Preuves
Paramètres effectifs et modèles non-linéaires

Alexandre CONSTANTIN[1] Rodrigo CABRAL FARIAS[2] Jean-Marc BROSSIER[1] Olivier MICHEL[1]

[1]Univ. Grenoble-Alpes, CNRS, Grenoble-INP, GIPSA Lab, 38000 Grenoble

[2]Univ. Côte d'Azur, CNRS, Laboratoire I3S, Sophia-Antipolis 06900

Résumé – Nous présentons une synthèse de la littérature autour des questions de degrés de liberté et en proposons un estimateur applicable à tout modèle paramétrique. Des résultats sur un perceptron avec une couche cachée montrent des performances équivalentes à l'algorithme de Ye pour une complexité numérique réduite.

Abstract – A summary of the litterature on degrees of freedom is presented and an estimator for non-linear models is proposed. Results on a one hidden layer perceptron show equivalent behavior as Ye's algorithm, with lower numerical complexity.

A Degrés de liberté en linéaire

Afin de montrer l'égalité (3) :

$$\mathrm{dP} = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \eta},$$

dans le cas linéaire, où $\{\lambda_i\}_{i=1,\dots,P}$ sont les valeurs propres de la Hessienne de la fonction de coût sans régularisation et η la force de la régularisation 'ridge', nous rappelons que :

$$\mathbf{S}_\eta = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \eta\mathbf{I}\right)^{-1}\mathbf{X}^\top,$$

avec $\mathbf{X} \in \mathbb{R}^{N\times P}$ et $\mathbf{I}$ la matrice identité, ici de taille P.

La matrice $\mathbf{X}^\top\mathbf{X}$ étant symmétrique définie positive, elle admet une décomposition en valeur propre $\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$ telle que $\boldsymbol{\Lambda}$ est la matrice diagonale contenant les P valeurs propres $\{\lambda_i\}_{i=1}^{P}$ de $\mathbf{X}^\top\mathbf{X}$. Il vient alors, en utilisant la linéarité cyclique et la décomposition en valeur propre :

$$\begin{aligned}
\mathrm{Tr}(\mathbf{S}_\eta) &= \mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}+\eta\mathbf{I}\right)^{-1}\mathbf{X}^\top\right)\\
&= \mathrm{Tr}\left(\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}+\eta\mathbf{I}\right)^{-1}\right)\\
&= \mathrm{Tr}\left(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}\left(\mathbf{P}(\boldsymbol{\Lambda}+\eta\mathbf{I})\mathbf{P}^{-1}\right)^{-1}\right)\\
&= \mathrm{Tr}\left(\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{P}\left(\boldsymbol{\Lambda}+\eta\mathbf{I}\right)^{-1}\mathbf{P}^{-1}\mathbf{P}\right)\\
&= \mathrm{Tr}\left(\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}+\eta\mathbf{I}\right)^{-1}\right)\\
&= \sum_{i=1}^{P}\frac{\lambda_i}{\lambda_i+\eta}.
\end{aligned}$$

Pour démontrer l'égalité entre la définition (4) des degrés de liberté et dP, nous insistons sur le fait que les échantillons doivent être indépendants, en effet :

$$\begin{aligned}
\mathrm{df} &= \frac{1}{\sigma_y^2}\sum_{i=1}^{N}\mathrm{cov}(y_i, \sum_{k=1}^{N}(\mathbf{S}_\eta)_{i,k}y_k)\\
&= \frac{1}{\sigma_y^2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\mathbf{S}_\eta)_{i,k}\mathrm{cov}(y_i, y_k)
\end{aligned}$$

en utilisant $\hat{\mathbf{y}} = \mathbf{S}_\eta\mathbf{y}$, la linéarité à droite de la covariance (linéarité d'une espérance) et $\mathrm{cov}(y_i, y_k) = 0$ pour $i \neq k$ par indépendance. Il en découle naturellement :

$$\begin{aligned}
\mathrm{df} &= \frac{1}{\sigma_y^2}\sum_{i=1}^{N}(\mathbf{S}_\eta)_{i,i}\mathrm{cov}(y_i, y_i)\\
&= \frac{1}{\sigma_y^2}\sum_{i=1}^{N}(\mathbf{S}_\eta)_{i,i}\sigma_y^2\\
&= \mathrm{Tr}(\mathbf{S}_\eta)\\
&= \mathrm{dP},
\end{aligned}$$

