1. $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$ ...... (1)

$$\underset{w}{\mathrm{argmin}}\; f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2} \| w - w^{(t)} \|^2 \quad (4)$$

Take derivative of (4) w.r.t. w, & set to 0.

$$0 = \cancel{\frac{\partial f(\vec{w}^{(t)})}{\partial \vec{w}}}^{\,0} + \frac{\partial \langle \vec{w} - \vec{w}(t), \nabla f(\vec{w}^{(t)}) \rangle}{\partial \vec{w}} + \frac{\partial}{\partial \vec{w}} \frac{\lambda}{2} \| \vec{w} - \vec{w}^{(t)} \|^2$$

$f(\vec{w}^{(t)})$ is not dependent of $\vec{w}$

$$0 = \nabla f(\vec{w}^{(t)})^T + \frac{\lambda}{\cancel{2}} \cdot \cancel{2} \cdot (\vec{w} - \vec{w}^{(t)})^T$$

$$\lambda w = \lambda w^{(t)} - \nabla f(w^{(t)})$$

$$w = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)}) \quad (A)$$

comparing (1) and (A),

$$\eta = \frac{1}{\lambda}. \quad (*)$$

The update rule moves toward the opposite direction of the gradient (page 100 of textbook). And when the gradient is zero, w is optimized.

η and λ have an inverse relationship. (*)

2) (eq 5) $$\sum_{t=1}^{T} \langle w^{(t)} - w^{*}, v_t \rangle \le \frac{\|w^{*}\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

$$\rightarrow \left[\sum \left(\langle w^{(t)}, v_t \rangle - \langle w^{*}, v_t \rangle \right) \right] 2\eta \le \|w^{*}\|^2 + \eta^2 \sum_{t=1}^{T} \|v_t\|^2$$

$= w^{(t+1)} + \eta v_t$ (update rule)

$$LHS = \left[\sum_{t=1}^{T} \langle w^{(t+1)}, v_t \rangle + \eta \langle v_t, v_t \rangle - \langle w^{*}, v_t \rangle \right] 2\eta$$

$$= 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^{*}, v_t \rangle + 2\eta^2 \sum_{t=1}^{T} \|v_t\|^2$$

$$\le RHS = \|w^{*}\|^2 + \eta^2 \sum_{t=1}^{T} \|v_t\|^2$$

move to LHS

$$RHS = \|w^{*}\|^2 \ge 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^{*}, v_t \rangle + \eta^2 \sum_{t=1}^{T} \|v_t\|^2 = LHS$$

Add $\sum_{t=1}^{T} \|w^{(t+1)} - w^{*}\|^2$ to both sides

$$\|w^{*}\|^2 + \sum_{t=1}^{T} \|w^{(t+1)} - w^{*}\|^2 \ge \sum_{t=1}^{T} \|w^{(t+1)} - w^{*}\|^2 + 2\eta \sum_{t=1}^{T} \langle w^{(t+1)} - w^{*}, v_t \rangle + \eta^2 \sum_{t=1}^{T} \|v_t\|^2$$

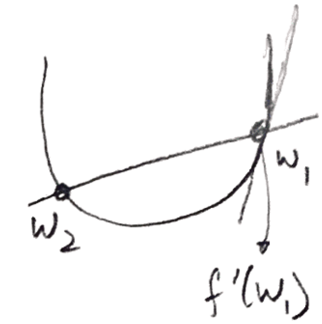
$$\|w^{*}\|^2 + \sum_{t=1}^{T} \|w^{(t+1)} - w^{*}\|^2 \ge \left(\sum \|w^{(t+1)} - w^{*}\|^2 + \eta \sum_{t=1}^{T} \|v_t\|^2 \right)^2$$

3] $f(w^{(1)}) - f(w^*) \le \langle w^{(1)} - w^*, f'(w^{(1)}) \rangle$

$f(w^{(2)}) - f(w^*) \le \langle w^{(2)} - w^*, f'(w^{(2)}) \rangle$

$\vdots$

$f(w^{(T)}) - f(w^*) \le \langle w^{(T)} - w^*, f'(w^{(T)}) \rangle$



for convex function instantaneous slope is greater than average slope

$$\frac{f(w_1)-f(w_2)}{w_1-w_2} < f'(w_1)$$

Add all LHS's and RHS's together.

$$\left(\sum_{t=1}^{T} f(w^{(t)})\right) - T \cdot f(w^*) \le \sum \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$$

divide by T, $\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$

by Jensen's Inequality, $f\left(\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})\right) \le \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})$

