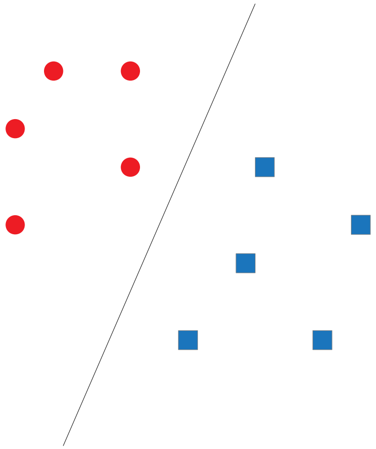
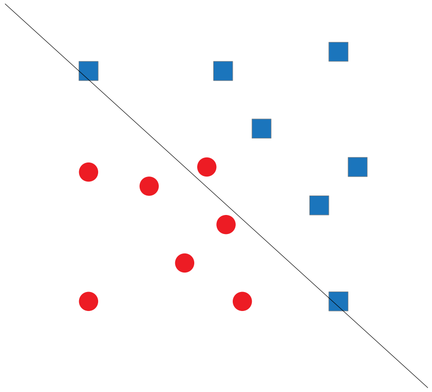
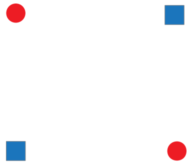
Nonlinear Models

Why Nonlinear Models?

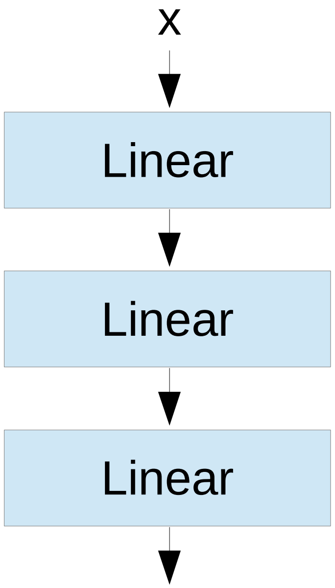




Example: XOR

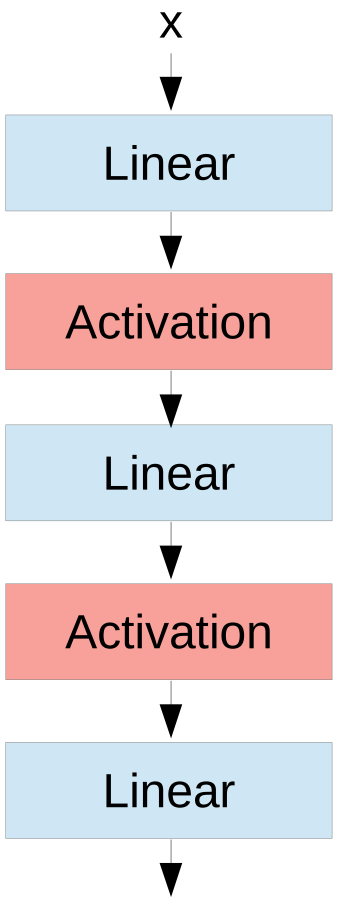


Extra Linear Layers Don’t Help



$$\begin{aligned} W_2(W_1 x + b_1) + b_2 &= W_2 W_1 x + W_2 b_1 + b_2 \\ &= (W_2 W_1) x + (W_2 b_1 + b_2) \end{aligned}$$

