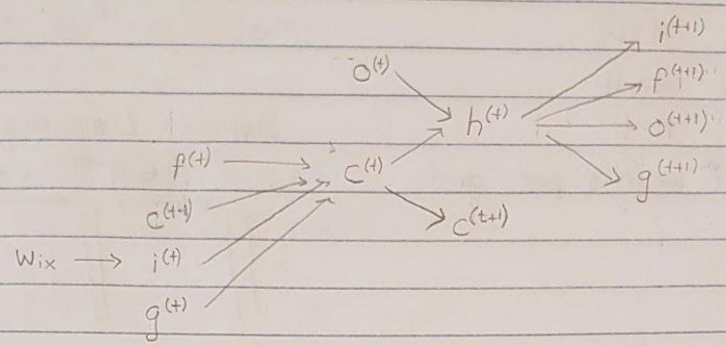
CSC421 HW4 2019年05月28日

1. a)



Fact: $\frac{d}{dx}\tanh(x) = 1-\tanh^2(x)$

$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$ } plug in accordingly.

$$\bar{h}^{(t)} = \bar{i}^{(t+1)}\sigma'(W_{ix}X^{(t)} + W_{ih}h^{(t)})W_{ih}$$
$$+ \bar{f}^{(t+1)}\sigma'(W_{fx}X^{(t)} + W_{fh}h^{(t)})W_{fh}$$
$$+ \bar{o}^{(t+1)}\sigma'(W_{ox}X^{(t)} + W_{oh}h^{(t)})W_{oh}$$
$$+ \bar{g}^{(t+1)}\frac{d}{dx}\tanh(W_{gx}X^{(t)} + W_{gh}h^{(t)})W_{gh}$$

$c^{(t+1)} = f^{(t+1)}c^{(t)} + i^{(t+1)}g^{(t+1)}$

$$\bar{c}^{(t)} = \bar{h}^{(t)}o^{(t)}(1-\tanh^2(c^{(t)})) + \bar{c}^{(t+1)}f^{(t+1)}$$
$$\bar{g}^{(t)} = c^{(t)}i^{(t)}$$
$$\bar{o}^{(t)} = \bar{h}^{(t)}\tanh(c^{(t)})$$
$$\bar{f}^{(t)} = \bar{c}^{(t)}c^{(t-1)}$$
$$\bar{i}^{(t)} = \bar{c}^{(t)}g^{(t)}$$

b) $\bar{W}_{ix} = \sum_t \bar{i}^{(t)}\sigma'(W_{ix}X^{(t)} + W_{ih}h^{(t-1)})X^{(t)}$

c) Suppose $f^{(t)} \approx 1$, $i^{(t)} \approx 0$ $o^{(t)} \approx 0$ $\forall t$

Then $\bar{c}^{(t)} \approx 0 \rightarrow \bar{f}^{(t)}, \bar{i}^{(t)} \approx 0$

$c^{(t)} \approx c^{(t-1)}$

$\bar{g}^{(t)} \approx 0$

$\bar{h}^{(t)} \approx \bar{o}^{(t+1)}\sigma'(W_{ox}X^{(t)} + W_{oh}h^{(t)})W_{oh}$

$\bar{o}^{(t)} \approx \bar{h}^{(t)}\tanh(c^{(t)})$

$\therefore \bar{h}^{(t)} \approx \bar{h}^{(t+1)}\tanh(c^{(t+1)}) \equiv \bar{h}^{(t+1)}\tanh(c)$

$\tanh(c) \in (-1,1)$ $\therefore$ This will not explode.

2. a) W_{in} : $H \times D$

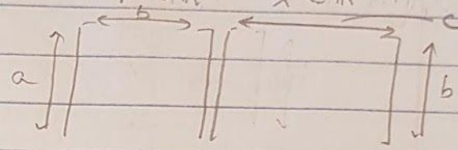
W_W : $H \times H$

W_N : $H \times H$

$DH + 2H^2$ weights per grid.

Suppose computing Ax.

$A \in \mathbb{R}^{a \times b}$ $x \in \mathbb{R}^{b \times c}$



each row: b ops

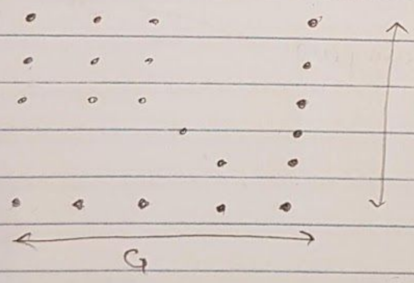
$\therefore abc$

#ops: W_{in} - HD

W_W - HH

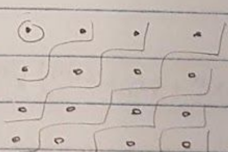
W_N - HH $\therefore$ $HD + 2H^2$ ops per grid.

b) Notice that top/left needs to be computed first.



At least G steps to get to right-most/bottom-most.

4x4 example:



Clearly, it will take ~G steps.

c) Advantage: higher capacity, can represent more fcts.

Disadvantage: more connections, backprop through time.

3. a) $p^{(K+1)} \leftarrow \beta p^{(K)} - \alpha \nabla F(\theta^{(K)})$

$\theta^{(K+1)} \leftarrow \theta^{(K)} + p^{(K+1)}$

$$\theta^{(K)} = \theta^{(K+1)} - p^{(K+1)}$$

$$p^{(K)} = \frac{1}{\beta}\left(p^{(K+1)} + \alpha \nabla F(\theta^{(K)})\right)$$

$$\therefore f^{-1}(s^{(K+1)}) = \begin{bmatrix} \theta^{k+1} - p^{k+1} \\ \frac{1}{\beta}\left(p^{k+1} + \alpha \nabla F(\theta^{k+1} - p^{k+1})\right) \end{bmatrix}$$

$$\frac{\partial s^{k+1}}{\partial s^{k}} = \begin{array}{cc} D & D \\ \left[\begin{array}{cc} I & 0 \\ -\alpha \nabla^2 F(\theta^{k}) & \beta I \end{array}\right. & \left.\right] \end{array} \begin{array}{l} \}D \\ \}D \end{array}$$

Lower triangular

$$\therefore \det\left(\frac{\partial s^{k+1}}{\partial s^{k}}\right) = \det(I)\det(\beta I) = \beta^{D}$$