymbol{y}^\top\right)_{ij} = x_iy_j.$$

6. Similarly, we can denote the tensor product of a vector $\boldsymbol{x} \in \mathbb{R}^p$ with a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ as

$$\boldsymbol{x} \otimes \boldsymbol{Y} = \begin{bmatrix} \left[x_1\boldsymbol{Y}^T\right] \\ \vdots \\ \left[x_p\boldsymbol{Y}^T\right] \end{bmatrix} =$$



$$\begin{matrix} x_1y_{11} & \cdots & x_1y_{n1} \\ \vdots & \ddots & \vdots \\ x_1y_{1m} & \cdots & x_1y_{nm} \end{matrix} \qquad \begin{matrix} x_py_{11} & \cdots & x_py_{n1} \\ \vdots & \ddots & \vdots \\ x_py_{1m} & \cdots & x_py_{nm} \end{matrix} \in \mathbb{R}^{p\times m\times n},$$

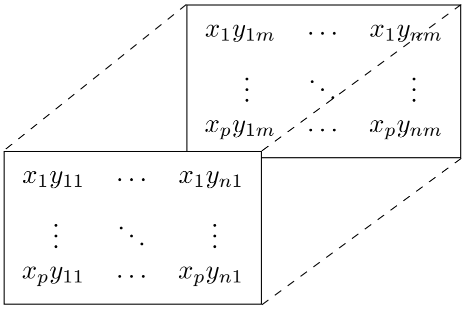
or the shorthand notation $(\boldsymbol{x}\otimes\boldsymbol{Y})_{ijk} = x_i y_{kj}$. Note that if $m = 1$, $\boldsymbol{Y} = \boldsymbol{y}$ (a column vector), and

$$x \otimes \boldsymbol{y} = \begin{bmatrix} [x_1 \boldsymbol{y}^T] \\ \vdots \\ [x_p \boldsymbol{y}^T] \end{bmatrix} \in \mathbb{R}^{p\times 1\times n} = \mathbb{R}^{p\times n},$$

which can be regarded as $\boldsymbol{x}\boldsymbol{y}^T \in \mathbb{R}^{p\times n}$.

7. The tensor product of a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$ with a vector $\boldsymbol{x} \in \mathbb{R}^p$ is given by the $m \times p \times n$ tensor

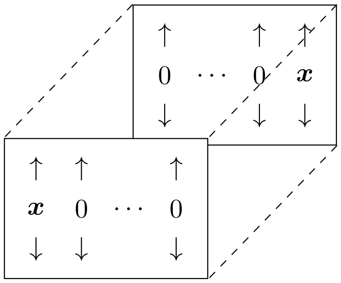
$\boldsymbol{Y} \otimes \boldsymbol{x} =$


x_1y_{1m} $\cdots$ x_1y_{nm}
$\vdots$ $\ddots$ $\vdots$
x_py_{1m} $\cdots$ x_py_{nm}
x_1y_{11} $\cdots$ x_1y_{n1}
$\vdots$ $\ddots$ $\vdots$
x_py_{11} $\cdots$ x_py_{n1}

$\in \mathbb{R}^{m\times p\times n}, \qquad (\boldsymbol{Y}\otimes\boldsymbol{x})_{ijk} = x_j y_{ki}.$

In particular, if $\boldsymbol{Y} = \boldsymbol{I}_m \in \mathbb{R}^{m\times m}$, the identity matrix, we have

$\boldsymbol{I}_m \otimes \boldsymbol{x} =$


$\uparrow$ $\uparrow$ $\uparrow$
0 $\cdots$ 0 $\boldsymbol{x}$
$\downarrow$ $\downarrow$ $\downarrow$
$\uparrow$ $\uparrow$ $\uparrow$
$\boldsymbol{x}$ 0 $\cdots$ 0
$\downarrow$ $\downarrow$ $\downarrow$

$\in \mathbb{R}^{m\times p\times m}.$

This tensor has the property that the vector-tensor product $\boldsymbol{y}^\top(\boldsymbol{I}_m \otimes \boldsymbol{x}) = \boldsymbol{x}\boldsymbol{y}^\top$ for any $\boldsymbol{y} \in \mathbb{R}^m$.

