VII.3) Backpropagation and the Chain Rule

When training a neural network, most of the time is being spent on computing the gradient of the loss function with respect to the many parameters of the neural net.

Finding an explicit formula for the gradient and writing the code to evaluate the gradient is not reasonable for such a large optimization problem.

There are two alternatives to finding the gradient explicitly.

One alternative is to use finite differences to approximate all the partial derivatives in the gradient:

$$\nabla f(x)_i = \frac{\partial f}{\partial x_i} \approx \frac{f(x + he_i) - f(x)}{h},$$

where e_i is the i^{th} column of the identity matrix. However, this would require us to evaluate the function $f : \mathbb{R}^n \to \mathbb{R}$ many times, and we would only get an approximation.

A better approach is to use AD (automatic differentiation) which produces a routine that will exactly evaluate the derivatives.

Two ways AD is implemented are:

(1) source code transformation (SCT),

(2) operator overloading (OO).

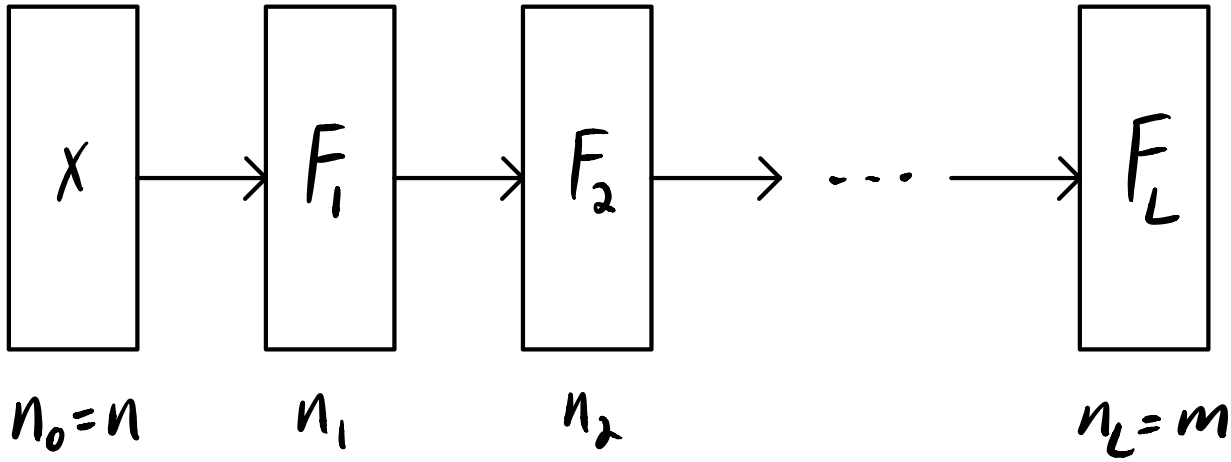
The Julia package Zygote.jl is an SCT implementation while the ForwardDiff.jl package uses OO.

Forward-mode AD

Suppose $F: \mathbb{R}^n \to \mathbb{R}^m$ is given by a composition of intermediate calculations:

$$F(x) = (F_L \circ \cdots \circ F_2 \circ F_1)(x)$$

$$= F_L(\cdots F_2(F_1(x))).$$



Here $F_i : \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}$, for $i = 1, \ldots, L$.

We use $F'(x) = \frac{\partial F}{\partial x}(x)$ to represent the $m \times n$ Jacobian matrix of $F : \mathbb{R}^n \to \mathbb{R}^m$:

$$F(x+d) \approx F(x) + F'(x)d, \quad \forall d \in \mathbb{R}^n \text{ small}.$$

Example: $F : \mathbb{R}^2 \to \mathbb{R}^3$ $\quad F'(x)$ is 3×2

$$F(x) = \begin{bmatrix} x_1 + x_2 \\ x_1 x_2 \\ x_1^2 - x_2 \end{bmatrix} \qquad F'(x) = \begin{bmatrix} 1 & 1 \\ x_2 & x_1 \\ 2x_1 & -1 \end{bmatrix}$$

