Perceptron Learning Criterion

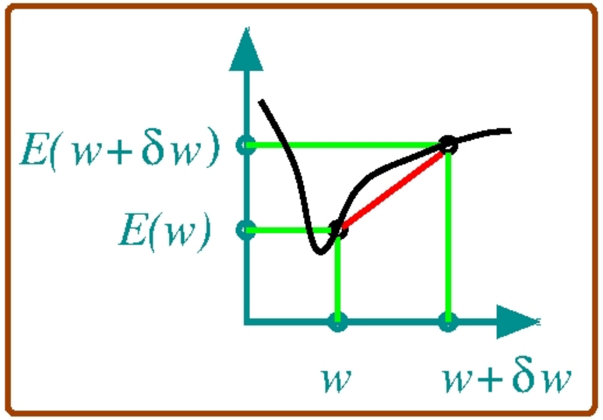
- 2-class classifier, i'th training point $\mathbf{x}^i$: $t_i = \pm 1$
- $y(\mathbf{x}^i) = h(\mathbf{w}^T \mathbf{x}^i) = h(a) = +1 (a \geq 0), = -1 (a < 0)$
- Declare Class $\mathscr{C}_1 (t_i = +1)$ if $\mathbf{w}^T \mathbf{x}^i \geq 0$
- Declare Class $\mathscr{C}_2 (t_i = -1)$ if $\mathbf{w}^T \mathbf{x}^i < 0$
- Combined: (SVM-like!) Correct, if $t_i \; \mathbf{w}^T \mathbf{x}^i \geq 0$
- Perceptron penalty: 0 if correct , else sum
- Simple criterion: $\boxed{E(\mathbf{w}) \stackrel{\triangle}{=} -\sum_{\forall i \in \mathscr{M}} \; t_i \; \mathbf{w}^T \mathbf{x}^i}$
- no penalty for 'how much' misclassification
- Solved numerically, Cauchy's weight update rule

Iterative Weight Update: Learning

https://upload.wikimedia.org/wikipedia/commons/d/d3/Augustin-Louis_Cauchy_1901.jpg

- A.-L. Cauchy [1789-1857]
- **Cauchy's Rule[1849]**
- $\boxed{\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E(\mathbf{w})}$

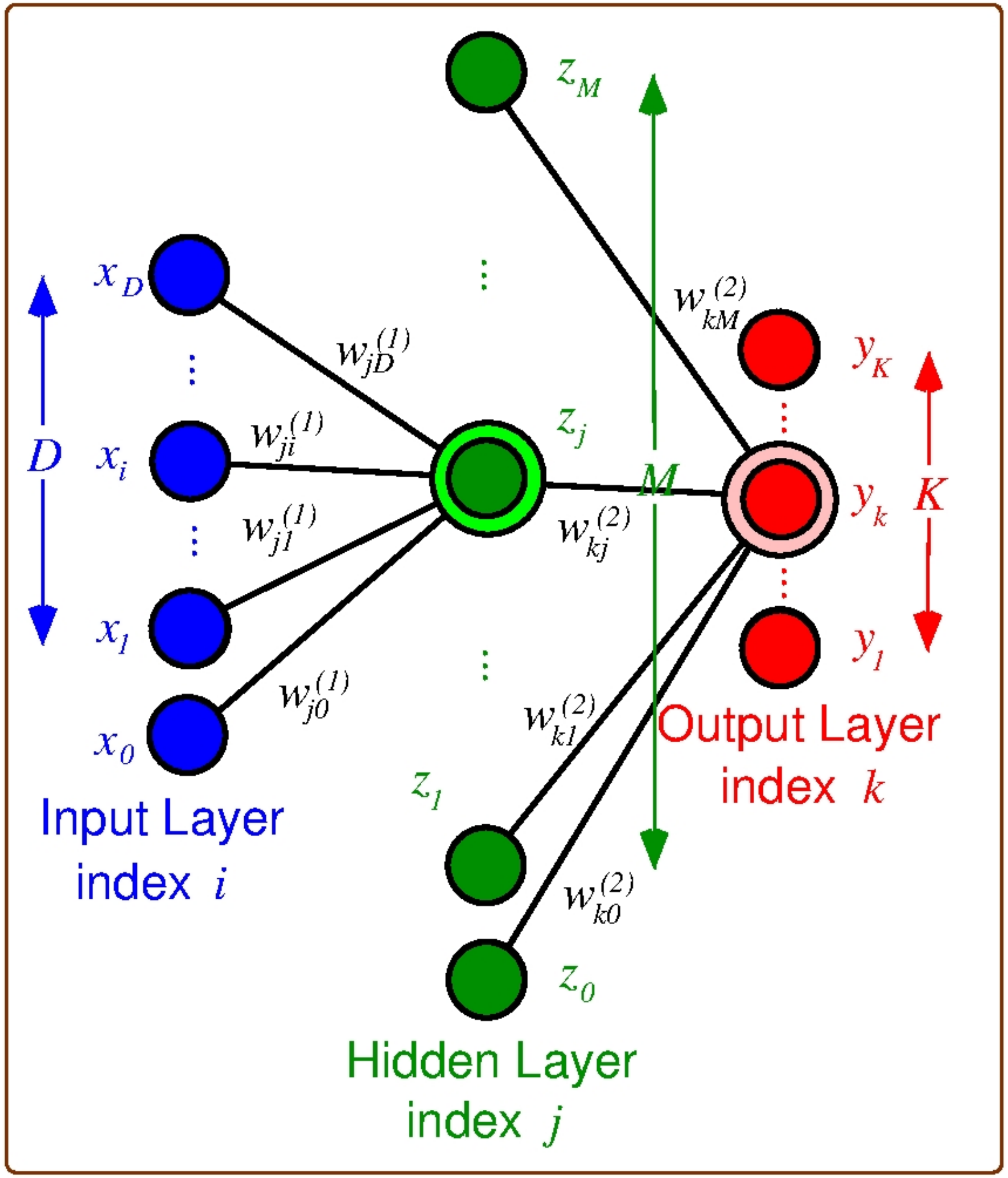
- dim cons?
- Step 'η'?
- why '-'?[00:44]



