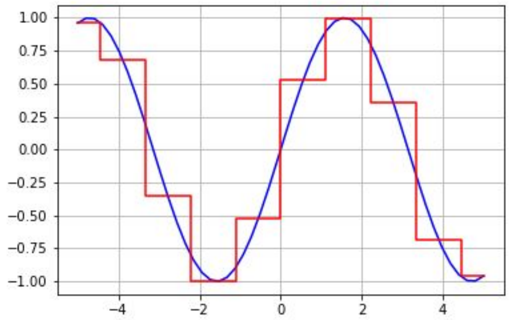
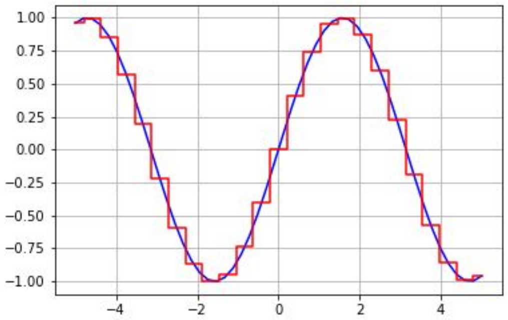
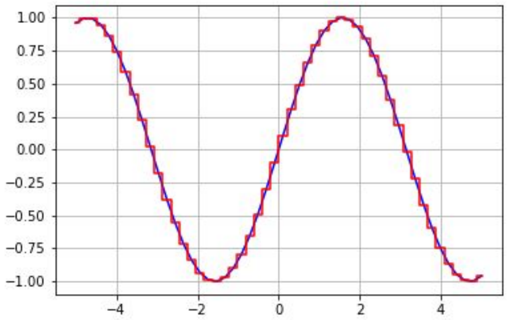
Week 5
Different Types of Activation Units

Universal function approximation

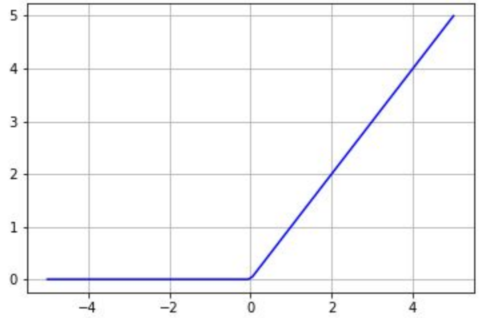
- Non-linear function enables approximation to any function
-
- Example: Unit step function
- Problem: The gradient of the unit step function is zero everywhere

