Optmization Methods for Machine Learning
Gradient method for 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

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

BP Gradient method

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

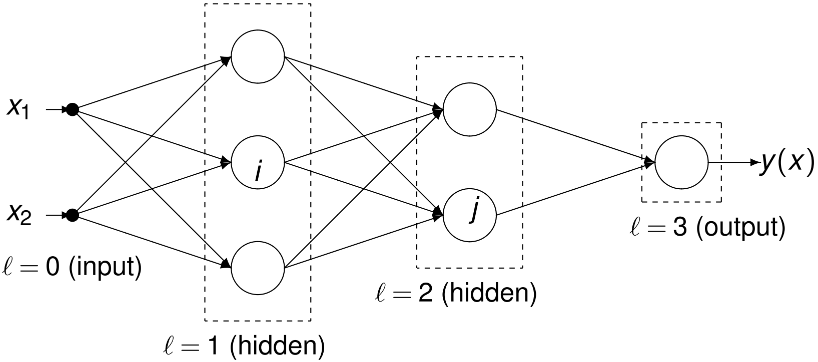
- BP *batch*, when the parameter are updated using all the samples in the training set T_t;

$$w^{k+1} = w^k - \eta \nabla E(w^k),$$

- BP *on-line*, when the parameters are updated using one sample of T_t at the time.

$$w^{k+1} = w^k - \eta \nabla E_{p(k)}(w^k).$$

SAPIENZA
UNIVERSITÀ DI ROMA



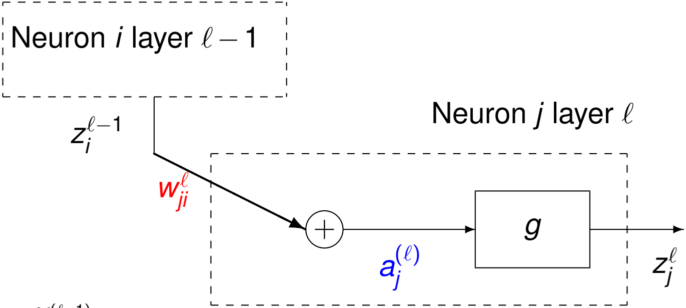
We assume that $g_j^{(\ell)}(\cdot) = g(\cdot)$ for all j, ℓ.

$$E(w) = \frac{1}{2}\sum_{p=1}^{P} E_p(w^k) = \frac{1}{2}\sum_{p=1}^{P} \|e^p(w^k)\|^2$$

where $e^p(w^k) = y(w^k; x^p) - y^p \in \mathbb{R}^K$

SAPIENZA UNIVERSITÀ DI ROMA

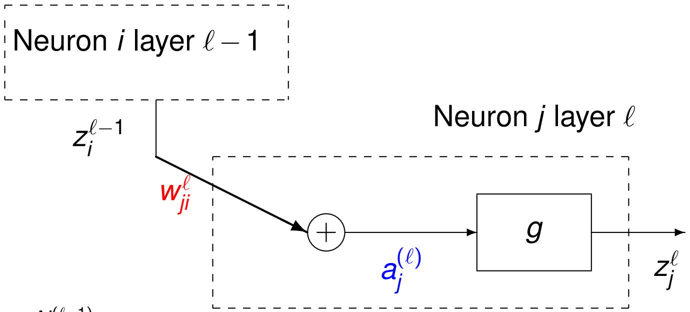
Forward computation



$$a_j^{(\ell)} = \sum_{k=0}^{N^{(\ell-1)}} w_{jk}^{(\ell)} z_k^{(\ell-1)}, \qquad z_j^{\ell} = g(a_j^{(\ell)}) = g(\cdots + w_{ji}{}^{(\ell)} z_i^{(\ell-1)} + \dots)$$

$$\frac{\partial E_p}{\partial w_{ji}{}^{\ell}} = \frac{\partial E_p}{\partial a_j^{(\ell)}} \cdot \frac{\partial a_j^{(\ell)}}{\partial w_{ji}{}^{\ell}} = \quad \frac{\partial E_p}{\partial a_j^{(\ell)}} \cdot z_i^{(\ell-1)}$$

