Lagrange

Thm (Le paysage des réseaux larges)

On suppose σ C^1 et que, pour tout $t \in \mathbb{R}$, $\sigma'(t) \neq 0$.
On suppose L est C^1, à valeur dans $\mathbb{R}^+$ telle que $L(0) = 0$.
On suppose $\nabla L(y) = 0 \iff y = 0$.
On considère $X \in \mathbb{R}^{n_0 \times n}$ et $A = \begin{pmatrix} X \\ \mathbb{1}_n^T \end{pmatrix}$.
On suppose $\mathrm{rk}(A) = n$.
Soit θ un point critique du premier ordre de E tel que pour tout $k = 1, \dots, H$ $\mathrm{rk}(W_k) = n_k$.
On a $E(\theta) = \hat{R}(f_\theta) = 0$.
θ est un minimiseur global.

---

preuve : On rappelle que pour tout $k \in \{1, \dots, H\}$, $i \in \{1, \dots, n_k\}$, $j \in \{1, \dots, n_{k-1}\}$ et $l \in \{1, \dots, n\}$, on a

$$\frac{\partial E}{\partial (W_k)_{ij}}(\theta) = \left[f_\theta^{k-1}(x^l) \right]_j \cdot \Delta_i^k(x^l)$$

$$\frac{\partial E}{\partial (b_k)_i}(\theta) = \Delta_i^k(x^l)$$

avec $\Delta_i^k(x^l) = \sigma'\left(\left[W_k f_\theta^{k-1}(x^l) + b_k \right]_i \right) \Delta_i^k(x^l)$

et

$$\Delta^k(x^l) = \begin{cases} \nabla L\left(f_\theta(x^l) - y^l \right) & \text{si } k = H \\ W_{k+1}^T \Delta^{k+1}(x^l) & \text{si } k = 0, \dots, H-1 \end{cases}$$

rappelons les formules de la backprop.

On va montrer par récurrence que pour tout $k = 1, \dots, H$ et tout $l = 1, \dots, n$ on a

$$\overline{\Delta^k}(x^l) = 0.$$

+ initialisation : $k = 1$, pour $i \in \{1, \dots, n_1\}$ et $j \in \{1, \dots, n_0\}$ on a, comme θ est un point critique de E,

$$(*) \begin{cases} 0 = \dfrac{\partial E}{\partial (W_1)_{ij}}(\theta) = \displaystyle\sum_{l=1}^{n} \left[f_{n_0}^{0}(x^l) \right]_j x_j^l \, \overline{\Delta^1}(x^l)_i \\ 0 = \dfrac{\partial E}{\partial (b_1)_{ij}}(\theta) = \displaystyle\sum_{l=1}^{n} \overline{\Delta^1}(x^l)_i \end{cases}$$

On a $\overline{\Delta^1}(x^l) \in \mathbb{R}^{n_1}$ pour tout $l \in \{1, \dots, n\}$

On note

$$\overline{\Delta^1} = \left(\overline{\Delta^1(x^1)} \quad \cdots \quad \overline{\Delta^1(x^n)} \right)^T = \begin{pmatrix} \overline{\Delta^1(x^1)}_1 & \cdots & \overline{\Delta^1(x^1)}_{n_1} \\ \vdots & & \vdots \\ \overline{\Delta^1(x^n)}_1 & \cdots & \overline{\Delta^1(x^n)}_{n_1} \end{pmatrix} \in \mathbb{R}^{n \times n_1}$$

et $A = \begin{pmatrix} x_1^1 & \cdots & x_1^n \\ \vdots & & \vdots \\ x_{n_0}^1 & \cdots & x_{n_0}^n \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(n_0+1) \times n}$

Les équations $(*)$ sont équivalentes à

$$A \, \overline{\Delta^1} = 0 \in \mathbb{R}^{(n_0+1) \times n_1}$$

Comme on a supposé A de rang colonne plein ($\mathrm{rk}(A) = n$)

l'équation $A \, \overline{\Delta^1} = 0$ implique $\overline{\Delta^1} = 0$.

(On multiplie à gauche par $(A^T A)^{-1} A^T$)

système linéaire en $\overline{\Delta^1}(x^l)$

exprime ce système sous une forme de multiplication matrice-matrice

$$\begin{pmatrix} & \\ & \end{pmatrix} \begin{matrix} x_1 \\ \vdots \\ x_n \end{matrix} = \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}$$

$A^T A$ inversible.

$\overline{\Delta^1} = 0$ est équivalent à $\overline{\Delta^1}(x^l) = 0 \quad \forall l \in \{1, \dots, n\}$.

+ récurrence : Supposons que pour $k \in \{1, \dots, H-1\}$, on ait $\overline{\Delta^k}(x^l) = 0 \quad \forall l \in \{1, \dots, n\}$.

Montrons que $\overline{\Delta^{k+1}}(x^l) = 0 \quad \forall l \in \{1, \dots, n\}$.

Soit $l \in \{1, \dots, n\}$, on sait que pour tout $i \in \{1, \dots, n_k\}$

$\overline{\Delta^k_i}(x^l) = \sigma'\left(\left[W_k \phi_\theta^{k-1}(x^l) + b_k\right]_i\right) \Delta^k_i(x^l) = 0$.

Comme on a supposé $\sigma'(t) \neq 0, \forall t \in \mathbb{R}$, on a donc $\Delta^k_i(x^l) = 0$ et donc $\overline{\Delta^k}(x^l) = 0$.

On en déduit que $W_{k+1}^T \overline{\Delta^{k+1}}(x^l) = 0$

Comme on a supposé $\mathrm{rk}(W_{k+1}) = n_{k+1}$, on a $\mathrm{rg}(W_{k+1}^T) = n_{k+1}$ et W_{k+1}^T est de rang colonne plein. On en déduit que

$\overline{\Delta^{k+1}}(x^l) = 0$.

Cela conclut la récurrence et on a montré que $\forall k \in \{1, \dots, H\}$ et tout $l \in \{1, \dots, n\}$, $\overline{\Delta^k}(x^l) = 0$

On en déduit que

$\overline{\Delta^H}(x^l)_i = \sigma'\left(\left[W_H \phi_\theta^{H-1}(x^l) + b_H\right]_i\right) \Delta^H(x^l)_i = 0$

et donc, comme $\sigma'(t) \neq 0 \ \forall t$, que $\Delta^H(x^l)_i = 0$

paysage large | 3 | 4

def de $\overline{\Delta^k_i}(x^l)$

$0 = \overline{\Delta^q}(x^q) = W_{q+1}^T \overline{\Delta^{q+1}}(x^q)$ ← formule de backprop.

$W_{k+1}^T \in \mathbb{R}^{n_k \times n_{k+1}}$

$\Delta^{k}(x^{l})$

On a donc $= \nabla L\left(f_\theta(x^l) - y^l\right) = 0$ backprop

qui implique (voir hypothèse sur L) que

$$f_\theta(x^l) = y^l.$$

On obtient finalement

$$L\left(f_\theta(x^l) - y^l\right) = 0 \qquad \forall l \in \{1, \dots, n\}$$

et donc $E(\theta) = \sum_{l=1}^{n} L\left(f_\theta(x^l) - y^l\right) = 0.$

Comme $\forall \theta'$, $E(\theta') \geq 0$, θ^* est bien un minimiseur global. □