ENS Paris-Saclay
L3 SAPHIRE

Année 2023-2024
Examen

Examen - Modélisation - Estimation Paramétrique

L'usage des calculatrices est interdit. Seule une feuille de notes manuscrites A4 est autorisée. Les parties sont indépendantes. Il est conseillé de lire attentivement le sujet avant de commencer. Les réponses devront être soigneusement argumentées.

1 Modèle auto-régressif

On considère le modèle à temps discret

$$y_{\mathrm{m}}(\boldsymbol{\theta}, k) = \theta_1 y^2(k-1) + \theta_2 y(k-2)$$

du système dynamique non-linéaire suivant

$$y(k) = \theta_1^* y^2(k-1) + \theta_2^* y(k-2) + \varepsilon(k)$$

où les $\varepsilon(k)$ sont des réalisations indépendantes et identiquement distribuées d'un bruit de moyenne nulle et $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ est un vecteur de paramètres dont la vraie valeur est $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)^T$. On cherche à déterminer l'estimée $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ à partir de N mesures de la sortie $y(1), \ldots, y(N)$ qui minimise l'erreur de prédiction

$$\begin{aligned} f(\boldsymbol{\theta}) &= \sum_{k=3}^{N} \left(y(k) - y_{\mathrm{m}}(\boldsymbol{\theta}, k)\right)^2 \\ &= \|\mathbf{y} - \mathbf{R}\boldsymbol{\theta}\|^2 . \end{aligned}$$

1. Donner l'expression de $\mathbf{y}$ et de $\mathbf{R}$.
2. Donner l'expression de $\widehat{\boldsymbol{\theta}}$ minimisant $f(\boldsymbol{\theta})$ en fonction de $\mathbf{y}$ et de $\mathbf{R}$.

2 Localisation d'une victime à l'aide d'une flotte de drones

Suite à un tremblement de terre, vous avez été chargé(e) par la sécurité civile d'une mission de localisation de survivants enfouis dans des décombres. Pour cela, vous allez exploiter le fait que les équipements électroniques (montres connectées, téléphones portables) des survivants vont continuer à émettre un rayonnement électromagnétique. pendant une ou deux journées après la survenue du désastre. Vous disposez d'une flotte de N drones équipés de réseaux d'antennes capables de déterminer l'angle d'incidence d'une onde électromagnétique par rapport à la normale au réseau. A partir de la mesure des angles d'incidence des ondes reçues par les réseaux d'antennes, vous devez déterminer la position d'une victime.

Nous considérons un repère orthonormé direct $\mathcal{R} = (\mathbf{O}, \mathbf{i}, \mathbf{j}, \mathbf{k})$ attaché à l'environnement. A un instant t_1 donné, la position des drones dans $\mathcal{R}$ est $\mathbf{x}_k(t_1)$, $k = 1, \ldots, N$. Nous supposerons que les altitudes des drones sont égales, c'est-à-dire que $x_{3,k}(t_1) = h$, $k = 1, \ldots, N$. Une victime se trouve à la position $\mathbf{x}$. Pour simplifier, nous supposerons que la victime se trouve à la même altitude que les drones ($x_3 = h$) et qu'elle est statique. Les drones perçoivent une onde électromagnétique avec un angle d'incidence $\theta_k(t_1)$, $k = 1, \ldots, N$ par rapport à $(\mathbf{O}, \mathbf{i})$, dans le plan $(\mathbf{O}, \mathbf{i}, \mathbf{j})$.

1. Donner l'expression du critère à minimiser pour réaliser une estimation au sens des moindres carrés $\widehat{\mathbf{x}}_{\mathrm{MC}}$ de $\mathbf{x}$ à partir de $\theta_k(t_1)$, $k = 1, \ldots, N$.
2. Est-il possible d'obtenir une expression explicite de $\widehat{\mathbf{x}}_{\mathrm{MC}}$? Si non, quel type d'algorithme pouvez-vous utiliser pour obtenir une expression approchée de $\widehat{\mathbf{x}}_{\mathrm{MC}}$?
3. Les drones se déplacent à altitude constante et à l'instant t_2, leurs positions sont $\mathbf{x}_k(t_2)$, $k = 1, \ldots, N$. Ils réalisent une nouvelle mesure d'angles $\theta_k(t_2), k = 1, \ldots, N$. Quelle est l'expression de l'estimée au sens des moindres carrés de $\mathbf{x}$ à partir de $\theta_k(t_1)$ et $\theta_k(t_2)$, $k = 1, \ldots, N$?
4. Nous supposerons maintenant que les bruits de mesure des angles sont décrits par des variables aléatoires Gaussiennes indépendantes et identiquement distribuées de moyenne nulle et de variance σ^2. Donner l'expression de l'estimée au sens du maximum de vraisemblance de $\mathbf{x}$ à partir de $\theta_k(t_1)$, $k = 1, \ldots, N$.

