CS 7643 : HW 1

Name: Niranjan Thakurdesai

Solutions discussed with Srijon Sood, Karan Shah, Nupur Chatterji

1. Gradient Descent

Let $g(w) = f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2} \|w - w^{(t)}\|^2$

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

As w^* is the solution which minimizes the above function,

$\nabla g(w^*) = 0$

$\therefore \nabla f(w^{(t)}) + \lambda (w^* - w^{(t)}) = 0$

$\therefore \boxed{w^* = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)})}$

We can observe that the solution takes the same form as the gradient descent update rule. The update rule can be seen as minimizing the regularized Taylor approximation of $f(w)$ at each step. Also, $\boxed{\lambda = \frac{1}{\eta}}$

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

Consider the update rule:

$$w^{(2)} = w^{(1)} - \eta v_1$$
$$= -\eta v_1 \qquad (\because w^{(1)} = 0)$$
$$w^{(3)} = w^{(2)} - \eta v_2$$
$$= -\eta v_1 - \eta v_2$$
$$w^{(4)} = w^{(3)} - \eta v_3$$
$$= -\eta v_1 - \eta v_2 - \eta v_3$$

$\vdots$

$$w^{(t+1)} = -\eta\sum_{i=1}^{t}v_t \qquad (2)$$

Now consider the first term on the RHS in (1).

$$\sum_{t=1}^{T}\langle w^{(t)},v_t\rangle = \sum_{t=1}^{T}\langle -\eta\sum_{i=1}^{t-1}v_i, v_t\rangle \qquad \text{(From (2))}$$
$$= -\eta\sum_{t=1}^{T}\langle\sum_{i=1}^{t-1}v_i, v_t\rangle$$
$$= -\eta\left(\frac{\langle\sum_{t=1}^{T}v_t,\sum_{t=1}^{T}v_t\rangle}{2} - \frac{\sum_{t=1}^{T}\|v_t\|^2}{2}\right)$$
$$= -\frac{\eta}{2}\left\|\sum_{t=1}^{T}v_t\right\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2$$
$$= -\frac{1}{2\eta}\|w^{(T+1)}\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 \qquad (3)$$

(From (2))

For the second term on the RHS in ①,

$$\left\langle w^*, \sum_{t=1}^{T} v_t \right\rangle = \left\langle w^*, -\frac{1}{\eta} w^{(T+1)} \right\rangle \quad \text{(From ②)}$$

$$= -\frac{1}{\eta} \langle w^*, w^{(T+1)} \rangle \qquad \text{— ④}$$

From ①, ③ and ④,

$$\cancel{\sum_{t=1}^{T} \langle w^{(t)}, v_t \rangle - \langle w^*,}$$

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

The second term on the RHS is already on the RHS in the inequality to be proven. So, we just need to prove that

$$-\frac{1}{2\eta} \|w^{(T+1)}\|^2 + \frac{1}{\eta} \langle w^*, w^{(T+1)} \rangle \le \frac{\|w^*\|^2}{2\eta}$$

~~LHS~~ Now,

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

$$= -\frac{1}{2\eta} \langle w^{(T+1)}, w^{(T+1)} \rangle + \frac{1}{2\eta} \langle w^*, w^{(T+1)} \rangle + \frac{1}{2\eta} \langle w^*, w^{(T+1)} \rangle$$

$$\langle w^*, w^{(T+1)} \rangle \le \langle w^{(T+1)}, w^{(T+1)} \rangle, \quad \langle w^*, w^{(T+1)} \rangle \le \langle w^*, w^* \rangle = \|w^*\|^2$$

Thus,

$$-\frac{1}{2\eta} \|w^{(T+1)}\|^2 + \frac{1}{\eta} \langle w^*, w^{(T+1)} \rangle \le \frac{1}{2\eta} \|w^*\|^2 \qquad \text{— ⑥}$$

as $-\frac{1}{2\eta} \langle w^{(T+1)}, w^{(T+1)} \rangle + \frac{1}{2\eta} \langle w^*, w^{(T+1)} \rangle \le 0$ and $\frac{1}{2\eta} \langle w^*, w^{(T+1)} \rangle \le \frac{1}{2\eta} \|w^*\|^2$

