Formation Saphire

211

Identification paramétrique

14 mai 2019

Note de Cours

Version du 14 mai 2019

Un cours de:
Michel Kieffer

Pris en note par:

Table des matières

1 Introduction **1**
- I Objectifs . . . 1
  - I.1 Modèle . . . 1
- II Système . . . 1

2 Structure de modèle **5**
- I Modèle de connaissance ou de comportement . . . 5
  - I.1 Modèle de connaissance . . . 5
  - I.2 Modèle de comportement . . . 5
  - I.3 En résumé . . . 6
- II Modèle linéaire et non linéaires. . . . 6
- III Modèle à temps continu ou à temps discret . . . 6
  - III.1 Modèle à temps continu . . . 6
  - III.2 Modèle à temps discret . . . 6
  - III.3 Comparaison . . . 7
- IV Prise en compte de perturbations . . . 7
  - IV.1 Modèle ARX (Auto-Régressif à entrée eXogène) . . . 7
  - IV.2 Modèle ARMAX (Auto-Régressif à Moyenne Ajustée sur l'entrée eXogène 7
- V Propriétés structurelles d'un modèle . . . 7
  - V.1 identifiabilité structurelle . . . 7
  - V.2 discernabilité structurelle . . . 8

3 Critères **9**
- I Critère des moindres carrés . . . 9
- II Critères des moindres carrés pondérés . . . 9
- III Critère des moindres valeurs absolues . . . 10
- IV Critère de maximum de vraisemblance . . . 10
- V Critère du maximum à posteriori . . . 11
- VI Critère du choix de complexité . . . 11
- VII Robustesse . . . 11

4 Optimisation **13**
- I Condition d'optimalité . . . 13
- II Cas particulier des modèles LP avec critère quadratiques . . . 13
- III Critère générique dérivable, algorithme de descente . . . 14
  - III.1 Méthode du gradient . . . 14
  - III.2 Méthode de Newton et ses variantes . . . 15
- IV Méthode de Gauss-Newton . . . 16

5 Gradient **17**
- I Mauvaise méthode par différence finie . . . 17

II utilisation des fonctions de sensibilité. 17
II.1 Modèle LE décrit par des EDOs 17
II.2 Modèle quelconque décrit par des EDOs 18

Chapitre 1

Introduction

I Objectifs

I.1 Modèle

un modèle est un ensemble de fonctions mathématiques permettant de décrire un système d'intéret. Il est construit dans un but précis :

- afin d'estimer des quantités
- optimiser un processus de fabrication
- diagnostic
- prédire le comportement d'un système
- Simulateurs
- ...

L'objectif doit être explicité lors de la construction d'un modèle.

II Système

Un système est une partie de l'univers que l'on considère de manière isolée et qui est l'objet de l'étude.

SCHEMA SYSTEME

- u : entrée maîtrisée du système
- y : sortie mesurée
- z : sortie non mesurée
- b : perturbation non maîtrisées

Modèles Équation , algorithme permettant de décrire ce système. On associe modèle et sytème en :

- parallèle :
  On applique la même entrée que sur le système et il fournit y_m et z_m.

- série (ou modèle inverse)
  Ce type de modèle permet de reconstruire l'entrée du système lorsque elle n'est pas maitrisée.

On peut aussi combiner plusieurs type de modèle.

Modèle paramétriques : modèle faisant intervenir des grandeurs de réglages, les paramètres on note $\mathbf{p}$ le vecteur des paramètres. On distingue bien :

- $\mathcal{M}(\cdot)$ la structure du modèle
- $\mathcal{M}(\mathbf{p})$ le modèle de paramètre $\mathbf{p}$

Le choix de la structure du modèle s'appelle la caractérisation. On cherche à ajuster les paramètres du modèle de manière à ce que la sortie du modèle soit proches des mesures effectuées sur le système.

Il faut introduire un critère ui permet de quantifier l'écart entre la sortie du système et la sortie du modèle.

Le critère $j\mathbf{p}$ est une fonction scalaire que l'on cherche à minimiser. On cherche l'argument du minimum de $j(\cdot)$. On obtient l'estimée : $\hat{\mathbf{p}}$ au sens du critère choisi.

On considère un modèle parallèle :

SCHEMA MODELE PARALLELE

Le critère $j(\mathbf{p})$ sera en général une fonction de l'erreur de sortie :

$$e(t, \mathbf{p}) = y(t) - y_m(t, \mathbf{p})$$

L'optimiseur aura pour objectif de rendre j le plus petit possible. La plupart des méthodes utilisée sont itératives.

Il est nécessaire d'avoir une condition d'arret. L'optimisation peut être :

- en ligne
- hors-ligne (batch) où l'optimisation débute quand toutes les mesures ont étés collectées.

Déterminer l'*incertitude* d'estimation des paramètres est indispensable pour pouvoir juger la qualité de l'estimée.

On peux chercher le protocole expérimental permettant de rendre cette *incertitude la plus faible possible* = c'est l'objectif de la *planification d'expérience*

À coût expérimental fini quelles sont les conditions qui vont conduire à l'estimée de f la plus précise possible ?

