Vanishing and exploding gradient problem

Suppose that a computational graph contains a path that consists of repeatedly multiplying by a matrix W. After t steps, this is equivalent to multiplying by W^t. Assuming W has an eigendecomposition $W = V \operatorname{diag}(\lambda) V^{-1}$ then.

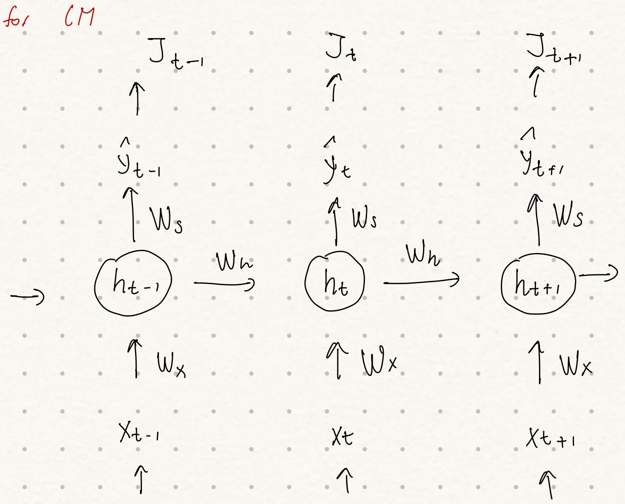
$$W^t = V \operatorname{diag}(\lambda)^t V^{-1}$$

Therefore, any eigenvalues λ_i that are not near 1 in absolute value will either explode if they are greater than 1 or vanish if they are less than 1 in magnitude.

Vanishing gradients make it difficult to know which direction the parameters should move to improve the cost function, while exploding gradients can make learning unstable.

This problem is more relevant in RNNs than FFN because RNNs use the same matrices of parameters at each time step but FFN don't.

RNN for LM



where $\quad h_t = \sigma(W_h h_{t-1} + W_x x_t)$

$$\hat{y}_t = \operatorname{softmax}(W_s h_t)$$

Cross-Entropy loss at t:
$$J^{(t)}(\theta) = -\sum_{j=1}^{|V|} y_{t,j} \times \log \hat{y}_{t,j} \overset{OHE}{=} -\log \tilde{y}_{t,j}$$

Total loss:
$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) \quad , \quad \theta \text{ contains all parameters.}$$

$$= -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} y_{t,j} \times \log \hat{y}_{t,j}$$

Backpropagation through time (BPTT)

Notation †

$$\frac{\partial J}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial J^{(t)}}{\partial \theta} \quad , \quad \frac{\partial J^{(t)}}{\partial \theta} = \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial \theta}$$

for all previous k time-steps

$$\frac{\partial J}{\partial \theta} = \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial \theta}$$

where $\quad † \; \dfrac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} \dfrac{\partial h_j}{\partial h_{j-1}} = \prod_{j=k+1}^{t} \dfrac{\partial h_j}{\partial \tilde{h}_j} \dfrac{\partial \tilde{h}_j}{\partial h_{j-1}} \quad , \quad \tilde{h}_j = W_h h_{j-1} + W_x x_j$

$$\frac{\partial h_t}{\partial h_k} = \prod_{j=k+1}^{t} W_h^T \operatorname{diag}(\sigma'(h_{j-1}))$$ and transport error from step t back to step k

BPTT

Since $h \in \mathbb{R}^{d_h}$, hence $\frac{\partial h_j}{\partial h_{j-1}}$ is the Jacobian matrix for h_j wrt all the units of h_{j-1}

Notice: product of $(t-k)$ Jacobian matrices → exploding grads / vanishing

Gradient of loss:

$$\frac{\partial J}{\partial \theta} = \sum_{t=1}^{T} \sum_{k=1}^{t} \frac{\partial J^{(t)}}{\partial \hat{y}_t} \frac{\partial \hat{y}_t}{\partial h_t} \left(\prod_{j=k+1}^{t} \frac{\partial h_j}{\partial h_{j-1}} \right) \frac{\partial h_k}{\partial \theta}$$

Assuming W^T and $\operatorname{diag}(\sigma'(h_{j-1}))$ have bounded norms given by β_w and β_h respectively.

$$\left\| \frac{\partial h_j}{\partial h_{j-1}} \right\| \le \|W_h^T\|_2 \, \|\operatorname{diag}(\sigma'(h_{j-1}))\|_2 \le \beta_w \beta_h$$

β_w and β_h bound the largest eigenvalue of the matrices.

