# Corrigé : Loi de Conservation et Régularisation Implicite dans les Réseaux Linéaires

**Rappel des notations :** $Y \in \mathbb{R}^{d_2 \times d_0}$, $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, $R = W_2 W_1 - Y \in \mathbb{R}^{d_2 \times d_0}$, et :

$$\mathcal{L}(W_1, W_2) = \frac{1}{2}\|R\|_F^2 = \frac{1}{2}\sum_{i=1}^{d_2}\sum_{k=1}^{d_0} R_{ik}^2$$

## Partie 1 : Du calcul de dérivées partielles aux gradients matriciels

**Question 1 : Expression explicite du résiduel**
**Solution :** Par définition du produit matriciel $(W_2W_1)_{ik} = \sum_{r=1}^{d_1}(W_2)_{ir}(W_1)_{rk}$, l'élément générique de la matrice résiduelle s'écrit :

$$R_{ik} = \sum_{r=1}^{d_1}(W_2)_{ir}(W_1)_{rk} - Y_{ik}$$

**Question 2 : Dérivée partielle par rapport à $W_2$**
**Solution :** Dérivons d'abord le terme résiduel par rapport au coefficient scalaire $(W_2)_{ab}$ pour $1 \le a \le d_2$ et $1 \le b \le d_1$. Seul le terme de la somme où l'indice de ligne vaut $i = a$ et l'indice intermédiaire vaut $r = b$ dépend de $(W_2)_{ab}$. En utilisant la fonction indicatrice $\mathbb{1}_{\{i=a\}}$, on obtient :

$$\frac{\partial R_{ik}}{\partial (W_2)_{ab}} = \frac{\partial}{\partial (W_2)_{ab}}\left(\sum_{r=1}^{d_1}(W_2)_{ir}(W_1)_{rk} - Y_{ik}\right) = \mathbb{1}_{\{i=a\}}(W_1)_{bk}$$

où $\mathbb{1}_{\{i=a\}} = 1$ si $i = a$, et 0 sinon.

En appliquant la règle de dérivation en chaîne à la fonction de perte $\mathcal{L} = \frac{1}{2}\sum_{i=1}^{d_2}\sum_{k=1}^{d_0} R_{ik}^2$ :

$$\frac{\partial \mathcal{L}}{\partial (W_2)_{ab}} = \sum_{i=1}^{d_2}\sum_{k=1}^{d_0} R_{ik}\frac{\partial R_{ik}}{\partial (W_2)_{ab}} = \sum_{i=1}^{d_2}\sum_{k=1}^{d_0} R_{ik}\mathbb{1}_{\{i=a\}}(W_1)_{bk} = \sum_{k=1}^{d_0} R_{ak}(W_1)_{bk}$$

Or, par transposition, on a $(W_1)_{bk} = (W_1^T)_{kb}$. L'expression devient :

$$\frac{\partial \mathcal{L}}{\partial (W_2)_{ab}} = \sum_{k=1}^{d_0} R_{ak}(W_1^T)_{kb} = (RW_1^T)_{ab}$$

Il s'agit exactement du coefficient $(a, b)$ du produit matriciel $RW_1^T$.

**Question 3 : Dérivée partielle par rapport à $W_1$**
**Solution :** De même, pour des indices $1 \le c \le d_1$ et $1 \le d \le d_0$, le coefficient $(W_1)_{cd}$ n'apparaît dans la somme de $R_{ik}$ que lorsque l'indice de colonne vaut $k = d$ et l'indice intermédiaire vaut $r = c$. On obtient donc :

$$\frac{\partial R_{ik}}{\partial (W_1)_{cd}} = (W_2)_{ic}\mathbb{1}_{\{k=d\}}$$

Par dérivation en chaîne :

$$\frac{\partial \mathcal{L}}{\partial (W_1)_{cd}} = \sum_{i=1}^{d_2}\sum_{k=1}^{d_0} R_{ik}(W_2)_{ic}\mathbb{1}_{\{k=d\}} = \sum_{i=1}^{d_2}(W_2)_{ic}R_{id} = \sum_{i=1}^{d_2}(W_2^T)_{ci}R_{id} = (W_2^T R)_{cd}$$

Ce résultat correspond bien au coefficient $(c, d)$ du produit matriciel $W_2^T R$.

