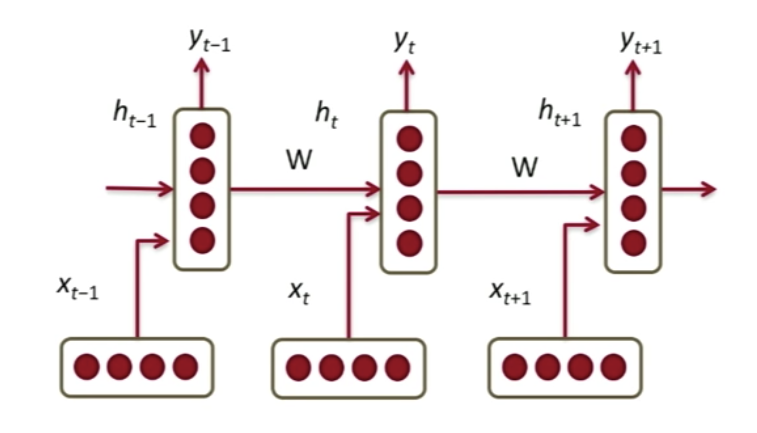
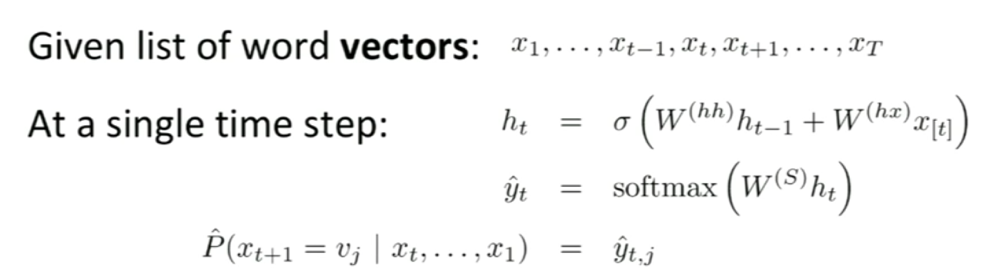
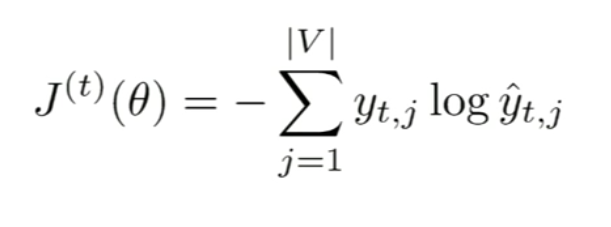
Lecture 8 | Recurrent Neural Networks and Language Models

* A language model computes a probability for a sequence of words:
  + P(w1, …. , wT)
  + Useful for machine translation
    - Word ordering
      * p(the cat is small) > p(small the is cat)
    - Word choice
      * p(walking home after school) > p(walking house after school)
* Recurrent Neural Networks
  + The NN ties the weights at each time step and condition the NN on all previous words
  + RAM requirement only scales with number of words