↑ 2-norm definition.

Therefore,

$$\left\| \frac{\partial h_t}{\partial h_k} \right\| \le (\beta_w \beta_h)^{t-k} \to 0 \text{ or } \infty$$ provided $t \gg k$ long term

Mikolov: sufficient $\lambda_1(W) < \frac{1}{\gamma} \Rightarrow$ vanishing ∇.

necessary $\lambda_1(W) > \frac{1}{\gamma} \Rightarrow$ exploding ∇

because $\frac{\partial J^{(t)}}{\partial h_t} \frac{\partial h_t}{\partial h_k} = \sum c_i \lambda_i^{\ell} q_i^T \approx c_j \lambda_j^{\ell} q_j^T$

$\left| \frac{\lambda_i}{\lambda_j} \right| < 1 \quad \forall i \neq j$

$\lambda_1 = \lambda_j$

Parameters update:

$$\theta \leftarrow \theta - \eta \nabla_\theta L \to \theta \text{ too large}$$

$\Rightarrow$ no learning

Gradient Explosion: detectable at run time → overflow

Vanishing Gradients: undetectable and reduces the learning quality of the model for far away words

$\Rightarrow$ model unable to capture relation between words at steps t and $t+k$

Solution for Gradient Explosion (Mikolov) Gradient Clipping

Clip the gradients to a small number whenever they explode

$$\hat{g} = \frac{\partial L}{\partial \theta} \to \frac{\varepsilon}{\|\hat{g}\|} \hat{g}$$

Solution for Vanishing Gradients

1) $W_h^{(0)} = I$ instead of a random matrix

2) $\sigma = \text{Relu}$ instead of sigmoid.

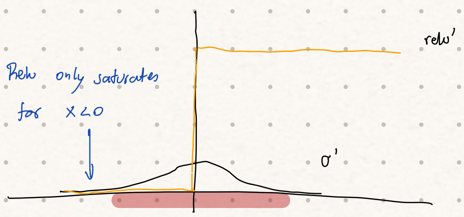
$$\text{Relu}(x) = \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases}$$

$$\frac{d\,\text{sigmoid}(x)}{dx} = \sigma(x)(1 - \sigma(x))$$ vanishes gradients for $\sigma(x)$ close to 0 or 1

$$\frac{d\,\text{Relu}(x)}{dx} = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases}$$

$$\frac{d\,\text{Relu}(x)}{dx} = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases} = \mathbb{1}_{\mathbb{R}^+}(x)$$

Obs: training w/ Relu is faster than sigmoid.



limited input interval w/ non-zero gradients ⇒ meaningless BP.

Mikolov explanation (On the difficulty of training RNNs) for w/ identity activation

Recall: $h^{(t)} = \sigma(W_h h_{t-1} + W_x x_t)$

If $\sigma(x) = x$ then:

$$\frac{\partial h_t}{\partial h_{t-1}} = W_h^T \operatorname{diag}(\sigma'(\underbrace{W_h h_{t-1} + W_x x_t}_{h_t}))$$

$$= W_h^T \cdot I = W_h^T$$

$\sigma(x) = x$

Hence,

$$\frac{\partial J^{(i)}(\theta)}{\partial h_j} = \frac{\partial J^{(i)}(\theta)}{\partial h_i} \prod_{t=j+1}^{i} \frac{\partial h_t}{\partial h_{t-1}}$$

$$= \frac{\partial J^{(i)}(\theta)}{\partial h_i} \left(W_h^T\right)^{\ell}, \quad \ell = i - j$$

$$= \sum c_i \lambda_i^{\ell} q_i \begin{matrix} \nearrow 0 \\ \searrow \infty \end{matrix}$$

λ_i eigenvalues of W_h

q_i eigenvectors.

Mikolov: in the linear case it is sufficient for the largest eigenvalue λ_1 to be smaller than 1 for long-term components to vanish ($t \to \infty$) and necessary for it to be larger than 1 for gradients to explode.