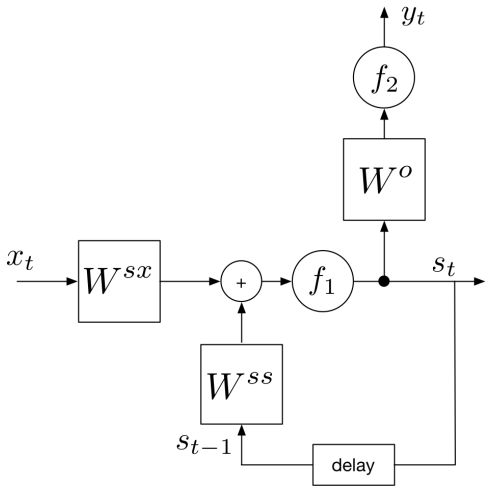
Name: _____

Double trouble

8. (7 points) One of the RNN architectures we studied was

$$s_t = f_1(W^{ss} s_{t-1} + W^{sx} x_t)$$
$$y_t = f_2(W^o s_t)$$

where W^{ss} is $m \times m$, W^{sx} is $m \times l$ and W^o is $n \times m$. Assume f_i can be any of our standard activation functions. We omit the offset parameters for simplicity (set them to zero).



(a) Suppose we modify the original architecture as follows:

$$s_t = f_1(W^{ss1} f_3(W^{ss2} s_{t-1}) + W^{sx} x_t)$$

i. Provide values for the original W^{ss} that make the original architecture equivalent to this one, or explain why none exist.

$W^{ss} =$ _____

ii. Provide values for W^{ss2}, f_3 and W^{ss1} that make this new architecture equivalent to the original, or explain why none exist.

$f_3 =$ _____

$W^{ss1} =$ _____

$W^{ss2} =$ _____

(b) Now, we'll consider two strategies for making the RNN generate two output symbols for each input symbol. Assume the symbols are drawn from a vocabulary of size n.

Model A: We use a separate softmax output for each symbol, so

$$y_t^1 = \text{softmax}(W^{o1} s_t)$$
$$y_t^2 = \text{softmax}(W^{o2} s_t)$$

where W^{o1} and W^{o2} are $n \times m$.

Model B: We use a single softmax output, but it ranges over n^2 possible pairs of symbols, so

$$y_t^1, y_t^2 = \text{softmax}(W^{o3} s_t)$$

i. What would the dimension of W^{o3} need to be?

ii. Which of the following is true:

- ◯ Models A and B can express exactly the same set of RNN models.
- ◯ Model A is more expressive than model B.
- ◯ Model B is more expressive than model A.