Note that $\boldsymbol{Y} \otimes \boldsymbol{x} \neq (\boldsymbol{x} \otimes \boldsymbol{Y})^\top$.

8. The element-wise product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is given by

$$\boldsymbol{x} \odot \boldsymbol{y} = \begin{bmatrix} x_1y_1 \\ \vdots \\ x_ny_n \end{bmatrix} \in \mathbb{R}^n.$$

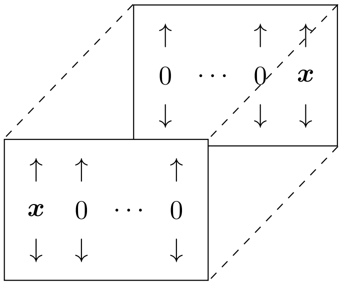
We may also write the element-wise product in terms of the diagonal operator

$$\boldsymbol{x} \odot \boldsymbol{y} = \operatorname{diag}(\boldsymbol{x})\boldsymbol{y} = \begin{bmatrix} x_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

2 Backpropagation of a Shallow Network

Proposition 1. *Consider $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{W} \in \mathbb{R}^{h \times d}, \boldsymbol{b} \in \mathbb{R}^h, \boldsymbol{z} \in \mathbb{R}^h$. Then*

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}} = \begin{bmatrix} \left[\frac{\partial z_1}{\partial \boldsymbol{X}}\right] \\ \vdots \\ \left[\frac{\partial z_h}{\partial \boldsymbol{W}}\right] \end{bmatrix} =$$



$$= \boldsymbol{I}_h \otimes \boldsymbol{x} \in \mathbb{R}^{h \times d \times h}, \quad \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} = \boldsymbol{I}_h.$$

Proof. Note the i^{th} component of $\boldsymbol{z}$ is

$$z_i = (\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_i = (\boldsymbol{W}\boldsymbol{x})_i + b_i = \boldsymbol{w}_i^\top \boldsymbol{x} + b_i,$$

where $\boldsymbol{w}_i$ is the i^{th} row vector of $\boldsymbol{W}$. Component-wise, we have the following:

$$\begin{aligned} \frac{\partial z_i}{\partial w_{jk}} &= \frac{\partial(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)}{\partial w_{jk}} = \frac{\partial(\sum_l w_{il} x_l)}{\partial w_{jk}} = \sum_l \frac{\partial w_{il}}{\partial w_{jk}} x_l = \sum_l \delta_{ij}\delta_{lk} x_l = \delta_{ij} x_k \\ \frac{\partial z_i}{\partial b_j} &= \frac{\partial(\boldsymbol{w}_i^\top \boldsymbol{x} + b_i)}{\partial b_j} = \frac{\partial b_i}{\partial b_j} = \delta_{ij}. \end{aligned}$$

Using the notations in Section 1, we complete the proof. □

Proposition 2. *Consider a shallow network*

$$\hat{y}(x; \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}_1, \boldsymbol{b}_2) = \boldsymbol{W}^{(2)} \sigma(\boldsymbol{W}^{(1)} \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2,$$

where $\boldsymbol{x} \in \mathbb{R}^d, \hat{y} \in \mathbb{R}^n, \boldsymbol{W}^{(1)} \in \mathbb{R}^{h \times d}, \boldsymbol{W}^{(2)} \in \mathbb{R}^{n \times h}, \boldsymbol{b}_1 \in \mathbb{R}^h$ and $\boldsymbol{b}_2 \in \mathbb{R}^n$. For a vector $\boldsymbol{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, denote $\widetilde{x} \in \mathbb{R}^{d+1}$ be the vector $\boldsymbol{x}$ appended by 1 at the end, $\widetilde{x} = (x_1, \ldots, x_d, 1)^T$.