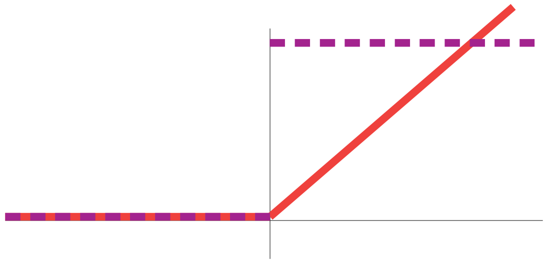
Activation Functions



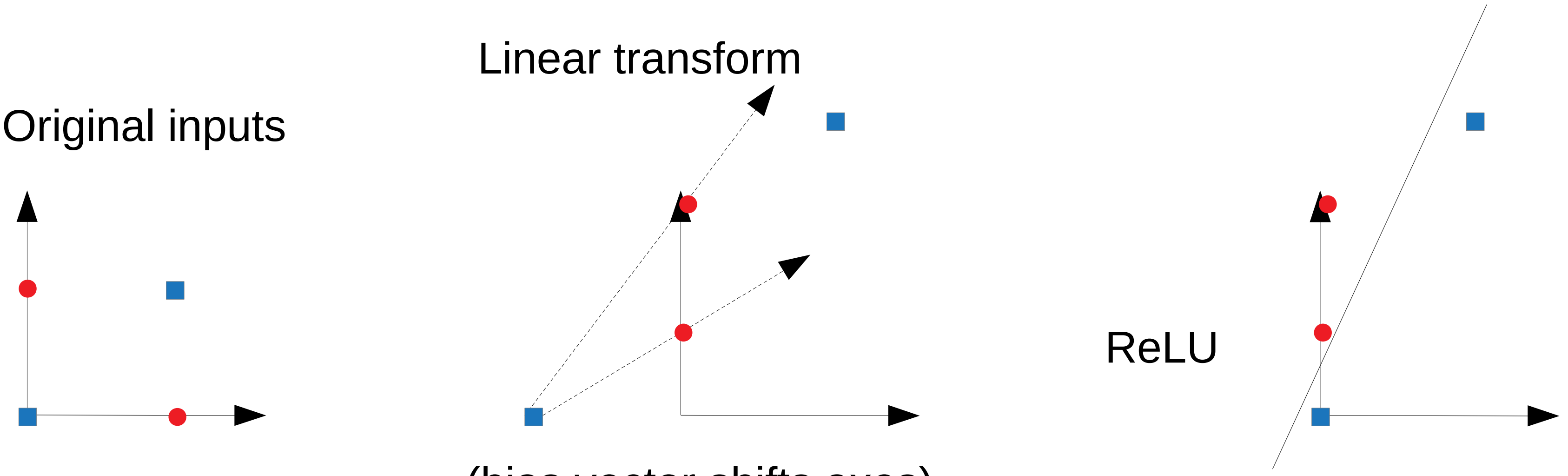
Nonlinear functions that are differentiable (almost) everywhere