Au cours du processus de modélisation, de nombreuse hypothèse sont faites :

- Le choix de la structure du modèle

- nature du bruit de mesure

Il faut les mettres à l'épreuve,chercher à invalider ces hypothèses à l'aide de tests. Si ces test ne sont pas concluants, il faut remettre en question des hypothèses. Sinon on a pas validé le modèle, on a juste réussi à ne pas l'invalidé.

Démarche de modélisation

1. Recueil des données du système
2. Choix d'une ou plusieurs structures du modèle
3. Choisir un critère pour obtenir le meilleur modèle et l'estimée du paramètre au sens du critère
4. Évaluer l'incertitude d'estimation
5. Analyse critique des résultats
6. Reboucler si nécessaire.

Chapitre 2

Structure de modèle

Choix d'une structure : c'est la caractérisation. On note :

- $\mathcal{M}(\cdot)$ la structure du modèle
- $\mathbf{p}$ le vecteur de paramètres
- $\mathcal{M}(\mathbf{p})$ le modèle pour la valeur $\mathbf{p}$ du vecteur des paramètres.
- $\mathbb{P}$ l'espace admissible du vecteurs des paramètres.

Le choix de $\mathcal{M}$ et $\mathbb{P}$ est délicat et déterminant pour la suite du processus de modélisation.

I Modèle de connaissance ou de comportement

I.1 Modèle de connaissance

Un modèle de connaissance implique les gds propres de la physique(équation de conservation, bilan) ce type de modèle se prête bien à la *prise en compte d'information à priori* sur les valeurs des paramètres et au *contrôle a posteriori*de la validité des paramètres estimées.

exemple : Réaction chimique Modèle de connaissance :

- Peut être de dimension élevée (discrétisation d'EDP)
- Comporter de nombreuse équations
- en général non linéaire
- plus difficile à simuler qu'un modèle de comportement
- approprié pour faire de la prédiction à long terme.

I.2 Modèle de comportement

Un modèle de comportement est un programme ou un ensemble d'équation faisant apparaitre un vecteur de paramètre dont l'objectif est de reproduire le comportement du système. Les paramètres n'ont pas de sens physique.

Exemple modèle polynomiale : $y_m(, \mathbf{p}, t) = p_0 + p_1 t + p_2 t^2 + \dots$ Il permet de reproduire tout type de signaux scalaire à condition que le degré soit assez élevé.

Exemple 2 :$y_m(,\mathbf{p},t) = \sum_{i=1}^{n} p_i h(t - z_i)$ avec $h(t)$ quelconque.

D'autre modèle sont possibles : splines, Réseau de neurone, Support Vector Machine.

I.3 En résumé

	Modèle de connaissance	**modèle de comportement**
paramètres	sens physique	pas de sens physique
simulation	plus difficile	aisée
info a priori	facile à intégrer	difficile à prendre en compte
domaine de validité	grand	limité

II Modèle linéaire et non linéaires.

Il existes deux type de linéarité :

- La linéarité par rapport à l'entrée (LE) $\forall \lambda, \mu$ et $u_1(t), u_2(t)$:

$$y_m(\lambda u_1 + \mu u_2, \mathbf{p}) = \lambda y_m(u_1, \mathbf{p}) + \mu y_m(u_1, \mathbf{p})$$

- La linéarité par rapport aux paramètres (LP). $\forall \lambda, \mu$ et $\mathbf{p_1}, \mathbf{p_2}$

$$y_m(t, \lambda \mathbf{p_1} + \mu \mathbf{p_2}) = \lambda y_m(t, \mathbf{p_1}) + y_m(t, \mu \mathbf{p_2})$$

La linéarité par rapport à l'entrée est souvent limité à des entrées de faibles amplitude. Elle permet d'appliquer des outils d'analyse, de stabilité et de synthèse de commande efficaces.

Les modèles LP facilites :

- l'estimation de leurs paramètres
- la synthèse de protocoles expérimentaux
- l'évaluation de l'incertitude d'estimation.

Exemple de modèle LP, LE ou non : Il est important de déterminer si un modèle est LP ce la facilite l'estimation . On peux dans certain cas transformer un modèle NLP en modèle LP.

III Modèle à temps continu ou à temps discret

III.1 Modèle à temps continu

La plupart des modèles de connaissances exploitant les lois de la physique sont à *temps continu*. On utilise des équation d'état. Ce type de modèle permet de décrire beaucoup de système physique. On peux souvent y appliquer la transformée de laplace et procédé à une résolution algébrique. Les simulations font appels au méthode de résolution d'équation différentielle (Euler , Runge-Kutta)

III.2 Modèle à temps discret

Les modèles à temps discret sont plus simple à implementer : on mets en œuvre des équation de récurrence. Dans le cas discret on utilise la transformée en Z.

III.3 Comparaison

	TC	TD
simulation	plus délicate	aisée
tps de mesure	quelconque	fixée par la période d'échantillonnage
valeurs des paramètres	fixe	peuvent dépendre de la période d'échantillonnage

IV Prise en compte de perturbations

La prise en compte de perturbation au niveau du modèle permet d'intégrer :

- des parties difficiles à décrire du système
- des incertitudes de mesures.

