CS7643: Deep Learning
Fall 2019
HW1 Solutions

Gaurav Pande

September 26, 2019

1 Gradient Descent:

Part1:

1.1 we have given the function:-

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

To optimize this, lets take the first order derivative and assume

$$w - w^t = x$$

$$f(x) = \arg\min f(w^t) + \langle x, \nabla f(w^t) \rangle + \frac{\lambda}{2} |x|^2$$

$$f'(x) = \nabla f(w^t) + \frac{\lambda}{2} \times 2 |x|$$

$$f'(x) = \nabla f(w^t) + \lambda |x|$$

$$f'(x) = 0$$

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

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

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

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

Now, we also know that:

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

If we compare the eqn (1) with above, we can say that

$$\boxed{\frac{1}{\lambda} = \eta}$$

from this we can say that learning rate is inversily proportional to the regularization.

What does it mean:

Regularization means controlling overfitting, and one of the way we can relate regularization with overfitting is by training the model over multiple iteration. The more iteration we have the less overfitting we have.

Now, our relationship tells us that, if we have:

(1) large η, then in few iteration (or with less regularization you can optimize your model.

(2) low η, then in large iteration you can optimize your model.

Part2:

(2) We have given:

a sequence of vectors $v_1, v_2 \dots v_T$

update eq^n

$$w^{t+1} = w^t - \eta v_t \quad \text{— (1)}$$

$$w^{(1)} = 0 \quad \text{— (2)}$$

To show:-

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

Fundamentals theorem of opt.?

We can write $\langle w^t - w^*, v_t \rangle$ as

$$\langle w^t - w^*, v_t \rangle = \frac{1}{\eta} \langle w^t - w^*, \eta v_t \rangle \quad \left[\text{multiply \& Dividing } \eta\right]$$

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

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

Now, we can sum the equation over t.

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

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

The first term on the right hand side collapses to

$$\|w^t - w^*\|^2 - \|w^{T+1} - w^*\|^2$$

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

$$\leq \frac{1}{2\eta} \|w^1 - w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

$$\boxed{\leq \frac{1}{2\eta} \|w^*\|^2 + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2}$$

Part3:

(3) To show $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

we have all the weight as $w_1, w_2 \ldots w^T$

let calculate their mean $\bar{w}$

$$\bar{w} = \frac{w_1 + w_2 + w_3 + \cdots w^T}{T}$$

$$\boxed{\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}}$$

Next:

To prove

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

let's first Jenson's inequality

it states that:

If $g(x)$ is a convex function on R_x, and $E(g(x))$ and $g(E(x))$ are finite, than

$$E(g(x)) \ge g(E(X))$$

How, do we come up with above eqn

we know that for every random variable X

$$V(X) = Ex^2 - (Ex)^2 \ge 0$$

$$\boxed{Ex^2 \ge (Ex)^2.}$$

Now, take any conven function, $g(x) = x^2$ [for e.g]

Jenson's inequality states that

$$\boxed{E[g(x)] \ge g(E(X))}$$

using the same above inequality, we can write that:

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

$$\le \frac{1}{T}\sum_{t=1}^{T} (f(w^t)) - f(w^*)$$

$$= \frac{1}{T}\sum_{t=1}^{T} (f(w^t) - f(w^*))$$

for every t, because of convexity, we have

$$f(w^t) - f(w^*) \le \langle w^t - w^*, \nabla f(w^t)\rangle$$

therefore

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

We have proved previously

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

Upperbounding $\|w^*\| \to B$, $\|v_t\| \to \rho$, $\eta \to \sqrt{\frac{B^2}{\rho^2 T}}$, dividing by T

$$\le \frac{1}{T}\left(\frac{B^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\rho^2\right)$$

$$\le \frac{1}{T}\left(\frac{B^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\rho^2\right)$$

$$\le \frac{1}{T}\left(\frac{B^2 + 2\eta^2 \sum_{t=1}^{T}\rho^2}{4\eta}\right)$$

$$\le \frac{1}{T}\frac{\left(2B^2 + \frac{2B^2}{\rho^2 T}\sum_{t=1}^{T}\rho^2\right)}{4 \times \left(\sqrt{\frac{B^2}{\rho^2 T}}\right)}$$

$$\propto \frac{1}{\sqrt{T}}$$

Part4:

④ Given the objective function :

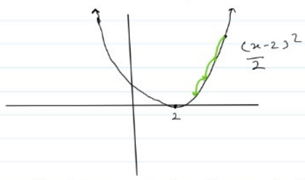
$$f(w) = \frac{1}{2}(w-2)^2 + \frac{1}{2}(w+1)^2$$