//

We can use the chain rule to take the derivative of a composition:

$$\boxed{F(x) = F_2(F_1(x)) \quad \Rightarrow \quad F'(x) = F_2'(F_1(x)) \cdot F_1'(x).}$$

With $F_1 : \mathbb{R}^n \to \mathbb{R}^{n_1}$, $F_2 : \mathbb{R}^{n_1} \to \mathbb{R}^m$, and $F : \mathbb{R}^n \to \mathbb{R}^m$, we have:

$$F'(x) = F_2'(F_1(x)) \cdot F_1'(x).$$

$m \times n \qquad m \times n_1 \qquad n_1 \times n$

Example: $F_1 : \mathbb{R}^2 \to \mathbb{R}^2$, $F_2 : \mathbb{R}^2 \to \mathbb{R}^2$

$$F_1(x) = \begin{bmatrix} x_1 - x_2 \\ x_1 x_2 \end{bmatrix}, \quad F_2(x) = \begin{bmatrix} x_1^2 \\ x_1 - x_2 \end{bmatrix}$$

$$\Rightarrow \quad F(x) = F_2(F_1(x)) = \begin{bmatrix} (x_1 - x_2)^2 \\ x_1 - x_2 - x_1 x_2 \end{bmatrix}.$$

$F : \mathbb{R}^2 \to \mathbb{R}^2$

$$F_1'(x) = \begin{bmatrix} 1 & -1 \\ x_2 & x_1 \end{bmatrix}, \quad F_2'(x) = \begin{bmatrix} 2x_1 & 0 \\ 1 & -1 \end{bmatrix}.$$

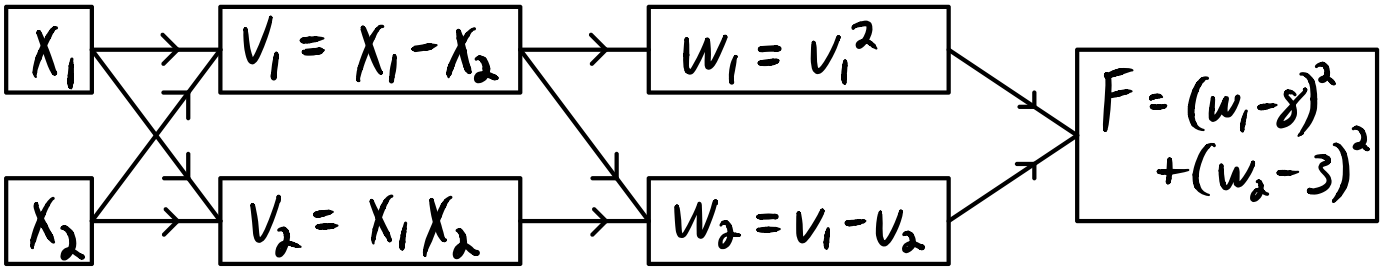
$$F'(x) = \begin{bmatrix} 2(x_1 - x_2) & -2(x_1 - x_2) \\ 1 - x_2 & -1 - x_1 \end{bmatrix}$$

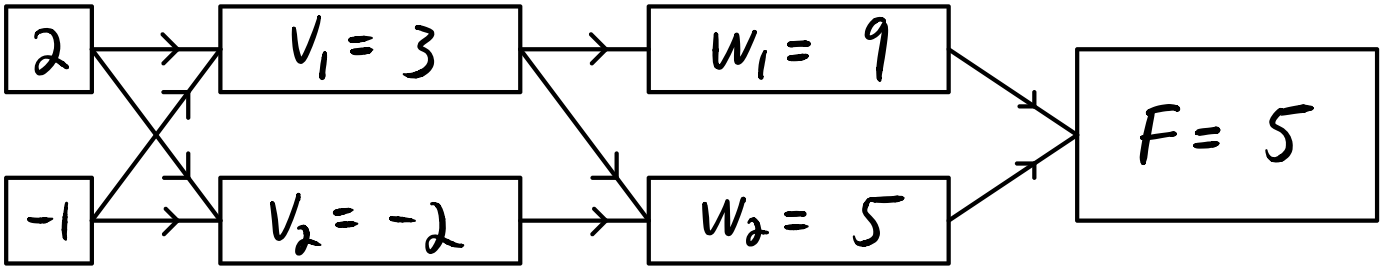
$$F_2'(F_1(x)) \cdot F_1'(x) = \begin{bmatrix} 2(x_1 - x_2) & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ x_2 & x_1 \end{bmatrix}$$

$$= \begin{bmatrix} 2(x_1 - x_2) & -2(x_1 - x_2) \\ 1 - x_2 & -1 - x_1 \end{bmatrix}. \checkmark$$

