1. a)

1. a) By Chain Rule

$$\overline{h^{(t)}} = \overline{i^{(t+1)}}\frac{\partial i^{(t+1)}}{\partial h^{(t)}} + \overline{g^{(t+1)}}\cdot\frac{\partial g^{(t+1)}}{\partial h^{(t)}} + \overline{f^{(t+1)}}\frac{\partial f^{(t+1)}}{\partial h^{(t)}} + \overline{o^{(t+1)}}\frac{\partial o^{(t+1)}}{\partial h^{(t)}}$$

$$= \overline{i^{(t+1)}}\frac{\partial i^{(t+1)}}{\partial W_{ix}x^{(t+1)}+W_{ih}h^{(t)}}\cdot\frac{\partial W_{ix}x^{(t+1)}+W_{ih}h^{(t)}}{\partial h^{(t)}}$$

$$+ \overline{g^{(t+1)}}\frac{\partial g^{(t+1)}}{\partial W_{gx}x^{(t+1)}W_{gh}h^{(t)}}\frac{\partial W_{gx}x^{(t+1)}+W_{gh}h^{(t)}}{\partial h^{(t)}}$$

$$+ \overline{f^{(t+1)}}\frac{\partial f^{(t+1)}}{\partial W_{fx}x^{(t+1)}W_{fh}h^{(t)}}\frac{\partial W_{fx}x^{(t+1)}+W_{fh}h^{(t)}}{\partial h^{(t)}}$$

$$+ \overline{o^{(t+1)}}\frac{\partial o^{(t+1)}}{\partial W_{ox}x^{(t+1)}W_{oh}h^{(t)}}\frac{\partial W_{ox}x^{(t+1)}+W_{oh}h^{(t)}}{\partial h^{(t)}}$$

$$= \overline{i^{(t+1)}}\,\sigma'(W_{ix}x^{(t+1)}+W_{ih}h^{(t)})W_{ih} + \overline{g^{(t+1)}}\,\tanh'(W_{gx}x^{(t+1)}+W_{gh}h^{(t)})\cdot W_{gh}$$

$$+ \overline{f^{(t+1)}}\,\sigma'(W_{fx}x^{(t+1)}+W_{fh}h^{(t)})W_{fh} + \overline{o^{(t+1)}}\,\sigma'(W_{ox}x^{(t+1)}+W_{oh}h^{(t)})W_{oh}$$

$$\overline{c^{(t)}} = \overline{h^{(t)}}\frac{\partial h^{(t)}}{\partial \tanh(c^{(t)})}\cdot\frac{\partial \tanh(c^{(t)})}{\partial c^{(t)}} + \overline{c^{(t+1)}}\frac{\partial c^{(t+1)}}{\partial c^{(t)}}$$

$$= \overline{h^{(t)}}\,o^{(t)}\cdot\tanh'(c^{(t)}) + \overline{c^{(t+1)}}\cdot f^{(t+1)}$$

$$\overline{g^{(t)}} = \overline{c^{(t)}}\frac{\partial c^{(t)}}{\partial g^{(t)}} = \overline{c^{(t)}}\cdot i^{(t)}$$

$$\overline{o^{(t)}} = \overline{h^{(t)}}\frac{\partial h^{(t)}}{\partial o^{(t)}} = \overline{h^{(t)}}\tanh(c^{(t)})$$

$$\overline{f^{(t)}} = \overline{c^{(t)}}\frac{\partial c^{(t)}}{\partial f^{(t)}} = \overline{c^{(t)}}\cdot c^{(t-1)}$$

$$\overline{i^{(t)}} = \overline{c^{(t)}}\cdot\frac{\partial c^{(t)}}{\partial i^{(t)}} = \overline{c^{(t)}}\cdot g^{(t)}$$

b)

b) $$\overline{W_{ix}} = \sum_t \overline{i^{(t)}}\frac{\partial i^{(t)}}{\partial W_{ix}}$$

$$= \sum_t \overline{i^{(t)}}\frac{\partial i^{(t)}}{\partial W_{ix}x^{(t)}+W_{ih}h^{(t-1)}}\cdot\frac{\partial W_{ix}x^{(t)}+W_{ih}h^{(t-1)}}{\partial W_{ix}}$$

$$= \sum_t \overline{i^{(t)}}\,\sigma'(W_{ix}x^{(t)}+W_{ih}h^{(t-1)})\cdot x^{(t)}$$

2. a)

2 a) Since we know the hiden unit dimension is H, then according to the function we know that

$\dim(W_{in}^T x^{(i,j)}) = \dim(W_{iv}^T h^{(i-1,j)}) = \dim(W_{h}^T h^{(i,j-1)}) = H$.

