Optmization Methods for Machine Learning
Multilayer Perceptron

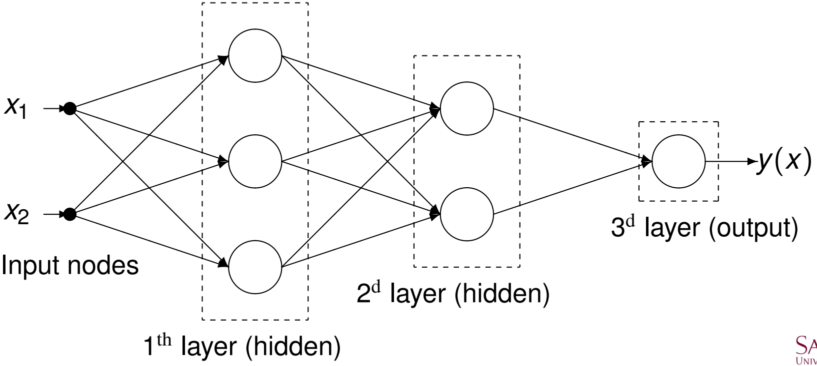
Laura Palagi

http://www.dis.uniroma1.it/∼palagi

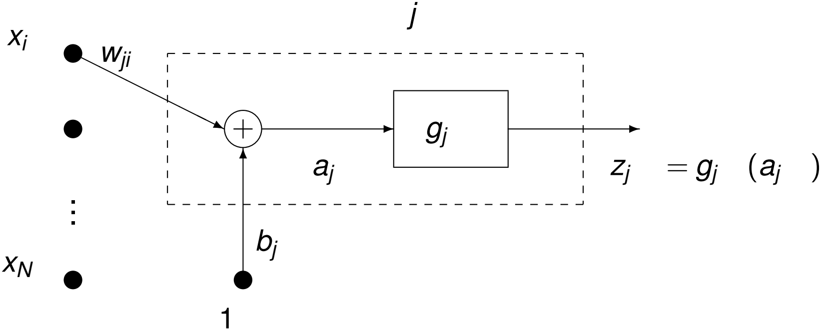
Dipartimento di Ingegneria informatica automatica e gestionale A. Ruberti
Sapienza Università di Roma

Via Ariosto 25

An MLP with input $x \in \mathbb{R}^2$ and a scalar output $y \in \mathbb{R}$

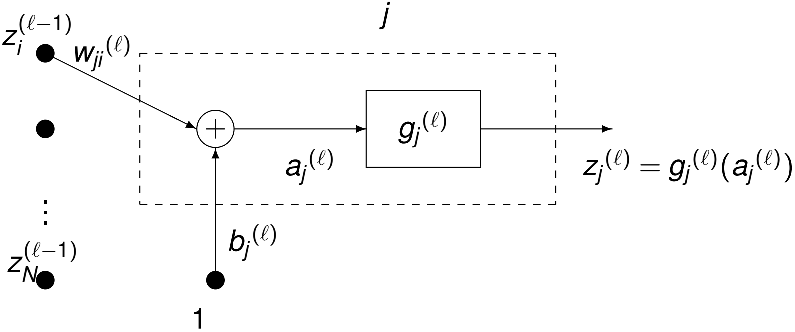


Internal structure of a single neuron j



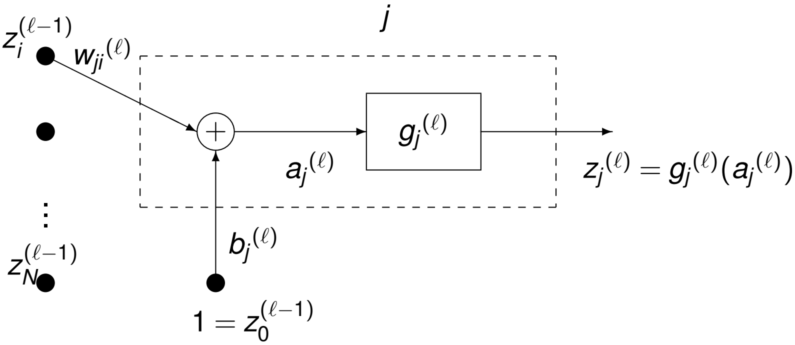
$$a_j \quad = \sum_{i=1}^{N} w_{ji} \quad x_i \qquad + b_j \quad =$$