Suppose $F_3(x) = (x_1 - 8)^2 + (x_2 - 3)^2$ and

$$F(x) = F_3(F_2(F_1(x))).$$





To compute F, we pass <u>forward</u> through the network. Now we want to compute

$F(x+d) \approx F(x) + \nabla F(x)^T d$

$$F'(x) = \begin{bmatrix} \frac{\partial F}{\partial x_1} & \frac{\partial F}{\partial x_2} \end{bmatrix} \quad (2 \times 1)$$

$F: \mathbb{R}^2 \to \mathbb{R}^1$

$$= \nabla F(x)^T.$$

First we compute $\frac{\partial F}{\partial x_1}$ with a <u>forward</u> pass:

$x_1 = 2$
$x_2 = -1$

$v_1 = x_1 - x_2$
$v_2 = x_1 x_2$

$\frac{\partial x_1}{\partial x_1} = 1$ → $\frac{\partial v_1}{\partial x_1} = \frac{\partial x_1}{\partial x_1} - \frac{\partial x_2}{\partial x_1} = 1$

$\frac{\partial x_2}{\partial x_1} = 0$ → $\frac{\partial v_2}{\partial x_1} = \frac{\partial x_1}{\partial x_1} x_2 + x_1 \frac{\partial x_2}{\partial x_1} = -1$

$1 \cdot (-1) + 2(0) = -1$

$v_1 = 3$
$v_2 = -2$

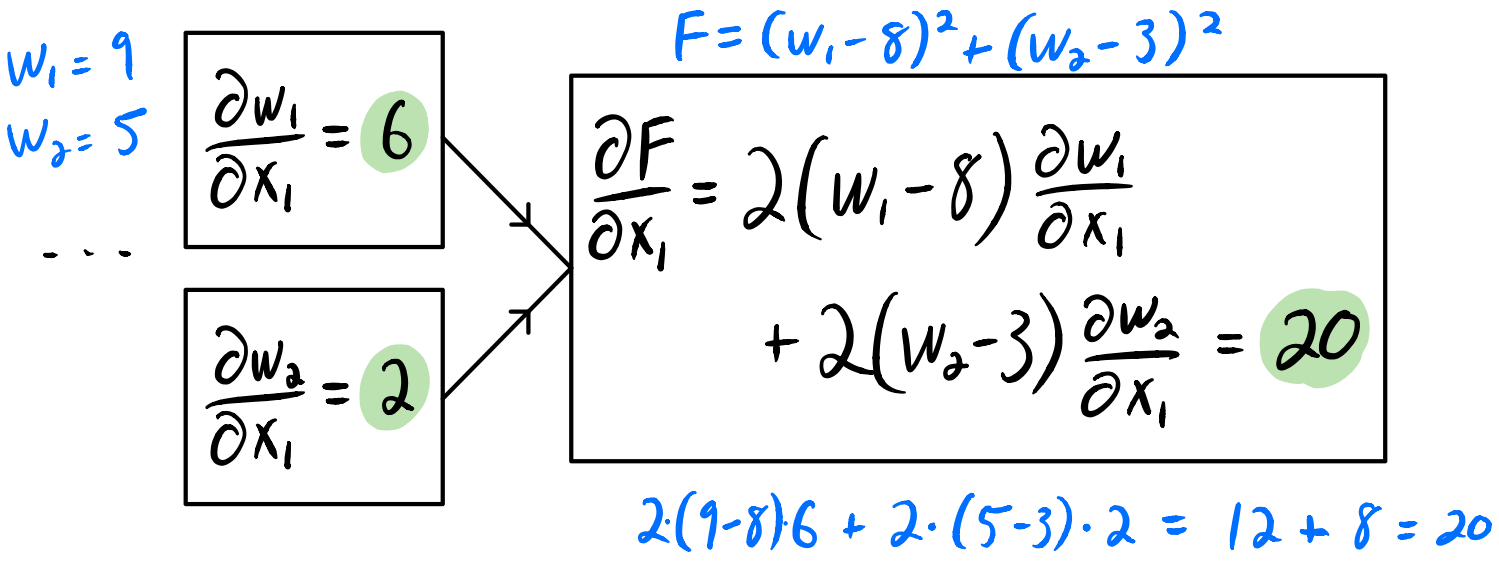
$\frac{\partial v_1}{\partial x_1} = 1$ → $\frac{\partial w_1}{\partial x_1} = 2 v_1 \frac{\partial v_1}{\partial x_1} = 6$

