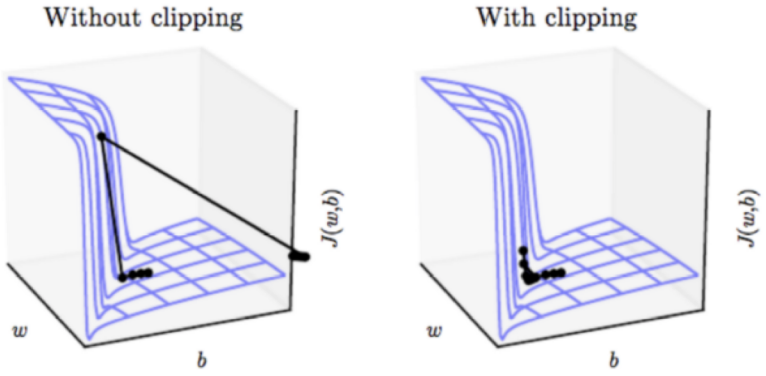
Dealing with the Vanishing Gradient

Heni Ben Amor, Ph.D.
Assistant Professor
Arizona State University

ASU Ira A. Fulton Schools of Engineering
Arizona State University

Modern Solutions: Gradient Clipping

- Clip the size of the gradient
- Evaluate norm and rescale to allowed threshold
- Avoids unfettered grows of the gradient



Algorithm 1 Pseudo-code for norm clipping the gradients whenever they explode

$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ **then**
  $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*. MIT Press.

Modern Solutions: Initialization

- Ensure that eigenvalues of the recurrent weight matrix are equal to one

$$\frac{\partial h_t}{\partial h_k} = \prod_{i=k}^{t} \frac{\partial h_i}{\partial h_{i-1}} = \prod_{i=k}^{t} W \operatorname{diag}(\sigma'(x_{i-1}))$$

Modern Solutions: Initialization

- The following matrices have eigenvalues equal to one.
- In practice, soft constraints imposed on matrices improve trainability of RNNs.

Identity Initialization

$$\mathbf{W}_{rec} = \begin{bmatrix} 1 & 0 & \dots \\ \vdots & \ddots & \\ 0 & & 1 \end{bmatrix}$$

Orthogonal Initialization

$$\mathbf{W}_{rec} = \begin{bmatrix} a_{11} & a_{12} & \dots \\ \vdots & \ddots & \\ a_{K1} & & a_{KK} \end{bmatrix}$$