Denote

$$\widetilde{\boldsymbol{W}}^{(k)} := [\boldsymbol{W}^{(k)}|\boldsymbol{b}_k] \in \mathbb{R}^{(\#rows\ of\ \boldsymbol{W}^{(k)})\times(1+\#cols\ of\ \boldsymbol{W}^{(k)})} \quad for \quad k=1,2.$$

$$\boldsymbol{z} := \boldsymbol{W}^{(1)}x + \boldsymbol{b}_1 \in \mathbb{R}^h, \quad \boldsymbol{s} := \sigma(\boldsymbol{z}) \in \mathbb{R}^h, \quad \boldsymbol{r} := \hat{y} \in \mathbb{R}^n,$$

The derivative of $\ell(\boldsymbol{x};\theta) := \frac{1}{2}\|\boldsymbol{r}\|_2^2$ *with respect to its inner and outer layer parameters is given by*

$$\frac{\partial \ell}{\partial \widetilde{\boldsymbol{W}}^{(1)}} = \widetilde{\boldsymbol{x}}\boldsymbol{r}^\top \boldsymbol{W}^{(2)} \operatorname{diag}(\sigma'(\boldsymbol{z})), \quad \frac{\partial \ell}{\partial \widetilde{\boldsymbol{W}}^{(2)}} = \widetilde{\boldsymbol{s}}\boldsymbol{r}^\top.$$

That is,

$$\begin{aligned} &\frac{\partial \ell}{\partial \boldsymbol{W}^{(1)}} := \boldsymbol{x}\boldsymbol{r}^\top \boldsymbol{W}^{(2)} \operatorname{diag}(\sigma'(\boldsymbol{z})) \in \mathbb{R}^{d\times h}, \quad \frac{\partial \ell}{\partial \boldsymbol{b}_1} = \boldsymbol{r}^\top \boldsymbol{W}^{(2)} \operatorname{diag}(\sigma'(\boldsymbol{z})) \in \mathbb{R}^{1\times h}, \\ &\frac{\partial \ell}{\partial \boldsymbol{W}^{(2)}} = \boldsymbol{s}\boldsymbol{r}^\top \in \mathbb{R}^{h\times n}, \quad \frac{\partial \ell}{\partial \boldsymbol{b}_2} = r^T \in \mathbb{R}^{1\times n}. \end{aligned}$$

Proof. Since $\ell = \frac{1}{2}\|r\|_2^2$, we have

$$\frac{\partial \ell}{\partial r_i} = \frac{\partial(\frac{1}{2}\sum_j r_j^2)}{\partial r_i} = \frac{1}{2}\sum_j \frac{\partial(r_j^2)}{\partial r_i} = \frac{1}{2}\sum_j 2r_j\frac{\partial r_j}{\partial r_i} = \sum_j r_j\delta_{ij} = r_i \Rightarrow \frac{\partial \ell}{\partial \boldsymbol{r}} = \boldsymbol{r}^T.$$

Since $\boldsymbol{r} = \boldsymbol{W}^{(2)}\boldsymbol{s} + \boldsymbol{b}^{(2)} \in \mathbb{R}^n$, from Proposition 1, we have

$$\frac{\partial r_i}{\partial W_{jk}^{(2)}} = \delta_{ij}s_k, \quad \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{W}^{(2)}} = \boldsymbol{I}_n \otimes \boldsymbol{s}.$$

$$\frac{\partial r_i}{\partial b_j^{(2)}} = \delta_{ij}, \quad \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{b}^{(2)}} = \boldsymbol{I}_n.$$

