XOR with intermediate ("hidden") units

- Intermediate units can re-represent input patterns as new patterns with altered similarities.
- Targets which are not linearly separable in the input space can be linearly separable in the intermediate representational space.
- Intermediate units are called "hidden" because their activations are not determined directly by the training environment (inputs and targets).

Hidden-to-output weights can be trained with the Delta rule:

\[ n_j = \sum_i a_i w_{ij} \]
\[ a_j = \frac{1}{1 + \exp(-n_j)} \]
Error \( E = \frac{1}{2} \sum_j (t_j - a_j)^2 \)

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

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

**Forward pass (⇑)**

\[ a_j = \frac{1}{1 + \exp(-n_j)} \]

\[ n_j = \sum_i a_i w_{ij} \]

\[ a_i = \frac{1}{1 + \exp(-n_i)} \]

**Backward pass (⇓)**

\[ \frac{\partial E}{\partial a_j} = -(t_j - a_j) \]

\[ \frac{\partial E}{\partial n_j} = \frac{\partial E}{\partial a_j} a_j (1 - a_j) \]

\[ \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial n_j} a_i \]

\[ \frac{\partial E}{\partial a_i} = \sum_j \frac{\partial E}{\partial n_j} w_{ij} \]

Accelerating learning: Momentum descent

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

**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)

“Aauto-encoder” network (4–2–4)

- Hidden representations “split the difference” between input and output similarity
Projections of error surface in weight space

- Asterisk: error of current set of weights
- Tick mark: error of next set of weights
- Solid curve (0): Gradient direction
- Solid curve (21): Integrated gradient direction (including momentum)
  - This is actual direction of weight step (tick mark is on this curve)
  - Number is angle with gradient direction
- Lens: “Grad lin” (gradient linearity) is (normalized) dot product of gradient direction and weight change direction
- Dotted curves: Random directions (each labeled by angle with gradient direction)

Epochs 1-2

Epochs 3-4

Epochs 5-6
<table>
<thead>
<tr>
<th>High momentum (epochs 1-2)</th>
<th>High learning rate (epochs 1-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td>High momentum (epochs 3-4)</td>
<td>High learning rate (epochs 3-4)</td>
</tr>
<tr>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
</tr>
</tbody>
</table>