Rectified Linear Unit – ReLU

$$\mathrm{ReLU}(x) = \max(x, 0)$$



XOR



Original inputs

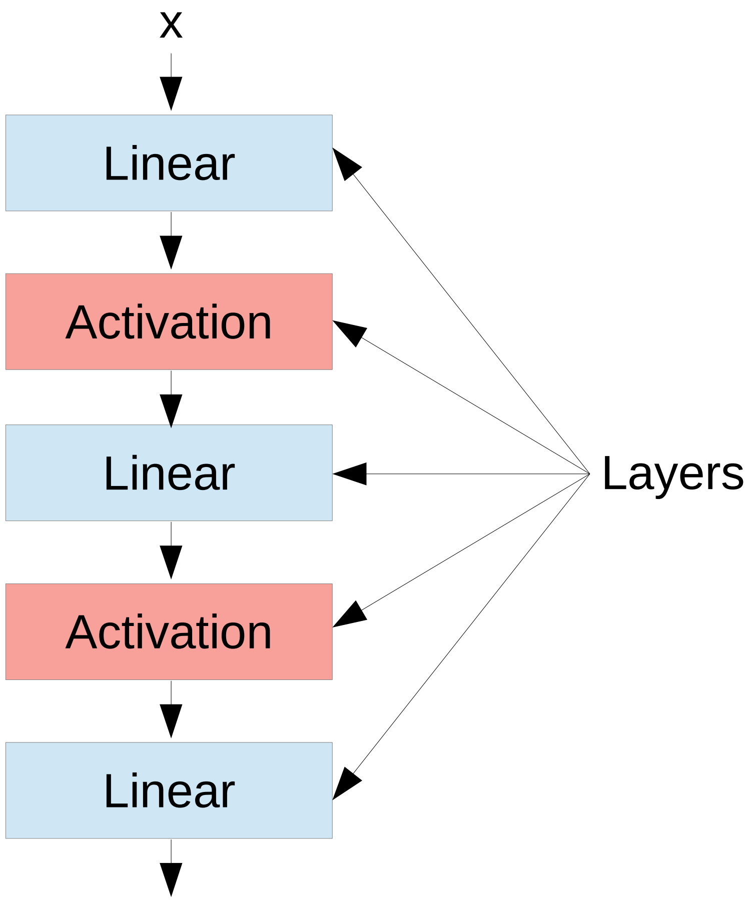
Linear transform

(bias vector shifts axes)

ReLU

Linearly seperable

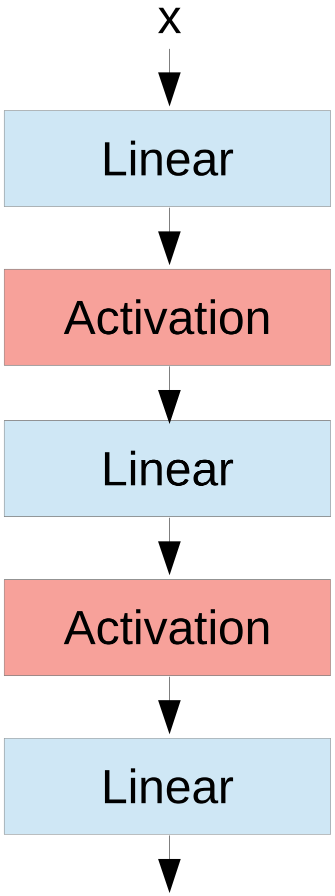
Layers



Only non-activation layers are
counted when we describe
the depth of a network

E.g., this is a three-layer network

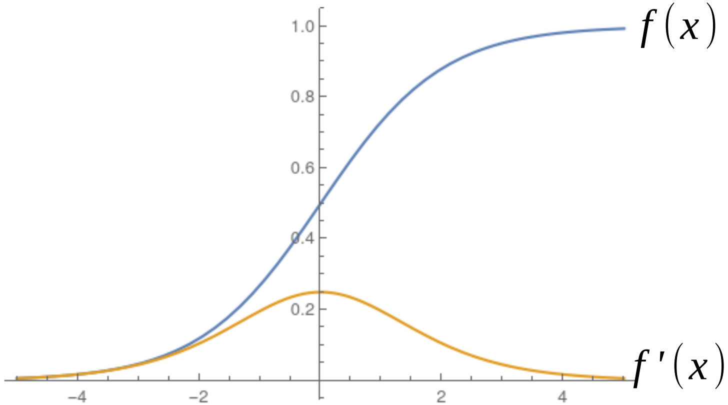
Deep Networks



- Alternation of linear layers with activation functions
- Can approximate any continuous function (assuming a sufficient number of layers and sufficient width)