Therefore,

$$\begin{aligned} \frac{\partial \ell}{\partial W_{jk}^{(2)}} &= \sum_{i=1}^n \frac{\partial \ell}{\partial r_i}\frac{\partial r_i}{\partial W_{jk}^{(2)}} = \sum_{i=1}^n r_i\ \delta_{ij}\ s_k = r_j s_k \\ \frac{\partial \ell}{\partial b_j^{(2)}} &= \sum_{i=1}^n \frac{\partial \ell}{\partial r_i}\frac{\partial r_i}{\partial b_j^{(2)}} = \sum_{i=1}^n r_i\delta_{ij} = r_j. \end{aligned}$$

Using the notations in Section 1 and the common matrix notations, we can rewrite the above derivatives as follows:

$$\frac{\partial \ell}{\partial \widetilde{\boldsymbol{W}}^{(2)}} = \frac{\partial \ell}{\partial \boldsymbol{r}}\frac{\partial \boldsymbol{r}}{\partial \widetilde{\boldsymbol{W}}^{(2)}} = \boldsymbol{r}^\top(\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{s}}) = \widetilde{\boldsymbol{s}}\boldsymbol{r}^\top \in \mathbb{R}^{(h+1)\times n}.$$

Let $\widetilde{\boldsymbol{w}}_i^\top$ be the i^{th} row of $\widetilde{\boldsymbol{W}}^{(1)}$.

$$\left(\frac{\partial \boldsymbol{s}}{\partial \boldsymbol{z}}\right)_{ij} = \frac{\partial s_i}{\partial z_j} = \frac{\partial(\sigma(\boldsymbol{z}))_i}{\partial z_j} = \frac{\partial \sigma(z_i)}{\partial z_j} = \sigma'(z_i)\frac{\partial z_i}{\partial z_j} = \sigma'(z_i)\delta_{ij}$$
$$\left(\frac{\partial \boldsymbol{z}}{\partial \widetilde{\boldsymbol{W}}^{(1)}}\right)_{ijk} = \frac{\partial z_i}{\partial \widetilde{W}_{kj}^{(1)}} = \frac{\partial(\widetilde{\boldsymbol{w}}_i^\top \widetilde{\boldsymbol{x}})}{\partial \widetilde{W}_{kj}^{(1)}} = \frac{\partial(\sum_l \widetilde{W}_{il}^{(1)} \widetilde{x}_l)}{\partial \widetilde{W}_{kj}^{(1)}} = \sum_l \frac{\partial \widetilde{W}_{il}^{(1)}}{\partial \widetilde{W}_{kj}^{(1)}} \widetilde{x}_l = \sum_l \delta_{ik}\delta_{lj}\widetilde{x}_l = \delta_{ik}\widetilde{x}_j.$$

We also have $\boldsymbol{r} = \boldsymbol{W}^{(2)}\boldsymbol{s} + \boldsymbol{b}^{(2)}$, so $\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{s}} = \boldsymbol{W}^{(2)}$. Using chain rule,

$$\begin{aligned}\frac{\partial \ell}{\partial \widetilde{\boldsymbol{W}}^{(1)}} &= \frac{\partial \ell}{\partial \boldsymbol{r}} \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{s}} \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \widetilde{\boldsymbol{W}}^{(1)}} \\ &= \boldsymbol{r}^\top \boldsymbol{W}^{(2)} \operatorname{diag}(\sigma'(\boldsymbol{z}))(\boldsymbol{I}_h \otimes \widetilde{\boldsymbol{x}}) \\ &= \widetilde{\boldsymbol{x}}\boldsymbol{r}^\top \boldsymbol{W}^{(2)} \operatorname{diag}(\sigma'(\boldsymbol{z})),\end{aligned}$$

