ML to DL

Simple linear model cannot be piecewise or curved, so there is Model Bias

New model use more features and new structure, therefore eliminate the model bias

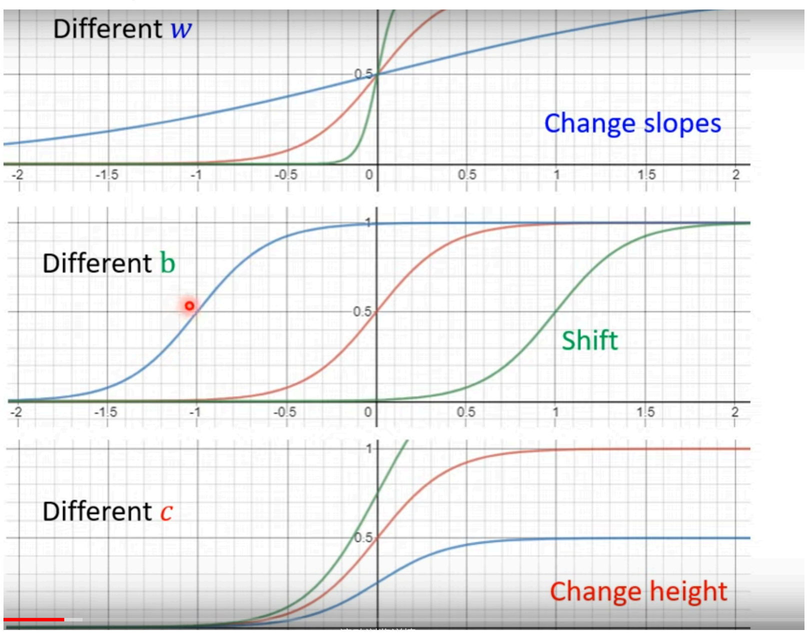
$$y = b + wx_1$$

$$y = b + \sum_i c_i \; sigmoid(b_i + w_i x_1)$$

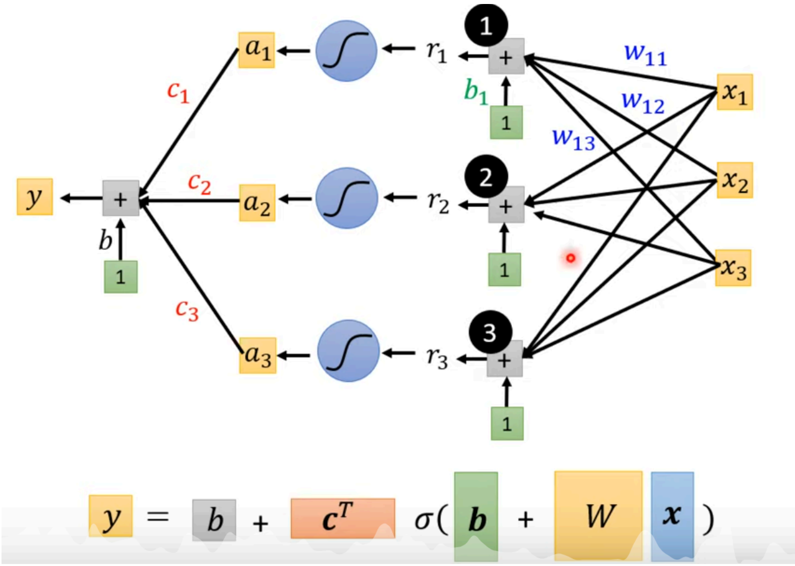
$$y = b + \sum_j w_j x_j$$

$$\boxed{y = b + \sum_i c_i \; sigmoid\left(b_i + \sum_j w_{ij} x_j\right)}$$

How the parameters effect the model



The model can be written in matrix multiplication



- x: features
- other parameters are unkown, set randomly, **but can be optimized**

Optimization

Put all unkown parameters (w_{ij}, b_i, c_i, b) into a column, θ
How do we measure the parameters are good or not?
-- By **Loss function** $L(\theta)$
To find the θ that minimize $L(\theta)$, we can use Gradient Descent