**Question 4 : Gradients matriciels et dimensions**
**Solution :** Puisque chaque composante scalaire du gradient coïncide avec le coefficient matriciel calculé :

$$\nabla_{W_2}\mathcal{L} = RW_1^T = (W_2W_1 - Y)W_1^T \in \mathbb{R}^{d_2 \times d_1}$$

$$\nabla_{W_1}\mathcal{L} = W_2^T R = W_2^T(W_2W_1 - Y) \in \mathbb{R}^{d_1 \times d_0}$$

Les dimensions concordent parfaitement avec les matrices respectives : $W_2 \in \mathbb{R}^{d_2 \times d_1}$ et $W_1 \in \mathbb{R}^{d_1 \times d_0}$.

## Partie 2 : Analyse en temps continu (Gradient Flow)

**Question 5 : Règle de dérivation matricielle**

**Solution :**

a) Considérons le coefficient $(i, j)$ de la matrice produit $C(t) = A(t)B(t)$, où $A(t) \in \mathbb{R}^{n\times p}$ et $B(t) \in \mathbb{R}^{p\times m}$ :

$$C_{ij}(t) = \sum_{k=1}^{p} A_{ik}(t)B_{kj}(t)$$

En appliquant la règle usuelle de dérivation d'un produit scalaire :

$$\frac{d}{dt}C_{ij}(t) = \sum_{k=1}^{p}\left(\dot{A}_{ik}(t)B_{kj}(t) + A_{ik}(t)\dot{B}_{kj}(t)\right) = \sum_{k=1}^{p}\dot{A}_{ik}(t)B_{kj}(t) + \sum_{k=1}^{p}A_{ik}(t)\dot{B}_{kj}(t)$$

On reconnaît directement la somme des coefficients $(\dot{A}(t)B(t))_{ij} + (A(t)\dot{B}(t))_{ij}$. En réécrivant sous forme matricielle :

$$\frac{d}{dt}\left(A(t)B(t)\right) = \left(\frac{dA(t)}{dt}\right)B(t) + A(t)\left(\frac{dB(t)}{dt}\right)$$

b) Par définition de la transposition, $(A(t)^T)_{ij} = A_{ji}(t)$. Dès lors :

$$\left(\frac{d}{dt}A(t)^T\right)_{ij} = \frac{d}{dt}A_{ji}(t) = \dot{A}_{ji}(t) = (\dot{A}(t)^T)_{ij} \implies \frac{d}{dt}\left(A(t)^T\right) = \left(\frac{dA(t)}{dt}\right)^T$$

## Partie 3 : Démonstration de la loi de conservation

**Question 6 : Dérivées temporelles de $W_2^TW_2$ et $W_1W_1^T$**

**Solution :** Sous les équations du gradient flow, les vitesses d'évolution sont données par :

$$\dot{W}_2(t) = -\nabla_{W_2}\mathcal{L} = -R(t)W_1(t)^T \quad \text{et} \quad \dot{W}_1(t) = -\nabla_{W_1}\mathcal{L} = -W_2(t)^TR(t)$$

a) En utilisant la dérivation d'un produit :

$$\frac{d}{dt}\left(W_2^TW_2\right) = \dot{W}_2^TW_2 + W_2^T\dot{W}_2$$

En substituant l'expression de $\dot{W}_2$ et en appliquant $(AB)^T = B^TA^T$ :

$$\dot{W}_2^T = \left(-RW_1^T\right)^T = -(W_1^T)^TR^T = -W_1R^T$$

Il vient alors immédiatement :

$$\frac{d}{dt}\left(W_2^TW_2\right) = -W_1R^TW_2 - W_2^TRW_1^T$$

b) De façon analogue pour la première couche :

$$\frac{d}{dt}\left(W_1W_1^T\right) = \dot{W}_1W_1^T + W_1\dot{W}_1^T$$

En substituant $\dot{W}_1 = -W_2^TR$ et $\dot{W}_1^T = -R^TW_2$ :

$$\frac{d}{dt}\left(W_1W_1^T\right) = -W_2^TRW_1^T - W_1R^TW_2$$

**Question 7 : Théorème de conservation**

**Solution :** Soustrayons les deux expressions obtenues à la question précédente :

$$\frac{d}{dt}\Delta(t) = \frac{d}{dt}\left(W_2(t)^TW_2(t) - W_1(t)W_1(t)^T\right) = \left(-W_1R^TW_2 - W_2^TRW_1^T\right) - \left(-W_2^TRW_1^T - W_1R^TW_2\right) = 0$$