3 Calcul de gradient

On considère un système décrit par le modèle à temps discret suivant

$$\begin{cases} x_1(k+1) = p_1 x_1^2(k) + p_2 x_2(k) \\ x_2(k+1) = -p_3 x_1(k)\, x_2(k) \end{cases}$$

avec $\mathbf{x}(0) = (1, 0)^T$. La sortie mesurée du modèle est

$$y_{\mathrm{m}}(\mathbf{p}, k) = x_1(k).$$

On souhaite réaliser une estimée au sens des moindres carrés du vecteur des paramètres $\mathbf{p} = (p_1, p_2, p_3)^T$ à partir de mesures $y(1), \ldots, y(N)$ obtenues sur le système. Pour cela, on cherche

$$\widehat{\mathbf{p}}_{\mathrm{MC}} = \arg\min_{\mathbf{p}} c(\mathbf{p})$$

avec

$$c(\mathbf{p}) = \sum_{k=1}^{N} \left(y(k) - y_{\mathrm{m}}(\mathbf{p}, k)\right)^2. \tag{1}$$

A l'aide d'une technique par état adjoint ou par code adjoint, proposez un algorithme ou un programme permettant de calculer le gradient du critère (1) de manière exacte.

4 Minimisation d'un critère quadratique

On cherche à calculer

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{H} \mathbf{x},$$

où $\mathbf{H}$ est une matrice symétrique définie positive. Pour cela, on utilise la méthode du *gradient* avec optimisation *exacte* du pas.

1. Montrez que $\mathbf{x} = \mathbf{0}$ est la solution unique de ce problème.
2. Rappelez la structure d'un algorithme d'optimisation par la méthode du gradient.
3. On fixe une direction de descente $\mathbf{p}$. Quel sous-problème doit-on résoudre pour effectuer l'optimisation exacte du pas α. Montrez que la longueur du pas α que l'on obtient en faisant une optimisation exacte du pas est donnée par

$$\alpha = -\frac{\mathbf{p}^{\mathrm{T}} \mathbf{g}}{\mathbf{p}^{\mathrm{T}} \mathbf{H} \mathbf{p}},$$

où $\mathbf{g}$ est le gradient de $f(\mathbf{x})$ en $\mathbf{x}$.

4. On considère le cas particulier où la matrice $\mathbf{H}$ est la matrice diagonale suivante

$$\mathbf{H} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix},$$

avec $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n > 0$. On considère un point de départ

$$\mathbf{x}_1 = \left(\frac{\sigma}{\lambda_1}, 0, \ldots, 0, \frac{1}{\lambda_n}\right)^{\mathrm{T}},$$

avec $\sigma = \pm 1$. Montrez qu'après la première itération de l'algorithme du gradient avec optimisation exacte du pas, on obtient

$$\mathbf{x}_2 = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\left(\frac{-\sigma}{\lambda_1}, 0, \ldots, 0, \frac{1}{\lambda_n}\right)^{\mathrm{T}}.$$

Montrez qu'après l'itération $k+1$, on obtient

$$\mathbf{x}_{k+1} = \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^k \left(\frac{(-1)^k \sigma}{\lambda_1}, 0, \ldots, 0, \frac{1}{\lambda_n}\right)^{\mathrm{T}}.$$

5. Que peut-on dire de la vitesse de convergence lorsque
   (a) $\lambda_1 = \lambda_n$,
   (b) $\lambda_1 >> \lambda_n$.