which completes the proof. □

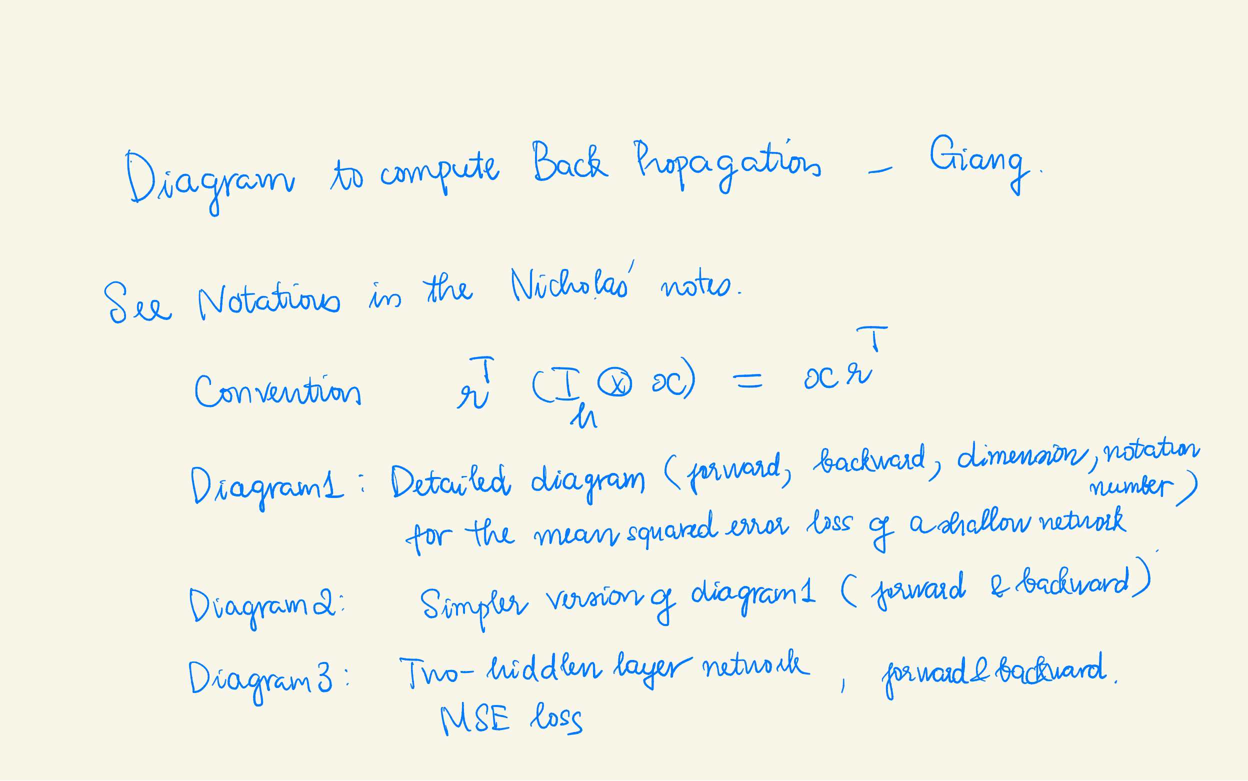


Diagram to compute Back Propagation - Giang.