La dérivée temporelle de $\Delta(t)$ étant identiquement nulle pour tout $t \geq 0$, la matrice est constante le long de la trajectoire d'optimisation :

$$W_2(t)^TW_2(t) - W_1(t)W_1(t)^T = C = W_2(0)^TW_2(0) - W_1(0)W_1(0)^T$$

Comme $(W_2^TW_2)^T = W_2^TW_2$ et $(W_1W_1^T)^T = W_1W_1^T$, la matrice $C$ est une différence de matrices symétriques, elle est donc elle-même symétrique ($C^T = C \in \mathbb{R}^{d_1\times d_1}$).

## Partie 4 : Conséquences géométriques et régularisation implicite

**Question 8 : Initialisation équilibrée**
**Solution :** L'hypothèse d'initialisation équilibrée se traduit par $C = \Delta(0) = 0$, d'où :

$$W_2(t)^T W_2(t) = W_1(t) W_1(t)^T, \quad \forall t \geq 0$$

Rappelons que la norme de Frobenius d'une matrice $M$ vérifie $\|M\|_F^2 = \mathrm{Tr}(M^T M) = \mathrm{Tr}(M M^T)$. En appliquant l'opérateur trace (qui est linéaire) :

$$\|W_2(t)\|_F^2 - \|W_1(t)\|_F^2 = \mathrm{Tr}\left(W_2(t)^T W_2(t)\right) - \mathrm{Tr}\left(W_1(t) W_1(t)^T\right) = \mathrm{Tr}\left(W_2(t)^T W_2(t) - W_1(t) W_1(t)^T\right) = \mathrm{Tr}(0) = 0$$

On en déduit que $\|W_2(t)\|_F^2 = \|W_1(t)\|_F^2$. Les normes étant strictement positives :

$$\|W_2(t)\|_F = \|W_1(t)\|_F \quad \forall t \geq 0$$

*Interprétation pédagogique : Le gradient descent empêche une couche d'exploser en norme pendant que l'autre tendrait vers zéro ; les échelles des couches restent automatiquement équilibrées.*

**Question 9 : Goulot d'étranglement unidimensionnel ($d_1 = 1$)**
**Solution :**

a) Lorsque $d_1 = 1$, la matrice $W_2$ se réduit à un vecteur colonne $u \in \mathbb{R}^{d_2 \times 1}$ et $W_1$ à un vecteur ligne $v^T$ avec $v \in \mathbb{R}^{d_0 \times 1}$. La dimension intermédiaire $d_1 \times d_1$ devient $1 \times 1$. On a alors :

$$W_2^T W_2 = u^T u = \|u\|_2^2 \in \mathbb{R} \quad \text{et} \quad W_1 W_1^T = v^T (v^T)^T = v^T v = \|v\|_2^2 \in \mathbb{R}$$

La matrice $\Delta(t)$ devient ainsi le scalaire :

$$\Delta(t) = \|u(t)\|_2^2 - \|v(t)\|_2^2 \in \mathbb{R}$$

b) D'après le théorème de conservation démontré en Question 7, cette quantité scalaire est constante :

$$\|u(t)\|_2^2 - \|v(t)\|_2^2 = c \quad \text{avec } c = \|u(0)\|_2^2 - \|v(0)\|_2^2 \in \mathbb{R}$$

En posant $x(t) = \|u(t)\|_2 \geq 0$ et $y(t) = \|v(t)\|_2 \geq 0$, la trajectoire dans le premier quadrant de $\mathbb{R}^2$ est contrainte par l'équation :

$$x^2 - y^2 = c$$

- Si $c \neq 0$, la courbe géométrique suivie est une branche d'**hyperbole équilatère**.
- Si $c = 0$ (cas équilibré), l'équation devient $x^2 = y^2$, ce qui correspond à la demi-droite bissectrice $y = x$.

---