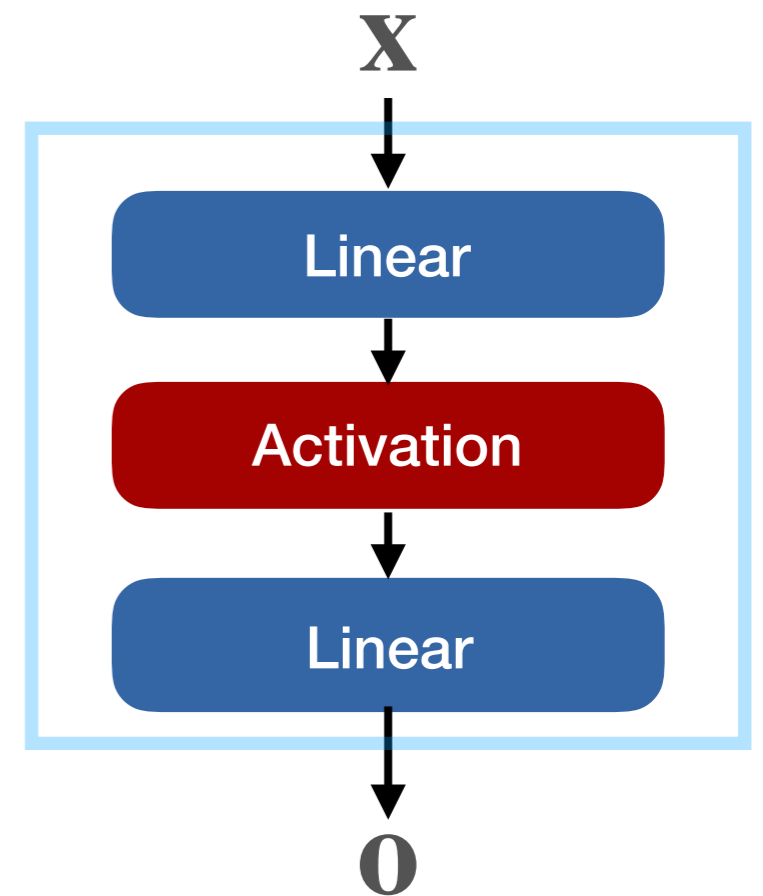


Activation functions

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Non-linearities

- Allow a deep network to model arbitrary differentiable functions



Zoo of activation functions

ReLU