On considère ici des modèles scalaires (décrit par des équation de récurrence) :

- $y(t)$ sortie
- $u(t)$ entrée
- $b(t)$ perturbation

IV.1 Modèle ARX (Auto-Régressif à entrée eXogène)

$$y(t)+a_1y(t-1)+\cdots+a_{n_a}y(t-n_a)=b_1u(t-n_r)+b_2u(t-n_r-1)+\cdots+b_{n_b}u(t-n_r-n_b)+\varepsilon(t)$$

$\varepsilon(t)$ sont des réalisations de variables indépendantes et identiquement distribuées (bruit blanc). Les paramètres de ce modèles sont $\mathbf{p}=(a_1,\ldots,a_{n_a},b_1\ldots b_{n_b}),n_r$.

IV.2 Modèle ARMAX (Auto-Régressif à Moyenne Ajustée sur l'entrée eXogène

$$y(t)+a_1(t-1)+\ldots a_{n_a}t(t-n_a)=b_1u(t-n_r)+\cdots+b_{n_b}u(t-n_r-n_b)+c_1\varepsilon(t)+\cdots+c_{n_c}\varepsilon(t-n_c+1)$$

avec $\varepsilon(t)$ une suite de VA indépendante et identiquement distribuée.

V Propriétés structurelles d'un modèle

On cherche à déterminer des propriétés structurelles càd liées à la structure du modèle et valable pour presque toutes les valeurs du vecteurs des paramètres. On s'intéresse à 2 propriétés :

V.1 identifiabilité structurelle

Une fois une structure de modèle choisie a-t-on une chance de pouvoir estimer $\mathbf{p}$ de manière non ambiguë ?

On se place dans une cadre idéalisé :

- Le système est décris par le modèle $\mathcal{M}(\mathbf{p}^*)$
- Il n'y a pas de bruit de mesure
- l'entrée u peux être choisie arbitrairement.

SCHEMA MODELE PARALLELE UNE ENTREE

Le paramètre p_i sera structurellement globalement identifiable (sgi) ssi$\mathbf{p} = \mathbf{p}^* \implies p_i = p_i^*$

Le modèle est sgi si pour presque toutes les valeurs de $\mathbf{p}$ $\mathbf{p} = \mathbf{p}^* \implies \mathbf{p} = \mathbf{p}^*$

Un paramètre est structurellement localement identifiable (sli) si pour presque tout $\mathbf{p}^*$ il existe un voisinage $V(\mathbf{p}^*)$ tq $\mathbf{p} \in V(\mathbf{p}^*)$ et $\mathbf{p} = \mathbf{p}^* \implies p_i = p_i^*$

il existe un ensemble dénombrable de p_i possibles.

Dans les autres cas, p_i sera structurellement non identifiable.

Méthode de test de l'identifiabilité structurelle pour des modèles LE de type

$$\begin{cases} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y} = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{cases}$$

Ou $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ dépendent de $\mathbf{p}$.

1. Calculer la FT de ce modèle $\mathbf{H}(s, \mathbf{p})$
2. Mettre les fractions rationnelles de $\mathbf{H}(s, \mathbf{p})$ sous formes réduites et canonique.
3. On forme un système d'équation en identifiant tous les termes de degrés similaire de $\mathbf{H}(s, \mathbf{p}) = \mathbf{H}(s, \mathbf{p}^*)$
4. Résoudre le système d'équation.

Une solution unique à ce système implique que $\mathcal{M}(\cdot)$ est sgi
Un ensemble dénombrable de solution implique que $\mathcal{M}(\cdot)$ est sli.

Même si un système est sgi il n'est pas certain qu'on puisse estimer $\mathbf{p}^*$ à cause du bruit.

V.2 discernabilité structurelle

On dispose de plusieurs structures de modèles de concurrents $\mathcal{M}_\infty(\cdot)$,$\mathcal{M}_\infty(\cdot), \ldots$ On suppose que le système est décrit par $\mathcal{M}(\mathbf{p}^*)$. On cherche à déterminer si on peux discerner $\mathcal{M}_2$ de $\mathcal{M}_1$.

- $\mathcal{M}_2$ est structurellement discernable de $\mathcal{M}_1$ si pour presque toutes les valeurs de $\mathbf{p}^*$, il n'existe pas de vecteur $\mathbf{q}$ tq $\mathcal{M}_2(\mathbf{q}) = \mathcal{M}_1(\mathbf{p}^*)$ (sortie identique pour toute entrée)
- cette propriété n'est pas symétrique systématiquement.

Méthode de test pour des modèles LE décrit par équation d'état

1. Écrire les matrices de transferts sous formes canoniques
2. Identifier termes à termes pour obtenir un système d'équation faisant intervenir les composantes de $\mathbf{p}$ et $\mathbf{q}$
3. Si on chervhe à déterminer si $\mathcal{M}_2$ est sd de $\mathcal{M}_1$ il faut chercher à exprimer $\mathbf{q}$ en fonction de $\mathbf{p}$.
4. si c'est possible, $\mathcal{M}_2$ n'est pas s.d de $\mathcal{M}_2$.

