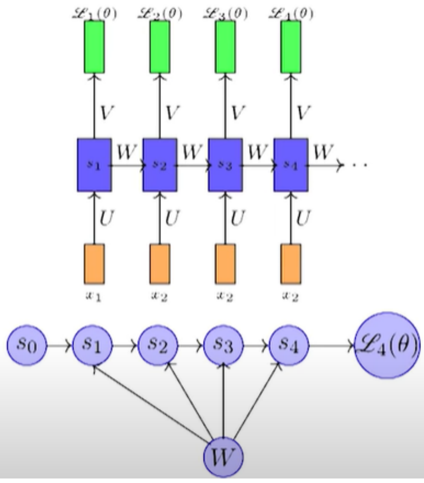
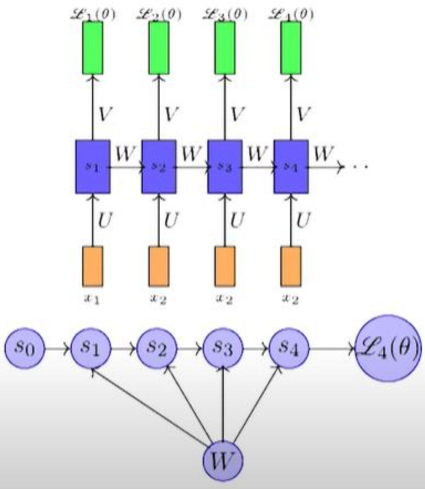
The problem of Exploding and Vanishing Gradients



- We will now focus on $\frac{\partial s_t}{\partial s_k}$ and highlight an important problem in training RNN's using BPTT



- We will now focus on $\frac{\partial s_t}{\partial s_k}$ and highlight an important problem in training RNN's using BPTT

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$
$$= \prod_{j=k}^{t-1} \frac{\partial s_{j+1}}{\partial s_j}$$

- Let us look at one such term in the product (i.e., $\frac{\partial s_{j+1}}{\partial s_j}$)

$\frac{\partial s_j}{\partial s_{j-1}}$

$$a_j = [a_{j1}, a_{j2}, a_{j3}, \dots a_{jd},]$$
$$s_j = [\sigma(a_{j1}), \sigma(a_{j2}), \dots \sigma(a_{jd})]$$

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \cdots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$
$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \ddots & \\ 0 & 0 & \dots & \sigma'(a_{jd}) \end{bmatrix}$$
$$= diag(\sigma'(a_j))$$

- We are interested in $\frac{\partial s_j}{\partial s_{j-1}}$

$$a_j = W s_{j-1} + b$$
$$s_j = \sigma(a_j)$$

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j} \frac{\partial a_j}{\partial s_{j-1}}$$
$$= diag(\sigma'(a_j))W$$

- We are interested in the magnitude of $\frac{\partial s_j}{\partial s_{j-1}}$ $\leftarrow$ if it is small (large) $\frac{\partial s_t}{\partial s_k}$ and hence $\frac{\partial \mathscr{L}_t}{\partial W}$ will vanish (explode)

$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| = \left\| diag(\sigma'(a_j))W \right\|$$
$$\leq \left\| diag(\sigma'(a_j)) \right\| \|W\|$$

$\because \sigma(a_j)$ is a bounded function (sigmoid, tanh) $\sigma'(a_j)$ is bounded

$$\sigma'(a_j) \leq \frac{1}{4} = \gamma \text{ [if } \sigma \text{ is logistic]}$$
$$\leq 1 = \gamma \text{ [if } \sigma \text{ is tanh]}$$
$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| \leq \gamma \|W\|$$
$$\leq \gamma\lambda$$

$$\left\| \frac{\partial s_t}{\partial s_k} \right\| = \left\| \prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}} \right\|$$
$$\leq \prod_{j=k+1}^{t} \gamma\lambda$$
$$\leq (\gamma\lambda)^{t-k}$$

- If $\gamma\lambda < 1$ the gradient will vanish
- If $\gamma\lambda > 1$ the gradient could explode

$$\left\|\frac{\partial s_j}{\partial s_{j-1}}\right\| = \left\|diag(\sigma'(a_j))W\right\|$$
$$\le \left\|diag(\sigma'(a_j))\right\| \|W\|$$

$\because \sigma(a_j)$ is a bounded function (sigmoid, tanh) $\sigma'(a_j)$ is bounded

$$\sigma'(a_j) \le \frac{1}{4} = \gamma \text{ [if } \sigma \text{ is logistic]}$$
$$\le 1 = \gamma \text{ [if } \sigma \text{ is tanh]}$$
$$\left\|\frac{\partial s_j}{\partial s_{j-1}}\right\| \le \gamma \|W\|$$
$$\le \gamma\lambda$$