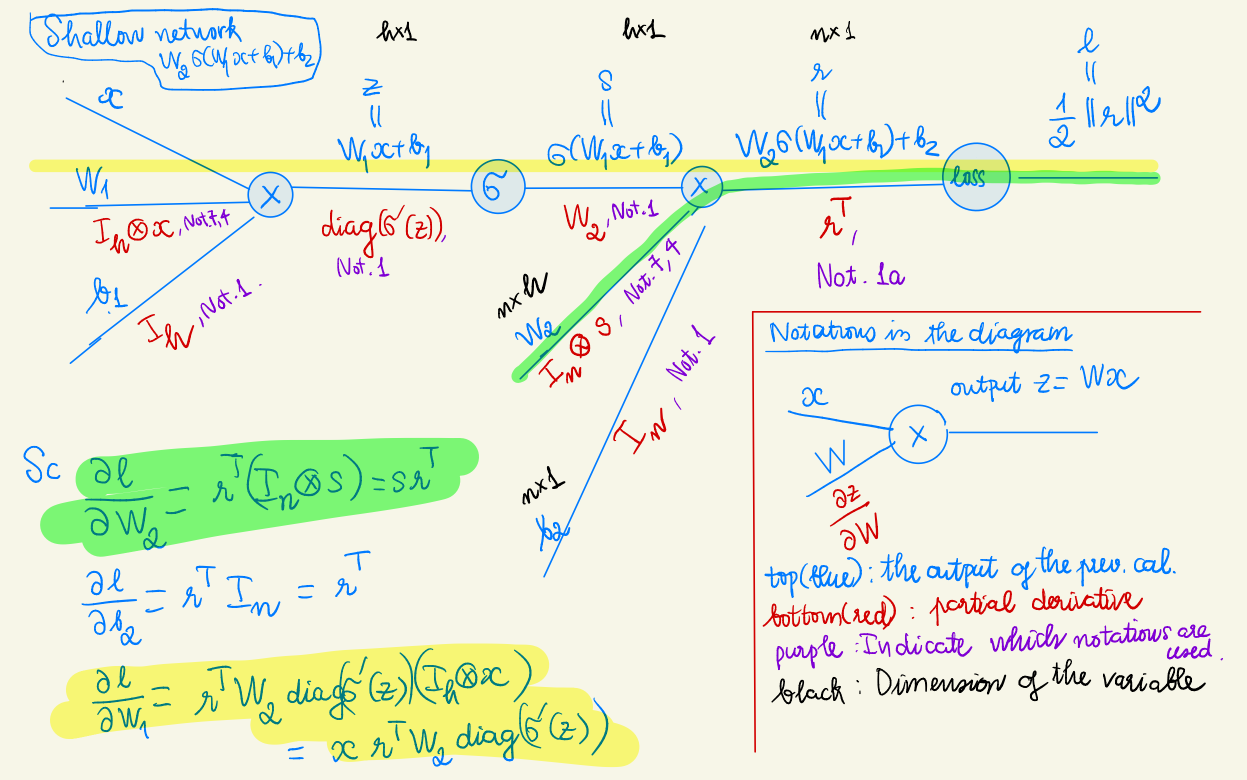
See Notations in the Nicholas' notes.

Convention $r^T (I_h \otimes x) = x r^T$

Diagram1: Detailed diagram (forward, backward, dimension, notation number) for the mean squared error loss of a shallow network

Diagram2: Simpler version of diagram1 (forward & backward)