Forward computation



$$a_j^{(\ell)} = \sum_{k=0}^{N^{(\ell-1)}} w_{jk}^{(\ell)} z_k^{(\ell-1)}, \qquad z_j^{\ell} = g(a_j^{(\ell)}) = g(\cdots + w_{ji}^{(\ell)} z_i^{(\ell-1)} + \dots)$$

$$\frac{\partial E_p}{\partial w_{ji}{}^{\ell}} = \frac{\partial E_p}{\partial a_j^{(\ell)}} \cdot \frac{\partial a_j^{(\ell)}}{\partial w_{ji}{}^{\ell}} = \underbrace{\frac{\partial E_p}{\partial a_j^{(\ell)}}}_{\delta_j^{(\ell)}} \cdot z_i^{(\ell-1)}$$

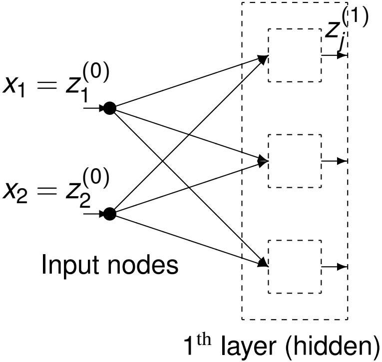
Forward propagation of the input

$x_1 = z_1^{(0)}$

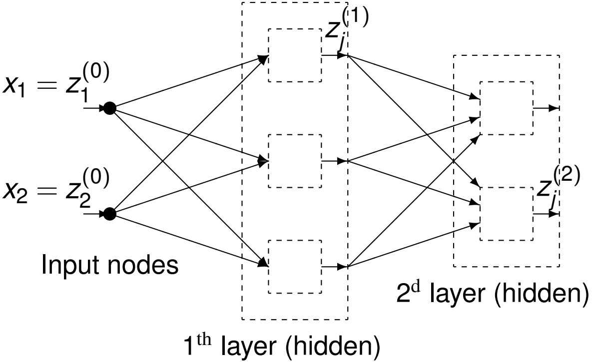
$x_2 = z_2^{(0)}$

Input nodes

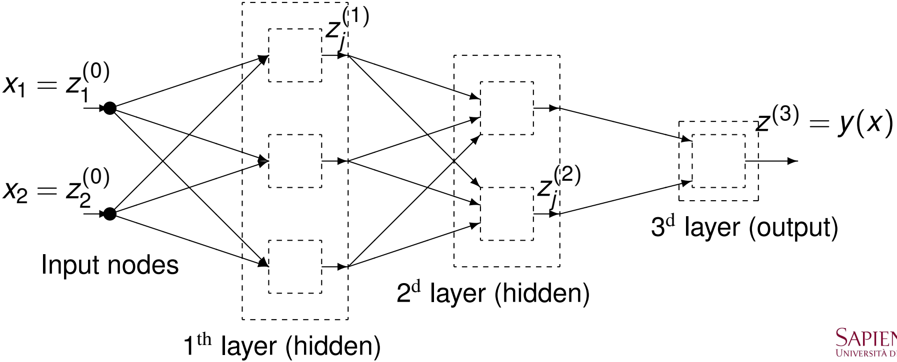
Forward propagation of the input



Forward propagation of the input

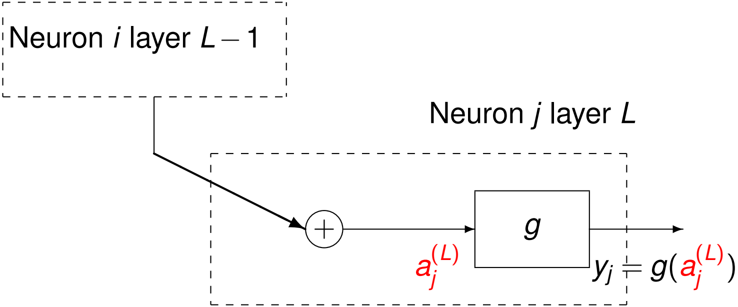


Forward propagation of the input



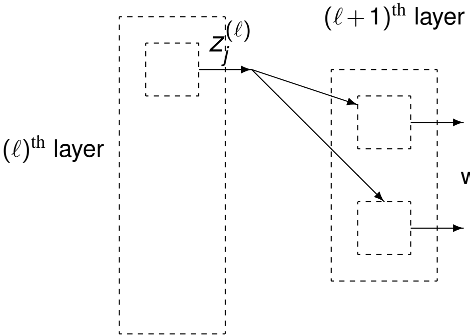
Back computation of errors $\delta_j^{(\ell)} = \frac{\partial E_p}{\partial a_j^{(\ell)}}$