Term1 = $\frac{1}{2}(w-2)^2$

Term2 = $\frac{1}{2}(w+1)^2$

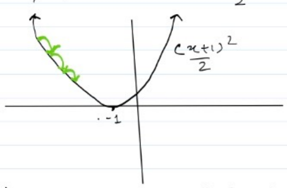
Prob. of chosing the term = $\frac{1}{2}$

Let's say we choose $\frac{1}{2}(w-2)^2$



Since it is a convex function, and it is given that η is small enough, that every update result in improvement,
Hence it is bound to converge at some point with small step-size.

similarly if we choose $\frac{1}{2}(w+1)^2$



with small step-size this is also bound to converge.

2: Automatic Differentiation
a) Computation Graphs

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

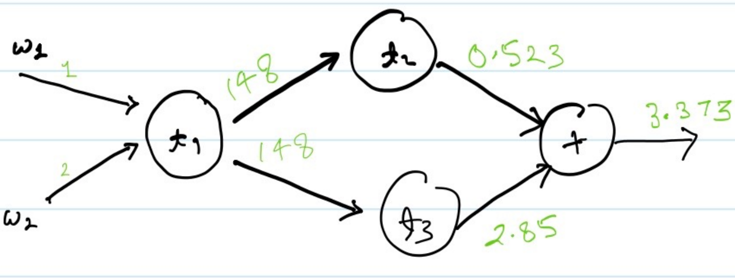
$e^{w_1 + 2w_2} = t_1 \Rightarrow t_1 = e^{1+4} = e^5 = 148.4$

$f_1 = e^{t_1} + \sin(t_1)$

$\sin(t_1) = t_2 \Rightarrow t_2 = \sin(e^5) = 0.523$

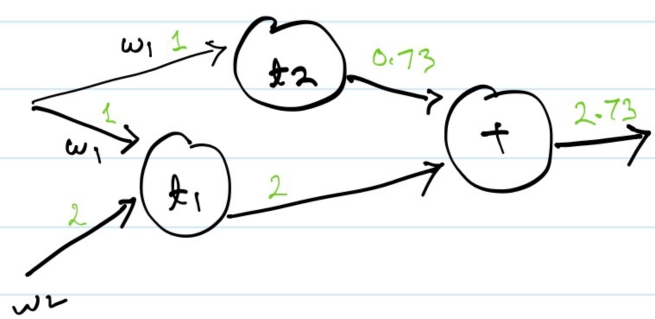
$e^{t_1} = t_3 \Rightarrow t_3 = e^{e^5} = 2.85$

$f_1 = t_2 + t_3 \Rightarrow e^{e^5} + \sin(e^5) = 3.373$



$f_2 = w_1 w_2 + \sigma(w_1)$

$w_1 w_2 = t_1$	$t_1 = 2$
$\sigma(w_1) = t_2$	$t_2 = 0.73$
$f_2 = t_1 + t_2$	$f_2 = 2.73$



b) Numerical Differentiation

(b) $f_1(\omega_1,\omega_2) = e^{e^{\omega_1}+e^{2\omega_2}} + \sin(e^{\omega_1}+e^{2\omega_2})$ | $f_2(\omega_1,\omega_2) = \omega_1\omega_2 + \sigma(\omega_1)$ | $(1,2)$

$$\frac{\partial f_1}{\partial \omega_1} = \lim_{\Delta\omega\to 0} \frac{f(\omega_1+\Delta\omega, \omega_2) - f(\omega_1,\omega_2)}{\Delta\omega}$$

$$= \lim_{\Delta\omega\to 0} \frac{\left[e^{e^{\omega_1+\Delta\omega}+e^{2\omega_2}} + \sin(e^{\omega_1+\Delta\omega}+e^{2\omega_2})\right] - \left[e^{e^{\omega_1}+e^{2\omega_2}} + \sin(e^{\omega_1}+e^{2\omega_2})\right]}{\Delta\omega}$$

$$= \lim_{\Delta\omega\to 0} \frac{\left[e^{e^{1.01}+e^4} + \sin(e^{1.01}+e^4) - e^{e^1+e^4} - \sin(e^1+e^4)\right]}{0.01}$$

$$= 2.1608 \times 10^{25}$$

$$\frac{\partial f_1}{\partial \omega_2} = \lim_{\Delta\omega\to 0} \frac{f(\omega_1, \omega_2+\Delta\omega) - f(\omega_1,\omega_2)}{\Delta\omega}$$