- sl = $\frac{E(w+\delta w) - E(w)}{(w+\delta w)-(w)} = \frac{E(w+\delta w) - E(w)}{\delta w}$
- $\lim_{\delta w \to 0} : \partial E / \partial w \Longrightarrow \nabla E(\mathbf{w})$
- $\mathbf{w}$: $[\ w_j\ ]$, $\nabla E(\mathbf{w}) = [\ \partial E / \partial w_j\ ]$
- '-' sign: go against the gradient!
- 'η': step size or learning rate, tuning (adaptive?)
- small η: small steps, longer attain local min
- large η: large steps, may miss local min

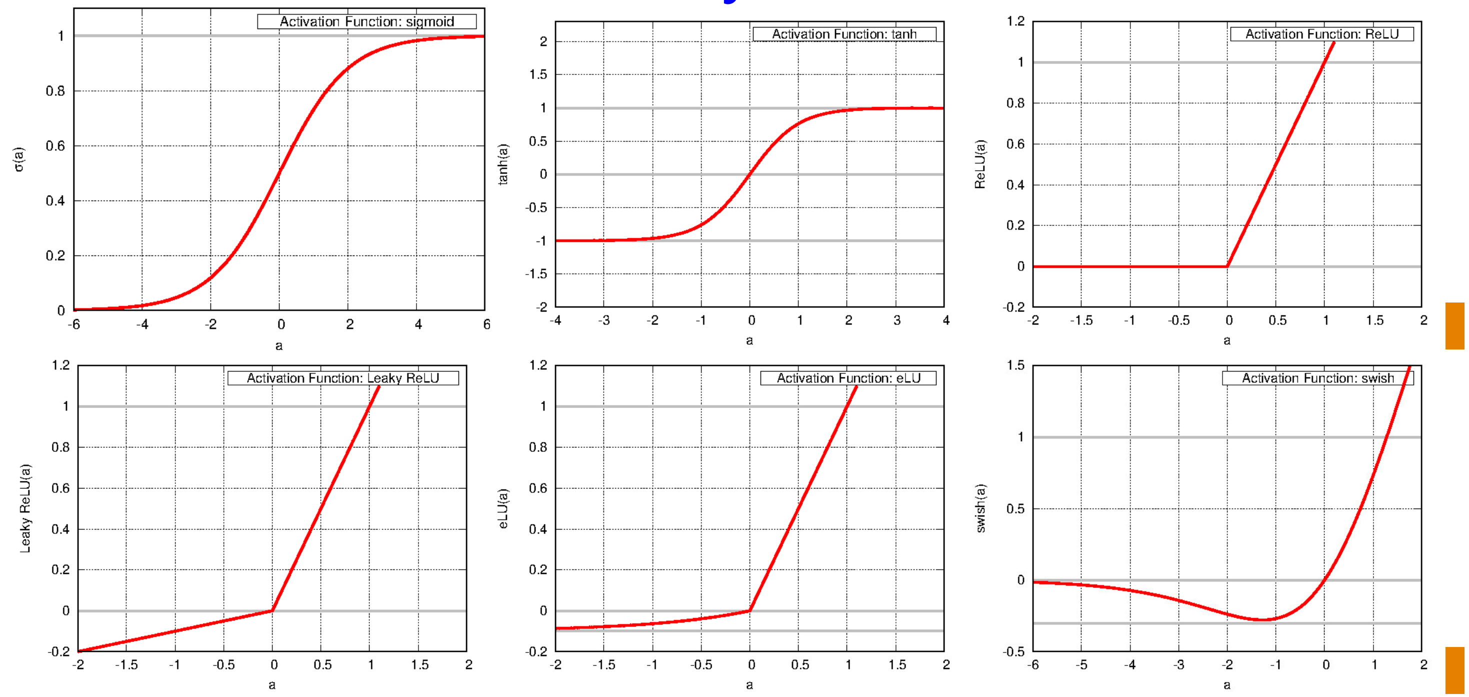
Home Page
Title Page
Contents
Page 8 of 30
Go Back
Full Screen
Close
Quit