$2 \cdot 3 \cdot 1 = 6$

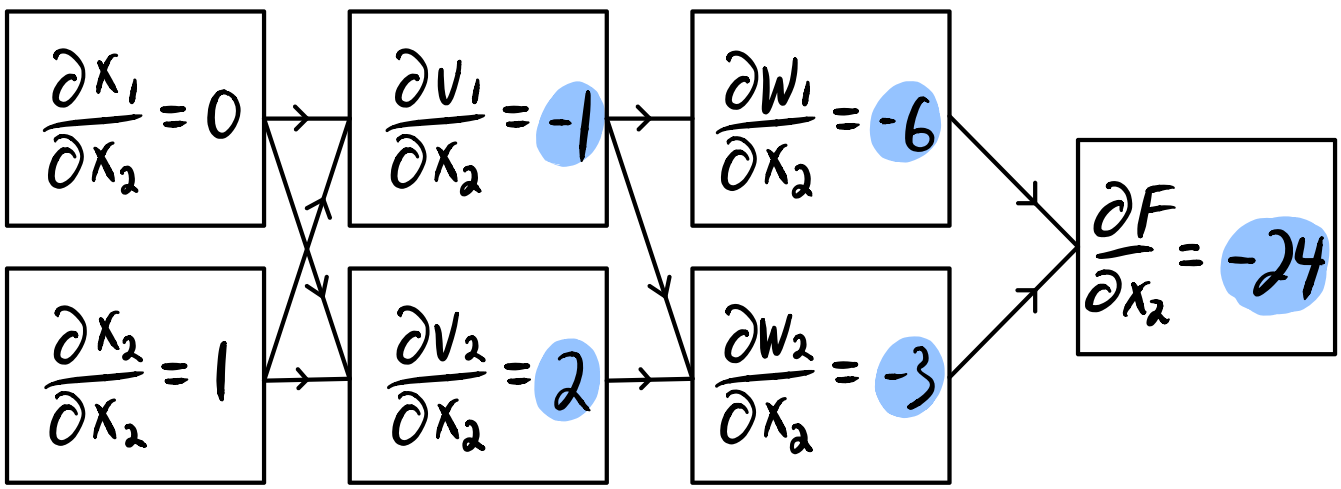
$w_1 = v_1^2$
$w_2 = v_1 - v_2$

$\frac{\partial v_2}{\partial x_1} = -1$ → $\frac{\partial w_2}{\partial x_1} = \frac{\partial v_1}{\partial x_1} - \frac{\partial v_2}{\partial x_1} = 2$

$1 - (-1) = 2$



Computing $\frac{\partial F}{\partial x_2}$ requires another <u>forward</u> pass through the network:



Thus, <u>forward-mode AD</u> requires a separate pass through the network for each variable. This will be a lot of work if there are many variables, like in neural nets.

Let

$$v = F_1(x) = \begin{bmatrix} x_1 - x_2 \\ x_1 x_2 \end{bmatrix}, \quad w = F_2(v) = \begin{bmatrix} v_1^2 \\ v_1 - v_2 \end{bmatrix},$$

and $F = F_3(w) = (w_1 - 8)^2 + (w_2 - 3)^2$.

