Recurrent Neural Networks

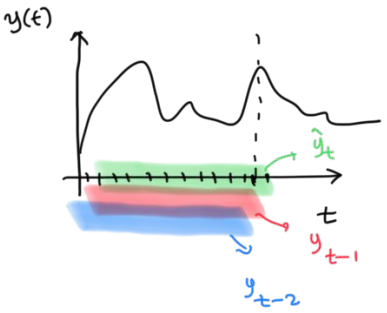
Setup: We have sequence data-set $\{ y_t : t = 1, 2, \dots, T \}$, $y_t \in \mathbb{R}^d$

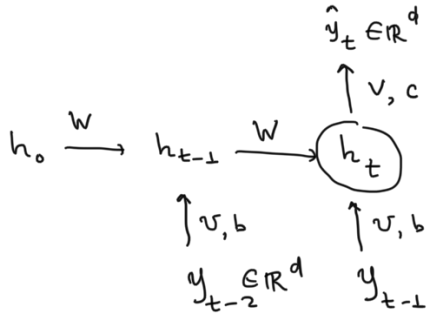
Goal: Predict the next sequence $\hat{y}_t$ as a fuction of previous values (lags), i.e. $\hat{y}_t = f_\theta (y_{t-1}, y_{t-2}, \dots, y_{t-L})$

L: # of lags

f_θ: recurrent neural net.

Example: Prediction using 2 lags, i.e. $\hat{y}_t \in \mathbb{R}^d = f_\theta (y_{t-1} \in \mathbb{R}^d, y_{t-2} \in \mathbb{R}^d)$





Params: $\theta := \{ U, b, W, V, c$

⊛ Notice how U, b, W are shared.

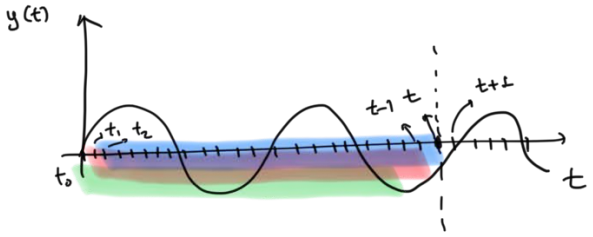
$$
\begin{cases}
\hat{y}_t = h_t V + c & (1 \times d)(q \times d) \to 1 \times d \\
h_t = \tanh (h_{t-1} W + y_{t-1} U + b) & (1 \times q)(q \times q) + (1 \times d)(d \times q) + 1 \times q \to 1 \times q \\
h_{t-1} = \tanh (h_0 W + y_{t-2} U + b) & \\
h_0 = 0 &
\end{cases}
$$

User defined hyper-F
1.) # of lags
2.) dimensi on the hidden state q

Training: Minimize an MSE loss:

$$\mathcal{L}(\theta) := \frac{1}{T-2} \sum_{t=3}^{T} (y_t - \hat{y}_t)^2 \qquad \hat{y}_t \in \mathbb{R}^d$$

Example: Predict the dynamics of a sine wave, i.e. $y(t) = \sin(T$



$\hat{y}_t = f_\theta(y_{t-1}, y_{t-2})$

$y_t \in \mathbb{R}^{N \times D}$ (here $D=1$ since we have a single state/time series)

$y_t := [\underbrace{y(2), y(3), \dots, y(t)}_{N}] \in \mathbb{R}^1$

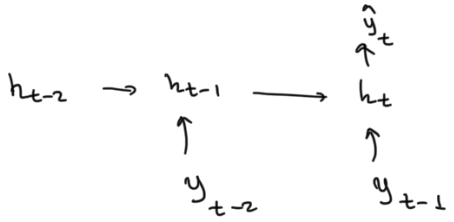
$y_{t-1} := [\underbrace{y(1), y(2), \dots, y(t-1)}_{N}] \in \mathbb{R}^1$

$y_{t-2} := [y(0), y(1), \dots, y(t-2)]$

Input: (y_{t-1}, y_{t-2})

X : $L \times N \times D$ (# of lags, # of measurements, dim of state)

Y : y_t, $N \times D$



$$\begin{cases} h_{t-2} = 0 \\ h_{t-1} = \tanh(h_{t-2} W + X[0,:,:] V + \\ h_t = \tanh(h_{t-1} W + X[1,:,:] V + \\ \hat{y}_t = h_t V + c \end{cases}$$

Long Short-Term Memory network :

In an LSTM one replaces the hidden units :

Standard RNN Cell update rule : $h_t = \tanh(h_{t-1} W + y_{t-1} V + b)$ $\{\theta := \{W, V,$

Replace it with :

Input : h_{t-1}, y_{t-1}

Output : h_t

$$\begin{cases} h_t = O_t \odot \tanh(S_t) & : \text{output vector} \\ O_t := \sigma(h_{t-1} W_o + y_{t-1} V_o + b_o) & : \text{output gat} \\ S_t := f_t \odot S_{t-1} + i_t \odot \tilde{S}_t & : \text{cell state} \end{cases}$$

($\odot$: element-wise multiplication)

$$
\begin{cases}
s_t := \tanh\left(h_{t-1} W_s + y_{t-1} U_s + b_s\right) \\
i_t := \sigma\left(h_{t-1} W_i + y_{t-1} U_i + b_i\right) : \text{external inp gate} \\
f_t := \sigma\left(h_{t-1} W_f + y_{t-1} U_f + b_f\right) : \text{forget ga}
\end{cases}
$$

Now we have more parameters:

$$
\theta := \{W_o, U_o, b_o, W_s, U_s, b_s, W_i, U_i, b_i, W_f, U_f, b_f, \underbrace{V, c}_{\text{final out layer param}}
$$