Multi-Layer Perceptron



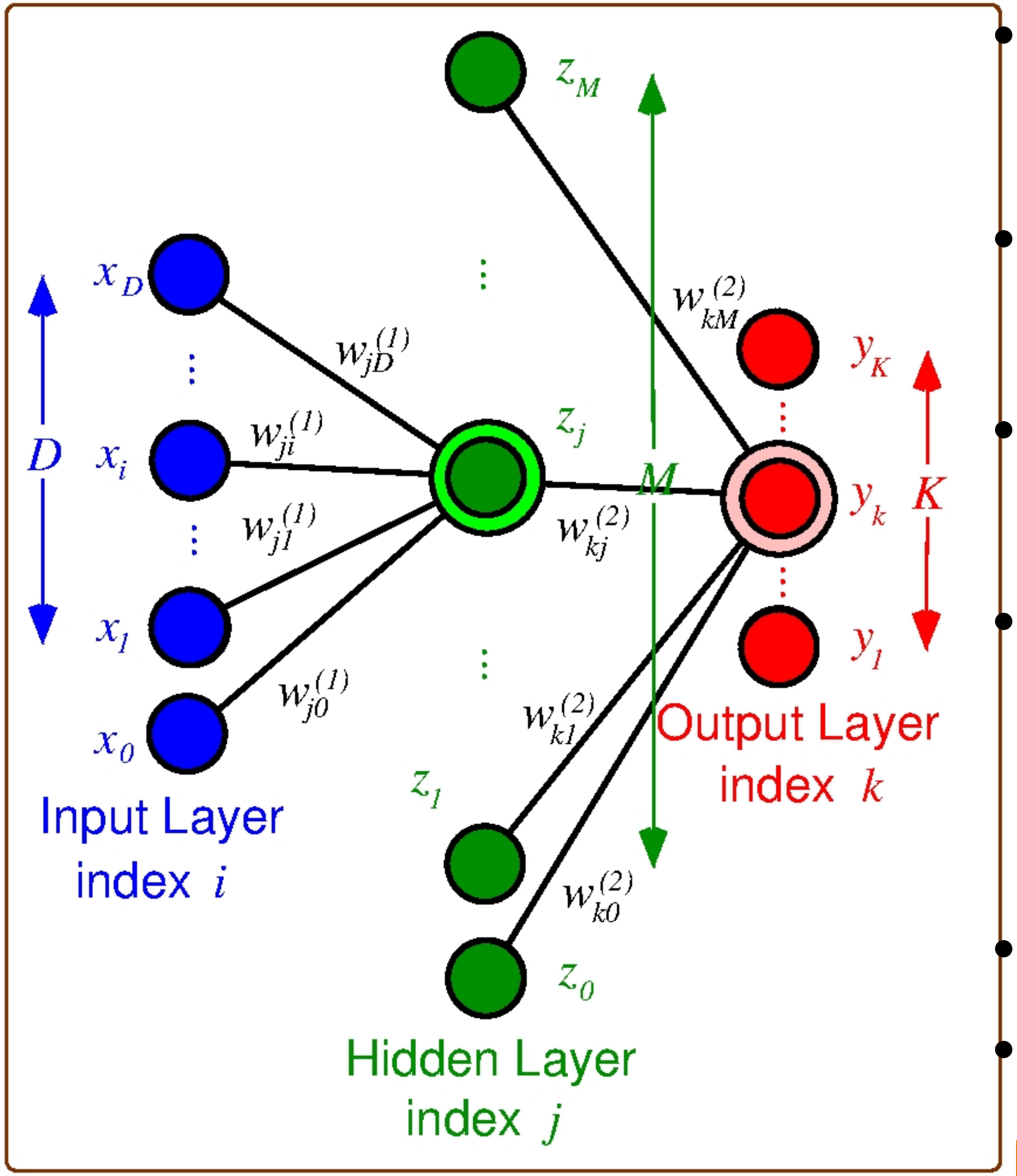
Activation Functions

- Neuron input: scalar, sum of weighted inputs
- Activation fn: possible non-linearity, scalar output
- sigmoid, tanh, RelU, Leaky ReLU, eLU, swish