Chapitre 3

Critères

On a choisit une structure de modèle. On a prélevé des mesures sur la sortie du système. Il s'agir de trouver une fonction scalaire permettant pour une valeur de $\mathbf{p}$ du vecteur des paramètres de déterminer si $\mathcal{M}(\mathbf{p})$ est nous satisfaisant et dans quelle mesure.

I Critère des moindres carrés

On suppose que des mesures y_i ont été prélevées à $t_i, i = 1 \dots N$. on note alors $\mathbf{y} = (y(t_1) \dots y(t_N)^T$ le vecteur des mesures. Et $y_m(\mathbf{p}) = (y_m(p_1) \dots y_m(p_N))^T$ le vecteur des sorties du modèle.

Le critère des moindres carrés est :

$$\begin{aligned} j_{MC}(\mathbf{p}) &= \sum_{i=1}^{N} (y_i - y_{m,i}(\mathbf{p}))^2 \\ &= \|\mathbf{y} - \mathbf{y_m}(\mathbf{p})\|_2^2 \\ &= (\mathbf{y} - \mathbf{y_m}(\mathbf{p}))^T (\mathbf{y} - \mathbf{y_m}(\mathbf{p})) \end{aligned}$$

C'est le critère à utiliser quand on ne sait rien sur les bruits de mesures et les valeurs que peut prendre $\mathbf{p}$. On obtient une estimée au sens des moindres carrés :

$$\hat{\mathbf{p_{MC}}} = \arg\min_{\mathbf{p}} (j_{MC}(\mathbf{p}))$$

II Critères des moindres carrés pondérés

On considère une matrice $\mathbf{W}$ symétrique et définie non négative. Le critère des moindres carrés pondérés est :

$$j_{MCP}(\mathbf{p}) = (\mathbf{y} - \mathbf{y_m}(\mathbf{p}))^T W (\mathbf{y} - \mathbf{y_m}(\mathbf{p}))$$

Si on choisit $\mathbf{W} = \mathrm{diag}(w_1 \dots w_N)$ on a alors :

$$j_{MCP}(\mathbf{p}) = \sum_{i=1}^{N} w_i (y_i - y_{m,i}(\mathbf{p}))^2$$

Les w_i permettent de donner plus ou moins d'importance à la i-ème mesure, mais le choix des poids n'est pas évident. Ce type de critère est très utilisé car il peux conduire à une optimisation relativement aisée (différentiabilité).

III Critère des moindres valeurs absolues

Le critère des moindres valeurs absolues est de la forme :

$$j_{MVA} = \|\mathbf{y} - \mathbf{y_M}(\mathbf{p})\|_1$$

Version pondérée par une matrice $\mathbf{W}$:

$$j_{MVAP} = \|vecW(\mathbf{y} - \mathbf{y_M}(\mathbf{p}))\|_1$$

Lorsque $\mathbf{W} = \mathrm{diag}(w_1 \dots w_N)$ on a :

$$j_{MVAP} = \sum_{i=1}^{N} w_i |y_i - y_{m,i}(\mathbf{p})|$$

Le critère de MVA n'est pas partout différentiable !

Dans le cas particulier des modèles LP ou affines $\mathbf{y_m}(\mathbf{p}) = \mathbf{r}^T\mathbf{p} + \mathbf{c_0}$ on peux réaliser aisément une estimée au sens des MVA en résolvant le problème équivalent :

$$(\hat{\mathbf{p}}, \hat{\alpha}) = \arg\min \sum_{i=1}^{N} \alpha_i \text{ sous la contrainte } |y_i - \mathbf{r}(\mathbf{t_i})^T\mathbf{p} - \mathbf{c_0}| \leq \alpha_i \forall i$$

C'est un programme d'optimisation sous contrainte qui peux être résolu aisément à l'aide d'outils de résolution de programme linéaire.

IV Critère de maximum de vraisemblance

On suppose connaitre dans ce cas la distribution de probabilité du bruit $\mathbf{b}$ qui est venu perturber les mesures et on fait l'hypothèse que $\mathbf{y} = \mathbf{y_m}(\mathbf{p}) + \mathbf{b}$. On note la distribution du bruit $\pi_B(\mathbf{b})$ (proba). Le critère du maximum de vraisemblance est alors :

$$j_{MV}(\mathbf{p}) = \pi_{y|p}(\mathbf{y}|\mathbf{p})$$

et l'estimée au sens du maximum de vraisemblance :

$$\hat{\mathbf{p_{MV}}} = \arg\max_{\mathbf{p}} \pi_{y|p}(\mathbf{y}|\mathbf{p})$$

La vraisemblance $\pi_{y|p}(\mathbf{y}|\mathbf{p})$ donne une mesure du caractère vraisemblable des mesures comptes tenus de la valeur $\mathbf{p}$ du vecteur de paramètres. Comment calculer cette vraisemblance ? on a besoin de $\mathbf{y}, \pi_B(\mathbf{b})$

alors :

$$\pi(\mathbf{p}|\mathbf{y}) = \pi(\mathbf{y}|\mathbf{y_m}(\mathbf{p}), \mathbf{p}) = \pi_B(\mathbf{y} - \mathbf{y_m}(\mathbf{p}))$$