B Preuve de la proposition 1

L'estimateur $\hat{\mathrm{df}}$ est un estimateur de :

$$\mathrm{df}' = \mathrm{Tr}\left[\mathbb{E}\left(-\mathbf{J}_{\boldsymbol{\theta}}[m_{\boldsymbol{\theta}}(\mathbf{y})]\left[\mathbf{H}_{\boldsymbol{\theta}}\mathcal{L}\right]^{-1}\mathbf{J}_{\mathbf{y}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]\right)\right],$$

où $\mathbf{J}_{\mathbf{x}}[\mathbf{y}] \in \mathbb{R}^{m\times n}$ est la matrice jacobienne telle que, pour $\mathbf{y} \in \mathbb{R}^m$ et $\mathbf{x} \in \mathbb{R}^n$, $(\mathbf{J}_{\mathbf{x}}[\mathbf{y}])_{i,j} = \partial y_i/\partial x_j$.

Nous voulons donc montrer que :

$$\mathrm{df} = \sum_{i=1}^{N}\mathbb{E}\left(\frac{\partial\hat{y}_i}{\partial y_i}\right) = \mathrm{df}'.$$

Or $\sum_{i=1}^{N}\mathbb{E}\left(\frac{\partial\hat{y}_i}{\partial y_i}\right) = \mathrm{Tr}\left[\mathbb{E}\left(\mathbf{J}_{\mathbf{y}}\hat{\mathbf{y}}\right)\right]$, c'est-à-dire, par égalité des dimensions des matrices, montrer que :

$$\mathbf{J}_{\mathbf{y}}[\hat{\mathbf{y}}] = -\mathbf{J}_{\boldsymbol{\theta}}[m_{\boldsymbol{\theta}}(\mathbf{y})]\left[\mathbf{H}_{\boldsymbol{\theta}}\mathcal{L}\right]^{-1}\mathbf{J}_{\mathbf{y}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}].$$

Pour cela nous rappelons que les sorties prédites sont fonction du modèle, à $\mathbf{X}$ fixé, tel que :

$$\hat{\mathbf{y}} = m_{\hat{\boldsymbol{\theta}}}(\mathbf{X}),$$

où $\hat{\boldsymbol{\theta}}$ sont les paramètres, au point de convergence, tel que $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Et $\hat{\boldsymbol{\theta}}$ est une fonction implicite, notée f, de $\mathbf{y}$ (à $\mathbf{X}$ fixé car seules les perturbations en $\mathbf{y}$ nous intéressent).

Or $\mathbf{J}_{\mathbf{y}}[\hat{\mathbf{y}}] = \mathbf{J}_{f(\mathbf{y})}[\hat{\mathbf{y}}]\mathbf{J}_{\mathbf{y}}[f(\mathbf{y})]$ par la composition des fonctions m et f de la matrice Jacobienne. D'une part, $\mathbf{J}_{f(\mathbf{y})}[\hat{\mathbf{y}}] =$

$\mathbf{J}_{\boldsymbol{\theta}}[m_{\hat{\boldsymbol{\theta}}}(\mathbf{y})]$ est évaluable en $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. D'autre part, par le théorème des fonctions implicites avec $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, il vient :

$$\mathbf{J}_{\mathbf{y}}[f(\mathbf{y})] = -\left[J_{f(\mathbf{y})}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]\big|_{(\mathbf{y},\hat{\boldsymbol{\theta}})} \right]^{-1} J_{\mathbf{y}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]\big|_{(\mathbf{y},\hat{\boldsymbol{\theta}})} .$$

Le terme $J_{\mathbf{y}}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]$ est évaluable de manière directe car le gradient est connu. Par définition $J_{f(\mathbf{y})}[\nabla_{\boldsymbol{\theta}}\mathcal{L}]$ est la matrice Hessienne de la fonction de coût évaluée à $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. D'où l'égalité $\mathrm{df}' = \mathrm{df}$.

