CS 344 – Review March 15 Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Suppose we have an MLP with 768 input features, 3072 hidden features, and 1 output feature. (This is the shape of a GPT-2 “fully-connected” layer, except it would normally have 768 output features.) Assume ReLU activation. All linear layers have bias terms.

1. Complete the following to write out the PyTorch module definition.

model = nn.Sequential(

nn.Linear(

Consider just the final linear layer of this model. Call its input *x* andits one output feature *y*.

1. Write out the mathematical expression that would compute y from x. (Recall x.shape = (3072,).) Assume that variables *w* and *b* are defined as needed.

y = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. y.shape = (1,). len(w.shape) = 1. w.shape = \_\_\_\_\_\_\_\_ b.shape = \_\_\_\_\_\_\_
2. a. What is the gradient of y with respect to *the first element of* w?

b. What is the gradient of y with respect to *the full vector* *w*?

1. Suppose the gradient of the loss with respect to *y* has already been computed and stored in the variable y\_grad. Compute the gradient with respect to *w*.

w\_grad =

1. Repeat for x\_grad.