EXEMPLE IMPORTANT A RECOPIER (bruit gaussien , laplacien)

V Critère du maximum à posteriori

On suppose connaitre la distribution de proba du bruit de mesure, une distribution à priori sur le vecteur de paramètre $\pi_0(\mathbf{p})$ De plus on considère que $\mathbf{y} = \mathbf{y_m}(\mathbf{p}^*) + \mathbf{b}$

L'estimation au sens du maximum a posteriori introduit le critère :

$$j_{MAP}(\mathbf{p}) = \pi(\mathbf{p}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\mathbf{p})\pi_0(\mathbf{p})}{\pi(\mathbf{y})}$$

et on cherche

$$\hat{\mathbf{p}}_{MAP} = \arg\max_{\mathbf{p}} \pi(\mathbf{p}|\mathbf{y}) = \arg\max \pi(\mathbf{y}|\mathbf{p})\pi_0(\mathbf{p})$$

plus on fait de mesure et moins l'a priori à d'importance et plus la vraisemblance à d'importance.

VI Critère du choix de complexité

L'objectif de ce type de critère est de choisir la complexité du modèle pour décrire le système. Il faut trouver un compromis entre :

- la précision de représentation des mesures par le modèle.
- la complexité du modèle qui ne doit pas être trop importante afin de ne pas sur-représenter les mesures (overfit).

On considère ici uniquement le critère Akaike (AIC)

$$j_{AIC}(\mathbf{p}, \mathcal{M}) = \frac{1}{n}\left(-\pi(\mathbf{y}|\mathbf{p}) + n_p\right)$$

où n_p est le nombre de paramètre et n le nombre de mesure.

VII Robustesse

En cas de présence de donnée aberrante (défaillance de capteur, mesure non expliqué par un modèle trop grossier...) il est utile d'utiliser des critères robustes. M estimateurs permettent de bénéficier des bonnes propriétés des deux critères. on a :

$$j_M(\mathbf{p}) = \sum_{i=1}^{M} \rho(e(i, \mathbf{p}))$$

Avec $e(i, \mathbf{p}) = y_i - y_{m,i}(\mathbf{p})$.

Critère de Huber

$$\rho_H = \begin{cases} \frac{1}{2}e^2 \text{ si } |e| < \delta \\ \delta|e| - \delta^2/2 \text{ sinon} \end{cases}$$

Critère de Tuckey

$$\rho_T = \begin{cases} \frac{1}{2}\left(e^2 - \frac{e^4}{\delta^2} + \frac{e^6}{3\delta^4}\right) \text{ si } |e| < \delta \\ \frac{\delta^2}{6} \text{ sinon} \end{cases}$$

GRAPHE des deux critère et leur dérivée

qualitativement ρ_T a une meilleur robustesse de que ρ_H mais des algorithme de d'optimisation peuvent êtres bloqués si l'initialisation est mauvaise.

Pour évaluer la robustesse d'un estimateur la notion de point de rupture a été introduite :

Si $\mathbf{y}$ est le vecteur des mesures, on positionne $\mathbf{y}$ en $\mathbf{y'}_{1-\alpha}$ et $\mathbf{y'}_{\alpha}$ avec $\mathbf{y'}_{1-\alpha}$ mesures originales et $\mathbf{y'}_{\alpha}$ mesures modifiées par un mauvais génie. α représente la proportion de mesures modifiées. biais de l'estimateur :

$$b(\alpha, \hat{\mathbf{p}}) = \max_{\mathbf{y'}_{\alpha}} \|\hat{\mathbf{p}}(\mathbf{y}) - \hat{\mathbf{p}}(\mathbf{y'}_{1-\alpha}, \mathbf{y'}_{\alpha})\|$$

Le point de rupture de l'estimateur est α tel que $b \to \infty$ Si on suppose que N mesure ont été réalisés : $PR(p_{MC}) = PR(p_{MVA}) = 1/N$

Chapitre 4

Optimisation

Un critère $j(\mathbf{p})$ a été choisi. Il s'agit de trouver :

- $\hat{\mathbf{p}} = \arg\min j(\mathbf{p})$ MC , MVA
- $\hat{\mathbf{p}} = \arg\max j(\mathbf{p})$ MV , MAP

Les deux problèmes sont résolus à l'aide des mêmes outils ($\max = -\min$)

I Condition d'optimalité

On considère $j : \mathbb{R}^n \to \mathbb{R}$

- $\mathbf{p}$ est un minimum local de j s'il existe $V(\mathbf{p})$ un voisinage de $\mathbf{p}$ tel que $\forall \mathbf{p} \in V, j(\mathbf{p}) \leq j(\mathbf{p}')$. $\mathbf{p}$ est un minimum local *strict* de j s'il existe $V(\mathbf{p})$ un voisinage de $\mathbf{p}$ tel que $\forall \mathbf{p} \in V, j(\mathbf{p}) < j(\mathbf{p}')$.

