Paysage

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 $A = \begin{pmatrix} X \\ 1_n^T \end{pmatrix}$.

On suppose $rk(A) = n$

Soit θ un point critique du premier ordre de E tel que pour tout $h = 1, \dots, H$ $rk(W_h) = n_h$.

On a $E(\theta) = \hat{R}(f_\theta) = 0$.

θ est un minimiseur global.

---

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

$$\frac{\partial E}{\partial (W_h)_{ij}}(\theta) = \left[f_\theta^{h-1}(x^l)\right]_j \cdot \bar{\Delta}_i^h(x^l)$$

$$\frac{\partial E}{\partial (b_h)_i}(\theta) = \bar{\Delta}_i^h(x^l)$$

avec $\bar{\Delta}_i^h(x^l) = \sigma'\left(\left[W_h f_\theta^{h-1}(x^l) + b_h\right]_i\right) \Delta_i^h(x^l)$

et

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

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

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

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

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

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

On note

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

et

$$A = \begin{pmatrix} x^1_1 & \cdots & x^n_1 \\ \vdots & & \vdots \\ x^1_{n_0} & \cdots & x^n_{n_0} \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{(n_0+1) \times n}$$

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 ($rk(A) = n$)

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

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

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

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

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

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

$$\overline{\Delta^h_i}(x^l) = \sigma'\left(\left[W_h f_\theta^{h-1}(x^l) + b_h\right]_i\right) \Delta^h_i(x^l) = 0.$$

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

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

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

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

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

On en déduit que

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

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

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

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

Comme $\forall \theta'$, $E(\theta') \geq 0$, θ^* est

bien un minimiseur global. $\square$