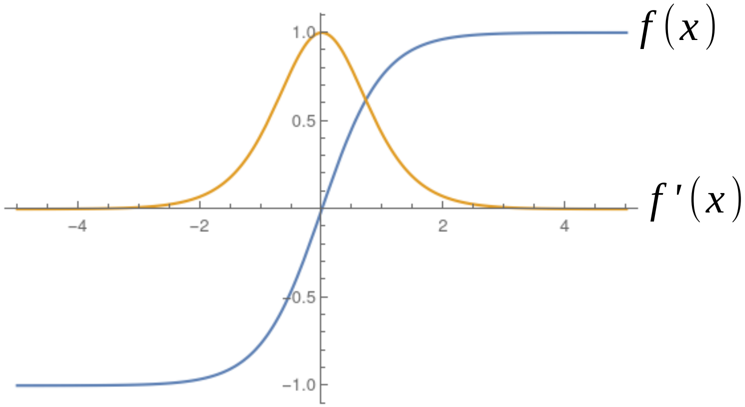
Activation Functions – Old

sigmoid



$$f(x)=\frac{1}{1+e^{-x}}$$

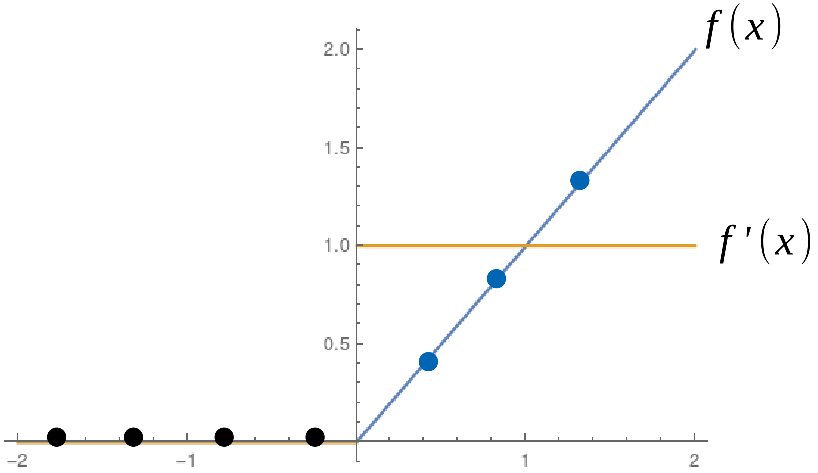
tanh



$$f(x)=\tanh(x)=\frac{e^{2x}-1}{e^{2x}+1}$$

Don’t use these!

Activation Functions – ReLU



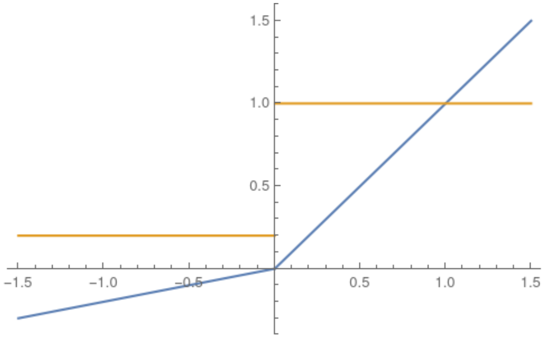
“Dead ReLU”

Initialize carefully

Smaller learning rate

Solving the “Dead ReLU” Problem

Leaky ReLU

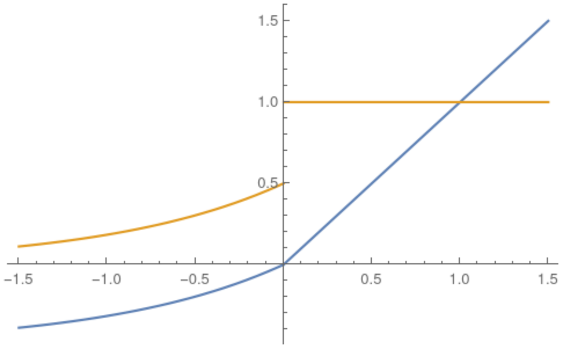


$$f(x) = \max(x, \alpha x)$$

$$0 < \alpha < 1$$

α can be a learned parameter:
“Parameterized ReLU (PReLU)”

Exponential Linear Unit (ELU)



$$f(x) = \begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

Choosing Activations

- Start with ReLU
  - Initialize carefully (we’ll talk about this more later)
  - Use a small learning rate
- If ReLU fails, try a leaky ReLU or PReLU
- Don’t use sigmoid or tanh