GRAPHE p15

Condition suffisante pour avoir un minimum local

$$\mathbf{g}(\mathbf{p}) = \left.\frac{\partial j}{\partial \mathbf{p}}\right|_{\hat{\mathbf{p}}} = 0 \quad et \quad \mathbf{H}(\hat{\mathbf{p}}) = \left.\frac{\partial^2 j}{\partial \mathbf{p} \partial \mathbf{p}^T}\right|_{\hat{\mathbf{p}}} \text{ est définie non négative}$$

pour avoir un minimum strict on veut $\mathbf{H}(\hat{\mathbf{p}})$ *définie positive.*

II Cas particulier des modèles LP avec critère quadratiques

Un modèle LP est de la forme :

$$y_m(\mathbf{p}) = \mathbf{R}\mathbf{p}$$

Avec $\mathbf{R}$ matrice de régresseurs.

On choisit un critère des moindres carrés pondérés.

$$j_{MCP}(\mathbf{p}) = (\mathbf{y} - \mathbf{y_m}(\mathbf{p}))^T W (\mathbf{y} - \mathbf{y_m}(\mathbf{p}))$$

ou $\mathbf{W}$ est une matrice de pondération symétrique définie positive. On cherche $\hat{\mathbf{p}} = \arg\min j_{MCP}(\mathbf{p})$

$$\begin{aligned} j_{MCP}(\mathbf{p}) &= (\mathbf{y} - \mathbf{Rp}))^T W(\mathbf{y} - \mathbf{Rp})) \\ &= \mathbf{y^T W y} - \mathbf{p^T R^T W y} - \mathbf{y^T W R p} + \mathbf{p^T R^T W R p} \end{aligned}$$

Pour calculer $\frac{\partial j_M CP}{\partial \mathbf{p}}$ on utilise :

- $\frac{\partial \alpha}{\partial \hat{\mathbf{p}}} = 0$
- $\frac{\partial \mathbf{p^T a}}{\partial \mathbf{p}} = \frac{\partial \mathbf{a^T p}}{\partial \mathbf{p}} = \mathbf{a}$
- $\frac{\partial \mathbf{p^T A p}}{\partial p} = 2\mathbf{Ap}$

Ainsi on a :

$$\frac{\partial j_{MCP}}{\partial \mathbf{p}} = 0 - \mathbf{R^T W y} - \mathbf{R^T W y} + \mathbf{2R^T W R \hat{p}} = 0 \implies \boxed{\hat{\mathbf{p}} = \left(\mathbf{R^T W R}\right)^{-1} \mathbf{R^T W y}}$$

De plus :

$$\frac{\partial^2 j_{MCP}(\mathbf{p})}{\partial \mathbf{p} \partial \mathbf{p^T}} = 2\mathbf{R^T W R} \quad \text{est definie positive}$$

si $\mathbf{R^T W R}$ est définie positive l'estimée est alors le *minimum global*. C'est le seul cas ou on est certain d'avoir un minimum global atteint , avec une expression explicite de l'estimée.

III Critère générique dérivable, algorithme de descente

on dispose d'un critère dérivable pour le minimiser on utilise un algorithme de descente :

1. Initialisation $j_0 = j(p_0), k = 1$
2. Faire
   $\mathbf{d_k}$ = recherche_dir_descente(p_k)
   α_k =opt_pas($\mathbf{p_k}, \mathbf{d_k}$)
   $p_{k+1} = p_{k-1} + \alpha_k \mathbf{d_k}$
3. Tant que critère_arret = Faux.

III.1 Méthode du gradient

La méthode du gradient considère un développement de Taylor du critère à l'ordre 1 :

$$j(\mathbf{p_k} + \mathbf{q}) = j(\mathbf{p_k}) + \mathbf{q}^T \underbrace{\left.\frac{\partial j}{\partial \mathbf{p}}\right|_{\mathbf{p_k}}}_{\mathbf{g_k}} + O(\|q\|^2)$$

n'importe quelle valeure de $\mathbf{q}$ telle que $\mathbf{q^T g_k} < 0$ permet localement de faire décroitre le critère (à condition que certaines propriétés sur $\frac{\partial j}{\partial \mathbf{p} \partial p}$ soient satisfaites).

La méthode du gradient considère $\mathbf{d_k} = -\mathbf{g_k}$.

GRAPHE p17

Une optimisation de pas est indispensable, on doit trouver $\hat{\alpha} = \arg\min \phi(\alpha)$ avec $\phi(\alpha) = j(\hat{p} + \alpha d_k)$ On peux faire une optimisation exacte du pas (coûteuse) ou utiliser une méthode de *Backtracking* :

1. Initialisation $\alpha_0 = 1, k = 1, \phi_0 = j(\mathbf{p})$.
2. $\phi_k = j(\mathbf{p} + \alpha_{k-1}\mathbf{d})$
3. Si $\phi_k > 0$

   $\alpha_k = z\alpha_{k-1}$

   aller en 2
4. renvoyer α_k

il faut que $z \in]0, 1[$.