From ⑤ and ⑥,

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

3. Using the first order definition of convexity,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \text{Domain}(f)$$

Substituting $x = w^{(t)}$ and $y = w^{(t)}$,

$$\cancel{f(w^{(t)}) \geq f(w^*) +}$$

$$f(w^*) \geq f(w^{(t)}) + \nabla f(w^{(t)})^T (w^* - w^{(t)})$$

$$\therefore f(w^{(t)}) - f(w^*) \leq \nabla f(w^{(t)})^T (w^{(t)} - w^*)$$

$$= \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

$$\therefore \frac{1}{T} \sum_{t=1}^{T} \left(f(w^{(t)}) - f(w^*) \right) \leq \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

$$\therefore \left(\frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) \right) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle \quad \text{—— } ①$$

Using the definition of convexity,

$$\cancel{f(\bar{w}) = f\left(\frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) \right)}$$

$$f(\bar{w}) = f\left(\frac{1}{T} \sum_{t=1}^{T} w^{(t)} \right) \leq \frac{1}{T} \sum_{t=1}^{T} f(w^{(t)}) \quad \text{——————— } ②$$

From ① and ②,

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

Using the result from part 2,

$$f(\bar{w}) - f(w^*) \leq \frac{\|w^*\|^2}{2\eta T} + \frac{\eta}{2T} \sum_{t=1}^{T} \|\nabla f(w^{(t)})\|^2$$

$$\leq \frac{B^2}{2\eta T} + \frac{\eta T \rho^2}{2T} \quad \text{(using the given bounds)}$$

$$= \frac{B^2}{2T\sqrt{\frac{B^2}{\rho^2 T}}} + \frac{\rho^2}{2}\sqrt{\frac{B^2}{\rho^2 T}}$$

$$= \frac{B\rho}{2\sqrt{T}} + \frac{B\rho}{2\sqrt{T}} = \frac{B\rho}{\sqrt{T}} \Rightarrow \text{Upper bound for } f(\bar{w}) - f(w^*) \propto \frac{1}{\sqrt{T}}$$

4. Let $w^{(1)} = 0$

$$f(w) = \frac{1}{2}(w-2)^2 + \frac{1}{2}(w+1)^2$$

$$= w^2 - w + \frac{5}{2}$$

$$f(w^{(1)}) = f(0) = \frac{5}{2}$$

Suppose the second term is picked for the next gradient descent step.

$$\therefore w^{(2)} = w^{(1)} - \eta(w^{(1)} + 1)$$

$$= -\eta \quad (\because w^{(1)} = 0)$$

$$f(w^{(2)}) = f(-\eta)$$

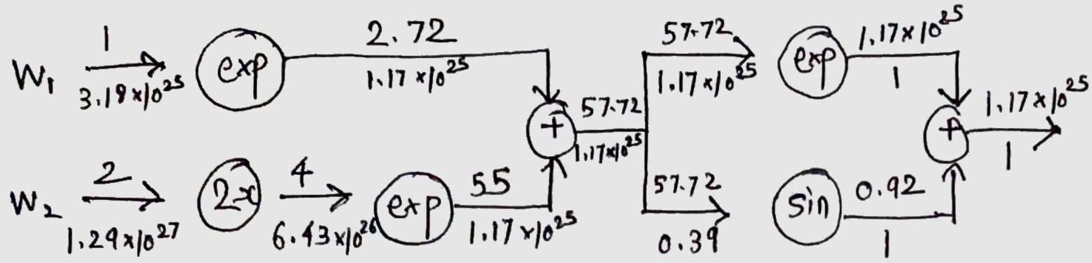
$$= \eta^2 + \eta + \frac{5}{2}$$

Thus,

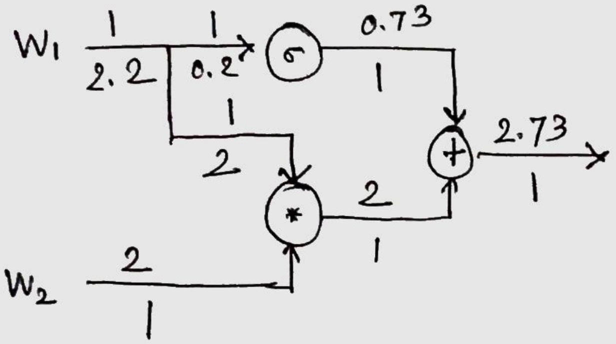
$$f(w^{(2)}) > f(w^{(1)})$$

$\Rightarrow$ SGD is not guaranteed to decrease the overall loss function in every iteration.