Leaky ReLU

PReLU

ELU

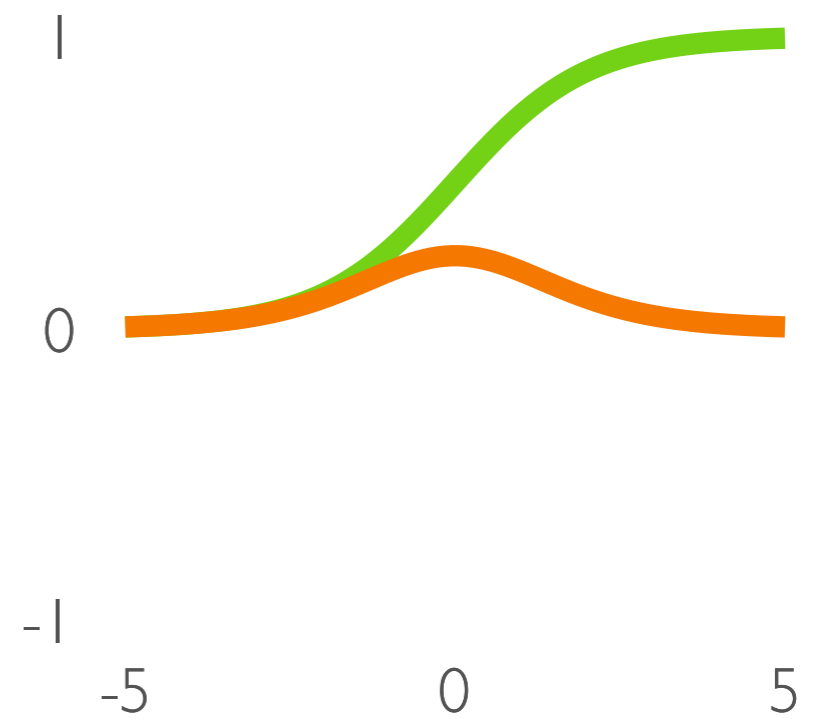
tanh



Sigmoid

Maxout

Sigmoid

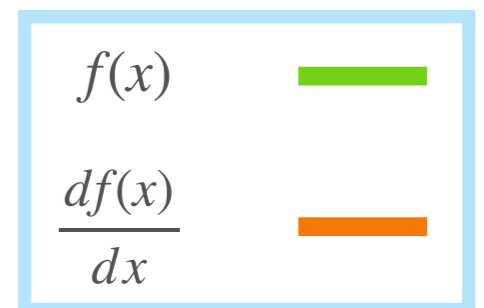
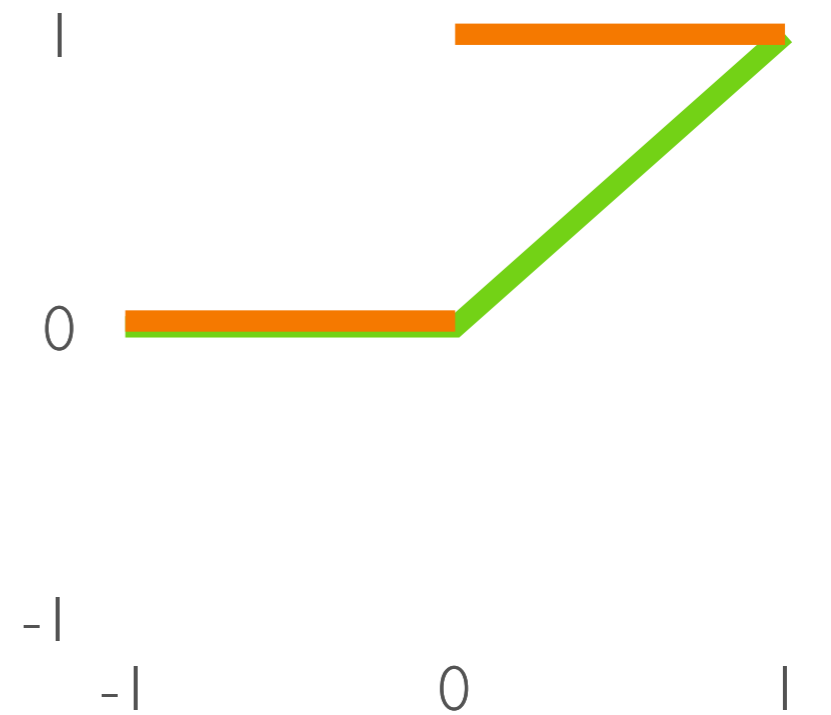
- $\frac{1}{1 + e^{-x}}$
- Same as tanh
- Do not use



