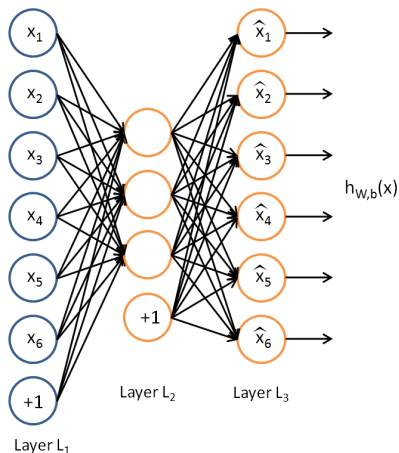


# Effects of Sparsity and Regularization Choice on Sparse Autoencoders

Autoencoders are neural networks that try to learn the identity function. Sparsity constraints force the learned representation to be nontrivial.



# Cost Function

Let the hidden layer have  $p$  units, and let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable activation function. The neural network is parameterized by terms weights  $W^{(1)} \in \mathbb{R}^{p \times n}$  and  $W^{(2)} \in \mathbb{R}^{n \times p}$  and bias terms  $b^{(1)} \in \mathbb{R}^p$  and  $b^{(2)} \in \mathbb{R}^n$ . The prediction on an input  $x \in \mathbb{R}^n$  is

$$h_{W,b} = f(W^{(2)} f(W^{(1)} x + b^{(1)}) + b^{(2)})$$

Given training examples  $x^{(1)}, \dots, x^{(m)} \in \mathbb{R}^n$  the objective function is

$$J(W, b) = \frac{1}{m} \sum_{i=1}^m \ell(h_{W,b}(x^{(i)}), x^{(i)}) + \lambda \psi(W, b) + \beta \sum_{j=1}^p \phi(\hat{p}_j)$$

where

$$\hat{p}_j = \frac{1}{m} \sum_{i=1}^m f\left((W_j^{(1)})^T x^{(i)} + b_j^{(1)}\right)$$

is the average activation of the  $j$ th hidden unit over the training set and  $\ell : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  is a loss function. The function  $\psi$  is a regularizer.

# Pictures