CS 6180 09/24

HW2 released tomorrow (will be a bit longer so you'll have 2 weeks instead of one)

Review from last time

* Vanishing / Exploding gradients are a main limitation of RNNs.
* limit the # of time steps we are considering in backpropagation (reduce the number of sigmoids we are multiplying)

  losing some connections between words at the beginning of a sentence and later ones.
* Orthogonal Init

  mainly helps with exploding gradients
  still have the issue of the products of sigmoids.

Long Short-term memory

main idea: store some info in a cell $\vec{c}^{(t)}$ (RAM in your computer)

to keep track of relevant info.

From one time step to another

* Forget gate: some info in the cell (RAM) will be forgotten

$$\vec{f}^{(t)} = \sigma(W_f \vec{h}^{(t-1)} + U_f \vec{e}^{(t)} + \vec{b}_f)$$

* Input gate: new info that will be stored in the cell.

$$\vec{i}^{(t)} = \sigma(W_i \vec{h}^{(t-1)} + U_i \vec{e}^{(t)} + \vec{b}_i)$$

* Output gate: new info that will be stored in the hidden state

$$\vec{o}^{(t)} = \sigma(W_o \vec{h}^{(t-1)} + U_o \vec{e}^{(t)} + \vec{b}_o)$$

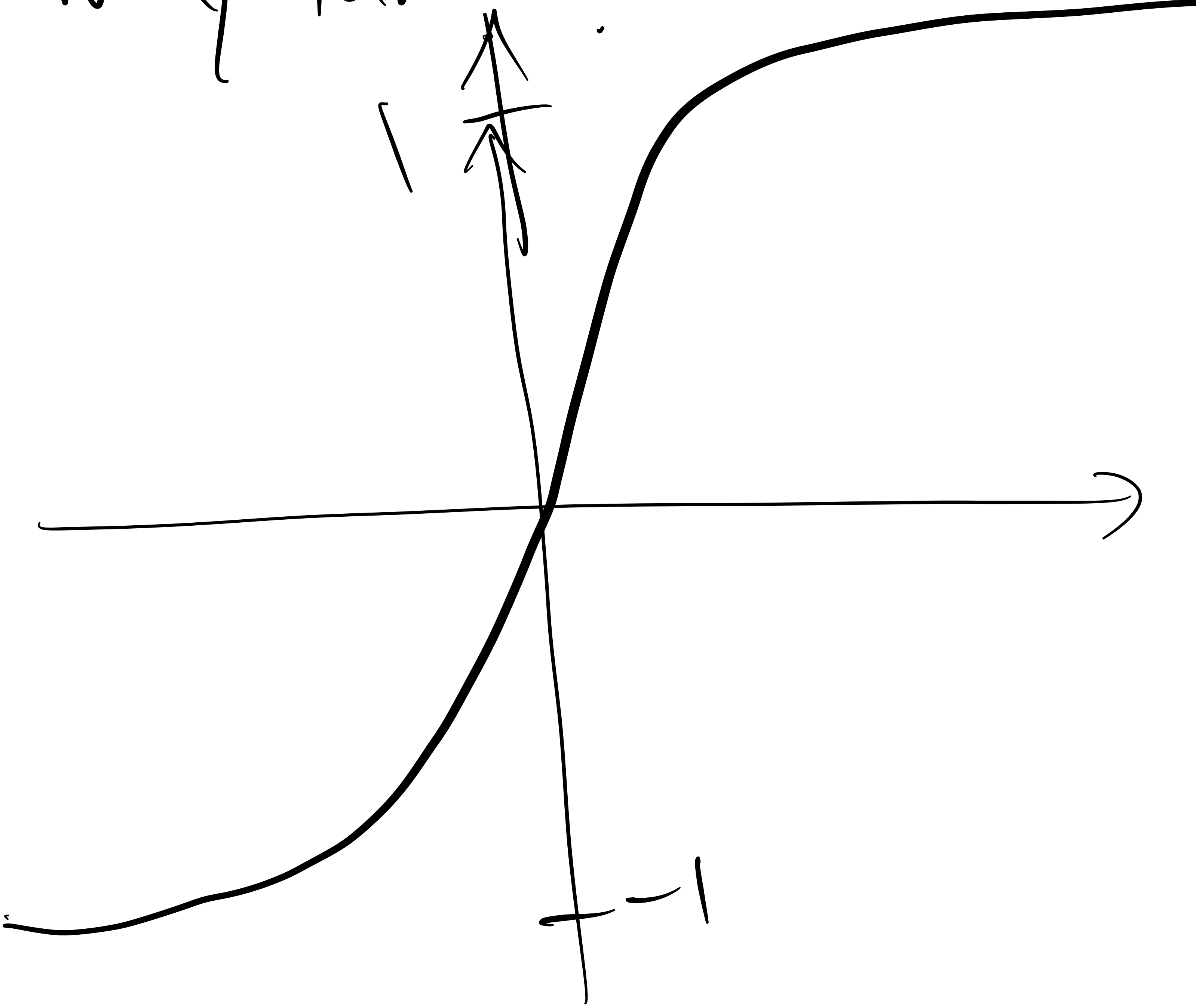
Let's first start by updating the cell

$$\vec{c}^{(t)} = \vec{f}^{(t)} * \vec{c}^{(t-1)} + i^{(t)} * \vec{\tilde{c}}^{(t)}$$

element wise multiplication

$$\vec{\tilde{c}}^{(t)} = \tanh(W_c \vec{h}^{(t-1)} + U_c \vec{e}^{(t)} + \vec{b}_c)$$

Why tanh?