$$= \lim_{\Delta\omega\to 0} \left[e^{e^{\omega_1}+e^{2\omega_2+2\Delta\omega}} + \sin(e^{\omega_1}+e^{2\omega_2+2\Delta\omega}) - e^{e^{\omega_1}+e^{2\omega_2}} - \sin(e^{\omega_1}+e^{2\omega_2})\right] / \Delta\omega$$

$$= \lim_{\Delta\omega\to 0} (e^{e^1+e^{4.02}} + \sin(e^1+e^{4.02}) - e^{e^1+e^4} - \sin(e^1+e^4)) / 0.01$$

$$= 1.57 \times 10^{27}$$

$$\frac{\partial f_2}{\partial \omega_1} = \lim_{\Delta\omega\to 0} ((\omega_1+\Delta\omega)\omega_2 + \sigma(\omega_1+\Delta\omega) - (\omega_1\omega_2) - \sigma(\omega_1))/\Delta\omega$$

$$= \lim_{\Delta\omega\to 0} ((1.01)(2) + \sigma(1.01) - 2 - \sigma(1))/0.01$$

$$= \lim_{\Delta\omega\to 0} (0.733 + 2.02 - 2 - 0.731)/0.01$$

$$= (0.002 + 0.02)/0.01$$

$$= 0.022/0.01$$

$$= 2.2$$

$$\frac{\partial f}{\partial \omega_2} = \lim_{\Delta\omega\to 0} (\omega_1(\omega_2+\Delta\omega) + \sigma(\omega_1) - \omega_1\omega_2 - \sigma(\omega_1))/\Delta\omega$$

$$= \lim_{\Delta\omega\to 0} (1\times(2+0.01) + \sigma(1) - 1\times 2 - \sigma(1))/0.01$$

$$= 0.01/0.01$$

$$= 1$$

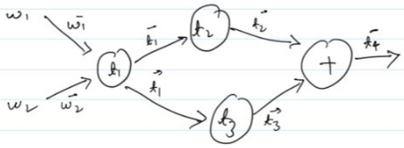
c) Forward Automatic Differentiation

ⓒ $f_1 = e^{e^{w_1}+e^{2w_2}} + \sin(e^{w_1}+e^{2w_2})$

$t_1 = e^{w_1} + e^{2w_2}$

$f_1 = \underbrace{e^{t_1}}_{t_2} + \underbrace{\sin t_1}_{t_3}$

$f_1 = t_2 + t_3 = t_4$



$$\frac{\partial f_1}{\partial \vec{w}} = \frac{\partial (f_1)}{\partial (w_1, w_2)} = \left| \frac{\partial f_1}{\partial w_1} \quad \frac{\partial f_1}{\partial w_2} \right|$$

jacobian

$$\bar{w}_1 = \frac{\partial w_1}{\partial w_1} = 1$$

$$\bar{w}_2 = \frac{\partial w_2}{\partial w_2} = 1$$

$$\vec{t}_1 = \frac{\partial (e^{w_1} + e^{2w_2})}{\partial w_1} = e^{w_1} = 2.713$$

$$\frac{\partial (e^{w_1} + e^{2w_2})}{\partial w_2} = 2e^{2w_2} = 109.19$$

$$\bar{t}_2 = \frac{\partial}{\partial w_1}(e^{t_1}) = \frac{\partial}{\partial t_1}(e^{t_1}) \times \frac{\partial t_1}{\partial w_1} = e^{t_1} \times e^{w_1} = \bar{t}_1 e^{w_1} = e^{w_1} \times e^{w_1} = 7.38$$

$$\frac{\partial}{\partial w_2}(e^{t_1}) = \frac{\partial}{\partial t_1}(e^{t_1}) \times \frac{\partial t_1}{\partial w_2} = e^{t_1} \times 2e^{2w_2} = 2\bar{t}_1 e^{2w_2} = 2 \times 2e^{2w_2} \times e^{2w_2} = 11923.83$$

$$\bar{t}_3 = \frac{\partial}{\partial w_1}(\sin t_1) = \frac{\partial}{\partial t_1} \sin t_1 \times \frac{\partial t_1}{\partial w_1} = \cos t_1 \times \frac{\partial}{\partial w_1} e^{w_1} + e^{2w_2} = \cos t_1 \times e^{w_1} = \cos e^{w_1} \times e^{w_1} = 2.7152$$

