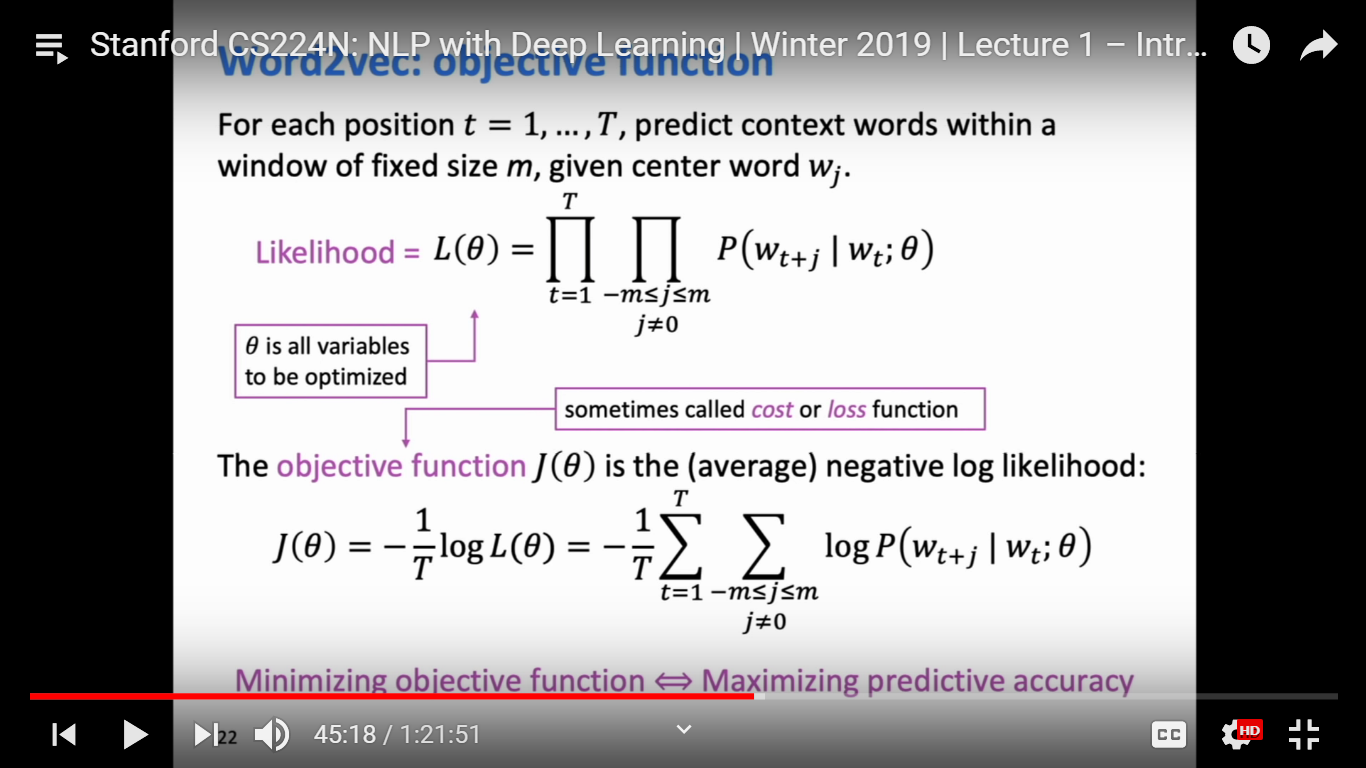
Full slide

<http://web.stanford.edu/class/cs224n/slides/cs224n-2019-lecture01-wordvecs1.pdf>



Context vector and target/output vector

Orange part: Dot product: same sign)+/-) and same position / one small and other big will decide how dot product will be big or small.