Notre estimateur est égal au degrés de liberté dans le cas linéaire (noté dP), car :

$$\begin{cases} \mathcal{L} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \eta\|\boldsymbol{\theta}\|_2^2 \\ \nabla_{\boldsymbol{\theta}}\mathcal{L} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + 2\eta\boldsymbol{\theta} \\ m_{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\theta}, \end{cases}$$

ce qui permet d'obtenir :

$$\hat{\mathrm{df}} = \mathrm{Tr}\left[-\mathbf{X}\left(2\mathbf{X}^\top\mathbf{X} + 2\eta\mathbf{I}\right)^{-1}\left(-2\mathbf{X}^\top\right) \right] = \mathrm{dP}.$$

C Cas d'un échantillon identiquement distribué mais non indépendant

Pour terminer, pour la généralisation au cas non-indépendant mais identiquement distribué, pour démontrer (7), nous avons :

$$\begin{aligned} \mathbb{E}(d_{\boldsymbol{\Sigma}}(\hat{\mathbf{y}}, \boldsymbol{\mu})) = \mathbb{E}(d_{\boldsymbol{\Sigma}}(\hat{\mathbf{y}}, \mathbf{y})) &+ \mathbb{E}(d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu})) \\ &+ 2\mathbb{E}\left((\hat{\mathbf{y}} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right), \end{aligned}$$

or, en utilisant la linéarité cyclique de la trace et la linéarité de l'espérance :

$$\begin{aligned} \mathbb{E}(d_{\boldsymbol{\Sigma}}(\mathbf{y}, \boldsymbol{\mu})) &= \mathbb{E}(\mathrm{Tr}[(\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})]) \\ &= \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^\top]) \\ &= \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) = N, \end{aligned}$$

et, de manière similaire :

$$\begin{aligned} &\mathbb{E}\left[(\hat{\mathbf{y}} - \mathbf{y})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right] \\ =&\mathbb{E}\left[\hat{\mathbf{y}}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y} - \hat{\mathbf{y}}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] - \mathbb{E}\left[\mathbf{y}^\top \boldsymbol{\Sigma}^{-1}\mathbf{y} - \mathbf{y}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] \\ =&\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\left[\mathbb{E}(\hat{\mathbf{y}}^\top\mathbf{y}) - \mathbb{E}(\hat{\mathbf{y}}^\top)\boldsymbol{\mu}\right]) \\ &\quad - \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\left[\mathbb{E}(\mathbf{y}^\top\mathbf{y}) - \mathbb{E}(\mathbf{y})^\top\boldsymbol{\mu}\right]) \\ =&\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathrm{cov}(\hat{\mathbf{y}}, \mathbf{y})) - \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) \\ =&\mathrm{df} - N. \end{aligned}$$

D Calcul de la Hessienne du perceptron une couche

Nous considérons le perceptron multi-couche décrit par la Figure 1. Il présente un vecteur d'entrées $\mathbf{x}$ et un scalaire en sortie $\hat{y}$, une fonction (non-linéaire) en couche cachée notée f_h et une fonction de transformation de la sortie (quelconque), notée f_o. Le vecteur de paramètres $\boldsymbol{\theta}$ est ordonné de la manière suivante :

$$\boldsymbol{\theta} = \left\{ \bigcup_{i=1}^{d} \mathbf{w}_i^{(1)} \right\} \cup \{\mathbf{w}^{(o)}\} \cup \{\mathbf{b}^{(1)}\} \cup \{b^{(2)}\}, \qquad (1)$$

où $\mathbf{w}_i^{(1)}, \mathbf{b}^{(1)} = [b_1^{(1)}, \ldots, b_d^{(1)}]^\top$ sont les poids et les biais de la couche cachée appliquées à l'entrée i. Les sorties après la non-linéarité de la couche cachée sont représentés par le vecteur $\mathbf{o}^{(1)} = [o_1^{(1)}, \ldots, o_d^{(1)}]^\top$. $\mathbf{w}^{(o)}$ sont les poids de la couche de sortie avec un biais $b^{(2)}$. Nous considérons aussi le cas d'un échantillon $y \in \mathbf{y}$, car la matrice Hessienne sur l'ensemble des N échantillons est la somme des N matrices Hessiennes.