According to question we know input dimension is D, then $\dim(x^{(i,j)}) = D$. To let $\dim(W_{in}^T x^{(i,j)}) = H$, we should let $\dim(W_{in}^T) = H \times D$.

Samilarly, we know $\dim(h^{(i-1,j)}) = \dim(h^{(i,j-1)}) = H$, then $\dim(W_{iv}^T) = \dim(W_{h}^T) = H \times H$

So the total number of weight is

$\dim(W_{in}^T) + \dim(W_{iv}^T) + \dim(W_{h}^T) = \boxed{H \times D + 2(H \times H)}$

Since we know the total number of weight is $HD + 2(HH)$; and for the function of $h^{(i,j)}$, each h need to compute $(HD + 2(HH))$ times, then for each hidden unit it takes $O(HD + 2(HH))$ time. According to the question, we know that $\dim(\text{grid}) = G \times G$, then for the total time, it should be

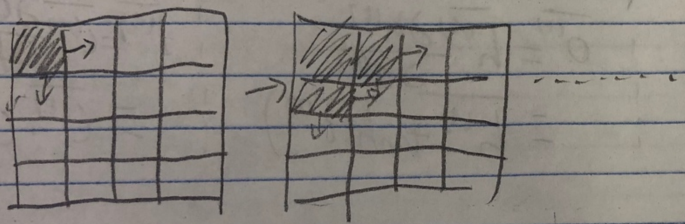
$O(HD + 2(HH)) \cdot GG$

$= \boxed{O((HD + 2H^2) G^2)}$

b)

b) It should be $O(2G-1)$ steps.

If we know $h^{(i,j)}$, then we can compute it's adjacents. If computing $h^{(i,j)}$ takes 1 time, then for all hidden activations, it should be $O(2G-1)$ steps.



c)

Advantage: MDRNNs are more robust to input warping than convolution networks (Multi-Dimensional Recurrent Neural Networks)[1] and capable of modeling long-term sequential dependencies. (Recent Advances in Recurrent Neural Networks)[2]

Disadvantage: It has higher computational complexity science it uses LSTM. (Recent Advances in Recurrent Neural Networks)

3. a)

3, a)

According to question, we know

$$s^{(k+1)} = f(s^{(k)})$$
$$= (\theta^{(k+1)}, p^{(k+1)})$$
$$= (\theta^{(k)} + p^{(k+1)}, \beta p^{(k)} - \alpha \nabla J(\theta^{(k)}))$$

We need $s^{(k)}$, and $s^{(k)} = f^{-1}(s^{(k+1)})$

$$= f^{-1}((\theta^{(k+1)}, p^{(k+1)}))$$

according to question we also know $s^{(k)} = (\theta^{(k)}, p^{(k)})$

So, from the function of $\theta^{(k+1)}, p^{(k+1)}$, we can get

$$\theta^{(k)} = \theta^{(k+1)} - p^{(k+1)}$$
$$p^{(k)} = \frac{p^{(k+1)} + \alpha \nabla J(\theta^{(k)})}{\beta}$$

1. https://arxiv.org/pdf/0705.2011.pdf
2. https://arxiv.org/pdf/1801.01078.pdf

b)

b) $\det \frac{\partial s^{(k+1)}}{\partial s^{(k)}}$

According to question, we can get .

$$\frac{\partial p^{(k+1)}}{\partial p^{(k)}} = \beta \qquad \frac{\partial \theta^{(k+1)}}{\partial p^{(k)}} = \beta$$

$$\frac{\partial p^{(k+1)}}{\partial \theta^{(k)}} = -\alpha \frac{\partial \nabla J(\theta^{k})}{\partial \theta^{(k)}} \qquad \frac{\partial \theta^{(k+1)}}{\partial \theta^{(k)}} = 1 - \alpha \frac{\partial \nabla J(\theta^{(k)})}{\partial \theta^{(k)}}$$

If we assume $\frac{\partial \nabla J(\theta^{(k)})}{\partial \theta^{(k)}} = 0$, then

$$\det \frac{\partial s^{(k+1)}}{\partial s^{(k)}} = \det \begin{pmatrix} \frac{\partial p^{(k+1)}}{\partial p^{(k)}} & \frac{\partial p^{(k+1)}}{\partial \theta^{(k)}} \\ \frac{\partial \theta^{(k+1)}}{\partial p^{(k)}} & \frac{\partial \theta^{(k+1)}}{\partial \theta^{(k)}} \end{pmatrix}$$

$$= \det \begin{pmatrix} \beta & 0 \\ \beta & 1 \end{pmatrix}$$

0 is zero matrix and I is identity matrix.

$$= \beta$$