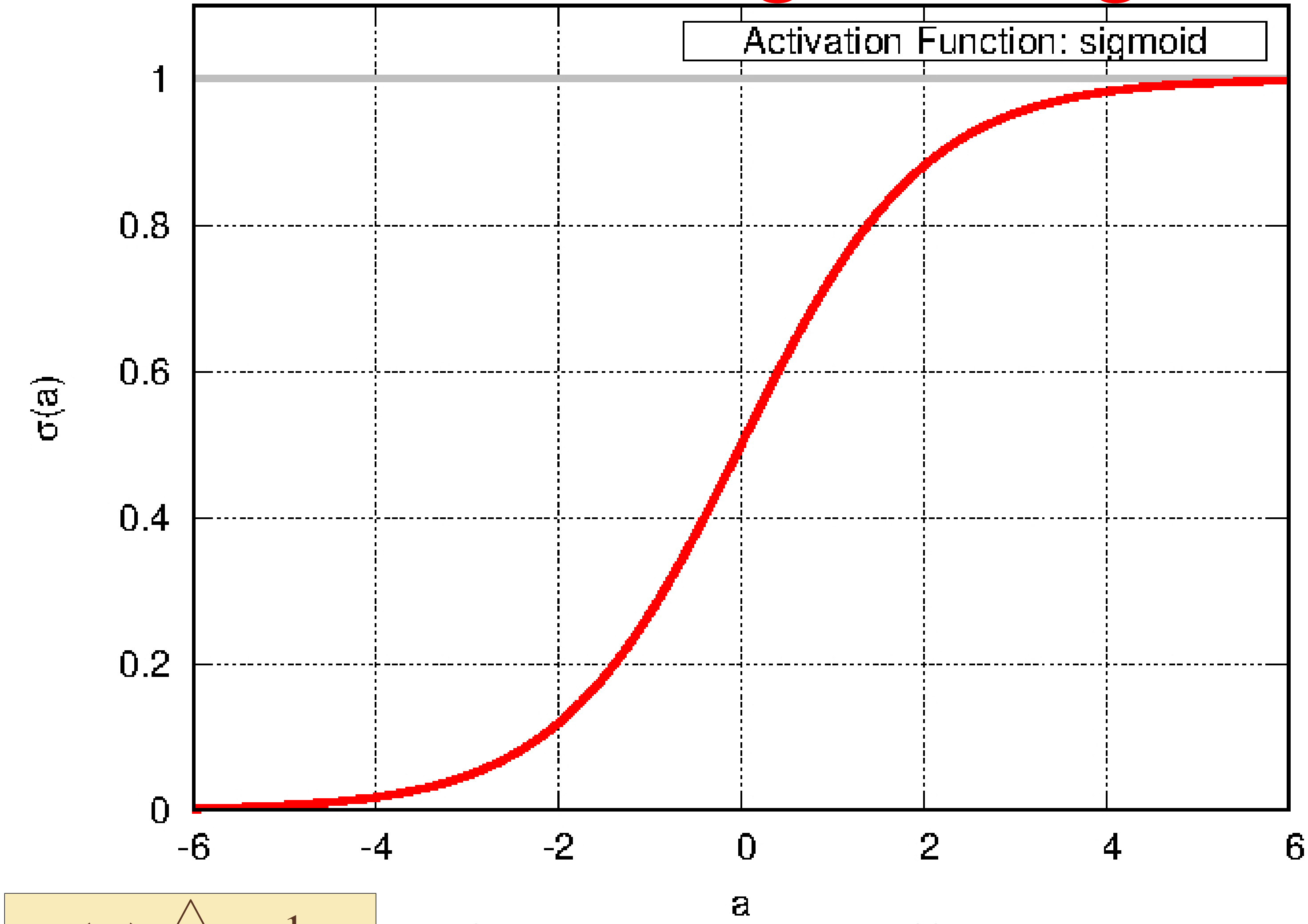
- $LeakyRelU(a,\alpha) \triangleq \max(\alpha a, a), \;\; \alpha \in (0,1)$
- $eLU(a,\alpha) \triangleq \left\{ \begin{array}{c} a, a > 0 \\ \alpha(e^a - 1), a \leq 0 \end{array} \right\}$ [00:50, 03:32]

Multi-Layer Perceptron



- h'layer activ'n fn: z_j: $h(a_j^{(1)})$: sigmoid/tanh
- o'layer activ'n fn: y_k: $\sigma(a_k^{(2)})$: prob specs
- Regression: Identity $y_k = a_k^{(2)}$
- Classification: sigmoid/softmax; sigmoid: 2-class softmax: multi-class
- softmax = $\exp / \sum \exp$
- Sgn: harsh $\tanh(\cdot)$; 0/1 step: harsh $\sigma(\cdot)$