Then

$$F'(x) = F_3'(w) \cdot F_2'(v) \cdot F_1'(x)$$

$$= \begin{bmatrix} 2(w_1 - 8) & 2(w_2 - 3) \end{bmatrix} \begin{bmatrix} 2v_1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ x_2 & x_1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 4 \end{bmatrix} \left(\begin{bmatrix} 6 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 2 & 4 \end{bmatrix} \begin{bmatrix} 6 & -6 \\ 2 & -3 \end{bmatrix}$$

first pass

second pass

$$= \begin{bmatrix} 20 & -24 \end{bmatrix}.$$

Forward-mode AD $\equiv$ mult. from the right.

We could compute $F'(x)$ more efficiently by multiplying from the left:

$$F'(x) = \left(\begin{bmatrix} 2 & 4 \end{bmatrix} \begin{bmatrix} 6 & 0 \\ 1 & -1 \end{bmatrix} \right) \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 16 & -4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 20 & -24 \end{bmatrix}$$

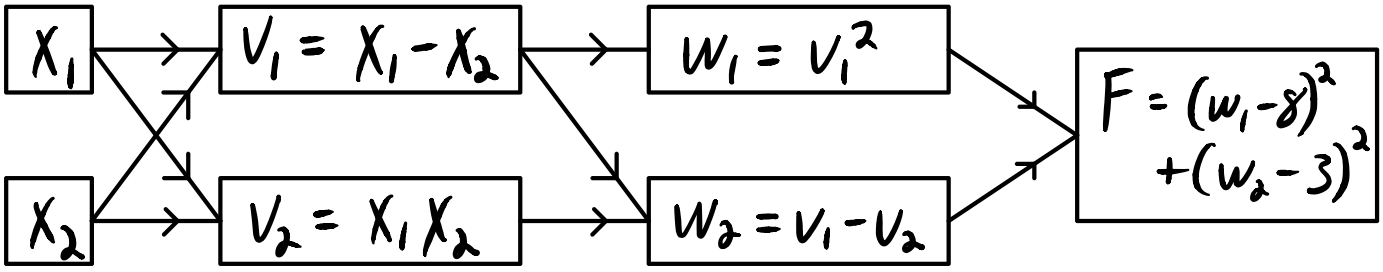
Notice that we only needed to do vector-matrix multiplication; we avoided doing the matrix-matrix multiplication. This can save us from doing a lot of extra work.

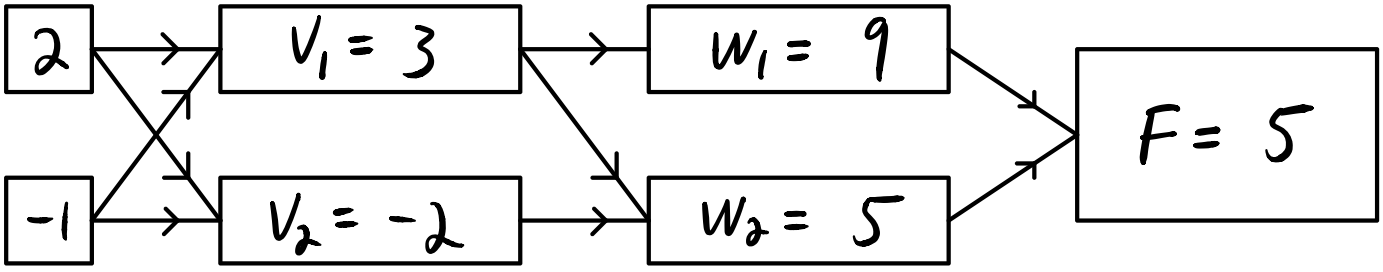
Mult. from the left $\equiv$ Backward-mode AD