III.2 Méthode de Newton et ses variantes

On effectue ici un développement à l'ordre 2 du critère :

$$j(\mathbf{p_k} + \mathbf{d}) = j(\mathbf{p}) + \mathbf{d}^T \underbrace{\left.\frac{\partial j}{\partial \mathbf{p}}\right|_{\mathbf{p_k}}}_{\mathbf{g_k}} + \frac{1}{2}\mathbf{d}^T \underbrace{\left.\frac{\partial^2 j}{\partial \mathbf{p}\partial \mathbf{p}^T}\right|_{\mathbf{p_k}}}_{\mathbf{H_k}} \mathbf{d} + O(\|d\|^2)$$

La méthode de newton cherche la direction de descente qui minimise $j(\hat{\mathbf{p}} + \mathbf{d})$. Il faut que $\mathbf{H_k d} = -\mathbf{d}$ Si on calcule $\mathbf{d_N^T g_k} = -\mathbf{d_N^T H_k d_n}$ $\mathbf{d_N}$ est une direction de descente uniquement si H_k est definie positive. On a alors $d_n = -H_k^- 1 g_k$

GRAPHE p18 .

La méthode de Newton est indépendante du choix des unités de $\mathbf{p}$.

Dans le bassin d'attraction la méthode de Newton converge en moins d'itération, mais il faut calculer une hessienne à chaque itération.

L'algorithme de Levenberg-Marquardt (LM) permet de combiner les avantages du gradient et de newton. Idée : utiliser une direction de descente $\mathbf{d}_{LM}$ telle que $(\mathbf{H_k} + \mu\mathbf{I})\mathbf{d}_{LM} = -\mathbf{g_k}$ avec μ tel que $(\mathbf{H_k} + \mu\mathbf{I})$ soit définie positive.

1. Initialisation $\mathbf{p}_0, \mu_0$
2. calcul de j
3. Faire
4. Calcul de $\mathbf{d}_{LM}^k = -(\mathbf{H}_k + \mu_k\mathbf{I})\mathbf{g_k}$

   Si $j(\mathbf{p}_k + \mathbf{d_k})^{-1}\mathbf{g_k}$

   $\mathbf{p_{k+1}} = \mathbf{p_k} + \mathbf{d_k}$

   $\mu_{k+1} = \alpha\mu_k$ (avec $\alpha \in]0, 1[$)

   Sinon

   $\mu_k = \beta\mu$ avec $\beta > 1$

   aller en 4
5. Tant que condition d'arret pas satisfaite.

Lorsque μ est petit $\mathbf{d_{LM}} \simeq \mathbf{d_N}$
Lorsque μ est grand $\mathbf{d_{LM}} \simeq \frac{1}{\mu}\mathbf{d_g}$.

IV Méthode de Gauss-Newton

Adaptée à des critère de la forme $j(\mathbf{p}) = \mathbf{f^T}(\mathbf{p})\mathbf{f}(\mathbf{p})$. Par exemple avec les moindres carrés :

$$j_{MC}(\mathbf{p}) = \sum_{i=1}^{N} (y_i - y_{m,i}(\mathbf{p}))^2$$

on pose alors

$$\mathbf{f}(\mathbf{p}) = \begin{pmatrix} y_1 - y_{m,1} \\ \vdots \\ y_N - y_{m,N} \end{pmatrix}$$

On fait une approximation linéaire de f et on a

$$\mathbf{f}(\mathbf{p} + \mathbf{d}) \simeq \mathbf{f}(\mathbf{p}) + \underbrace{\frac{\partial \mathbf{f}}{\partial \mathbf{p}^T}(\mathbf{p})}_{\mathbf{J}} \mathbf{d}$$

Où $\mathbf{J}$ est la jacobienne de $\mathbf{f}$.

On peux alors obtenir une approximation quadratique de $j(\mathbf{p})$ à l'aide $\mathbf{J}(\mathbf{p})$

$$\begin{aligned} j(\mathbf{p}) &\simeq (\mathbf{f}(\mathbf{p}) + J(p)d)^T (f(p) + J(p)d) \\ &\simeq \mathbf{f}(\mathbf{p})^\mathbf{T}\mathbf{f}(\mathbf{p}) + \mathbf{d^T}\mathbf{J^T}(\mathbf{p})\mathbf{F}(\mathbf{p}) + \mathbf{f^T}\mathbf{J}(\mathbf{p})\mathbf{d} + \mathbf{d^T}\mathbf{J^T}(\mathbf{p})\mathbf{J}(\mathbf{p})\mathbf{d} \\ &= j_Q(\mathbf{d}) \end{aligned}$$

On cherche la direction de descente minimisant l'approximation quadratique.

$$\frac{\partial j_Q(\mathbf{d})}{\partial \mathbf{d}} = 2\mathbf{J^T}(\mathbf{p})\mathbf{F}(\mathbf{p}) + 2\mathbf{J^T}(\mathbf{p})\mathbf{J}(\mathbf{p})\mathbf{d}$$

et

$$\mathbf{d}_{GN} = -(\mathbf{J^T}(\mathbf{p})\mathbf{J}(\mathbf{p}))^{-1}\mathbf{J}(\mathbf{p})^\mathbf{T}\mathbf{F}(\mathbf{p})$$

On a donc $\mathbf{J^T}(\mathbf{p})\mathbf{J}(\mathbf{p})$ une approximation de la hessienne, définie non négative. On a une convergence proche de la méthode de Newton sans avoir à calculer de Hessienne.