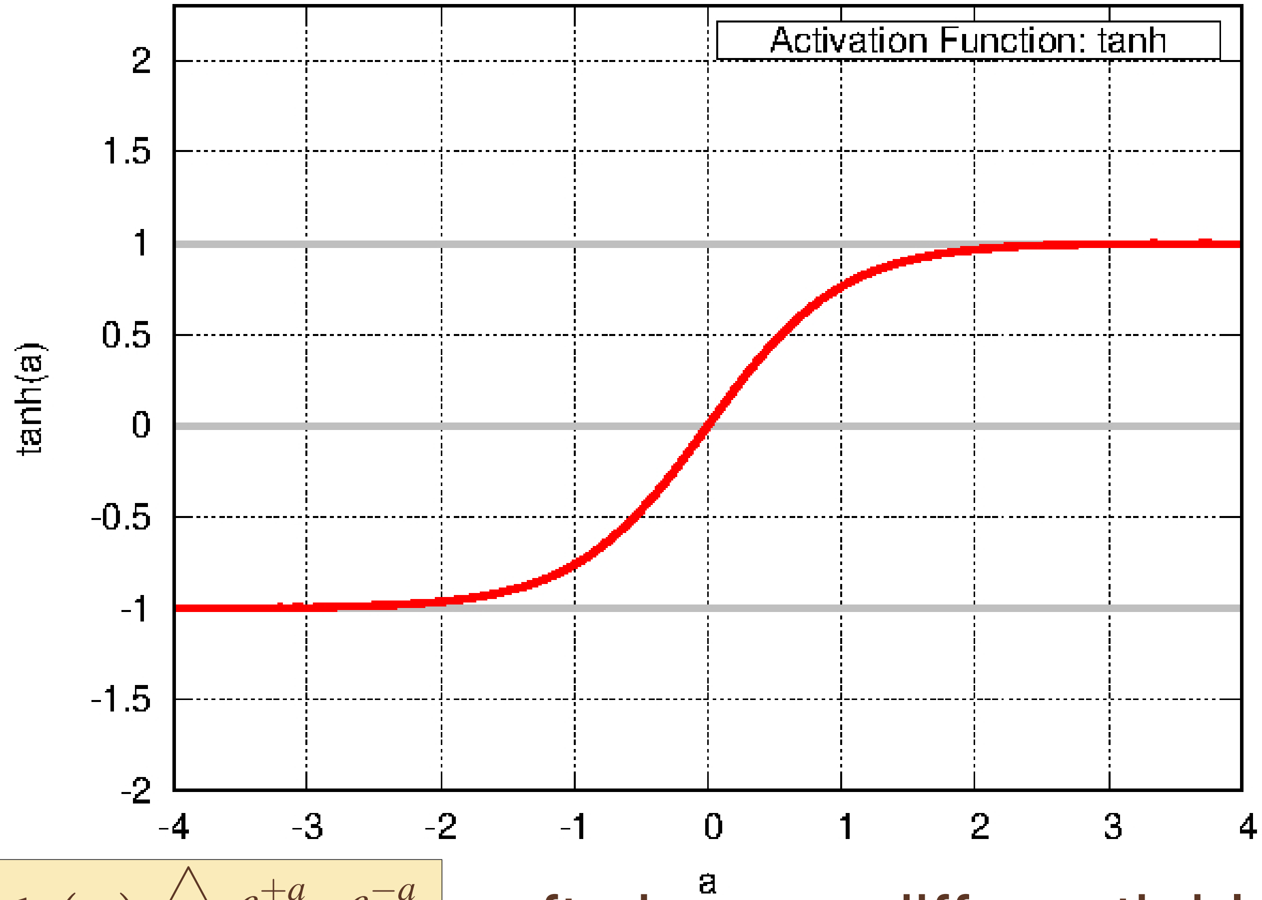
Activation Fns: Logistic Sigmoid



- $\boxed{\sigma(a) \stackrel{\triangle}{=} \frac{1}{1+e^{-a}}}$ softer unit step; differentiable
- $a \to -\infty, \sigma(a) \to 0$; $a \to +\infty, \sigma(a) \to 1$;
  $a = 0, \sigma(a) = 0.5$
- (-) Computation with exponentials is difficult!