5 Evaluation du gradient d'une fonction coût par rétro-propagation dans un réseau de neurones

Un réseau de neurones tel que celui représenté sur la figure 1 peut être considéré comme une manière d'implanter la composition de fonctions

$$F^L\left(F^{L-1}\left(\ldots F^1\left(\mathbf{x}\right)\right)\right)$$

où F^ℓ représente la couche $\ell = 1, \ldots, L$ du réseau de neurones. Chaque fonction F^ℓ consiste en une partie linéaire impliquant une matrice $\mathbf{W}^\ell$ et une fonction d'activation non-linéaire f^ℓ appliquée élément par élément. L'entrée de la fonction d'activation de la couche ℓ est appelé vecteur de pré-activation $\mathbf{z}^\ell$ et est le résultat de la multiplication de l'entrée de la couche par $\mathbf{W}^{\ell T}$. La fonction d'activation f^ℓ est appliquée élément par élément aux coefficients de $\mathbf{z}^\ell$ pour obtenir le vecteur d'activation $\mathbf{a}^\ell$. La couche 1 est la couche d'entrée alimentée par le vecteur d'entrée $\mathbf{x}$ et la couche L est la couche de sortie fournissant le vecteur de sortie $\mathbf{y}$.

5.1 Propagation vers l'avant

La propagation vers l'avant de la couche $\ell - 1$ vers la couche ℓ des vecteurs d'activation s'effectue par le biais de la fonction F^ℓ telle que

$$\mathbf{a}^\ell = F^\ell\left(\mathbf{a}^{\ell-1}\right)$$

où F^ℓ peut être décomposée en une partie linéaire et une partie non-linéaire. La partie linéaire évalue

$$\mathbf{z}^\ell = \mathbf{W}^{\ell T}\mathbf{a}^{\ell-1} + \mathbf{b}^\ell, \tag{2}$$

où $\mathbf{W}^\ell$ est une matrice de poids de taille $n_{\ell-1} \times n_\ell$ pour la couche ℓ et où $\mathbf{b}^{(\ell)}$ est un vecteur de biais de taille $n_\ell \times 1$.

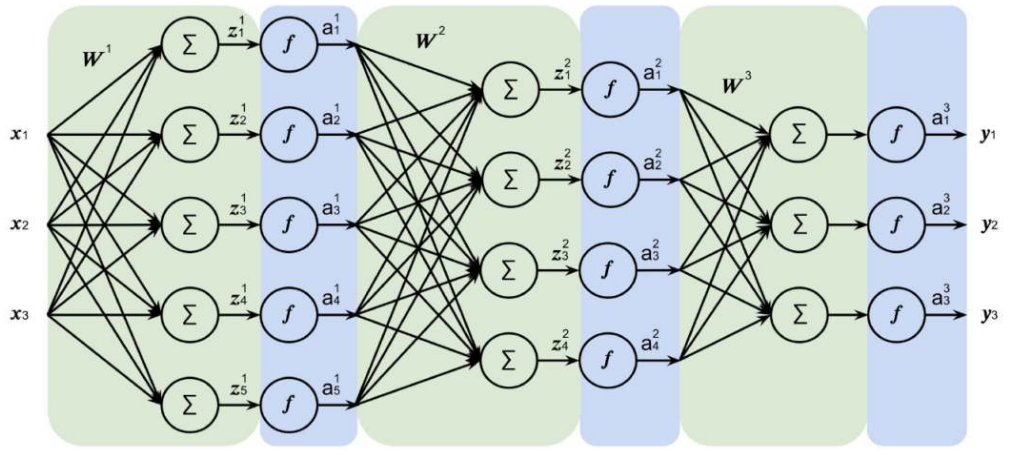

FIGURE 1 – Exemple de réseau de neurones à trois couches entièrement connecté. Chaque couche du réseau se compose d'une partie linéaire et d'une partie non linéaire. L'entrée est le vecteur $\mathbf{x} = (x_1, x_2, x_3)^T$ de taille 3×1, qui est multiplié par la matrice $\mathbf{W}^{1T}$ de taille 5×3 pour produire le vecteur de pré-activation $\mathbf{z}^1 = \left(z_1^1, \ldots, z_5^1\right)^T = \mathbf{W}^{1T}\mathbf{x}$ de taille 5×1. Une fonction d'activation non linéaire f^1 est appliquée à chaque élément z_i^1 du vecteur de pré-activation afin d'obtenir un vecteur d'activation $\mathbf{a}^1 = \left(a_1^1, \ldots, a_5^1\right)^T$ tel que $a_i^1 = f\left(z_i^1\right)$. Le vecteur d'activation $\mathbf{a}^1$ est l'entrée de la deuxième couche. Des opérations similaires sont effectuées aux deuxième et troisième couches. La sortie du réseau $\mathbf{y}$ est le vecteur d'activation de la troisième couche $\mathbf{y} = \mathbf{a}^3$.

