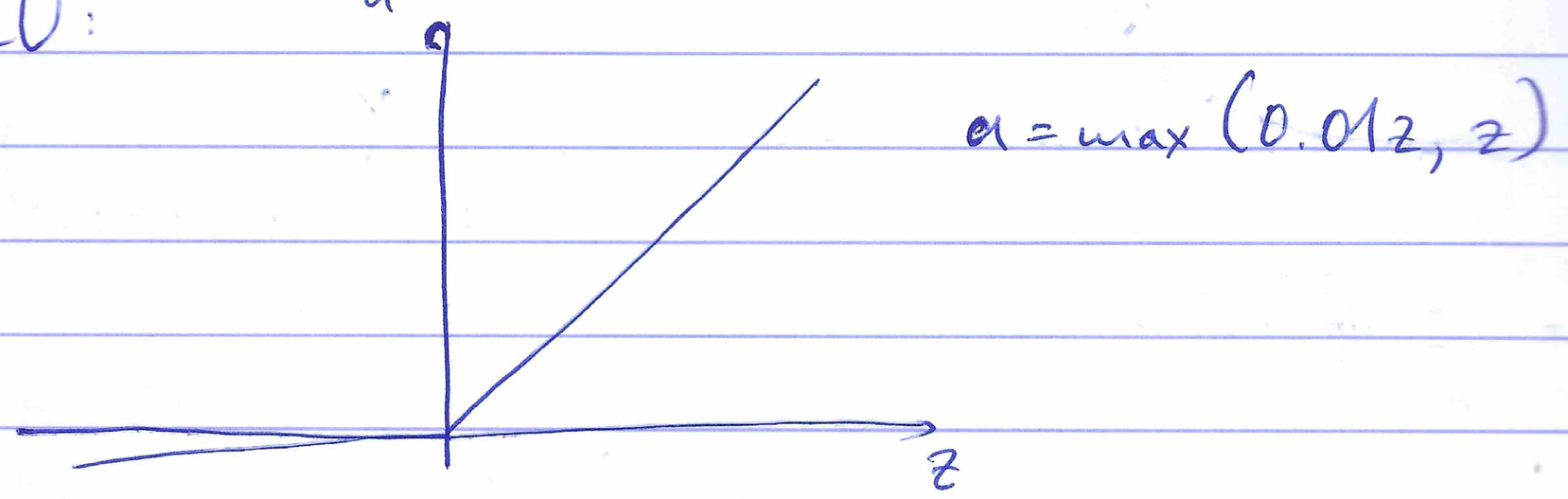
For a lot of the space of z, the slope of g is very different from 0, so the net will learn much fa

Leaky ReLU:



Why do you need non-linear activation functions?

Say $g(z) = z$ (linear activ. func.):

$$z^{[1]} = W^{[1]}x + b^{[1]}$$
$$a^{[1]} = z^{[1]}$$
$$z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a^{[2]} = z^{[2]}$$

$$a^{[1]} = z^{[1]} = W^{[1]}x + b^{[1]}$$
$$a^{[2]} = z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$
$$a^{[2]} = W^{[2]}\underbrace{(W^{[1]}x + b^{[1]})}_{z^{[1]}} + b^{[2]}$$
$$= \underbrace{(W^{[2]}W^{[1]})}_{W'}x + \underbrace{(W^{[2]}b^{[1]} + b^{[2]})}_{b'}$$
$$= W'x + b'$$

A linear hidden layer is practically useless. The combo of 2 linear function is a non-linear function.

This might only be useful if you're doing ML on regression problems - maybe the output layer (if y ∈ ℝ, $0 ... $1,000,000 Choosing prices?)