GD *vs* BGD: BGD speeds up the optimization
NOTE: 1 epoch = see all the batches once

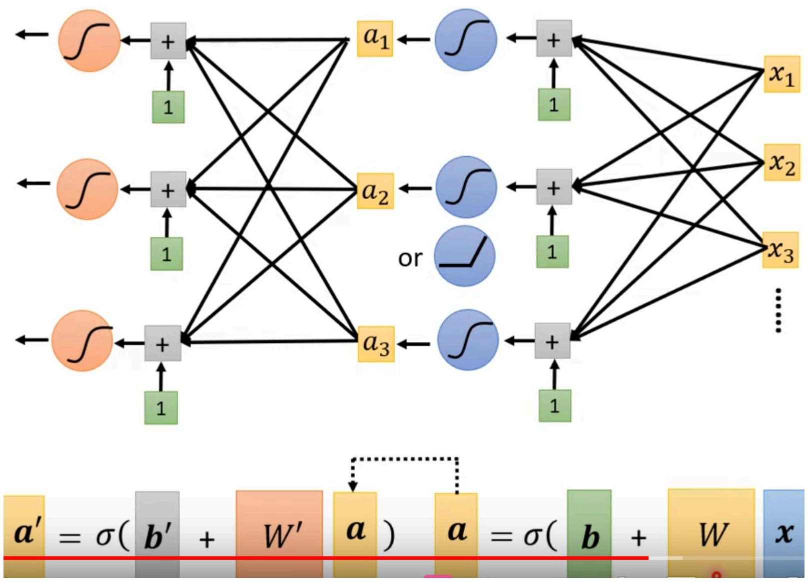
Activation Function

Remember that Sigmoid was used to introduce non-linearity into the model.
In fact, there are other functions that have the same effect, called **Activation Function**

- Sigmoid
- ReLU
- ...

More Layers

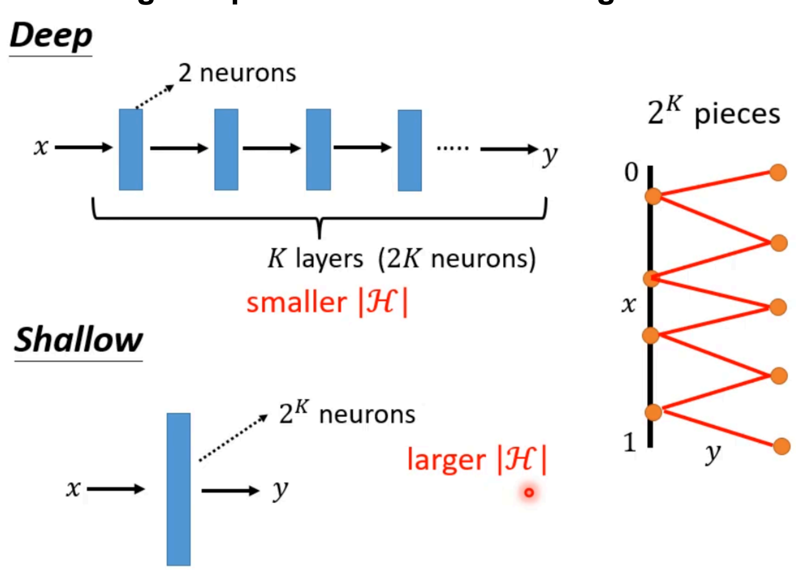
After applying the Activation function, the output can be seen as new features, thus our model would be complex.



Deep Learning: "deep" indicates many hidden layers

Why *Deep*, not *Fat* network?

Given the same number of parameters, fat + short v.s. thin + tall, which one is better? -- **Thin + tall wins.**
Deep network uses parameters effciently by addressing compound information layer by layer. Thus, deep network is **more efficient with same parameters** and it is **less likely to be overfitting compared to shallow learning**.



Deep learning yields highly **complex and structured** models (as depicted in the picture), which excel in tasks requiring complex and structured function mappings (e.g. image, speech recognition...)

Hyperparameters

- Learning rate
- Batch size

- number of neurons per layer (number of activation functions)
- type of activation function
- number of layers
- ...