1. Afin de simplifier les notations, construire une matrice $\overline{\mathbf{W}}^\ell$ à partir de $\mathbf{W}^\ell$ et de $\mathbf{b}^\ell$ et construire un vecteur $\overline{\mathbf{a}}^{\ell-1}$ déduit de $\mathbf{a}^{\ell-1}$ tel que $\mathbf{z}^\ell = \overline{\mathbf{W}}^{\ell T}\overline{\mathbf{a}}^{\ell-1}$.

Dans ce qui suit, le vecteur de biais n'est plus pris en compte dans (2). En outre, pour alléger les notations, $\mathbf{W}^\ell$ est utilisé à la place de $\overline{\mathbf{W}}^\ell$ et $\mathbf{a}^{\ell-1}$ à la place de $\overline{\mathbf{a}}^{\ell-1}$. La fonction d'activation non linéaire f^ℓ pour la couche ℓ opère sur chaque élément de $\mathbf{z}^\ell$ pour obtenir

$$\mathbf{a}^\ell = f^\ell\left(\mathbf{z}^\ell\right).$$

2. Écrire un algorithme simple effectuant la propagation vers l'avant de l'entrée $\mathbf{x}$ à travers le réseau de neurones pour obtenir la sortie $\mathbf{y}$.
3. La fonction d'activation non linéaire considérée dans ce qui suit est la sigmoïde

$$f(z) = \frac{1}{1 + e^{-z}}. \tag{3}$$

Évaluer la dérivée par rapport à z de $f(z)$ donnée par (3) et montrer qu'elle peut être exprimée comme une fonction de $f(z)$.

5.2 Fonction coût

En apprentissage supervisé, un ensemble de d'apprentissage $\mathcal{T}$, contenant des paires $\mathcal{T} = \left\{\left(\mathbf{x}_{(i)}, \mathbf{y}_{(i)}\right)_{i=1,\ldots,N}\right\}$, est utilisé pour ajuster les paramètres du réseau de neurones de manière à minimiser une certaine fonction coût. Soit $\widehat{\mathbf{y}}_{(i)} = F\left(\mathbf{x}_{(i)}, \mathcal{W}\right)$ la sortie du réseau de neurones lorsqu'il est alimenté par $\mathbf{x}_{(i)}$. L'ensemble $\mathcal{W} = \left\{\mathbf{W}^1, \ldots, \mathbf{W}^L\right\}$ contient tous les paramètres du réseau de neurones.

En considérant une paire $\left(\mathbf{x}_{(i)}, \mathbf{y}_{(i)}\right)$ et la fonction coût quadratique, on obtient

$$\begin{aligned}\mathcal{L}\left(\mathbf{y}_{(i)}, \widehat{\mathbf{y}}_{(i)}\right) &= \left\|\mathbf{y}_{(i)} - \widehat{\mathbf{y}}_{(i)}\right\|^2 \\ &= \left\|\mathbf{y}_{(i)} - F\left(\mathbf{x}_{(i)}, \mathcal{W}\right)\right\|^2.\end{aligned} \tag{4}$$

4. Expliquer pourquoi l'ajustement des paramètres du réseau de neurones est généralement effectué en considérant un ou plusieurs sous-ensembles de l'ensemble d'apprentissage au lieu d'une seule paire dans l'évaluation de la fonction coût

$$\mathcal{L}\left(\mathbf{y}_{\mathcal{M}}, \widehat{\mathbf{y}}_{\mathcal{M}}\right) = \frac{1}{m} \sum_{i \in \mathcal{M}} \left\| \mathbf{y}_{(i)} - F\left(\mathbf{x}_{(i)}, \mathcal{W}\right) \right\|^2 \tag{5}$$

où $\mathcal{M}$ est un sous-ensemble de $m < N$ éléments de $\{1, \ldots, N\}$.

5.3 Rétro-propagation des gradients

La minimisation de (5) par descente de gradient nécessite l'évaluation du gradient de $\mathcal{L}$ par rapport à $\mathbf{W}^\ell$, $\ell = 1, \ldots, L$. Cette opération peut être réalisée efficacement par rétro-propagation.