Backward-mode AD

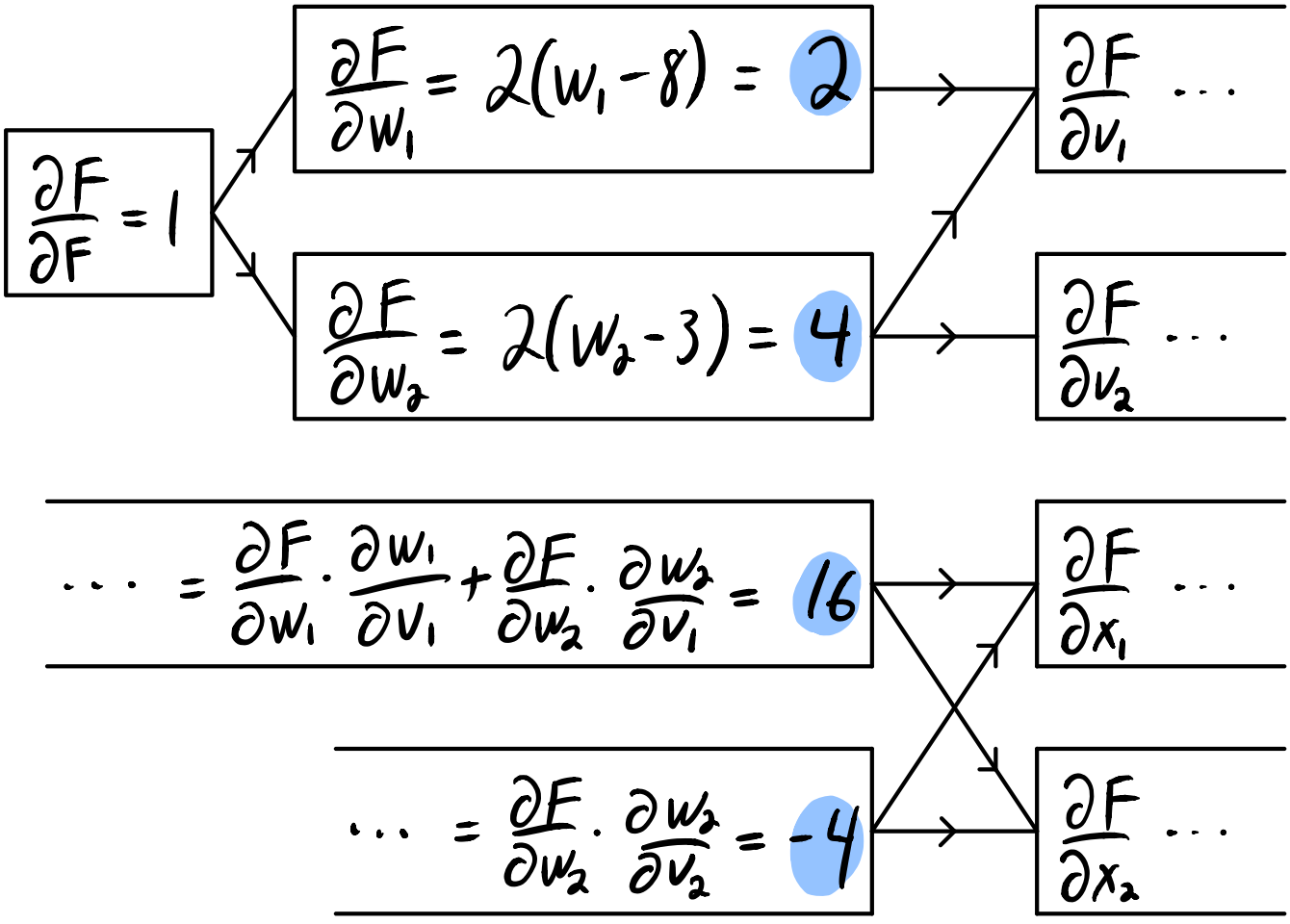
This mode is also called reverse-mode AD or backpropagation.

We still begin with a forward pass to compute $F(x)$.





