> Note: Here we use the identity
> $\sigma'(x) = \sigma(x)(1-\sigma(x))$

Backpropagate (y, A, η, error function)

- Given: weight matrices $W^2, \ldots, W^L$
  bias vectors $\vec{b}^2, \ldots, \vec{b}^L$
  at each layer
- Declare vars:
  - $\vec{d}_{\vec{a}} C$, deriv. of error wrt current layer's activations
  - $\vec{d}_{\vec{z}} C$, same but wrt current layer's z values
  - $\Delta\vec{b}$, change in $\vec{b}$ at current layer
  - ΔW, same for W
  - $\vec{a}$, current layer's activations
  - $\vec{a}'$, prev. layer's activations
  - $\vec{y}$, vector corresponding to classification y
- $\vec{y} = \hat{e}_y$
- $\vec{a}$ = last element of A
- If error function is Euclidean Distance:
  - $\vec{d}_{\vec{a}} C = \vec{a} - \vec{y}$
  - $\vec{d}_{\vec{z}} C = \vec{d}_{\vec{a}} C \odot \left[\vec{a} \odot (\vec{1} - \vec{a})\right]$
- Else if error function is Cross Entropy:
  - $\vec{d}_{\vec{z}} C = \vec{a} - \vec{y}$
- 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 ($\vec{x}_1, \ldots, \vec{x}_n, y_1, \ldots, y_n, \eta, T$, error function):

- 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, error function)