Chapitre 5

Gradient

I Mauvaise méthode par différence finie

On considère $j(\mathbf{p})$ différentiable , $\frac{\mathrm{d}j}{\mathrm{d}p_i} \simeq \frac{j(\mathbf{p}+\varepsilon\mathbf{e_i})-j(\mathbf{p})}{\varepsilon}$

Si $\mathbf{p} \in \mathbb{R}^{n_p}$ il faut réaliser $n_p + 1$ calculs de j pour obtenir le gradient en $\mathbf{p}$ (en fait une *valeur approchée* dépendante de ε) De plu sis epsilon est trop petit, $1 + \varepsilon = 1$ numériquement et les calculs sont aberrants.

II utilisation des fonctions de sensibilité.

on considère un critère $j(\mathbf{p}) = \sum_{i=1}^{N} e(\mathbf{p}, i)^2$. On calcul son gradient :

$$\frac{\partial j(\mathbf{p})}{\partial \mathbf{p}} = 2\sum_{i=1}^{N} e(\mathbf{p}, i)\frac{\partial e(\mathbf{p}, i)}{\partial \mathbf{p}}$$

On introduit la fonction de sensibilité de l'erreur par rapport à $\mathbf{p}$

$$\mathbf{s}^e(\mathbf{p}, i) = \frac{\mathrm{d}e(\mathbf{p}, i)}{\mathrm{d}\mathbf{p}}$$

Si en plus on a $e(\mathbf{p}, i) = y_i - y_{m,i}(\mathbf{p})$ alors $\mathbf{s}^e(\mathbf{p}, i) = -\frac{\partial y_{m,i}(\mathbf{p})}{\partial \mathbf{p}} = -\mathbf{s}^{y_m}(\mathbf{p}, i)$

$\mathbf{s}^{\mathbf{y}}(\mathbf{p}, i)$ est la fonction de sensibilité du modèle par rapport à $\mathbf{p}$. ses composantes indiquent si $y_m(\mathbf{p}, i)$ dépend peu ou beaucoup des p_i. *Comment la calculer ?*

II.1 Modèle LE décrit par des EDOs

On considère l'exemple suivant :

$$\frac{\partial^2 y_m}{\partial t^2} + p_1\frac{\partial y_m}{\partial t} + p_2 y_m = p_3\frac{\mathrm{d}u}{\mathrm{d}t} + p_4 u(t)$$

Avec $y_m(\mathbf{p}, 0) = 0$ et $\frac{\partial y_m(\mathbf{p},t)}{\partial t}(0) = 0$

Pour trouver une EDO satisfaite par $\mathbf{s}_1^y(\mathbf{p},t)$ on dérive l'equation du modèle par rapport à p_1 et on inverse les dérivées.

$$\frac{\partial^2}{\partial t^2}\underbrace{\frac{\partial y_m}{\partial p_1}}_{s_1^y}+\frac{\partial y_m}{\partial t}p_1\frac{\partial}{\partial t}\underbrace{\frac{\partial y_m}{\partial p_1}}_{s_1^y}+p_2\underbrace{\frac{\mathrm{d}y_m}{\mathrm{d}p_1}}_{s_1^y}=0$$

on dérive aussi les conditions initiales et on a $s_1^y(\mathbf{p},0)=0$ et $\frac{\partial}{\partial t}s_1^y(\mathbf{p},t)=0$

On fait de même pour les autres paramètres et on obtient 5 EDO du deuxième ordre. Sans réécriture du système, on ne gagne rien en temps de calcul par rapport au gradient, par contre on n'a pas à choisir de paramètre ε , on a une meilleure précision.

II.2 Modèle quelconque décrit par des EDOs

On considère la famille de modèles :

$$\begin{cases}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}=\mathbf{f}(\mathbf{x},\mathbf{p},\mathbf{u})\\ \mathbf{y_m}(\mathbf{t})=\mathbf{h}(\mathbf{x},\mathbf{p})\end{cases}$$

On cherche à calculer

$$\frac{\partial y_m(\mathbf{p},i)}{\partial \mathbf{p}^T}=\frac{\partial \mathbf{h}}{\partial x^T}\frac{\partial \mathbf{x}^T}{\partial \mathbf{p}}+\frac{\partial \mathbf{h}}{\partial \mathbf{p}^T}$$

$\mathbf{S}^x(\mathbf{p})=\frac{\partial \mathbf{x}^T}{\partial \mathbf{p}}$ est la mtrice de sensibilité de l'état par rapport aux paramètres.

On peux trouver une EDO satisfaire par chaque composante de $\mathbf{S}^x(\mathbf{p})$ en se servant de l'équation du modèle, on dérive par rapport au paramètres et en inversant les dérivées on obtient :

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathbf{x}}{\partial p_i}=\frac{\partial \mathbf{f}}{\partial \mathbf{x^T}}\frac{\partial \mathbf{x}^T}{\partial p_i}+\frac{\partial f}{\partial p_i}$$