Then we compute all partial derivatives with a single backward pass.



$$\cdots = \frac{\partial F}{\partial v_1} \cdot \frac{\partial v_1}{\partial x_1} + \frac{\partial F}{\partial v_2} \cdot \frac{\partial v_2}{\partial x_1} = 20$$

Single pass

$$\cdots = \frac{\partial F}{\partial v_1} \cdot \frac{\partial v_1}{\partial x_2} + \frac{\partial F}{\partial v_2} \cdot \frac{\partial v_2}{\partial x_2} = -24$$

$$F'(x) = \left(\begin{bmatrix} 2 & 4 \end{bmatrix} \begin{bmatrix} 6 & 0 \\ 1 & -1 \end{bmatrix} \right) \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 16 & -4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 20 & -24 \end{bmatrix}$$

Note: If $F: \mathbb{R}^n \to \mathbb{R}^m$, then using forward-mode AD to compute $F'(x)$ requires n forward passes through the computational graph, while backward-mode AD requires m backward passes. Therefore, if $n < m$, forward-mode should be used and if $n > m$, backward-mode is preferred. Training a neural net requires minimizing a loss function $F: \mathbb{R}^n \to \mathbb{R}$ where n, the number of parameters, is very large; thus, backward-mode AD will be far more efficient than forward-mode AD.

Backpropagation on a neural net

Example:

$$A = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$x = (A, b, c)$ are the parameters

$$v \to y = Av + b \to z = [y]_+ \to w = c^T z \to \ell = (w - \bar{w})^2$$

Data: $v = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $\bar{w} = 1$

$$v = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \to y = \begin{bmatrix} 3 \\ -5 \end{bmatrix} \to z = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \to w = 3 \to \ell = 4$$

$$\frac{\partial \ell}{\partial w} = 2(w - \bar{w}) = 4 \qquad \left(\ell = (w - \bar{w})^2\right)$$

$$\frac{\partial \ell}{\partial c_1} = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial c_1} = 4 z_1 = 12$$

$$\frac{\partial \ell}{\partial c_2} = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial c_2} = 4 z_2 = 0$$

$$\left(w = c_1 z_1 + c_2 z_2\right)$$

$$\frac{\partial \ell}{\partial z_1} = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial z_1} = 4 c_1 = 4$$

$$\frac{\partial \ell}{\partial z_2} = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial z_2} = 4 c_2 = 4$$

$$\left(w = c_1 z_1 + c_2 z_2\right)$$

$$\frac{\partial \ell}{\partial y_1} = \frac{\partial \ell}{\partial z_1} \cdot \frac{\partial z_1}{\partial y_1} = 4 \cdot 1 = 4 \qquad (z_1 = y_1)$$

$$\frac{\partial \ell}{\partial y_2} = \frac{\partial \ell}{\partial z_2} \cdot \frac{\partial z_2}{\partial y_2} = 4 \cdot 0 = 0 \qquad (z_2 = 0)$$

$$\frac{\partial \ell}{\partial a_{11}} = \frac{\partial \ell}{\partial y_1} \cdot \frac{\partial y_1}{\partial a_{11}} = 4v_1 = 8$$

$$\frac{\partial \ell}{\partial a_{12}} = \frac{\partial \ell}{\partial y_1} \cdot \frac{\partial y_1}{\partial a_{12}} = 4v_2 = -4 \qquad (y_1 = a_{11}v_1 + a_{12}v_2 + b_1)$$

$$\frac{\partial \ell}{\partial b_1} = \frac{\partial \ell}{\partial y_1} \cdot \frac{\partial y_1}{\partial b_1} = 4 \cdot 1 = 4$$

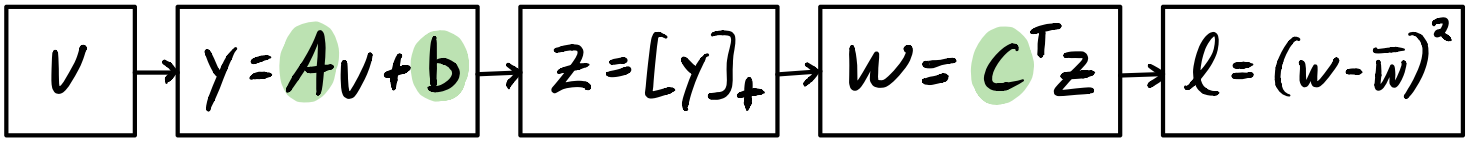
$$\frac{\partial \ell}{\partial a_{21}} = \frac{\partial \ell}{\partial y_2} \cdot \frac{\partial y_2}{\partial a_{21}} = 0 \cdot v_1 = 0$$