So,

$$\underbrace{f\left(\frac{1}{T}\sum_{t=1}^{T} f(w^{(t)})\right)}_{f(\bar{w})} - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f(w^*)$$

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, f'(w^{(t)}) \rangle$$

take result from #2, and divide by T.

$$\frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \le \frac{1}{T}\left(\frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|v_t\|^2\right) \le \cdots \to$$

$$\dashrightarrow \quad \leq \frac{1}{T}\left(\frac{B^2}{\eta} + \frac{\eta}{2} \cdot T\rho^2\right)$$

Now, rewrite (eq 6) using this inequality

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\left(\frac{B^2}{2\eta} + \frac{\eta}{2} \cdot T\rho^2\right)$$

if $\eta = \sqrt{\frac{B^2}{\rho^2 T}} = \frac{B}{\rho} \frac{1}{\sqrt{T}}$ then

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\left(\frac{B^2}{2} \cdot \frac{\rho\sqrt{T}}{B} + \frac{1}{2} \frac{B}{\rho} \frac{1}{\sqrt{T}} T\rho^2\right)$$

$$\leq \frac{BP}{2} \cdot \left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{T}}\right)$$

$$\leq \frac{BP}{2} \cdot \frac{2}{\sqrt{T}}$$

$$f(\bar{w}) - f(w^*) \leq \frac{BP}{\sqrt{T}}$$

Note, $\frac{BP}{\sqrt{T}} \propto \frac{1}{\sqrt{T}}$

4 $f_1(w) = -\ln\left(1 - \frac{1}{1+e^{-w}}\right) = -\ln\left(\frac{e^{-w}}{1+e^{-w}}\right)$

$$\frac{d\, f_1(w)}{dw} = (-1) \cdot \left(\frac{1+e^{-w}}{e^{-w}}\right) \cdot \frac{e^{-w}(-1) \cdot (1+e^{-w}) - e^{-w} \cdot e^{-w}(-1)}{(1+e^{-w})^2}$$

$$= \frac{-\cancel{(1+e^{-w})}}{\cancel{e^{-w}}} \cdot \frac{-\cancel{e^{-w}} - \cancel{e^{-2w}} + \cancel{e^{-2w}}}{\cancel{(1+e^{-w})^2}_1}$$

$$= \frac{1}{1+e^{-w}} > 0$$

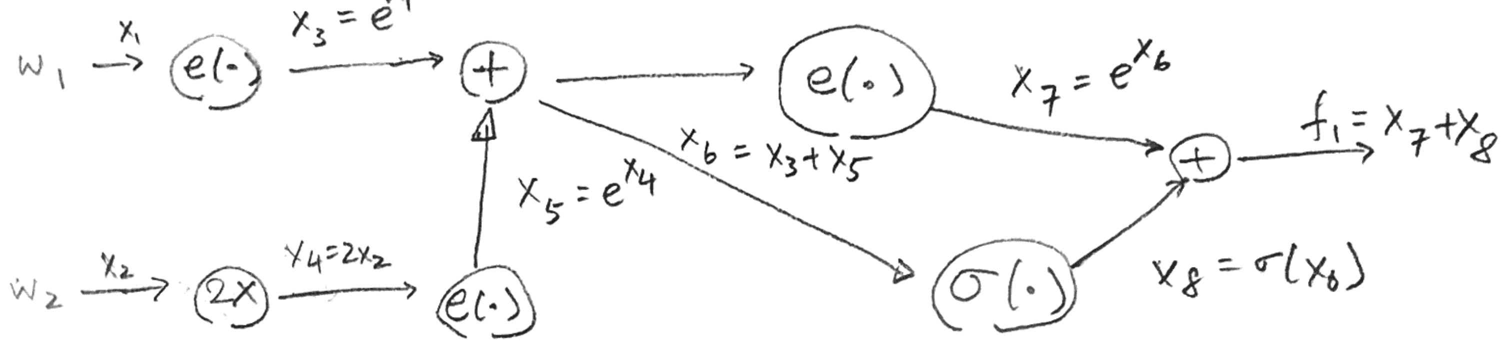
$$f_2(w) = -\ln\left(\frac{1}{1+e^{-w}}\right)$$

$$\frac{d\, f_2(w)}{dw} = (-1)(1+e^{-w}) \cdot \frac{(-1)(e^{-w})(-1)}{(1+e^{-w})^2}$$

$$= \frac{-e^{-w}}{1+e^{-w}} < 0$$

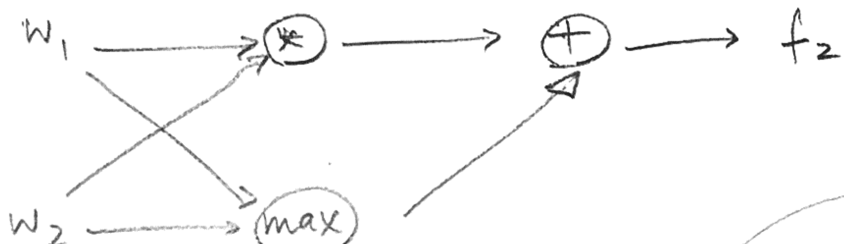
There is no guarantee, since the gradients, depending on which term is picked, is in opposite directions.