De plus, nous avons la décomposition suivante pour une fonction de coût $\mathcal{L}$ étant la somme des erreurs quadratiques, pour l'échantillon $y \in \mathbf{y}$:

$$\mathbf{H} = \frac{1}{N}\left(\nabla_{\boldsymbol{\theta}}\hat{y}(\nabla_{\boldsymbol{\theta}}\hat{y})^\top + (\hat{y} - y)\mathbf{H}^{yy}\right), \qquad (2)$$

où $\mathbf{H}^{yy}$ est la hessienne de la sortie du modèle et $\nabla_{\boldsymbol{\theta}}\hat{y}$ le gradient en sortie du modèle.

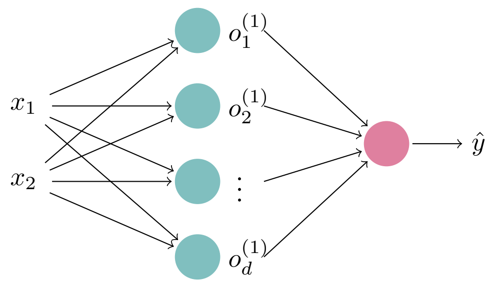

FIGURE 1 : Perceptron avec $d > 1$ neurones.

Le gradient en sortie est exprimé par :

$$\begin{cases} \dfrac{\partial y}{\partial b^{(2)}} = \alpha_o \\ \nabla_{\mathbf{w}^{(o)}} y = \alpha_o \mathbf{o}^{(1)} \\ \nabla_{\mathbf{b}^{(1)}} y = \alpha_o \boldsymbol{\alpha}_h^{(1)} \odot \mathbf{w}^{(o)} \\ \nabla_{\mathbf{w}_i^{(1)}} y = \alpha_o \alpha_{h,i} w_i^{(o)} \mathbf{x}, \qquad i = 1, \ldots, d; \end{cases} \qquad (3)$$

en introduisant $\boldsymbol{\alpha}_h^{(1)} = [\alpha_{h,1}^{(1)}, \ldots, \alpha_{h,d}^{(1)}]^\top$, tel que $\alpha_{h,i}^{(1)} = f_h'(b_i^{(1)} + \langle \mathbf{w}_i^{(1)}, \mathbf{x}\rangle), i = 1, \ldots, d$ et $\alpha_o = f_o'(b^{(2)} + \langle \mathbf{w}^{(o)}, \mathbf{o}^{(1)}\rangle)$. Par ailleurs, $\odot$ dénote le produit de Hadamard.

Nous définissons aussi, pour la suite, les quantités $\beta_{h,i}^{(1)} = f_h''(b_i^{(1)} + \langle \mathbf{w}_i^{(1)}, \mathbf{x}\rangle), i = 1, \ldots, d$, le vecteur $\boldsymbol{\beta}_h^{(1)} = [\beta_{h,1}^{(1)}, \ldots, \beta_{h,d}^{(1)}]^\top$ et $\beta_o = f_o''(b^{(2)} + \langle \mathbf{w}^{(o)}, \mathbf{o}^{(1)}\rangle)$. La hessienne, en $\boldsymbol{\theta}$ (1), se décompose alors suivant les blocs représentés dans la Figure 2.