exponential will do always +ive

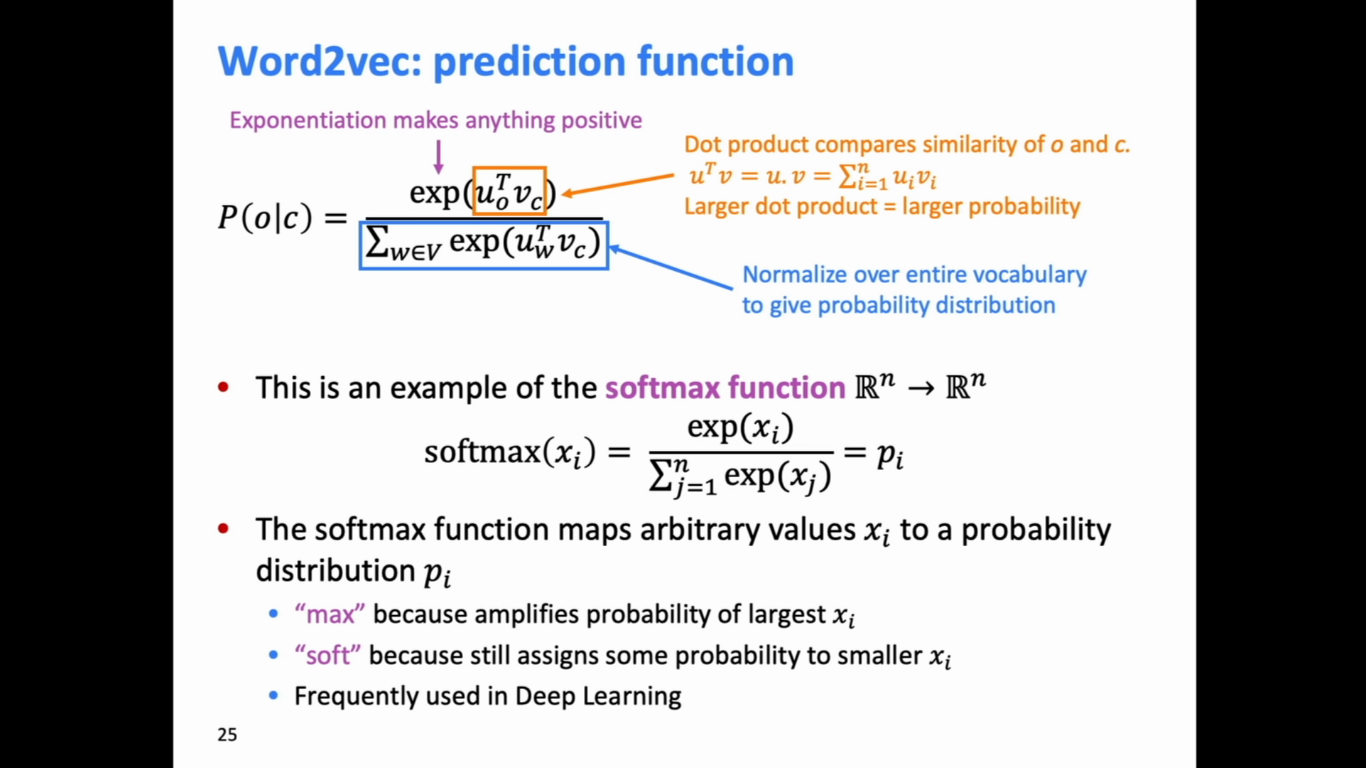
Sum all and devide

Softmax vs hard max:

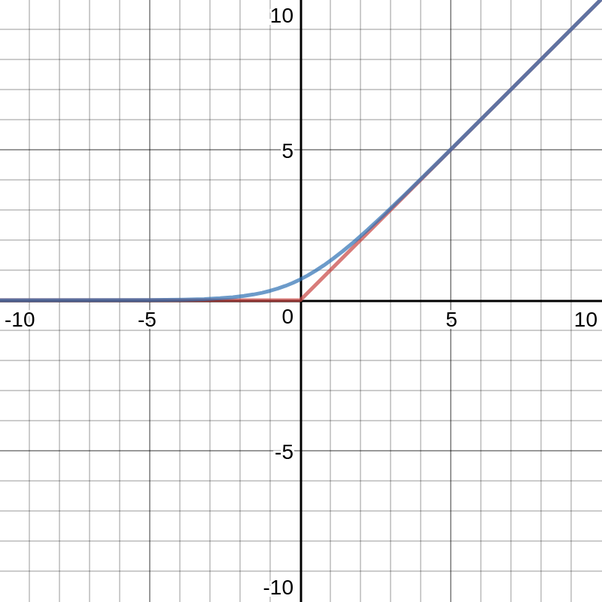
Max: make big number bigger

Soft: distribute probability everywhere

<https://www.quora.com/Why-is-softmax-activate-function-called-softmax>



(blue).



**Why is it called Softmax?**

* It is an approximation of Max.
* It is a *soft/smooth* approximation of max. Notice how it approximates the sharp corner at 0 using a smooth curve.

**What is the purpose of Softmax?**

Softmax gives us the [differentiable](https://en.wikipedia.org/wiki/Differentiable_function) approximation of a [non-differentiable](https://math.stackexchange.com/questions/1329252/is-max0-x-a-differentiable-function) function max. Why is that important? *For optimizing models, including machine learning models, it is required that functions describing the model be differentiable. So if we want to optimize a model which uses the max function then we can do that by replacing the max with softmax****.***

Understand below part timing and

Differentiation