5) a) $f_1(w_1, w_2) = e^{e^{w_1} + e^{2w_2}} + \sigma(e^{w_1} + e^{2w_2})$



$$f_1(1,-1) = e^{(e + e^{-2})} + \sigma(e + e^{-2}) = 18.30$$

$f_2(w_1, w_2) = w_1 w_2 + \max(w_1, w_2)$



$f_2(1,-1) = -1 + 1 = 0$

$$\therefore f = \begin{bmatrix} 18.30 \\ 0 \end{bmatrix}$$

b) Jacobian $\Rightarrow J_{i,j} = \frac{\partial}{\partial x_j} f_i(x)$

$$\begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix} = \begin{pmatrix} 48.2 & 4.76 \\ 0 & 1 \end{pmatrix}$$

* these values are computed from the following ...→

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Set $h = \Delta w = 0.01, \quad \vec{w} = (1, -1)$

$$\frac{\partial f_1}{\partial w_1} = \frac{f_1(1.01, -1) - f_1(1, -1)}{0.01} = 48.192$$

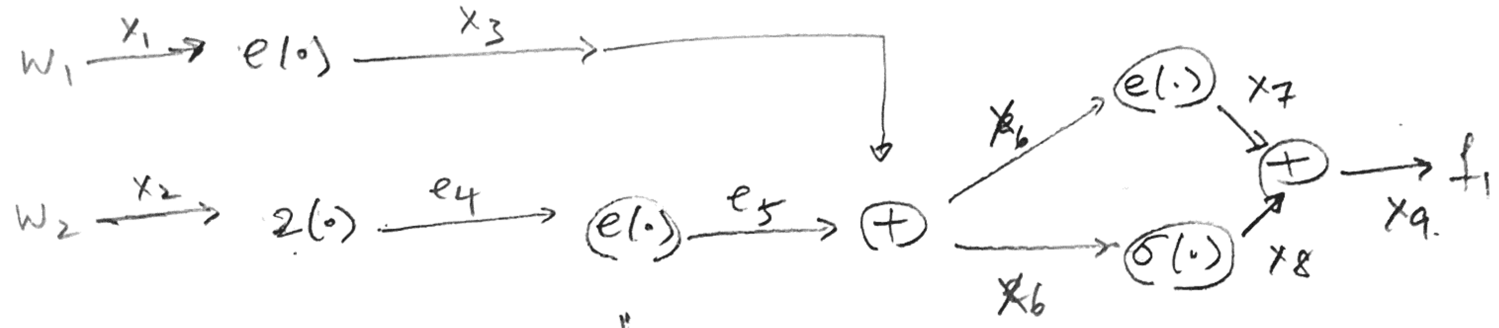
$$\frac{\partial f_1}{\partial w_2} = \frac{f_1(1, -0.99) - f_1(1, -1)}{0.01} = 4.764$$

$$\frac{\partial f_2}{\partial w_1} = \frac{f_2(1.01, -1) - f_2(1, -1)}{0.01} = 0$$

$$\frac{\partial f_2}{\partial w_2} = \frac{f_2(1, -0.99) - f_2(1, -1)}{0.01} = 1$$

---

5) c) F_1, forward, so start from x_1, x_2.



equations $\boxed{\text{Table 1}}$	$\frac{\partial f_1}{\partial w_1}$	$\frac{\partial f_1}{\partial f_2}$
$x_1 = w_1$ $x_2 = w_2$	$\dot{x}_1 = 1$ $\dot{x}_2 = 0$	$\dot{x}_1 = 0$ $\dot{x}_2 = 1$
$x_3 = e^{x_1}$	$\dot{x}_3 = e^{x_1} \cdot \dot{x}_1$	
$x_4 = 2x_2$	$\dot{x}_4 = 2\dot{x}_2$	
$x_5 = e^{x_4}$	$\dot{x}_5 = e^{x_4} \cdot \dot{x}_4$	
$x_6 = x_3 + x_5$	$\dot{x}_6 = \dot{x}_3 + \dot{x}_5$	
$x_7 = e^{x_6}$	$\dot{x}_7 = e^{x_6} \cdot \dot{x}_6$	
$x_8 = \sigma(x_6)$	$\dot{x}_8 = \sigma(x_6)(1 - \sigma(x_6)) \cdot \dot{x}_6$	
$x_9 = x_7 + x_8$	$\dot{x}_9 = \dot{x}_7 + \dot{x}_8$	

