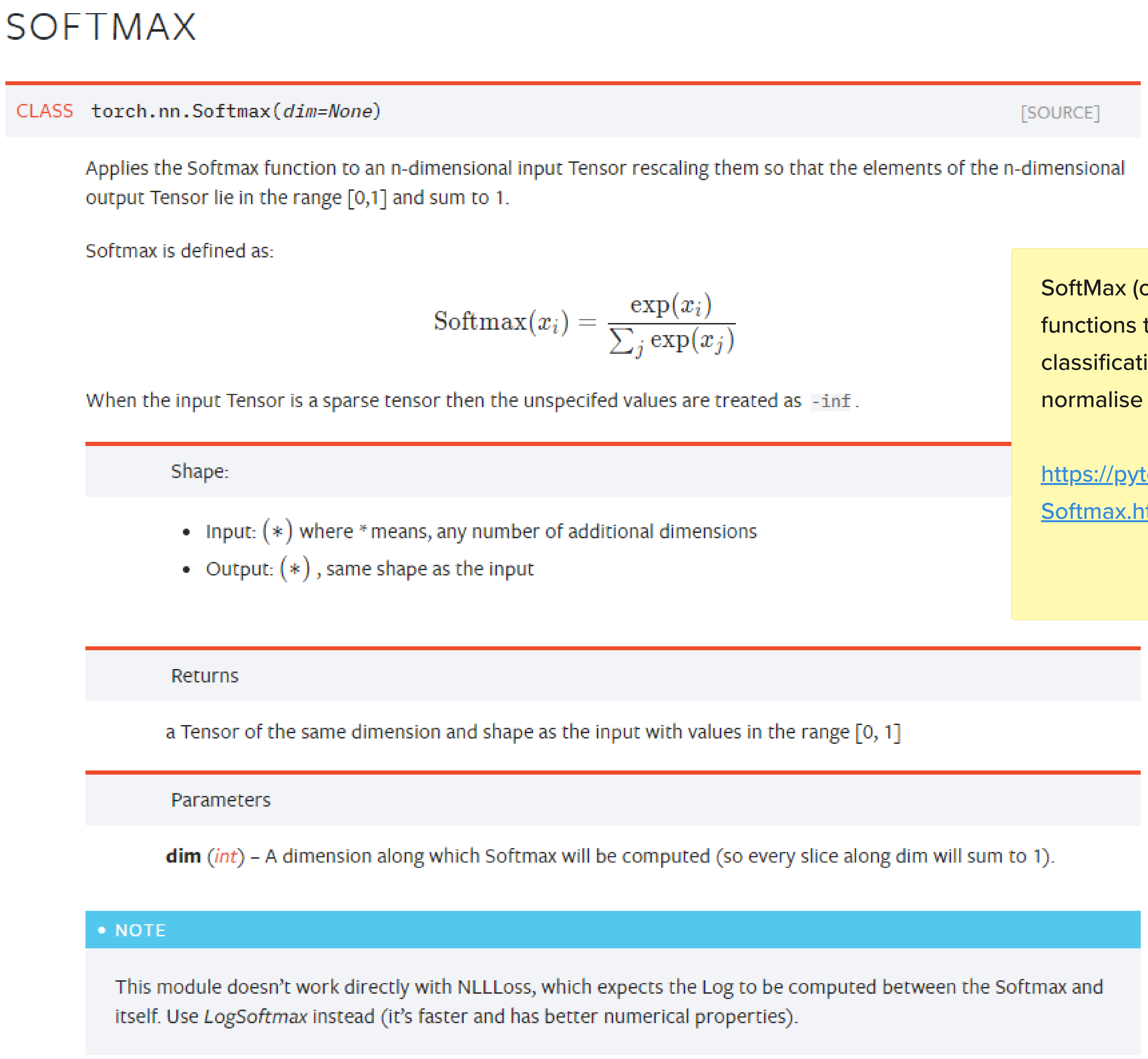
Lecture 4 Q&A

23 March 2021

What is "Softmax"?