$$= \frac{\partial}{\partial w_2}(\sin t_1) = \frac{\partial}{\partial t_1} \sin t_1 \times \frac{\partial t_1}{\partial w_2} = \cos t_1 \frac{\partial}{\partial w_2}(e^{w_1} + e^{2w_2}) = 2\cos t_1\, e^{2w_2} = 2\cos 2e^{2w_2} e^{2w_2} = -35.904$$

$$\vec{t}_4 = \frac{\partial}{\partial w_1}(t_2 + t_3) = \frac{\partial}{\partial w_1} t_2 + \frac{\partial t_3}{\partial w_1} = \bar{t}_2 + \bar{t}_3 = e^{w_1} \times e^{w_1} + \cos e^{w_1} e^{w_1} = 10.10$$

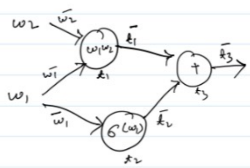
$$= \frac{\partial}{\partial w_2}(t_2 + t_3) = \frac{\partial}{\partial w_2} t_2 + \frac{\partial t_3}{\partial w_2} = \bar{t}_2 + \bar{t}_3 = 2 \times 2e^{2w_2} \times e^{2w_2} + 2\cos 2e^{2w_2} e^{2w_2} = 11887.9$$

$f_2 = w_1 w_2 + \sigma(w_1)$

$t_1 = w_1 w_2$

$t_2 = \sigma(w_1)$

$t_3 = t_1 + t_2 = f_2$



$\bar{w_1} = \frac{\partial w_1}{\partial w_1} = 1, \quad \frac{\partial w_1}{\partial w_2} = 0$

$\bar{w_2} = \frac{\partial w_2}{\partial w_2} = 1, \quad \frac{\partial w_2}{\partial w_1} = 0$

$\bar{t_1} = \frac{\partial w_1 w_2}{\partial w_1} = w_2$

$\frac{\partial w_1 w_2}{\partial w_2} = w_1$

$\bar{t_2} = \frac{\partial \sigma(w_1)}{\partial w_1} = \sigma(w_1)(1-\sigma(w_1))$

$= \frac{\partial \sigma(w_1)}{\partial w_2} = 0$

$\bar{t_3} = \frac{\partial}{\partial w_1}(t_1 + t_2) = \frac{\partial t_1}{\partial w_1} + \frac{\partial t_2}{\partial w_1} = \bar{t_1} + \bar{t_2} = w_2 + \sigma(w_1)(1-\sigma(w_1))$

$= \frac{\partial}{\partial w_2}(t_1 + t_2) = \frac{\partial t_1}{\partial w_2} + \frac{\partial t_2}{\partial w_2} = \bar{t_1} + \bar{t_2} = w_1 = 1$

d) Backward Automatic Differentiation

a.

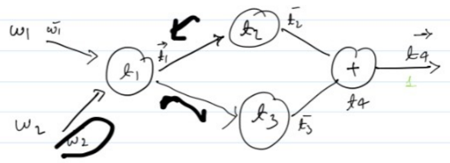
$$f_1 = e^{e^{w_1}+e^{2w_2}} + \sin(e^{w_1}+e^{2w_2})$$

$$t_1 = e^{w_1+2w_2}$$

$$t_2 = e^{t_1}$$

$$t_3 = \sin(t_1)$$

$$t_4 = t_2 + t_3$$



In reverse mode, we go from the right

$$\vec{t_4} = \frac{\partial f}{\partial t_4} = 1$$

$$\vec{t_2} = \frac{\partial f}{\partial t_2} = \frac{\partial f}{\partial t_4} \times \frac{\partial t_4}{\partial t_2} = 1 \times \frac{\partial}{\partial t_2}(t_2+t_3) = 1$$

$$\bar{t_3} = \frac{\partial f}{\partial t_3} = \frac{\partial f}{\partial t_4} \times \frac{\partial t_4}{\partial t_3} = 1 \times \frac{\partial}{\partial t_3}(t_2+t_3) = 1$$

$$\bar{t_1} = \frac{\partial f}{\partial t_1} = \frac{\partial f}{\partial t_3} \times \frac{\partial t_3}{\partial t_1} = 1 \times \frac{\partial}{\partial t_1}(\sin t_1) = \cos t_1$$

$$\bar{t_1} = \frac{\partial f}{\partial t_1} = \frac{\partial f}{\partial t_2} \times \frac{\partial t_2}{\partial t_1} = 1 \times \frac{\partial}{\partial t_1} e^{t_1} = e^{t_1}$$