Consider $\frac{\partial f_1}{\partial w_1}$, $\dot{x}_1 = 1$, $\dot{x}_2 = 0$ (from chart above ↑)

$x_1 = w_1 = 1$, $x_2 = w_2 = -1$

$$\frac{\partial f_1}{\partial w_1} = \dot{x}_9 = \dot{x}_7 + \dot{x}_8 = \left(e^{x_6} \cdot \dot{x}_6\right) + \left[\sigma(x_6)(1-\sigma(x_6))\, \dot{x}_6\right]$$

$$\dot{x}_6 = \dot{x}_3 + \dot{x}_5 = \left(e^{x_1} \cdot \dot{x}_1\right) + \left(e^{x_4} \cdot \dot{x}_4\right)$$

$$= \left(e^{x_1} \cdot \dot{x}_1\right) + \left(e^{x_4} \cdot 2\dot{x}_2\right)$$

$$= \left(e^{x_1} \cdot 1\right) + \left(e^{x_4} \cdot 2(0)\right) = e^{x_1}$$

So, $\frac{\partial f_1}{\partial w_1} = e^{x_6}\left(e^{x_1}\right) + \left[\sigma(x_6)(1-\sigma(x_6))\right] e^{x_1}$

$x_1 = w_1$

(☆☆) $x_6 = x_3 + x_5 = e^{x_1} + e^{x_4} = e^{x_1} + e^{2x_2} = e^{w_1} + e^{2w_2}$

(A) So, $\frac{\partial f_1}{\partial w_1} = e^{(e^{w_1} + e^{2w_2})} \cdot e^{w_1} + \sigma\left(e^{w_1} + e^{2w_2}\right)\left(1 - \sigma\left(e^{w_1} + e^{2w_2}\right)\right) e^{w_1}$

Consider $\frac{\partial f_1}{\partial w_2}$, $\dot{x}_1 = 0$, $\dot{x}_2 = 1$

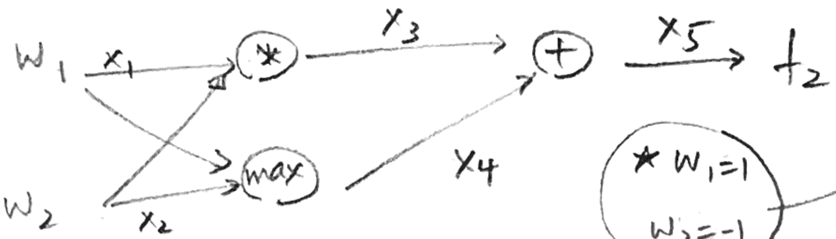
$\dot{x}_3 = 0$, $\dot{x}_4 = 2$, $\dot{x}_5 = 2e^{x_4}$, $\dot{x}_6 = \dot{x}_5$, $\dot{x}_7 = e^{x_6} \cdot 2e^{x_4}$

$\dot{x}_8 = \sigma(x_6)(1-\sigma(x_6)) \cdot 2e^{x_4}$

$\dot{x}_9 = e^{x_6} \cdot 2e^{x_4} + \sigma(x_6)(1-\sigma(x_6)) \cdot 2e^{x_4}$

(B) ∴ $\frac{\partial f_1}{\partial w_2} = e^{(e^{w_1} + e^{2w_2})} \cdot 2e^{2w_2} + \sigma\left(e^{w_1} + e^{2w_2}\right)\left(1 - \sigma\left(e^{w_1} + e^{2w_2}\right)\right) 2e^{2w_2}$

5] c] continued. F2, forward



$\star\ w_1 = 1$
$w_2 = -1$

Table 2

equation	$\frac{\partial f_{w_2}}{\partial w_1}$	$\frac{\partial f_2}{\partial w_2}$
$x_1 = w_1$ $x_2 = w_2$	$\dot{x}_1 = 1$ $\dot{x}_2 = 0$	$\dot{x}_1 = 0$ $\dot{x}_2 = 1$
$x_3 = x_1 \cdot x_2$	$\dot{x}_3 = \dot{x}_1 x_2 + x_1 \dot{x}_2$	
$x_4 = \max(x_1, x_2)$	$\dot{x}_4 = \begin{cases} \dot{x}_1 & \text{if } x_1 \geq x_2 \\ \dot{x}_2 & \text{o.w.} \end{cases}$	
$x_5 = x_3 + x_4$	$\dot{x}_5 = \dot{x}_3 + \dot{x}_4$	