— "is fantastically bad"

t-1 t

$$c^{(t-1)} = 0.8$$

positive sentiment

$$\tilde{c}^{(t)} = -0.6$$

$$f^{(t)} = 1$$

$$i^{(t)} = 1$$

$$c^{(t)} = f^{(t)} * c^{(t-1)} + i^{(t)} * \tilde{c}^{(t)}$$

$$= 1 * 0.8 + 1 * (-0.6)$$

$$= 0.2$$

we need these negative values to represent an idea of negation or reversal.

Remaining step:

Update the hidden state

$$\vec{h}^{(t)} = \vec{o}^{(t)} * \tanh(C_t)$$

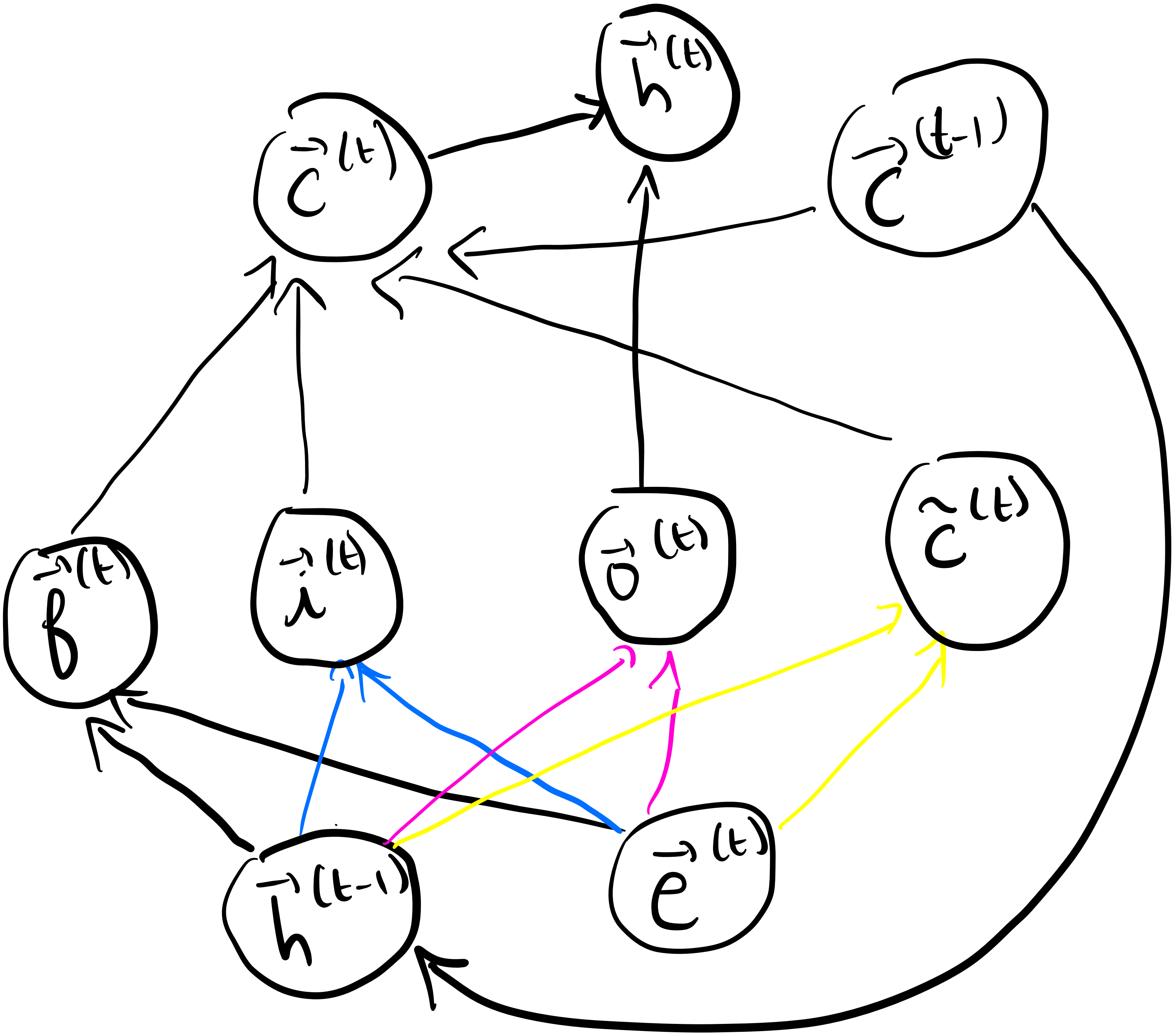
$f^{(t)}$ forget gate

$i^{(t)}$ input gate

$o^{(t)}$ output gate

$\hat{c}^{(t)}$ new info (tanh)

$$c^{(t)} = f * c + i * c$$
$$h^{(t)} = o^{(t)} * \tanh(c^{(t)})$$



Exercises:

Show

$$\frac{\partial \mathcal{L}}{\partial \vec{C}^{(t-1)}} = \frac{\partial \mathcal{L}}{\partial \vec{C}^{(t)}} \cdot \vec{f}^{(t)}$$

then show

$$\frac{\partial \mathcal{L}}{\partial \vec{C}^{(t-k)}} = \frac{\partial \mathcal{L}}{\partial \vec{C}^{(t)}} \cdot \prod_{j=0}^{k-1} \vec{f}^{(t-j)}$$