Internal structure of a single neuron j at layer ℓ



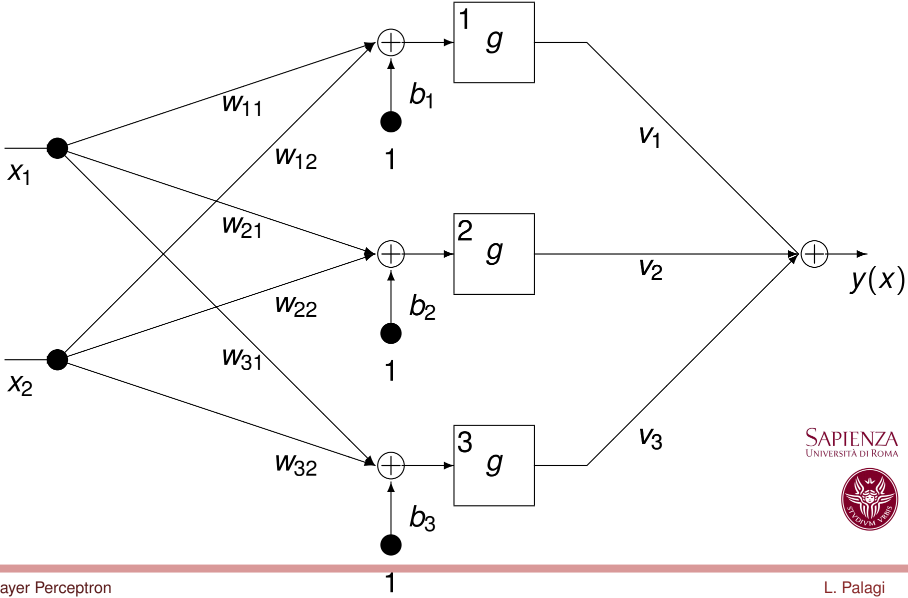
$$a_j^{(\ell)} = \sum_{i=1}^{N^{(\ell-1)}} w_{ji}^{(\ell)} \quad z_i^{(\ell-1)} + b_j^{(\ell)} =$$

Internal structure of a single neuron j at layer ℓ



$$a_j^{(\ell)} = \sum_{i=1}^{N^{(\ell-1)}} w_{ji}^{(\ell)} \, z_i^{(\ell-1)} + b_j^{(\ell)} = \sum_{i=0}^{N^{(\ell-1)}} w_{ji}^{(\ell)} x_i z_i^{(\ell-1)}$$

Two layer MLP



Two layer MLP

- N: number of neurons of the hidden layer;
- w_{ji}: weight of the arc connecting input node i with neuron j of the hidden layer;
- b_j: threshold of hidden neuron j;
- v_j: weight of the arc connecting hidden neuron j to the output;
- g: activation function of the hidden neurons;
- the activation function of the output neuron is a linear function of the inputs.

Then we can write

$$y(x) = \sum_{j=1}^{N} v_j g\left(\sum_{i=1}^{n} w_{ji} x_i + b_j\right) = \sum_{j=1}^{N} v_j g\left(w_j^T x + b_j\right)$$

where

$$w_j = (w_{j1}, \ldots, w_{jn})^T.$$

Interpolation property of MLP

Given p distinct points in R^n:

$$X = \{\overline{x}^i \in R^n,\ i = 1, \dots, p\},$$

and a corresponding set of real numbers

$$Y = \{\overline{y}^i \in R,\ i = 1, \dots, p\}.$$

The interpolation problem consists in finding a function $f : R^n \to R$, in a given class of real functions $\mathscr{F}$, which satisfies:

$$f(\overline{x}^i) = \overline{y}^i \qquad i = 1, \dots, P. \tag{1}$$

Theorem (Pinkus 1999)

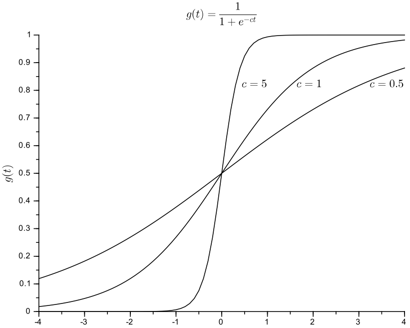
Let $g \in C(R)$ not polynomial. Then $w^j \in R^n$, and $v^j, b^j \in R$, for $j = 1, \dots P$ exist s.t.

$$\sum_{j=1}^{P} v^j g({w^j}^T \overline{x}^i - b^j) = \overline{y}^i, \quad i = 1, \dots, p.$$

Logistic

$$g(t) = \frac{1}{1+e^{-ct}} \qquad \dot{g}(t) = \frac{ce^{-ct}}{(1+e^{-ct})^2}, \quad c > 0$$

It is a differentiable approximation of a *threshold function* (or *Heaviside step function*), which is obtained, in the limit, for $c \to \infty$.



Hyperbolic tangent

$$g(t) \equiv \tanh(t/2) = \frac{1-e^{-ct}}{1+e^{-ct}}, \quad c > 0 \qquad \dot{g}(t) = \frac{2ce^{-ct}}{(1+e^{-ct})^2}$$

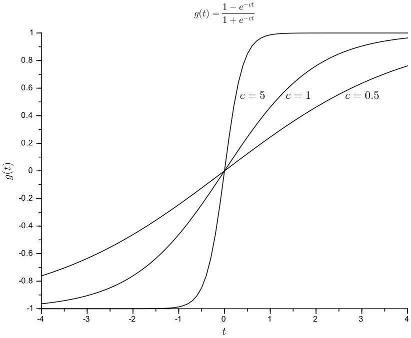


Figure : tanh(t/2), for $c = 5, 1, 0.5$

Unconstrained problem

$$\min_{w,b} E(w, b)$$

- Existence of a global solution
- Optimality conditions (for a point to be a local solution)
- Definition of an iterative algorithm

$$\begin{pmatrix} w^{k+1} \\ b^{k+1} \end{pmatrix} = \begin{pmatrix} w^k \\ b^k \end{pmatrix} + \alpha^k d^k$$

- Convergence