$$\frac{\partial \ell}{\partial a_{22}} = \frac{\partial \ell}{\partial y_2} \cdot \frac{\partial y_2}{\partial a_{22}} = 0 \cdot v_2 = 0 \qquad (y_2 = a_{21}v_1 + a_{22}v_2 + b_2)$$

$$\frac{\partial \ell}{\partial b_2} = \frac{\partial \ell}{\partial y_2} \cdot \frac{\partial y_2}{\partial b} = 0 \cdot 1 = 0$$

$$\therefore \nabla \ell(A, b, c) = \left(\begin{bmatrix} 8 & -4 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 12 \\ 0 \end{bmatrix} \right)$$

Using matrices to compute $\nabla \ell$:



$$\frac{\partial \ell}{\partial c}(\Delta c) = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial c}(\Delta c)$$

$w + \Delta w = (c + \Delta c)^T z$

$\Delta w = z^T \Delta c$

$$= 2(w - \bar{w})\, z^T \Delta c = 4 \begin{bmatrix} 3 & 0 \end{bmatrix} \Delta c$$

$$= \begin{bmatrix} 12 & 0 \end{bmatrix} \Delta c = \left\langle \begin{bmatrix} 12 \\ 0 \end{bmatrix}, \Delta c \right\rangle \quad \therefore \nabla_c \ell = \begin{bmatrix} 12 \\ 0 \end{bmatrix}$$

$$\frac{\partial \ell}{\partial A}(\Delta A) = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial z} \cdot \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial A}(\Delta A)$$

$y + \Delta y = (A + \Delta A)v + b$

$\Delta y = \Delta A v$

$$= 2(w - \bar{w}) \begin{bmatrix} c_1 & c_2 \end{bmatrix} \begin{bmatrix} [y_1]_+' & \\ & [y_2]_+' \end{bmatrix} (\Delta A v)$$

$$= 4 \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} (\Delta A v) = \begin{bmatrix} 4 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} (\Delta A v)$$

$$= \begin{bmatrix} 4 & 0 \end{bmatrix} \Delta A \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 4\Delta a_{11} & 4\Delta a_{12} \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$= 8\Delta a_{11} - 4\Delta a_{12} = \left\langle \begin{bmatrix} 8 & -4 \\ 0 & 0 \end{bmatrix}, \Delta A \right\rangle$$

$$\therefore \nabla_A \ell = \begin{bmatrix} 8 & -4 \\ 0 & 0 \end{bmatrix}.$$

$$\frac{\partial \ell}{\partial b}(\Delta b) = \frac{\partial \ell}{\partial w} \cdot \frac{\partial w}{\partial z} \cdot \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial b}(\Delta b)$$

$y + \Delta y = Av + b + \Delta b$

$\Delta y = \Delta b$

$$= 4 \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Delta b = \begin{bmatrix} 4 & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \Delta b$$

$$= \begin{bmatrix} 4 & 0 \end{bmatrix} \Delta b = \left\langle \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \Delta b \right\rangle$$

$$\therefore \nabla_b \ell = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

$$\therefore \nabla \ell(A, b, c) = \left(\begin{bmatrix} 8 & -4 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 4 \\ 0 \end{bmatrix}, \begin{bmatrix} 12 \\ 0 \end{bmatrix} \right)$$

$\ell : \mathbb{R} \to \mathbb{R}$ $\qquad f : \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \to \mathbb{R}^2$

$w : \mathbb{R}^2 \to \mathbb{R}$

$z : \mathbb{R}^2 \to \mathbb{R}^2$

$Y : \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \to \mathbb{R}^2$ $\qquad \frac{\partial Y}{\partial A} : \mathbb{R}^{2\times 2} \to \mathbb{R}^2$

$$\frac{\partial Y}{\partial A}(A) = \begin{bmatrix} \left\langle \begin{bmatrix} v_1 & v_2 \\ 0 & 0 \end{bmatrix}, A \right\rangle \\ \left\langle \begin{bmatrix} 0 & 0 \\ v_1 & v_2 \end{bmatrix}, A \right\rangle \end{bmatrix} = Av$$