



Using stochastic gradient descent and the *backpropagation algorithm* (just repeated application of the chain rule)

One output unit per class

x_i = total input to output unit i

$$f(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{1000} \exp(x_j)}$$

We maximize the log-probability of the correct label, $\log f(x_t)$

