Delta rule as gradient descent in error (sigmoid units)

\[ n_j = \sum_i a_i w_{ij} \]
\[ a_j = \frac{1}{1 + \exp(-n_j)} \]

Error \( E = \frac{1}{2} \sum_j (t_j - a_j)^2 \)

Gradient descent:
\[ \Delta w_{ij} = -\epsilon \frac{\partial E}{\partial w_{ij}} \]
\[ = -\epsilon \frac{\partial E}{\partial a_j} \frac{\partial a_j}{\partial n_j} \frac{\partial n_j}{\partial w_{ij}} \]
\[ = -\epsilon (t_j - a_j) a_j (1 - a_j) a_i \]

Generalized Delta rule ("back-propagation")

\[ n_j = \sum_i a_i w_{ij} \]
\[ a_j = \frac{1}{1 + \exp(-n_j)} \]

Error \( E = \frac{1}{2} \sum_j (t_j - a_j)^2 \)

Gradient descent:
\[ \Delta w_{ij} = -\epsilon \frac{\partial E}{\partial w_{ij}} \]
\[ = -\epsilon \frac{\partial E}{\partial n_j} \frac{\partial n_j}{\partial a_i} \frac{\partial a_i}{\partial a_j} \frac{\partial a_j}{\partial n_j} \frac{\partial n_j}{\partial w_{ij}} \]
\[ = -\epsilon (t_j - a_j) a_j (1 - a_j) a_i \]

Back-propagation

Forward pass (⇑)
\[ a_j = \frac{1}{1 + \exp(-n_j)} \]
\[ n_j = \sum_i a_i w_{ij} \]

Backward pass (⇓)
\[ \frac{\partial E}{\partial a_j} = -(t_j - a_j) \]
\[ \frac{\partial E}{\partial n_j} = \frac{\partial E}{\partial a_j} a_j (1 - a_j) \]
\[ \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial n_j} \frac{\partial n_j}{\partial w_{ij}} \]
\[ \frac{\partial E}{\partial a_i} = \sum_j \frac{\partial E}{\partial n_j} \frac{\partial n_j}{\partial a_i} \frac{\partial a_i}{\partial w_{ij}} \]
What do hidden representations learn?

Plaut and Shallice (1993)

- Mapped orthography to semantics (unrelated similarities)
- Compared similarities among hidden representations to those among orthographic and semantic representations (over settling)

Hidden representations “split the difference” between input and output similarity

Accelerating learning: Momentum descent

\[ \Delta w_{ij}[t] = -\epsilon \frac{\partial E}{\partial w_{ij}} + \alpha (\Delta w_{ij}[t-1]) \]

“Aauto-encoder” network (4–2–4)

Projections of error surface in weight space
High momentum

High learning rate

![Graphs comparing weight change factor and total error for different momentum and learning rate settings.](image-url)