> **Note:** $\sigma'(x) = \sigma(x)(1-\sigma(x))$
> We use this substitution in the algorithms.

Backpropagate (y, A, η)

- Given weight matrices $W^2, \dots, W^L$ and bias vectors $\vec{b}^2, \dots, \vec{b}^L$ at each layer,

- Declare vars:
  - $\vec{d}_{\vec{a}} C$, the derivative of cost w.r.t current layer's activations
  - $\vec{d}_{\vec{z}} C$, same w.r.t current layer's $\vec{z}$ values
  - $\Delta\vec{b}$, change in $\vec{b}$ at current layer
  - ΔW, same for W
  - $\vec{a}$, current layer's activations
  - $\vec{a}'$, previous layer's activations
  - $\vec{y}$, vector corresponding to classification y

- $\vec{y} = \hat{e}_y$
- $\vec{a}$ = last element of A
- $\vec{d}_{\vec{a}} C = \vec{a} - \vec{y}$
- $\vec{d}_{\vec{z}} C = \vec{d}_{\vec{a}} C \odot \left[\vec{a} \odot (\vec{1} - \vec{a}) \right]$

- For each layer l from L down to 2:
  - $\vec{a} = l^{th}$ elem. of A
  - $\vec{a}' = (l-1)^{th}$ elem. of A
  - $\Delta\vec{b} = -\eta\, \vec{d}_{\vec{z}} C$
  - $\Delta W = -\eta\, \vec{d}_{\vec{z}} C \otimes \vec{a}'$
  - $\vec{d}_{\vec{a}} C = (W^l)^t\, \vec{d}_{\vec{z}} C$
  - $\vec{d}_{\vec{z}} C = \vec{d}_{\vec{a}} C \odot \left[\vec{a}' \odot (\vec{1} - \vec{a}') \right]$
  - $\vec{b}^l \mathrel{+}= \Delta\vec{b}$
  - $W^l \mathrel{+}= \Delta W$

Train $(\vec{x}_1, \dots, \vec{x}_n, y_1, \dots, y_n, \eta, T)$

- Declare vars:
  - A, the list of activations returned by Feedforward
- Repeat T times:
  - For each $\vec{x}, y$:
    - A = Feedforward $(\vec{x})$
    - Backpropagate (y, A, η)