Typically: RNN has recurrence in the hidden layers

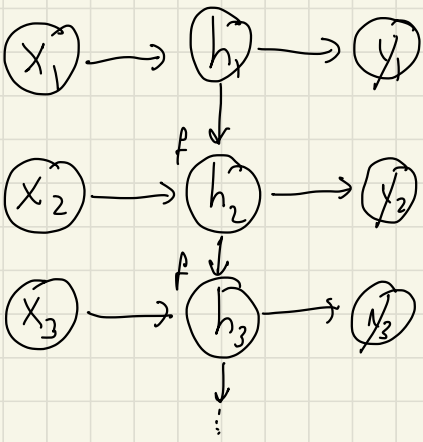
$$h_t = f(h_{t-1}, x_t, \beta)$$

↑ usually decomposed into an activation function with weights + bias

Goal: The RNN can remember a "summary" of the relevant aspects of the past

Visually:

Unfolds as:



Initialize $h_0 = 0$

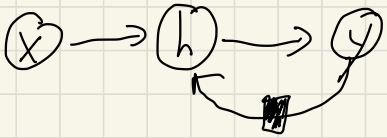
Often only run forward p steps so we still have a finite memory

Many possible recurrent structures exist

2 most common are:

1) recurrence between hidden nodes (displayed above)

2) recurrence from output at one time step to the hidden units at the next time step



Note: RNNs typically assume homoscedastic (iid) errors

There is recent research on heteroscedastic RNNs

Sometimes called Generalized Recurrent Neural Networks (GRNNs)

Gated RNNs and Long Short-Term Memory (LSTM) Networks

Gating allows for dynamic control of the flow of past information to the current state at each time instant

ex: α-RNN

$$y_t = W_y \hat{h}_t + b_y$$
$$\hat{h}_t = g(U_h \tilde{h}_{t-1} + W_h x_t + b_h)$$
$$\tilde{h}_t = \alpha \hat{h}_t + (1-\alpha)\tilde{h}_{t-1} \quad \leftarrow \text{smoothing}$$

If $\alpha = 1$: this is a typical (plain) RNN

If $\alpha \neq 1$: smooth the hidden layer to retain infinite memory (even if we restart with hidden node at p steps in the past)

LSTM Networks

LSTM networks use "cells" that have an internal recurrence + outer recurrence

↳ similar to (plain) RNN

↳ use an additional "state" s_t which is used to remember the past

Mathematically: 3 Gates

Forget Gate

$$f_t = \sigma\left(U_f h_{t-1} + W_f x_t + b_f\right)$$

h_{t-1}: hidden layer [the output of the LSTM]

Sets value between 0 & 1

Determines how much to remember / forget about the current state

External Input Gate

$$g_t = \sigma\left(U_g h_{t-1} + W_g x_t + b_g\right)$$

The updated state is then constructed as:

$$s_t = f_t s_{t-1} + g_t \sigma\left(U h_{t-1} + W x_t + b\right)$$

$f_t s_{t-1}$: amount to recall from prior state

$g_t \sigma(\cdots)$: "new" information for the state

Output Gate

$$h_t = \tanh(s_t)\, q_t$$

$$q_t = \sigma(U_o h_{t-1} + W_o x_t + b_o)$$

Extensions exist

ex: use s_{t-1} as an additional input to each gate

Visually