$$\bar{w_1} = \frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial t_1} \times \frac{\partial t_1}{\partial w_1} = \cos t_1 \times \frac{\partial}{\partial w_1}(e^{w_1}+e^{2w_2}) = e^{w_1}\cos t_1 = e^{w_1} \cos e^{w_1}+e^{2w_2}$$

$$= \frac{\partial f}{\partial t_1} \times \frac{\partial t_1}{\partial w_1} = e^{t_1} \times \frac{\partial}{\partial w_1}(e^{w_1}+e^{2w_2}) = e^{w_1} \times e^{t_1} = e^{w_1} \times e^{e^{w_1}+e^{2w_2}}$$

$$\bar{w_2} = \frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial t_1} \times \frac{\partial t_1}{\partial w_2} = \cos t_1 \times \frac{\partial}{\partial w_2}(e^{w_1}+e^{2w_2}) = \cos t_1 \times 2e^{2w_2} = \cos(e^{w_1}+e^{2w_2}) \times 2e^{2w_2}$$

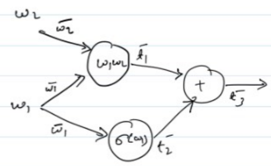
$$= e^{t_1} \times \frac{\partial}{\partial w_2}(e^{w_1}+e^{2w_2}) = e^{t_1} \times 2e^{2w_2} = e^{e^{w_1}+e^{2w_2}} \times 2e^{2w_2}$$

$f_2 = w_1 w_2 + \sigma(w_1)$

$t_1 = w_1 w_2$

$t_2 = \sigma(w_1)$

$t_3 = t_1 + t_2$



$$\bar{t}_3 = \frac{\partial f}{\partial t_3} = \frac{\partial t_3}{\partial t_3} = 1$$

$$\bar{t}_1 = \frac{\partial f}{\partial t_1} = \frac{\partial f}{\partial t_3} \times \frac{\partial t_3}{\partial t_1} = 1 \times \frac{\partial}{\partial t_1}(t_1 + t_2) = 1$$

$$\bar{t}_2 = \frac{\partial f}{\partial t_2} = \frac{\partial f}{\partial t_3} \times \frac{\partial t_3}{\partial t_2} = 1 \times \frac{\partial}{\partial t_2}(t_1 + t_2) = 1$$

$$\bar{w}_1 = \frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial t_2} \times \frac{\partial t_2}{\partial w_1} = 1 \times \frac{\partial}{\partial w_1}\sigma(w_1) = \sigma(w_1)(1 - \sigma(w_1))$$

$$= \frac{\partial f}{\partial t_1} \times \frac{\partial t_1}{\partial w_1} = 1 \times \frac{\partial}{\partial w_1}(w_1 w_2) = w_2 \quad = 2$$

$$\bar{w}_2 = \frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial t_1} \times \frac{\partial t_1}{\partial w_2} = 1 \times \frac{\partial}{\partial w_2}(w_1 w_2) = w_1 \quad = 1$$

(e) Yes, it is pretty easy with software libraries.

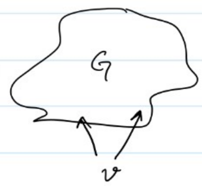
e): Yes it is pretty easy with software libraries

3 Directed Acyclic Graphs:

(4): If a graph G is a DAG, then G has a topological ordering

Proof: (By induction)

Base Case: If we have only 1 node, than G has a topological ordering.



Hypothesis: If G' has $\leq n$ nodes, than G' has topological ordering.

Consider G with n+1 nodes,
we know G is a DAG, & we added one extra node v, we add v with no incoming edges.
G- {v} is a DAG, since deleting v cannot create any cycles.

So, In short:

- Place v first; then add topological ordering of G- {v}.
- v has no incoming edges, hence first point is valid.

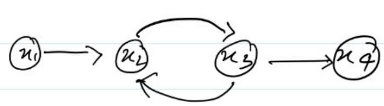
By induction the lemma is proven.

(5): If G has a topological ordering, than G is a DAG.

Proof: By contradiction

Let's suppose G has topological ordering $x_1, x_2, x_3 \ldots x_n$

Suppos G has a directed cycle like below



from the above figure x_2 has lower index than x_3 since $2 < 3$

Does Node x_2 should come before x_3, if they are topological order.

But the edge $x_3 \to x_2$, and $x_1, x_2 \ldots x_n$ is a topological order, we must have $3 < 2$, which is a contradiction.

So G has no cycle. G is a DAG.