Consider the case of the output layer $\ell = L$



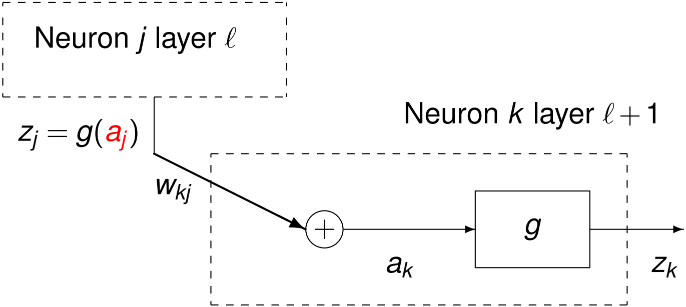
$$\delta_j^L = \frac{\partial E_p}{\partial a_j^{(L)}} = \frac{\partial E_p}{\partial y_j} \cdot \frac{\partial y_j}{\partial a_j^{(L)}} = \underbrace{\frac{\partial E_p}{\partial y_j}}_{e_j^p(w^k) \text{ analytic from expression of } E_p} \cdot \dot{g}(a_j^{(L)})$$

Hidden layer: $\frac{\partial E_p(w^k)}{\partial a_j^{(\ell)}}$



we consider $z_k^{(\ell+1)}$ for all k neurons

Back computation of errors hidden layer



$$a_k^{(\ell+1)} = \dots + w_{kj}\, g(a_j(\ell)) + \dots \qquad \text{for all } k \text{ in layer } \ell+1$$

$$\delta_j^{\ell} = \frac{\partial E_p}{\partial a_j{}^{(\ell)}} = \sum_{k=1}^{N^{(\ell+1)}} \frac{\partial E_p}{\partial a_k^{\ell+1}} \cdot \frac{\partial a_k^{\ell+1}}{\partial a_j{}^{\ell}} = \sum_{k=1}^{N^{(\ell+1)}} \delta_k^{(\ell+1)} \cdot \frac{\partial a_k^{(\ell+1)}}{\partial a_j{}^{(\ell)}}$$

Backpropagation gradient evaluation

$$\frac{\partial E_p}{\partial w_{ji}{}^{\ell}} = \delta_j^{\ell} \cdot z_i^{\ell-1} \qquad \ell = 1, \ldots, L$$

1. Compute FORWARD

$$z_i^{\ell} \qquad \ell = 1, \ldots, L$$

2. Compute BACKWARD For $k = 1, \ldots, K$ set

$$\delta_k^{(L)} = \frac{\partial E_p}{\partial a_k^{(L)}} = e_k^p \cdot \dot{g}(a_k^{(L)})$$

$$\delta_j^{\ell} = \sum_{k=1}^{N^{\ell+1}} \delta_k^{\ell+1} \cdot w_{kj}^{\ell+1} \dot{g}(a_j{}^{\ell}) \qquad \ell = L-1, \ldots, 1$$

Convergence

Theorem

Assume that a scalar $L > 0$ exists such that for each $w, u \in R^m$ we have:

$$\|\nabla E(w) - \nabla E(u)\| \leq L\|w - u\|$$

(Lipschtizt continuity of the gradient). Let $\{w^k\}$ be the sequence generated by

$$w^{k+1} = w^k - \eta \nabla E(w^k)$$

with $\varepsilon \leq \eta \leq \bar{\eta}_L - \varepsilon$, and $\varepsilon > 0$, Assume $\nabla E(w^k) \neq 0$ for all k, then every accumulation point of $\{w^k\}$ is a stationary point for E. (If the level set $\mathscr{L}^0$ is compact, there exists accumulation point of $\{w^k\}$).

Momentum modification

$$w^{k+1} = w^k - \eta \nabla E(w^k) + \beta(w^k - w^{k-1}),$$

where $\beta > 0$ is a given scalar with (typical values $= 0.8 \pm 0.9$).

SAPIENZA
UNIVERSITÀ DI ROMA