Consider $\frac{\partial f_2}{\partial w_1}$, $\dot{x}_1 = 1$, $\dot{x}_2 = 0$

$\dot{x}_3 = x_2$, $\dot{x}_4 = \dot{x}_1 = 1$

$\frac{\partial f_2}{\partial w_1} = \dot{x}_5 = x_2 + 1 = w_2 + 1$ (C)

Consider $\frac{\partial f_2}{\partial w_2}$, $\dot{x}_1 = 0$, $\dot{x}_2 = 1$

$\dot{x}_3 = x_1 = w_1$

$\dot{x}_4 = \dot{x}_1 = 0$

$\frac{\partial f_2}{\partial w_2} = \dot{x}_5 = w_1 + 0 = w_1$ (D)

from (A), $\frac{\partial f_1}{\partial w_1} = 47.3$

from (B), $\frac{\partial f_1}{\partial w_2} = 4.71$

from (C), $\frac{\partial f_2}{\partial w_1} = 0$

from (D), $\frac{\partial f_2}{\partial w_2} = 1$

So, Jacobian is:

$$\begin{pmatrix} 47.3 & 4.71 \\ 0 & 1 \end{pmatrix}$$

5] d] F1, backward (so start at $\dot{x}_9$).

$\dot{X}_9 = 1$, $\quad X_9 = X_7 \oplus X_8$, so pass gradient to both.

$\therefore \dot{x}_8 = \dot{x}_9 = 1$ and $\dot{x}_7 = \dot{x}_9 = 1$

Since x_6 goes onto 2 nodes, $\dot{x}_6$ must sum both gradients.

$$\therefore \dot{x}_6 = e^{x_6} \cdot \dot{x}_7 + \sigma(x_6)(1-\sigma(x_6))\dot{x}_8$$

$$= e^{x_6} + \sigma(x_6)(1-\sigma(x_6))$$

$\dot{x}_5 = \dot{x}_6$, and $\dot{x}_3 = \dot{x}_6$, since $x_6 = x_3 \oplus x_5$

$$\dot{x}_4 = e^{x_4} \cdot \dot{x}_5 = e^{x_4}\dot{x}_6 = e^{x_4}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

$$\dot{x}_2 = 2\dot{x}_4 = 2e^{x_4}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

$$\dot{x}_1 = e^{x_1} \cdot \dot{x}_3 = e^{x_1} \cdot \dot{x}_6 = e^{x_1}\left(e^{x_6} + \sigma(x_6)(1-\sigma(x_6))\right)$$

(a) $$\frac{\partial f_1}{\partial w_1} = \dot{x}_1 = e^{w_1}\left(e^{(e^{w_1}+e^{2w_2})} + \sigma(e^{w_1}+e^{2w_2})\left(1-\sigma(e^{w_1}+e^{2w_2})\right)\right)$$

$\boxed{\text{from } (\star\star\star) \text{ of part c, we know } x_6.}$

(b) $$\frac{\partial f_1}{\partial w_2} = \dot{x}_2 = 2e^{2w_2}\left(e^{(e^{w_1}+e^{2w_2})} + \sigma(e^{w_1}+e^{2w_2})\left(1-\sigma(e^{w_1}+e^{2w_2})\right)\right)$$

5) d) continued. Backward, F_2.

$$\mathring{x}_5 = 1, \quad \dot{x}_4 = \dot{x}_5 = 1, \quad \dot{x}_3 = \dot{x}_5 = 1$$

$$\dot{x}_1 = x_2 \dot{x}_3 + \dot{x}_4 = x_2 = w_2 + 1$$

$$\mathring{x}_2 = x_1 \dot{x}_3 + 0 = w_1$$

(c) $\frac{\partial f_2}{\partial w_1} = w_2 + 1 = \mathring{x}_1$, (d) $\frac{\partial f_2}{\partial w_2} = w_1 = \mathring{x}_2$.

Note a, b, c, d of part d = A, B, C, D of part c.

i.e. Forward & Backward produce same formula for each entry in ~~Gradient~~. Jacobian

So,

$$J = \begin{pmatrix} 47.3 & 4.71 \\ 0 & 1 \end{pmatrix}$$ "same as part c"

---

5) e) Yes! I do!