Diagram3: Two-hidden layer network, forward & backward. MSE loss



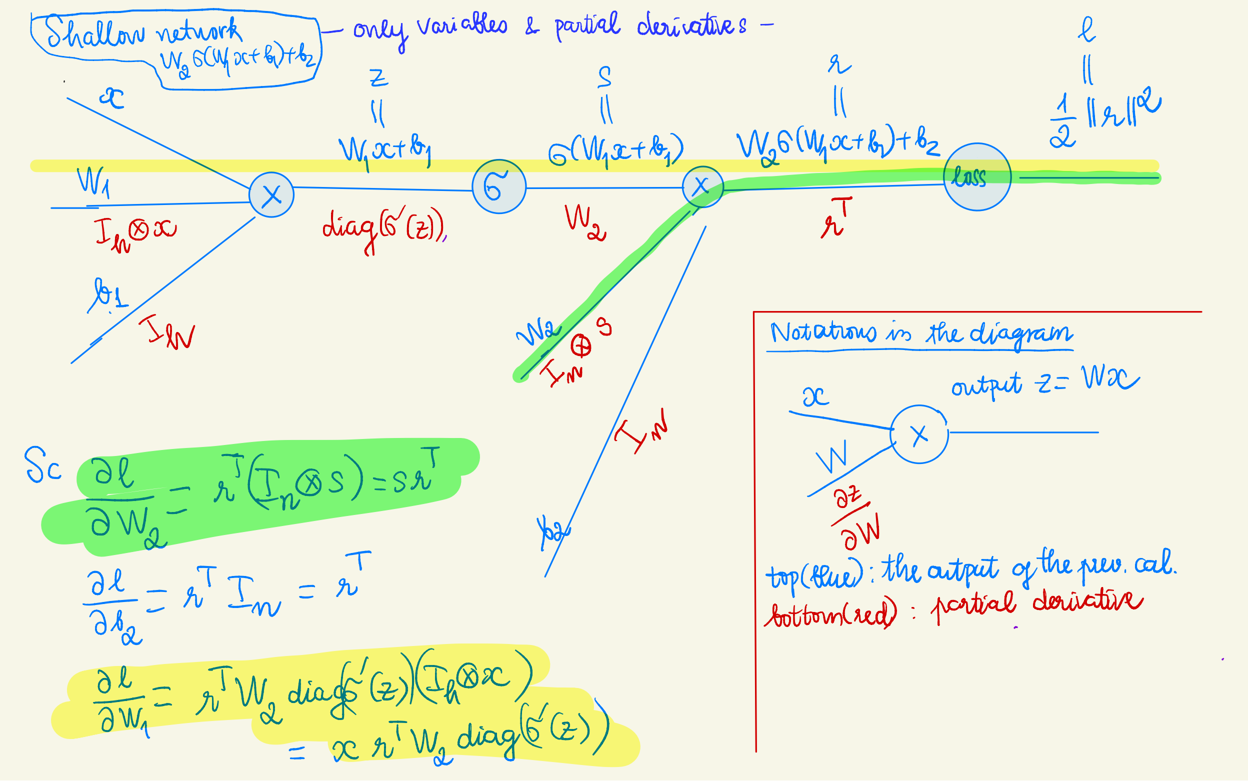
Sc $$\frac{\partial \ell}{\partial W_2} = r^T(I_n\otimes s) = sr^T$$

$$\frac{\partial \ell}{\partial b_2} = r^T I_n = r^T$$

$$\frac{\partial \ell}{\partial W_1} = r^T W_2\, \mathrm{diag}(\sigma'(z))(I_h\otimes x)$$
$$= x\, r^T W_2\, \mathrm{diag}(\sigma'(z)))$$

Notations in the diagram

top(blue): the output of the prev. cal.
bottom(red): partial derivative
purple: Indicate which notations are used.
black: Dimension of the variable



Shallow network
$W_2\sigma(W_1x+b_1)+b_2$
— only variables & partial derivatives —
x
W_1
$I_h \otimes x$
b_1
I_N
$z = W_1x+b_1$
$\mathrm{diag}(\sigma'(z))$,
σ
$s = \sigma(W_1x+b_1)$
W_2
$r = W_2\sigma(W_1x+b_1)+b_2$
r^T
loss
$\ell = \frac{1}{2}\|r\|^2$
W_2
$I_n \otimes s$
b_2
I_N