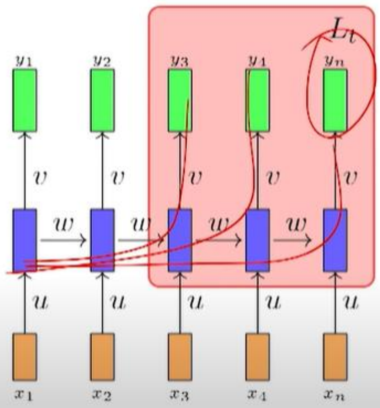
$$\left\|\frac{\partial s_t}{\partial s_k}\right\| = \left\|\prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}}\right\|$$
$$\le \prod_{j=k+1}^{t} \gamma\lambda$$
$$\le (\gamma\lambda)^{t-k}$$

$\to T$

T'

- If $\gamma\lambda < 1$ the gradient will vanish
- If $\gamma\lambda > 1$ the gradient could explode
- This is known as the problem of vanishing/ exploding gradients

8:53 / 11:03



- One simple way of avoiding this is to use truncated backpropogation where we restrict the product to $\tau(< t - k)$ terms

$$\frac{\nabla \omega}{\|\nabla \omega\|}$$ $< k$

Extended details:

$$\underbrace{\frac{\partial \mathscr{L}_t(\theta)}{\partial W}}_{\in \mathbb{R}^{d\times d}} = \underbrace{\frac{\partial \mathscr{L}_t(\theta)}{\partial s_t}}_{\in \mathbb{R}^{1\times d}} \sum_{k=1}^{t} \underbrace{\frac{\partial s_t}{\partial s_k}}_{\in \mathbb{R}^{d\times d}} \underbrace{\frac{\partial^+ s_k}{\partial W}}_{\in \mathbb{R}^{d\times d\times d}}$$

- We know how to compute $\frac{\partial \mathscr{L}_t(\theta)}{\partial s_t}$ which is the derivative of $\mathscr{L}_t(\theta)$ (scalar) w.r.t. last hidden layer (vector) using backpropagation
- We just saw a formula for $\frac{\partial s_t}{\partial s_k}$ (derivative of a vector w.r.t. a vector)
- $\frac{\partial^+ s_k}{\partial W}$ is a tensor $\in \mathbb{R}^{d\times d\times d}$, the derivative of a vector $\in \mathbb{R}^d$ w.r.t. a matrix $\in \mathbb{R}^{d\times d}$
- How do we compute $\frac{\partial^+ s_k}{\partial W}$? Let us see

---

- We just look at one element of this $\frac{\partial^+ s_k}{\partial W}$ tensor
- $\frac{\partial^+ s_{kp}}{\partial W_{qr}}$ is the (p, q, r)-th element of the 3d tensor

$a_k = W s_{k-1} + b$ $+ Ux_k.$

$s_k = \sigma(a_k)$

$$a_k = W s_{k-1}$$

$$\begin{bmatrix} a_{k1} \\ a_{k2} \\ \vdots \\ a_{kp} \\ \vdots \\ a_{kd} \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & \dots & W_{1d} \\ & & & \\ \vdots & \vdots & \vdots & \vdots \\ W_{p1} & W_{p2} & \dots & W_{pd} \\ \vdots & \vdots & \vdots & \vdots \\ & & & \end{bmatrix} \begin{bmatrix} s_{k-1,1} \\ s_{k-1,2} \\ \vdots \\ s_{k-1,p} \\ \vdots \\ s_{k-1,d} \end{bmatrix}$$

$$a_{kp} = \sum_{i=1}^{d} W_{pi} s_{k-1,i}$$

$$s_{kp} = \sigma(a_{kp})$$

$$\frac{\partial s_{kp}}{\partial W_{qr}} = \frac{\partial s_{kp}}{\partial a_{kp}} \frac{\partial a_{kp}}{\partial W_{qr}}$$

$$= \sigma'(a_{kp}) \frac{\partial a_{kp}}{\partial W_{qr}}$$

$$\frac{\partial a_{kp}}{\partial W_{qr}} = \frac{\partial \sum_{i=1}^{d} W_{pi} s_{k-1,i}}{\partial W_{qr}}$$

$= s_{k-1,i}$ if $p = q$ and $i = r$

$= 0$ otherwise

$\frac{\partial s_{kp}}{\partial W_{qr}} = \sigma'(a_{kp}) s_{k-1,r}$ if $p = q$ and $i = r$

$= 0$ otherwise