$f(x)$	
$\frac{df(x)}{dx}$	

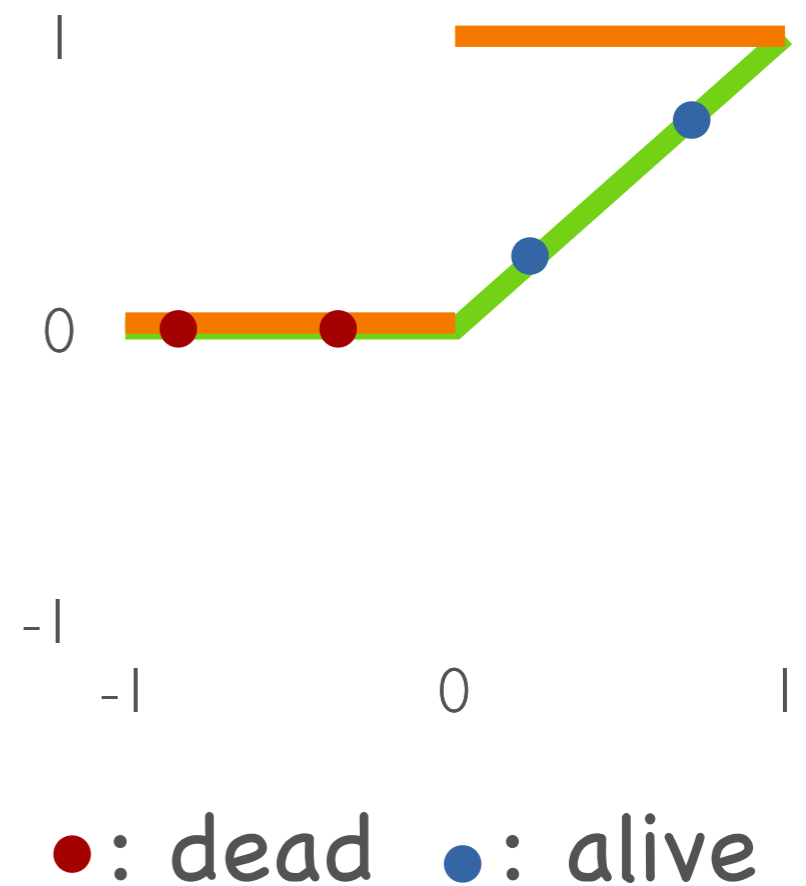
ReLU

- $\max(x, 0)$



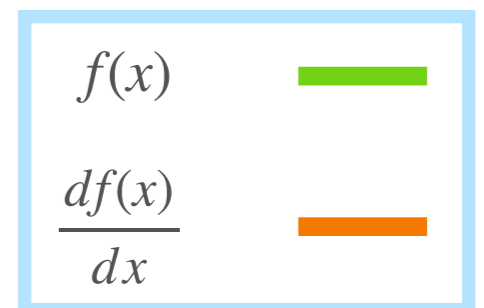
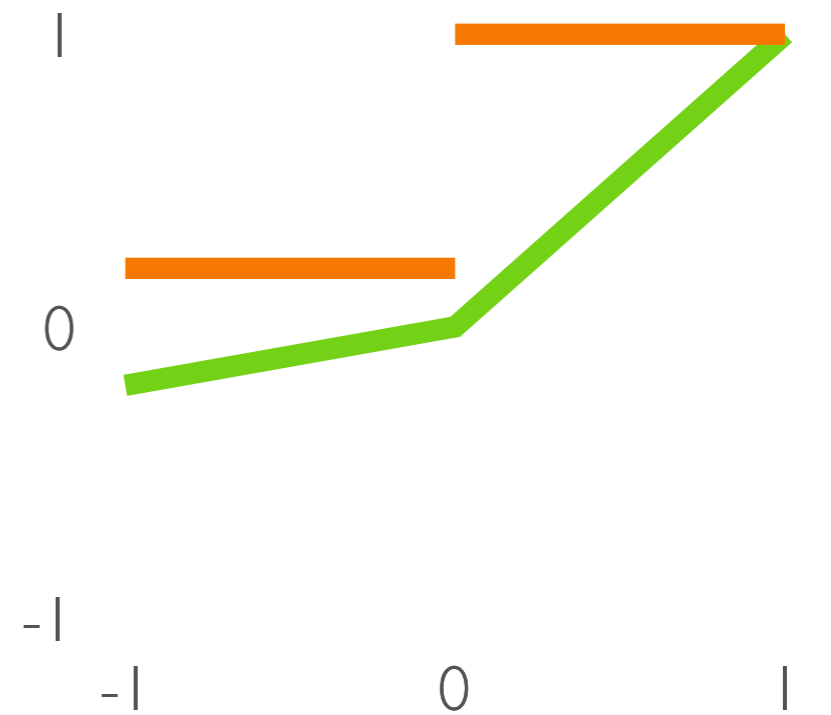
Dead ReLUs

- Prevent dead ReLUs:
- Initialize Network carefully
- Decrease the learning rate