So
$$\frac{\partial \ell}{\partial W_2} = r^T(I_n \otimes s) = s r^T$$

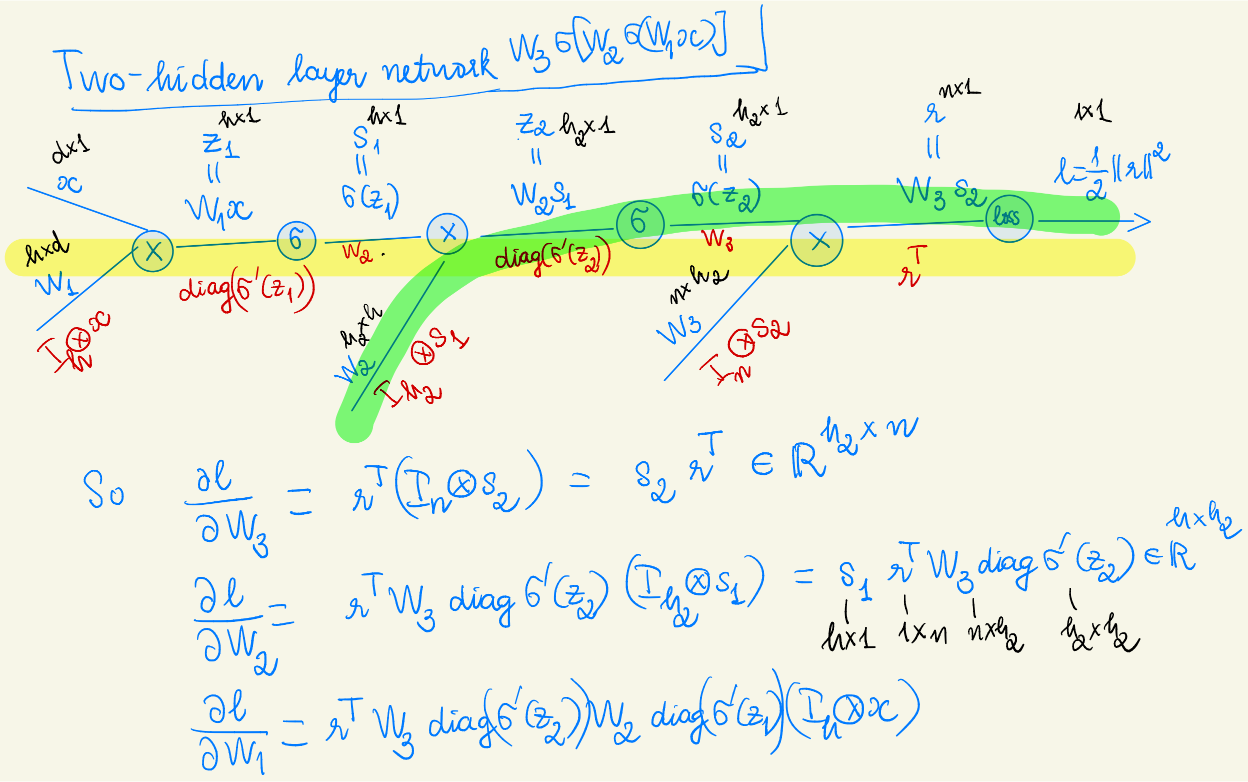
$$\frac{\partial \ell}{\partial b_2} = r^T I_n = r^T$$

$$\frac{\partial \ell}{\partial W_1} = r^T W_2\, \mathrm{diag}(\sigma'(z))(I_h \otimes x) = x\, r^T W_2\, \mathrm{diag}(\sigma'(z))$$

Notations in the diagram

x
W
output $z = Wx$
$\frac{\partial z}{\partial W}$

top(blue): the output of the prev. cal.
bottom(red): partial derivative



Two-hidden layer network $W_3\sigma[W_2\sigma(W_1x)]$

So $$\frac{\partial l}{\partial W_3} = r^T(I_n \otimes s_2) = s_2 r^T \in \mathbb{R}^{h_2\times n}$$

$$\frac{\partial l}{\partial W_2} = r^T W_3 \,\mathrm{diag}\, \sigma'(z_2)\,(I_{h_2}\otimes s_1) = \underbrace{s_1}_{h\times 1}\ \underbrace{r^T}_{1\times n}\ \underbrace{W_3}_{n\times h_2}\ \mathrm{diag}\, \underbrace{\sigma'(z_2)}_{h_2\times h_2} \in \mathbb{R}^{h\times h_2}$$

$$\frac{\partial l}{\partial W_1} = r^T W_3 \,\mathrm{diag}(\sigma'(z_2)) W_2 \,\mathrm{diag}(\sigma'(z_1))(I_h \otimes x)$$

$$= \underbrace{x}_{d\times 1}\ \underbrace{r^T}_{1\times n}\ \underbrace{W_3}_{n\times h_2}\ \underbrace{\mathrm{diag}(\sigma'(z_2))}_{h_2\times h_2}\ \underbrace{W_2}_{h_2\times h}\ \underbrace{\mathrm{diag}(\sigma'(z_1))}_{h\times h} \in \mathbb{R}^{d\times h}$$