Dans ce qui suit, nous supposons que $\mathcal{L}$ représente $\mathcal{L}\left(\mathbf{y}_{(i)}, \widehat{\mathbf{y}}_{(i)}\right)$. Supposons que $\frac{\partial \mathcal{L}}{\partial \mathbf{a}^\ell}$ soit disponible à la couche ℓ. L'objectif est d'évaluer $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^\ell}$ ainsi que $\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{\ell-1}}$.

5. En utilisant la règle de dérivation des fonctions composées

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell} = \frac{\partial \mathbf{a}^{\ell T}}{\partial \mathbf{z}^\ell} \frac{\partial \mathcal{L}}{\partial \mathbf{a}^\ell},$$

montrer que

$$\frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell} = f^{\ell\prime}\left(\mathbf{z}^\ell\right) \odot \frac{\partial \mathcal{L}}{\partial \mathbf{a}^\ell},$$

où $f^{\ell\prime}\left(\mathbf{z}^\ell\right)$ est la dérivée de f^ℓ appliquée à chaque élément de $\mathbf{z}^\ell$ et où $\odot$ est le produit élément par élément de deux vecteurs

$$\mathbf{u} \odot \mathbf{v} = \begin{pmatrix} u_1 v_1 \\ \vdots \\ u_n v_n \end{pmatrix}.$$

6. En utilisant à nouveau la règle de dérivation des fonctions composées, montrer que

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{\ell-1}} = \mathbf{W}^\ell \frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell}.$$

Pour évaluer $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^\ell}$, on applique une nouvelle fois la règle de dérivation des fonctions composées pour obtenir

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^\ell} = \frac{\partial \mathbf{z}^{\ell T}}{\partial \mathbf{W}^\ell} \frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell},$$

où

$$\frac{\partial \mathbf{z}^{\ell T}}{\partial \mathbf{W}^\ell} = \left[\frac{\partial z_1^\ell}{\partial \mathbf{W}^\ell}, \ldots, \frac{\partial z_{n_\ell}^\ell}{\partial \mathbf{W}^\ell} \right]$$

est un vecteur de matrices et

$$\frac{\partial \mathbf{z}^{\ell T}}{\partial \mathbf{W}^\ell} \frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell} = \sum_{i=1}^{n_\ell} \frac{\partial z_i^\ell}{\partial \mathbf{W}^\ell} \frac{\partial \mathcal{L}}{\partial z_i^\ell}$$

est une matrice.

7. Montrer que

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^\ell} = \mathbf{a}^{\ell-1} \left(\frac{\partial \mathcal{L}}{\partial \mathbf{z}^\ell} \right)^T.$$

Finalement, la rétro-propagation des gradients est initialisée par

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}^L} = \frac{\partial \mathcal{L}}{\partial \widehat{\mathbf{y}}_{(i)}}.$$

8. Evaluer $\frac{\partial \mathcal{L}}{\partial \widehat{\mathbf{y}}_{(i)}}$ pour la fonction coût quadratique (4).
9. Concevoir un algorithme pour effectuer l'évaluation de la fonction coût et la rétro-propagation des gradients.

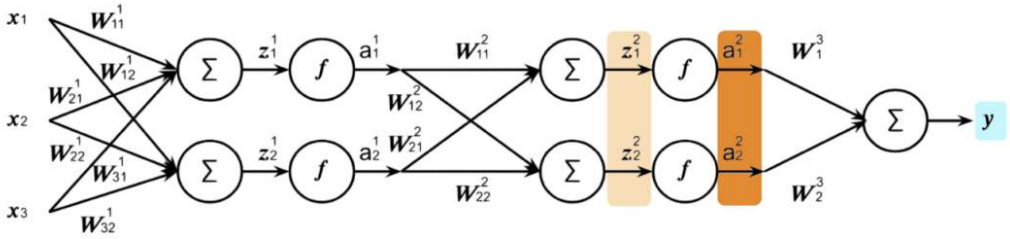

FIGURE 2 – Exemple de réseau de neurones à trois couches entièrement connecté.

10. Appliquer l'algorithme précédant au réseau de neurones de la figure 2 en considérant une fonction coût quadratique.

——oOo——