Pour $i, j = 1, \ldots, d$, nous retrouvons les quantités suivantes en dérivant les dérivées (3) par rapport à l'autre variable d'intérêt :

$$\begin{aligned} \mathbf{H}^{yy}_{\mathbf{w}_i^{(1)}\mathbf{w}_j^{(1)}} &= \left(\beta_o(\alpha_{h,i}^{(1)})^2(\mathbf{w}^{(o)})_i(\mathbf{w}^{(o)})_i + \alpha_o\beta_{h,i}^{(1)}(\mathbf{w}^{(o)})_i\right)\mathbf{x}\mathbf{x}^\top \\ &\quad \text{si } i = j; \\ &= \beta_o\alpha_{h,i}^{(1)}\alpha_{h,j}^{(1)}(\mathbf{w}^{(o)})_i(\mathbf{w}^{(o)})_j\, \mathbf{x}\mathbf{x}^\top \text{ sinon,} \\ \mathbf{H}^{yy}_{\mathbf{w}^{(o)}\mathbf{w}^{(o)}} &= \beta_o\mathbf{o}^{(1)}\mathbf{o}^{(1),\top}, \\ \mathbf{H}^{yy}_{\mathbf{b}_i^{(1)}\mathbf{b}_j^{(1)}} &= \beta_o(\alpha_{h,i}^{(1)})^2(\mathbf{w}^{(o)})_i^2 + \alpha_z\beta_{h,i}^{(1)}(\mathbf{w}^{(o)})_i \text{ si } i = j \\ &= \beta_o\alpha_{h,i}^{(1)}\alpha_{h,j}^{(1)}(\mathbf{w}^{(o)})_i(\mathbf{w}^{(o)})_j \text{ sinon,} \end{aligned}$$

$$\mathbf{H}^{yy} = \begin{pmatrix} \mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{w}_1^{(1)}} & \cdots & \mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{w}_d^{(1)}} & \mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{w}^{(o)}} & \mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{b}^{(1)}} & \mathbf{h}^{yy}_{\mathbf{w}_1^{(1)}b^{(2)}} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{w}_1^{(1)}} & \cdots & \mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{w}_d^{(1)}} & \mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{w}^{(o)}} & \mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{b}^{(1)}} & \mathbf{h}^{yy}_{\mathbf{w}_d^{(1)}b^{(2)}} \\ (\mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{w}^{(o)}})^\top & \cdots & (\mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{w}^{(o)}})^\top & \mathbf{H}^{yy}_{\mathbf{w}^{(o)}\mathbf{w}^{(o)}} & \mathbf{H}^{yy}_{\mathbf{w}^{(o)}\mathbf{b}^{(1)}} & \mathbf{h}^{yy}_{\mathbf{w}^{(o)}b^{(2)}} \\ (\mathbf{H}^{yy}_{\mathbf{w}_1^{(1)}\mathbf{b}^{(1)}})^\top & \cdots & (\mathbf{H}^{yy}_{\mathbf{w}_d^{(1)}\mathbf{b}^{(1)}})^\top & (\mathbf{H}^{yy}_{\mathbf{w}^{(o)}\mathbf{b}^{(1)}})^\top & \mathbf{H}^{yy}_{\mathbf{b}^{(1)}\mathbf{b}^{(1)}} & \mathbf{h}^{yy}_{\mathbf{b}^{(1)}b^{(2)}} \\ (\mathbf{h}^{yy}_{\mathbf{w}_1^{(1)}b^{(2)}})^\top & \cdots & (\mathbf{h}^{yy}_{\mathbf{w}_d^{(1)}b^{(2)}})^\top & (\mathbf{h}^{yy}_{\mathbf{w}^{(o)}b^{(2)}})^\top & (\mathbf{h}^{yy}_{\mathbf{b}^{(1)}b^{(2)}})^\top & h_{b^{(2)}} \end{pmatrix},$$

FIGURE 2 : Matrice Hessienne de la sortie du modèle par bloc.

$$\mathbf{H}^{yy} = c \begin{pmatrix} \mathrm{diag}(\boldsymbol{\beta}_h^{(1)} \odot \mathbf{w}^{(o)}) \otimes \mathbf{x}\mathbf{x}^\top & \mathrm{diag}(\alpha_{h,1}^{(1)}, \ldots, \alpha_{h,d}^{(1)}) \otimes \mathbf{x} & \mathrm{diag}(\boldsymbol{\beta}_h^{(1)} \odot \mathbf{w}^{(o)}) \otimes \mathbf{x} & \mathbf{0}_{dp} \\ & \mathbf{O}_d & \mathrm{diag}(\alpha_{h,1}^{(1)}, \ldots, \alpha_{h,d}^{(1)}) & \mathbf{0}_d \\ & & \mathrm{diag}(\boldsymbol{\beta}_h^{(1)} \odot \mathbf{w}^{(o)}) & \mathbf{0}_d \\ & & & 0 \end{pmatrix},$$