Home Page
Title Page
Contents
Page 11 of 30
Go Back
Full Screen
Close
Quit

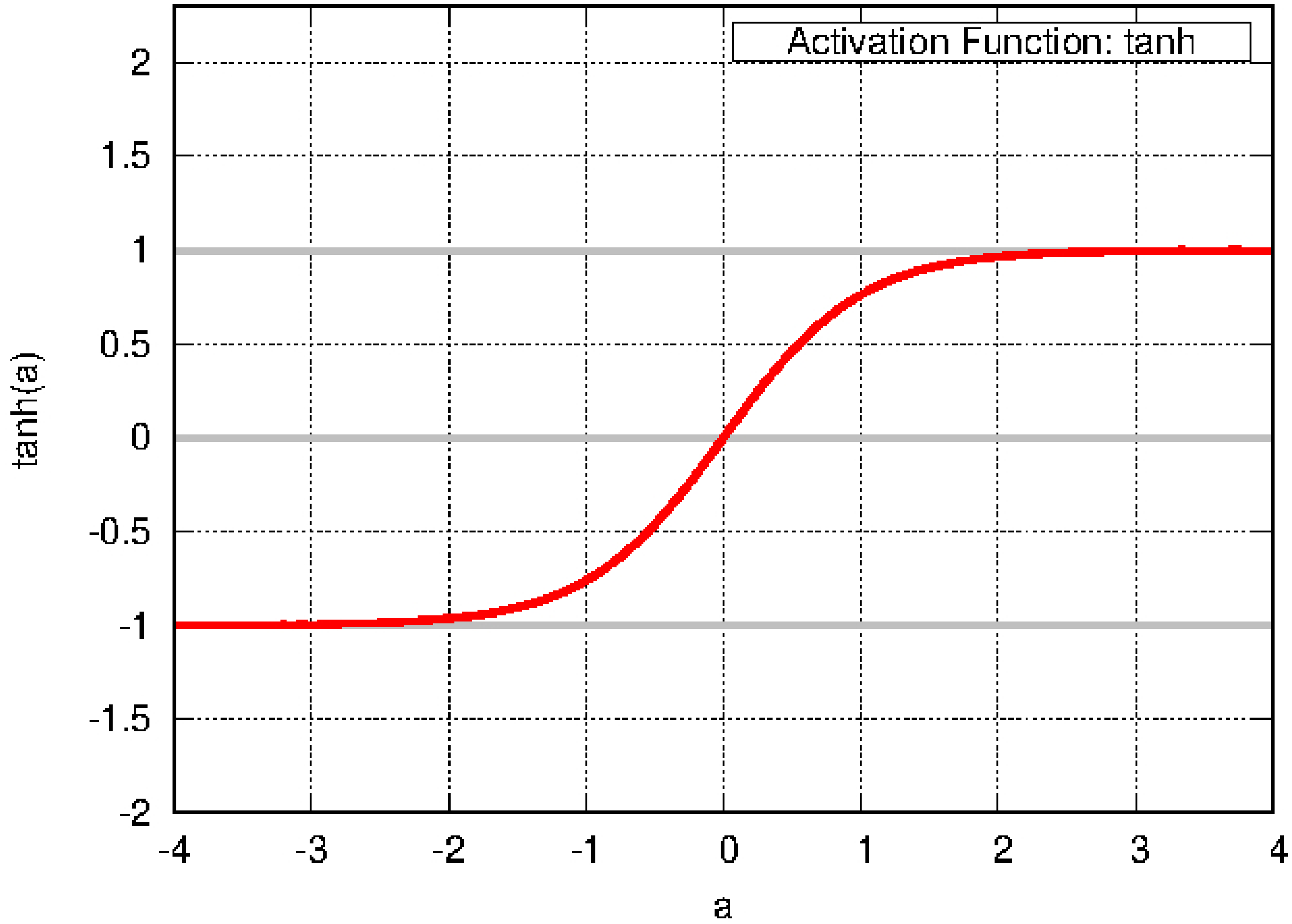
Activation Fns: tanh



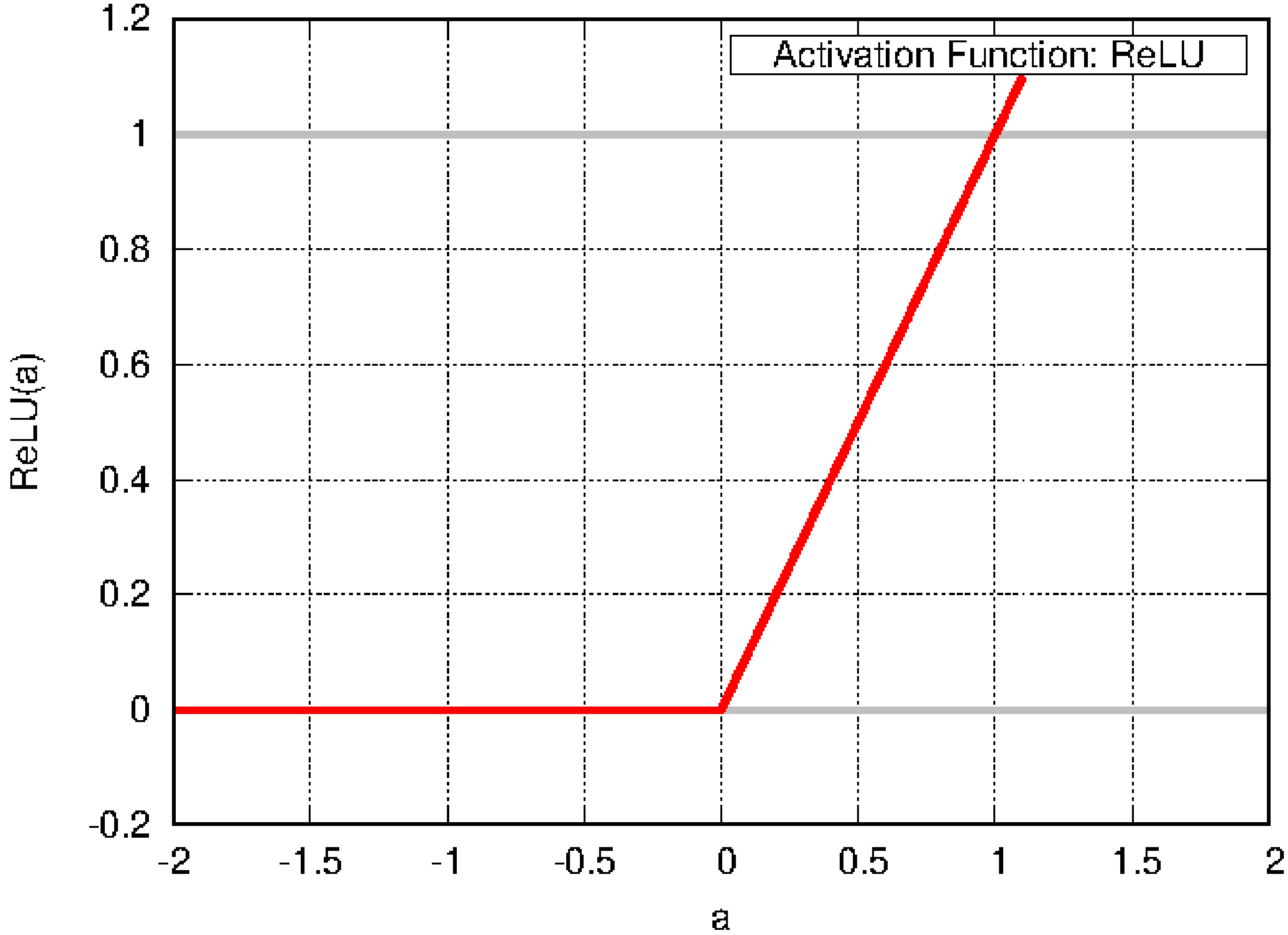
- $\tanh(a) \stackrel{\triangle}{=} \frac{e^{+a} - e^{-a}}{e^{+a} + e^{-a}}$ soft signum; differentiable
- $a \to -\infty, \sigma(a) \to -1$; $a \to +\infty, \sigma(a) \to +1$; $a = 0, \sigma(a) = 0$
- (-) computation with exponentials is difficult!
- (-) grad $\to 0$ as curve saturates! vanishing grad