5. Forward pass and reverse mode auto-differentiation for f_1:



Forward pass and reverse mode auto-differentiation for f_2:



→ Using backward mode auto-diff.,

$$\boxed{\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} 3.18 \times 10^{25} & 1.29 \times 10^{27} \\ 2.2 & 1 \end{bmatrix}}$$

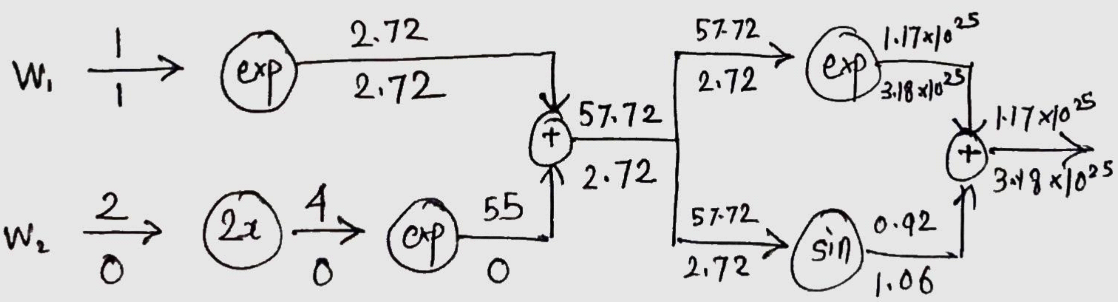
$$\boxed{\vec{f}(w_1, w_2) = [1.17 \times 10^{25}, 2.73]}$$

Using numerical differentiation,

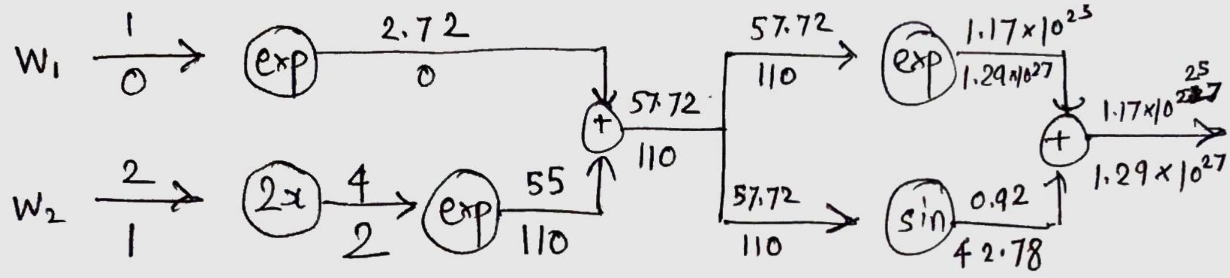
$$\boxed{\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} 2.16 \times 10^{25} & 1.57 \times 10^{27} \\ 2.2 & 1 \end{bmatrix}}$$

→ computed using forward differences

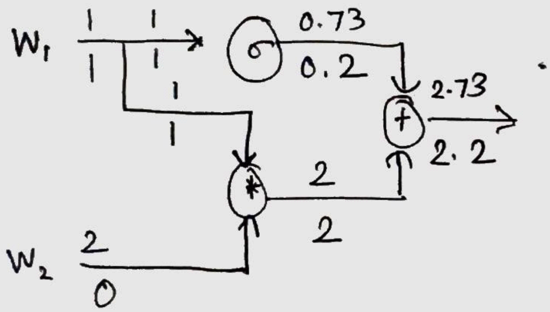
Forward mode auto-differentiation for f_1 wrt w_1:



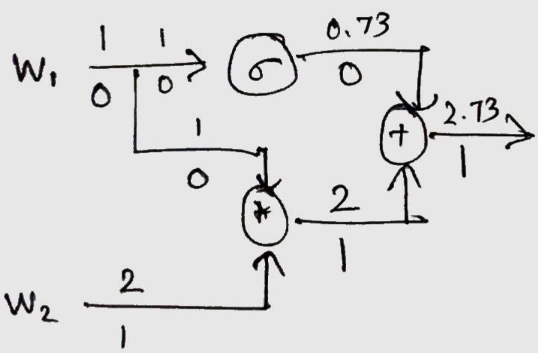
Forword mode auto-differentiation for f_1 wrt w_2:



Forword mode auto-differentiation for f_2 wrt w_1:



Forward mode auto-differentiation for f_2 wrt w_2:



$\Rightarrow$ Using forword mode auto-differentiation,

$$\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} 3.18 \times 10^{25} & 1.29 \times 10^{27} \\ 2.2 & 1 \end{bmatrix}$$

Backword mode ~~a~~ auto-differentiation is shown ~~on~~ in the first two graphs of the previous page. Using that,

$$\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} 3.18 \times 10^{25} & 1.29 \times 10^{27} \\ 2.2 & 1 \end{bmatrix}$$