FIGURE 3 : Matrice Hessienne de la sortie du modèle, par bloc, pour une sortie linéaire.

et, pour le dernier terme diagonal :

$$h_{b^{(2)}} = \beta_o.$$

Les autres termes sont :

$$\begin{aligned} \mathbf{H}^{yy}_{\mathbf{w}_i^{(1)}(\mathbf{b}^{(1)})_j} &= \frac{\partial}{\partial \mathbf{b}_j^{(1)}} \nabla_{\mathbf{w}_i^{(1)}} y \\ &= \beta_o (\alpha_{h,i}^{(1)})^2 (\mathbf{w}^{(o)})_i^2 \, \mathbf{x} + \alpha_o \beta_{h,i}^{(1)} (\mathbf{w}^{(o)})_i \, \mathbf{x} \\ &\quad \text{si } i = j; i, j = 1, \ldots, d \\ &= \beta_o \alpha_{h,j}^{(1)} (\mathbf{w}^{(o)})_j \alpha_{h,i}^{(1)} (\mathbf{w}^{(o)})_i \, \mathbf{x} \text{ sinon,} \\ \mathbf{H}^{yy}_{\mathbf{w}_i^{(1)}(\mathbf{w}^{(o)})_j} &= \frac{\partial}{\partial \mathbf{w}_j^{(o)}} \nabla_{\mathbf{w}_i^{(1)}} y \\ &= \left(\beta_o \alpha_{h,i}^{(1)} (\mathbf{w}^{(o)})_i o_i^{(1)} + \alpha_o \alpha_{h,i}^{(1)} \right) \mathbf{x} \\ &\quad \text{si } i = j \\ &= \beta_o \alpha_{h,i}^{(1)} (\mathbf{w}^{(o)})_i o_j^{(1)} \, \mathbf{x} \text{ sinon,} \\ \left(\mathbf{H}^{yv}_{\mathbf{w}^{(o)}\mathbf{b}^{(1)}}\right)_{i,j} &= \frac{\partial^2 y}{\partial \mathbf{w}_i^{(o)} \partial \mathbf{b}_j^{(1)}} \\ &= \beta_o (\mathbf{w}^{(o)})_i \alpha_{h,i}^{(1)} o_i^{(1)} + \alpha_o \alpha_{h,i}^{(1)} \\ &\quad \text{si } i = j \\ &= \beta_o (\mathbf{w}^{(o)})_j \alpha_{h,j}^{(1)} o_i^{(1)} \text{ sinon.} \end{aligned}$$

Finallement les dernières quantités valent :

$$\begin{aligned} \mathbf{h}^{yy}_{\mathbf{w}_i^{(1)}b^{(2)}} &= \beta_o \alpha_{h,i}^{(1)} (\mathbf{w}^{(o)})_i \, \mathbf{x}, \\ \mathbf{h}^{yy}_{\mathbf{w}^{(o)}b^{(2)}} &= \beta_o \mathbf{o}^{(1)}; \\ \left(\mathbf{h}^{yy}_{\mathbf{b}^{(1)}b^{(2)}}\right)_i &= \beta_o \alpha_{h,i}^{(1)} (\mathbf{w}^{(o)})_i. \end{aligned}$$

En utilisant l'hypothèse de linéarité de la couche de sortie

Nous considérons maintenant le problème de regression, c'est-à-dire $f_o(x) = cx, \forall x \in \mathbb{R}$, on a alors $\beta_o = 0$ et $\alpha_o = c$).

Les dérivées secondes (matrice symmétrique) est donné dans la Figure 3 avec, pour rappel, respectivement $\odot$ et $\otimes$ les produits d'Hadamard et de Kronecker respectivement.

Il ne reste alors plus qu'à recombiner (3) et les équations de la Figure 3 dans (2).