* Vanishing & Exploding Grads. in RNNs -

① In RNNs info. flows from 1 timestep to next. For training, we use Backprop Through Time (BPTT) to calc. gradient of err wrt. each wt. in the NW.

② During BPTT gradients are calculated at diff layers & propagated backward through each time step. This allows RNN to learn from prev. i/p in the sequence.

③ However, this also means that for long sequences, gradients are passed through many layers & the process of repeatedly multiplying these gradients can cause them to shrink (vanish) or grow (explode) dramatically.

④ Gradients are scaled during backprop. by wt. matrices of each layer.

⑤ If you consider a long seq. with many time steps, the tot. gradient for each param. is effectively a product of derivatives from each layer. This repeated multiplicatn can either diminish/amplify gradients based on the values of these derivatives.

- If derivative (or weight value) is small (<1), repeated multiplicatn causes gradient to approach zero → Vanishing gradients.
- If derivative (or weight value) is large (>1), repeated multiplicatn causes the gradient to grow exponentially. - exploding gradients.

A) Vanishing Gradients -

⊛ He Happen when the gradients or error signals become very small as they are backpropagated through layers. In RNNs, common issue bcoz N/w repeatedly applies same wts. over time steps (seq. length).

Why it happens:

① RNNs update wts. based on the gradients from backprop. When same wts. are multiplied across many time steps, gradients shrink exponentially.

② This makes gradients for earlier layers (or time steps) very small, almost zero.

③ As a result, N/w learns very slowly or not at all, especially for info. from earlier in the seq.

Consequence:

RNNs fail to learn long term dependencies (connections b/w distant time steps in the seq.) as it forgets the impact of earlier time steps.

B) Exploding Gradients -

Occur when the gradients become extremely large as they backprop. through the layers of NN.

Why it Happens:

① When networks wts. are large, multiplying these wts. across time steps causes the gradient to ↑ exponentially instead of L.Sng.

② This can happen due to instability in wt. values or

errors compounding over long sequences.

Consequence:

① Can cause wt updates to become very large causing leading to drastic changes in NNs wts.

② Model's training process becomes unstable & the NNs predictns might oscillate wildly or fail to converge.

c) Solutns for Vanishing & Exploding Gradients -

① Gradient Clipping:
For exploding gradients, applying gradient clipping limits the size of gradients to a max. threshold, preventing gradients from growing too large.

② Batch Normalizatn:
Help stabilize the gradients by normalizing the i/p to each layer, thus minimizing vanishing & exploding gradient problems.

③ Use of LSTM & GRU cells-
Long short term memory (LSTM) & Gated Recurrent Unit (GRU) are specialized RNN architectures designed to address manage these issues. They incorporate mechanisms to preserve info. over longer sequences, reducing vanishing grad. issues.

④ Proper Weight Initializatn -
Use of careful wt. initializatn techniques can help keep grads gradients in reasonable range initially, which helps prevent these issues.