Modern Solutions: Activations

Sigmoid Activation

$$\sigma = \frac{1}{1 - e^{-x}}$$

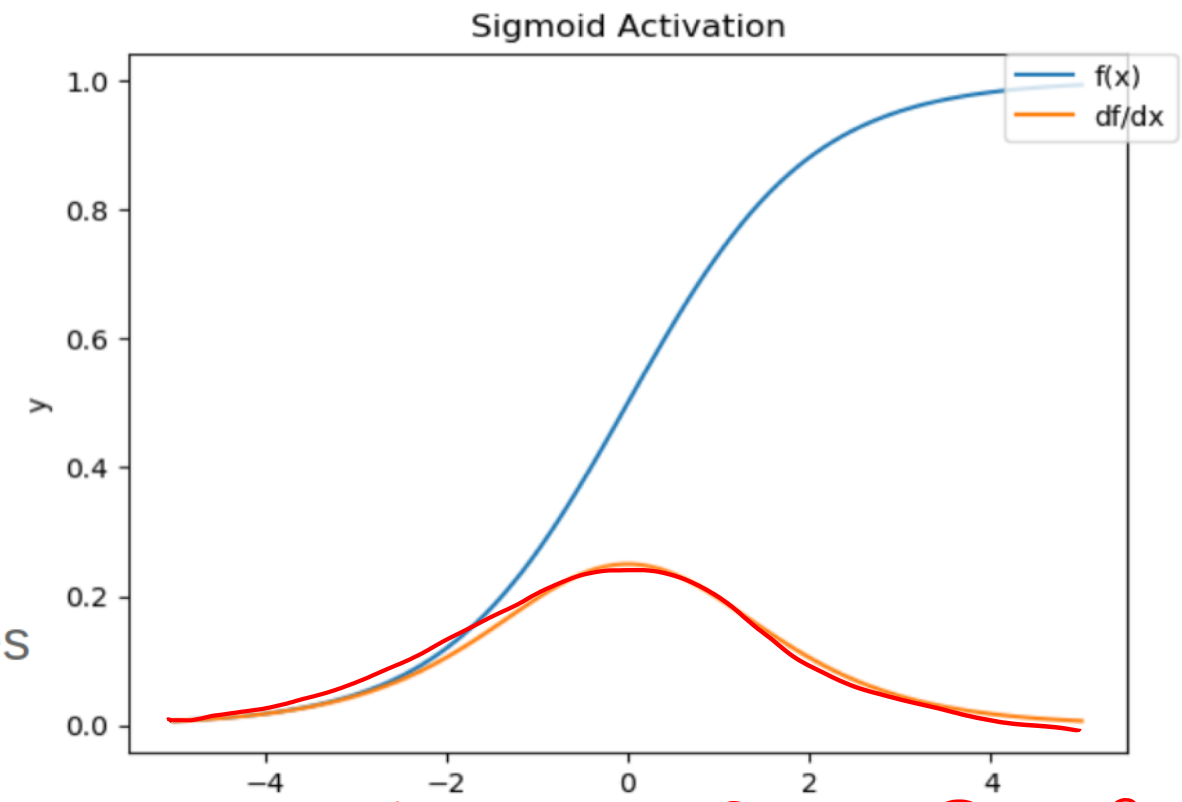
$$\sigma' = (1 - \sigma)\sigma$$

Max value of df/dx is .25

Temporal gradient vanishes quickly with this activation function

$$.25^2 = .0625$$

$$.25^5 = 0.00097$$



0.25 · 0.25 ○ — ○ — ○ — ○

t-2 t-1 t

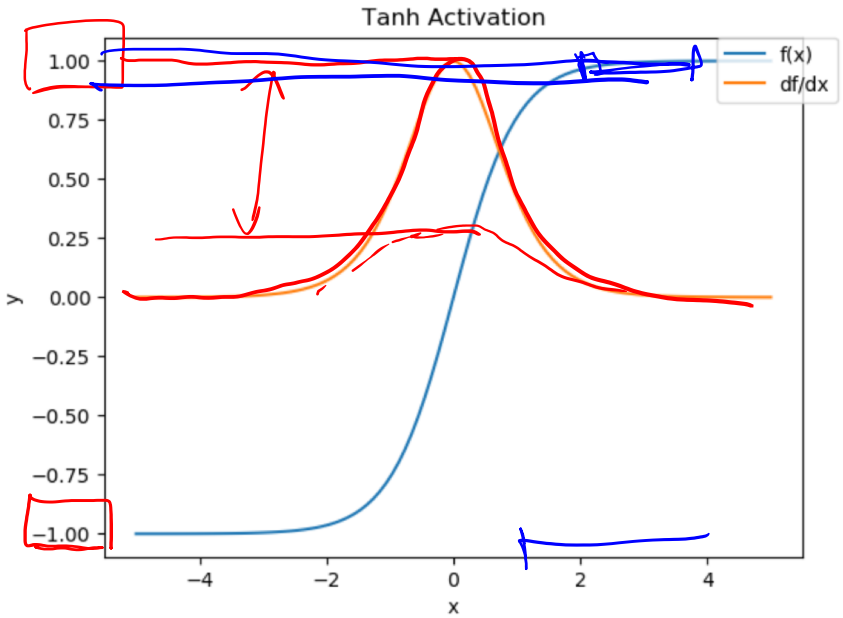
Modern Solutions: Activations

Tanh Activation

$$tanh = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad tanh' = 1 - tanh^2(x)$$

Heavily used in modern recurrent architectures

The gradient vanishes more quickly the further x deviates from 0



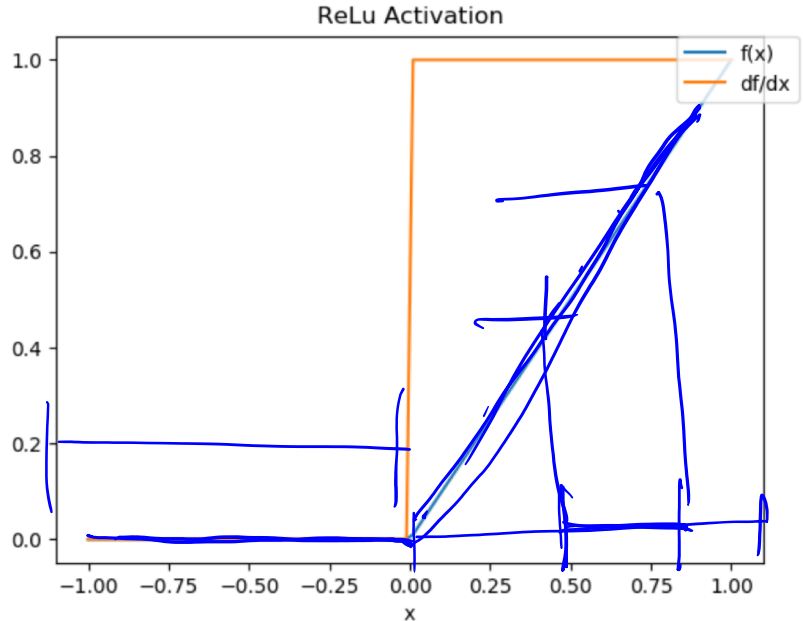
Rectified linear Unit

ReLu Activation

$$ReLu = max(0, x) \quad ReLu' = \begin{cases} x > 0, & 1 \\ x <= 0, & 0 \end{cases}$$

ReLu activation has desirable gradient behavior for values of x > 0

For x < 0 the temporal gradient does not exist



Model	Description
LSTM	Most ubiquitous RNN architecture today. Adds gated computations and cell memory state for long term memory.
LSTM Forget Gates	Adds new gate to LSTM architecture that focuses on “forgetting” long-term dependencies that are no longer relevant.
Peephole LSTM	Uses previous cell state for gate computations instead of hidden state; accesses constant error carousel.
GRU	Combines input and forget gates into single update gate and combines the cell and hidden memory states.
IndRNN	Forces the recurrent weight matrix to be a vector that is multiplied element-wise by the previous hidden state.
UGRNN RNN+	Modern architectures made to enhance trainability of deeply-stacked (RNN+) and shallow (UGRNN) models.

Modern RNN architectures have been proposed to address the vanishing and exploding gradient problem.