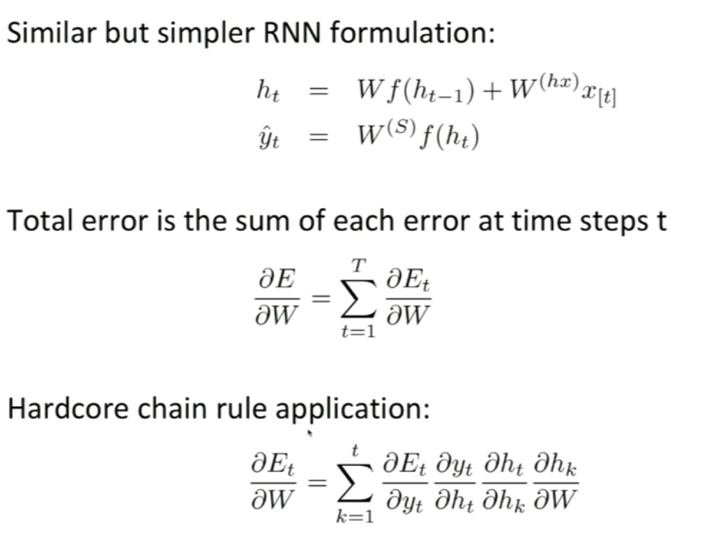


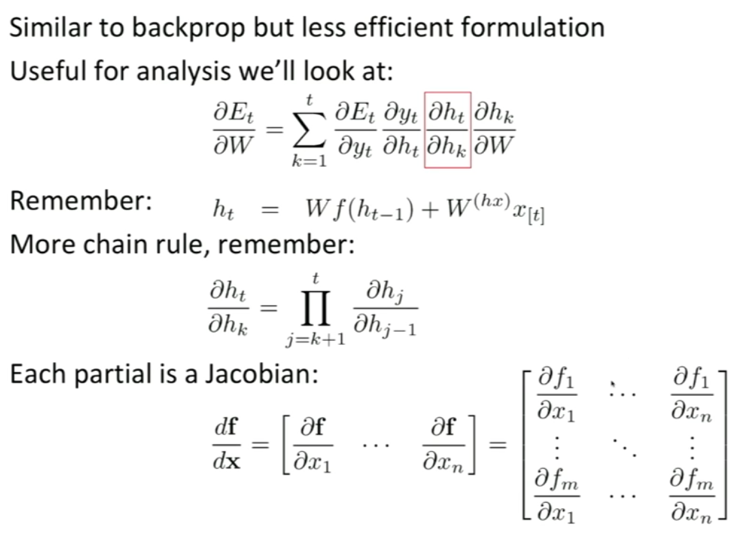


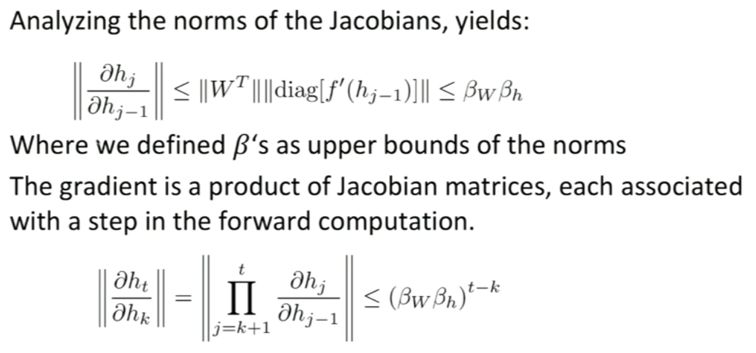
* + H0 is some initialisation vector
  + Y\_hat is a probability distribution over the vocabulary
  + We have the same cross entropy loss function but we are predicting words now instead of classes



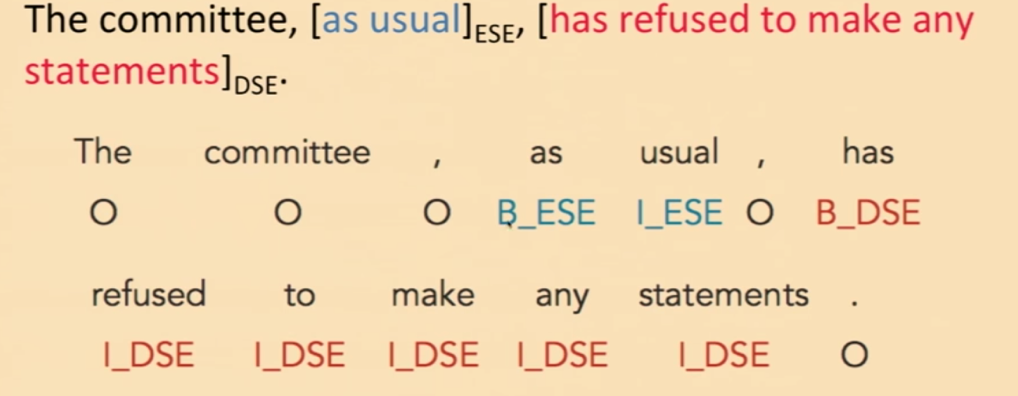
* + We face vanishing/exploding gradient problem if very deep neural networks



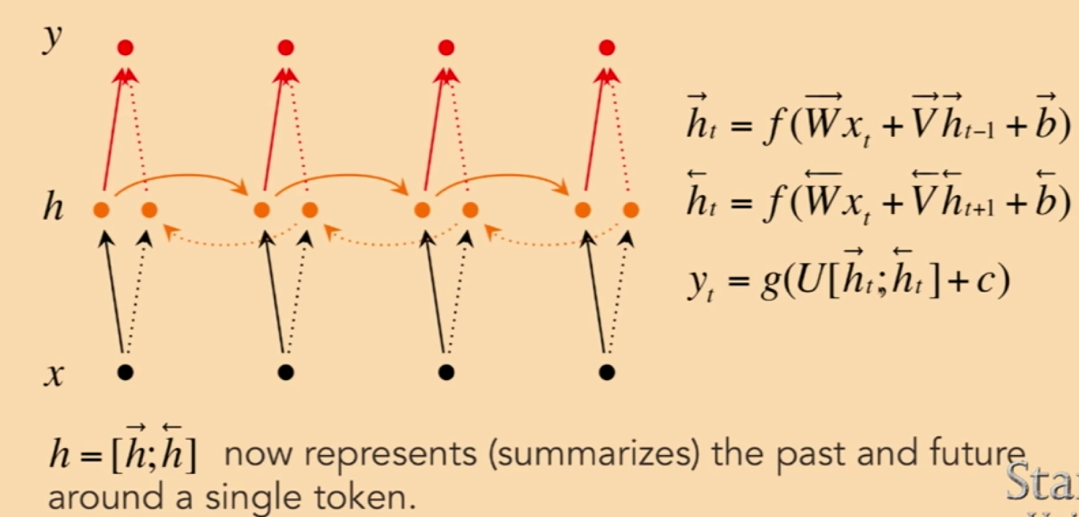




* + - Therefore, as the sequence gets longer and longer (i.e. t gets larger and larger), depending on the value of beta (if it’s greater or less than 1) we could be facing exploding or vanishing gradient
    - Vanishing gradient is a problem because it prevents us from allowing a word that appears very much in the past to have any influence on what we are trying to break in terms of the next word
  + Trick for exploding gradient
    - Clipping, where you clip gradients to a maximum value and this makes a big difference in RNNs
  + Trick for vanishing gradient
    - Initialisation of W using the identity matrix + ReLUs
  + Applications:
    - NER
    - Entity level sentiment in context
    - Opinionated expressions
      * Opinion mining where you try to classify each word as either direct subjective expressions (DSEs) or expressive subjective expressions (ESEs)
      * DSE : explicit mentions of private states or speech events expressing private states
      * ESE : expressions that indicate sentiment, emotion etc without explicitly conveying them



* Bidirectional RNNs
  + For classification you want to incorporate information from words both preceding and following



* Evaluation of RNNs: Use F1 score, harmonic mean of precision and recall