Recurrent Nets (03/26/2021)

Homework #4 due @ 04/02 11:30 PM

Our previous work, focused on training a feedforward neural net. That is given a point $x \in \mathbb{R}^D$, the sample was passed through the net to get $\hat{y}$. In this context, we are classifying a single point. However, what if we have a sequence? $x_1, x_2, \ldots, x_t$

- In some settings we have a history

Example: Language Model

Consider the task of making a prediction on the next word given the current word $w(t)$ and a history $h(t-1)$. To do this prediction task, we need to know some context. So given $w(t)$ and $h(t-1)$, what is

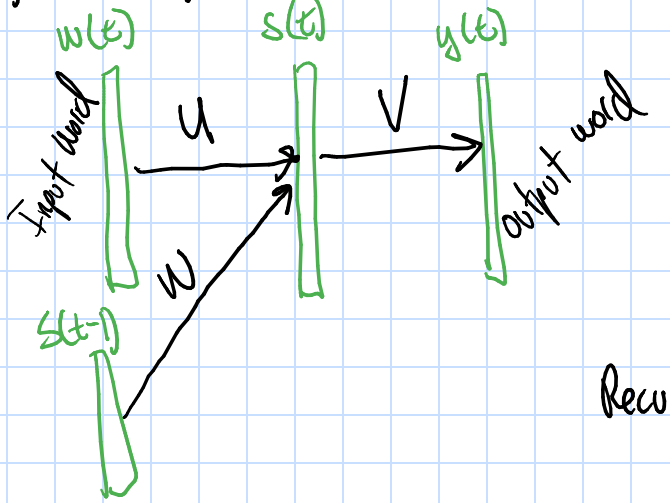
$$\Pr\{w(t+1) \mid w(t), h(t-1)\}$$

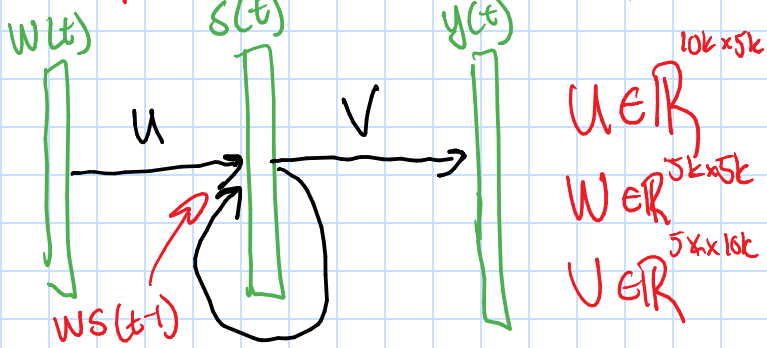
Can you pick up milk at the ___ ?

$w(t)$ → $w(t+1)$

$$s(t) = Uw(t) + Ws(t-1)$$
$$y(t) = Vs(t)$$

$h(t-1)$ → or some representation of this part



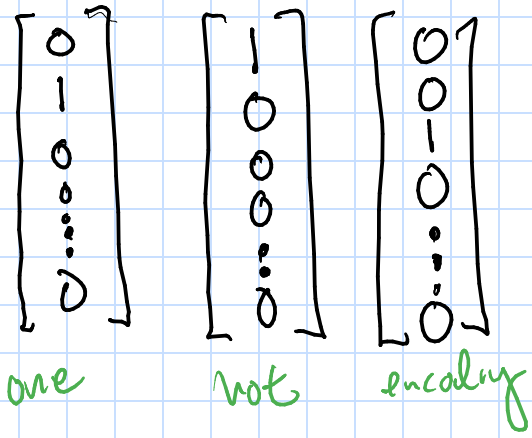


Recurrent Neural Net W

D = 10k
H = 5k

How do we encode w(t) if it is a word?