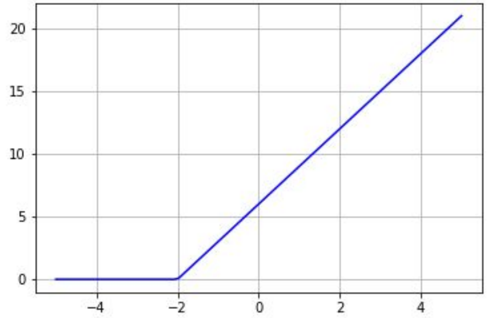


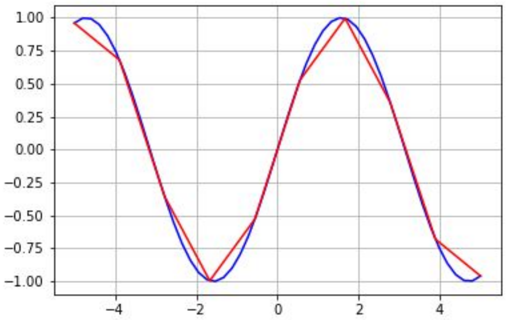


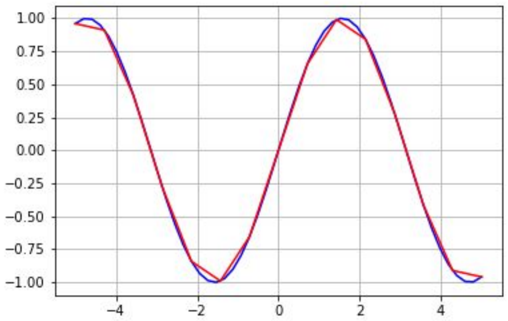


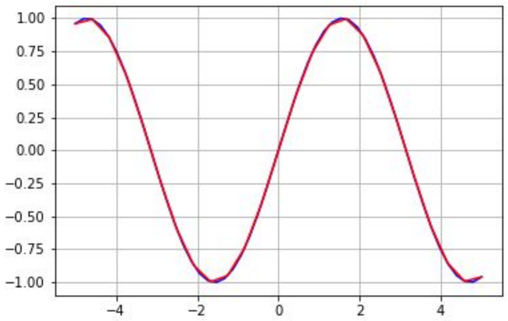
Ex: ReLU


0 1 2 3 4 5
−4 −2 0 2 4


0 5 10 15 20
−4 −2 0 2 4


1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00
−4 −2 0 2 4

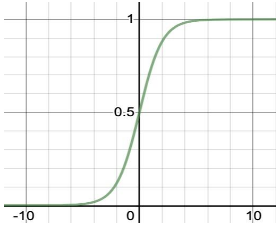

1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00
−4 −2 0 2 4


1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00
−4 −2 0 2 4

Comparison of Activation Units

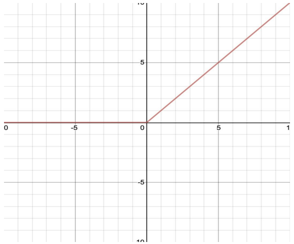
- Sigmoid

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



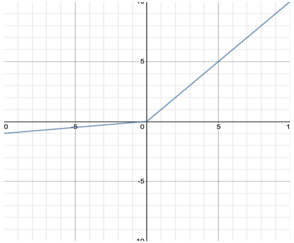
- ReLU

$$f(z) = \begin{cases} z & if \quad z \geq 0, \\ 0 & else. \end{cases}$$



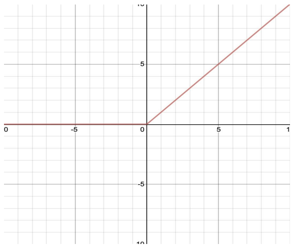
- Parametric ReLU

$$f(z) = \begin{cases} z & if \quad z \geq 0, \\ az & else. \end{cases}$$

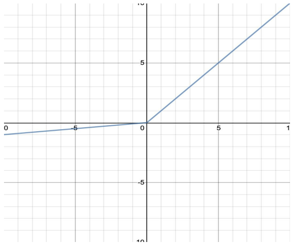


ReLU

- ReLU doesn’t saturate when z approaches infinity
  - First derivative has constant 1 when z > 0
  - Less likely to have vanishing gradient
- ReLU dies when z < 0
  - Doesn’t solve the vanishing gradient problem in the z<0 region



- Parametric ReLU
  - First derivative has non zero value everywhere
  - Solves the dying ReLU problem

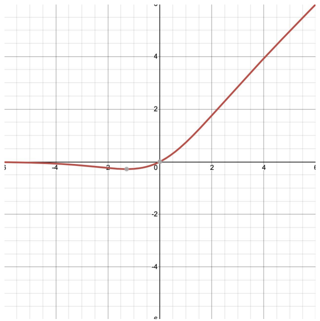


Swish and Mish

- Latest state of the art activation functions
- Swish

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

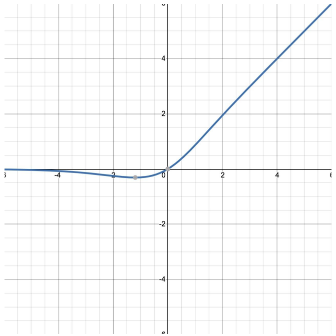
1. Unbounded above, bounded below
2. Non-monotonicity
3. Differentiable everywhere



- Mish -- Improve upon swish

$$f(z) = z \tanh\left(\ln\left(1 + e^{z}\right)\right)$$

1. First derivative is preconditioned



Gradient Vanishing/Exploding

- Recall back-propagation

$$\nabla_{\mathbf{W}^{(l)}}\mathcal{L} = \delta^{(l)}\mathbf{a}^{(l-1)T}$$

- Gradients are proportional to the multiplication of derivatives of activation functions and weight matrices in the following layers

$$\begin{aligned}\delta^{(l)} &= \frac{\partial\mathcal{L}}{\partial\mathbf{z}^{(l)T}} \\ &= \mathrm{diag}(\sigma'(\mathbf{z}^{(l)}))\mathbf{W}^{(l+1)T}\mathrm{diag}(\sigma'(\mathbf{z}^{(l+1)}))\cdots\mathbf{W}^{(L)T}\mathrm{diag}(\sigma'(\mathbf{z}^{(L)}))\frac{\partial\mathcal{L}}{\partial\mathbf{y}^T}\end{aligned}$$

Gradient Vanishing/Exploding

- Happens when
  - the there are too many cascaded layers (0.9^**100** ≈ 0 or 1.1^**100** ≈ ∞)
  - the model is poorly initialized (**0.1**^10 ≈ 0 or **1.1**^100 ≈ ∞)
  - nonlinear functions are inappropriate (**0.1**^10 ≈ 0 or **1.1**^100 ≈ ∞)
  - activations / inputs are inappropriate (**0.1**^10 ≈ 0 or **1.1**^100 ≈ ∞)
  - …
- Solve it from the source
  - use fewer layers (?)
  - design suitable initialization methods (Xavier, Kaiming, etc)
  - change activation units
  - batch normalization
  - ...