Activation Fns: tanh: Development

- $2\sigma(a) - 1$: stretch to [0, 2], then shift down by 1
- $\frac{2}{1+e^{-a}} - 1 = \frac{2-1-e^{-a}}{1+e^{-a}} = \frac{(1-e^{-a})e^{+a/2}}{(1+e^{-a})e^{+a/2}} = \frac{e^{+a/2}-e^{-a/2}}{e^{+a/2}+e^{-a/2}} = \tanh(\frac{a}{2})$

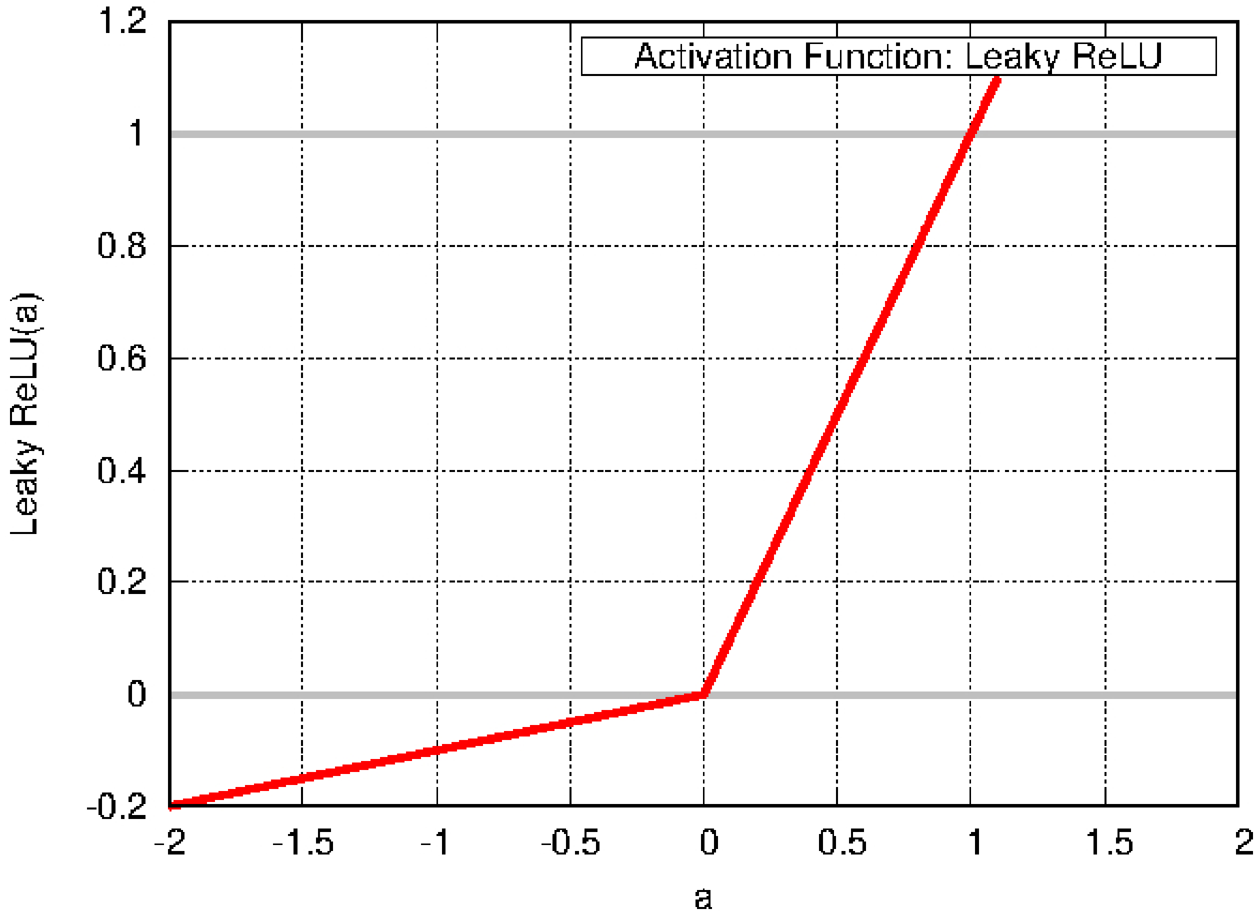


Activation Fns: ReLU



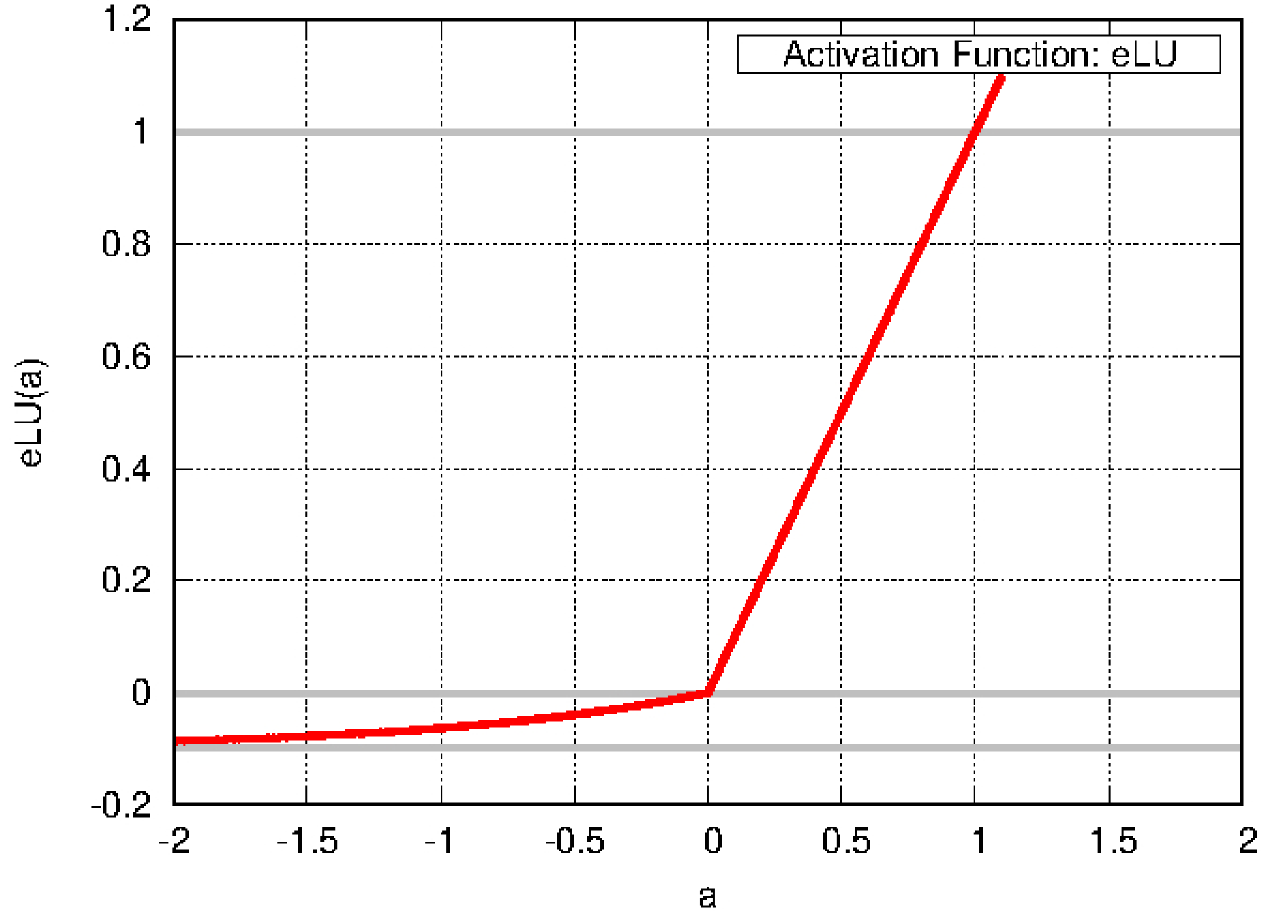
- $ReLU(a) \triangleq a, a \geq 0; 0,$ otherwise Easy to compute
- (+) no vanishing gradient as no saturation!
- (-) negative inputs, no gradient

Activation Fns: Leaky ReLU



- $\boxed{LeakyRelU(a,\alpha) \stackrel{\triangle}{=} \max(\alpha a, a),\ \alpha \in (0,1)}$
- $a \geq 0$: $\max(\alpha a, a) = a$ $(\alpha \in (0,1))$
- $a < 0$: αa: fraction of more neg: $\alpha a > a$, slope < 1
- (-) Not diff at $a = 0$: handled algorithmically

Activation Fns: eLU



- $\boxed{eLU(a,\alpha) \triangleq \left\{ \begin{array}{c} a, a > 0 \\ \alpha(e^a - 1), a \leq 0 \end{array} \right\}}$ $a = 0 : eLU = 0$
- $a \to -\infty : eLU \to \alpha(\frac{1}{e^\infty} - 1) \to -\alpha$ (α: deg of sat'n)
- (-) Computation with exponentials is difficult!