SOFTMAX

CLASS `torch.nn.Softmax(dim=None)` [SOURCE]

Applies the Softmax function to an n-dimensional input Tensor rescaling them so that the elements of the n-dimensional output Tensor lie in the range [0,1] and sum to 1.

Softmax is defined as:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

When the input Tensor is a sparse tensor then the unspecifed values are treated as `-inf`.

Shape:

- Input: $(*)$ where * means, any number of additional dimensions
- Output: $(*)$, same shape as the input

Returns

a Tensor of the same dimension and shape as the input with values in the range [0, 1]

Parameters

dim (*int*) – A dimension along which Softmax will be computed (so every slice along dim will sum to 1).

• NOTE

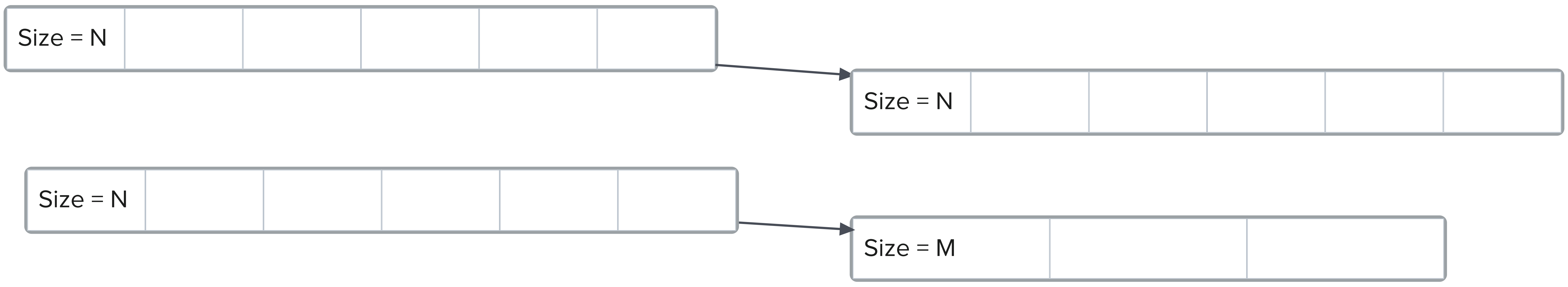
This module doesn't work directly with NLLLoss, which expects the Log to be computed between the Softmax and itself. Use *LogSoftmax* instead (it's faster and has better numerical properties).

SoftMax (or SoftArgMax) is one of activation functions that can be used for multi-class classification problems solving model. As SoftMax normalise vector to have a probability distribution.

https://pytorch.org/docs/stable/generated/torch.nn.Softmax.html

What the difference between "N to N" and "N to M" problem?

In "N to N" problem, the number of elements of the input sequence is the same as the number in the output sequence, whereas "N to M" problem has different length of input and output sequence length.



What is "Xn"?

Xn is the n-th element of the input sequence, the n-th token in one input sample

How to define the index for w, are there any rules?

There is no "index" for w, here we just use W_hh and W_xh to represent to two different weight matrix

Can we choose other activation functions?

Yes, you can, but you need to choose right function for the specific purpose. Normally we use different functions for hidden layers and output layer. Most of activation functions are non-linear.

Why do we need data transformation?

Sometimes you would like to transform data to make the data in different format (or structure). Word2Vec can also also be called as "Data Transformation". However, more typical examples would be:
To create features for images: for image captioning or visual question answering domain (in NLP related area)

During training of RNN, does each input need to be the same sentence length? (i.e. same number of vector inputs)

In practice, we often pad with empty tokens to make all input sequences the same length - we will see this in the lab materials

Are we sharing parameters for each RNN sequence?

Each layer time shares parameters. Main purpose of sharing parameters are reducing the number of parameters to be trained.