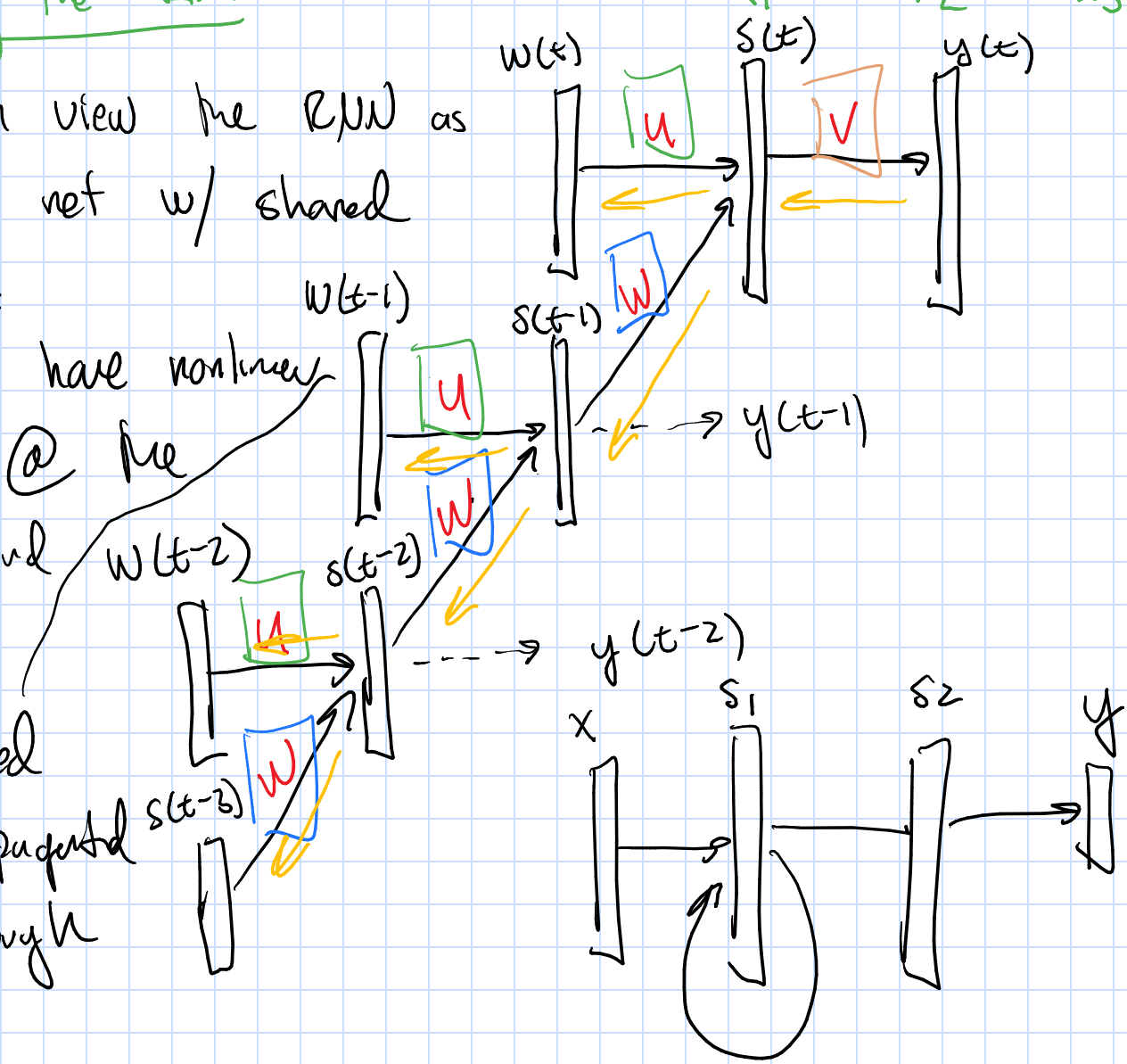
- "One hot encoding" Dictionary has D words
- Word embeddings (word2vec) Mikolov (2013) BERT



N-grams {a, g, c, t}

Unraveling the RNN

- We can view the RNN as a deep net w/ shared weights
- We still have nonlinear functions @ the hidden and outputs
- Errors need to be propagated back through time



Def:

$e_o(t) = d(t) - y(t)$ {error}

$\alpha \rightarrow$ learning rate

$\beta \rightarrow L_2$ regularization

$$e_h(t) = d_n(e_o^T V)$$

$$d_{hj}(x,t) = x\, s_j(t)(1 - s_j(t))$$

$$V(t+1) = V(t) + \alpha\, s(t)\, e_o(t)^T - \beta V(t)$$

$$U(t+1) = U(t) + \alpha\, w(t)\, e_h(t)^T - \beta U(t)$$

$$W(t+1) = W(t) + \alpha\, s(t-1)\, e_h(t)^T - \beta W(t)$$

$$s(t) = f(U w(t) + W s(t-1))$$

$$y(t) = g(V s(t))$$