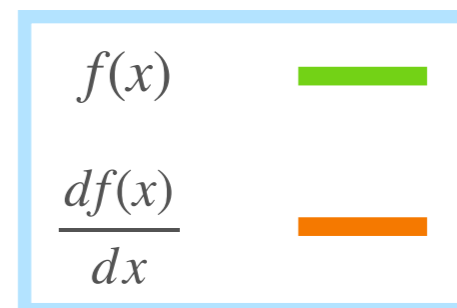
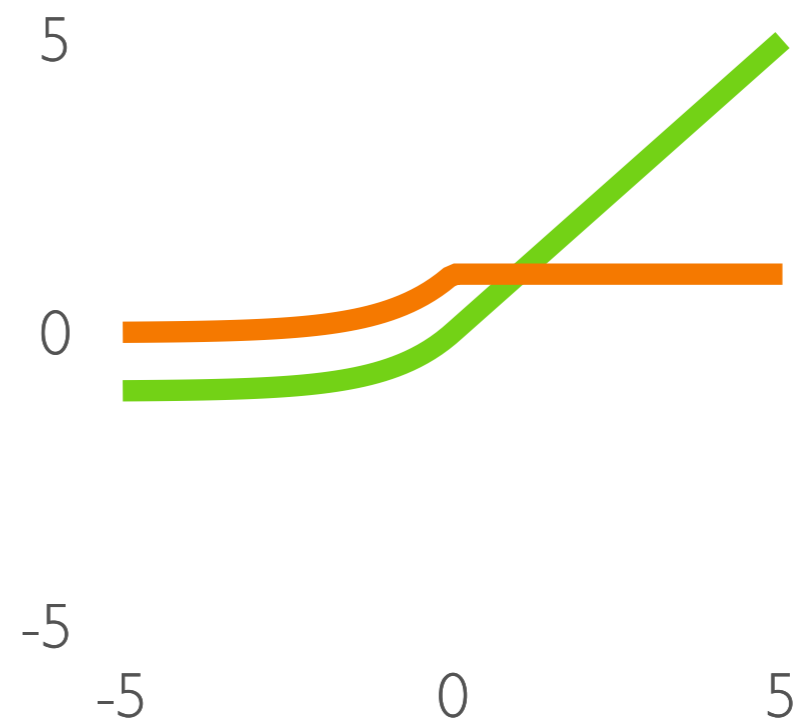
Leaky ReLU

- $\max(x, \alpha x)$
- For $0 < \alpha < 1$
- Called PReLU if α is learned



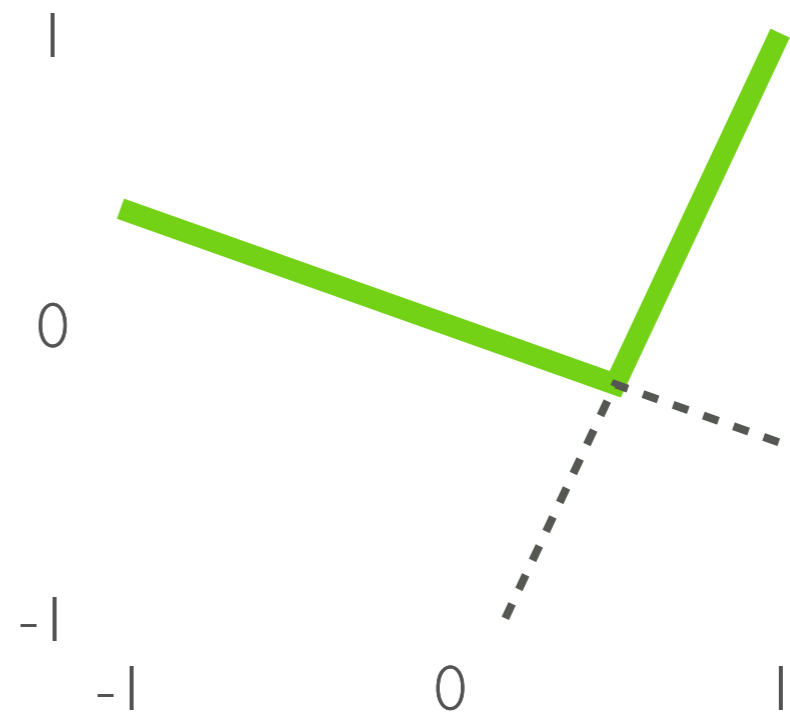
ELU

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



Maxout

- $\max(x_1, x_2)$



Which activation to choose?

- ReLU
 - Carefully initialize
 - Small enough learning rate
- If ReLU fails, try:
 - Leaky ReLU / PReLU
- Avoid sigmoid and tanh