Réalisations optimales pour l'implantation de systèmes LTI paramétrés

Thibault Hilaire [*] **Philippe Chevrel** [**]

[*] *Laboratoire d'Informatique de Paris 6 (LIP6),*
Université Pierre et Marie Curie, Paris.
thibault.hilaire@lip6.fr
[**] *Institut de Recherche en Communication et Cybernétique (IRCCyN),*
École des Mines de Nantes.
philippe.chevrel@emn.fr

Résumé : Cet article s'intéresse à la résilience des filtres/régulateurs LTI paramétrés. On entend ici par résilience la robustesse vis-à-vis de l'implantation en précision finie. Précisément, on s'intéresse à l'impact de la quantification des coefficients représentés en virgule fixe. La nouveauté tient au fait que les coefficients implantés ne sont pas supposés définis de manière définitive lors de l'étape d'initialisation du code embarqué, mais calculés à partir de paramètres dont la valeur n'est pas figée *a priori*.
Ces paramètres sont supposés pouvoir être modifiés ultérieurement, à l'intérieur d'une plage prédéfinie, sans reconception du code embarqué. C'est le cas souvent, par exemple dans le domaine automobile, où un dernier réglage ou une adaptation peuvent être réalisés très tard dans le cycle de conception.
Nous formaliserons l'erreur occasionnée par une erreur de quantification sur ces paramètres, avant d'étudier le cas d'un filtre du $2^{\text{ème}}$ ordre dont les gain, amortissement et pulsation naturelle ne sont pas supposés fixés *a priori*.

Mots-clés: Implantation, virgule fixe, système LTI, système paramétré, réalisation optimale.

1. INTRODUCTION

La phase d'implantation d'une loi de contrôle-commande dans un système embarqué numérique est une tâche difficile. En effet, que ce soit avec un processeur généraliste ou spécialisé (DSP), un ASIC ou un FPGA, la précision finie des calculs amène une modification de la relation entrée-sortie du régulateur ou filtre, qu'il faut nécessairement maîtriser. Cette dégradation possède deux origines : la quantification des coefficients utilisés et les arrondis qui interviennent dans les calculs. Nous nous intéresserons ici principalement à la première ; et ceci dans un contexte d'arithmétique en virgule fixe (arithmétique entière pour approcher les réels), courante pour les systèmes embarqués fortement contraints en ressources et puissance de calcul.
Ce problème de maîtrise de la précision des calculs lors de l'implantation de filtres ou régulateurs est abordé sous trois angles différents, par trois communautés différentes : l'automatique au travers de la recherche de la réalisation optimale de systèmes dynamiques, le traitement du signal pour les bruits de calcul, et l'arithmétique des ordinateurs dans un cadre plus général, principalement en flottant.
Le problème qui nous intéresse est un peu différent de celui traité classiquement : il s'agit de l'implantation de systèmes LTI, dont les coefficients, inconnus lors de l'implantation du code, sont définis et calculés *in-situ* par le calculateur au moment de son initialisation. Ces coefficients sont déduits de paramètres dont on connait simplement, lors de la phase d'implantation, la plage des valeurs admissibles. Ces paramètres sont utilisés principalement pour régler/calibrer le comportement du régulateur *a posteriori*, sur la base d'informations acquises plus tardivement dans le cycle de conception, voire de données expérimentales. Dans ce cadre, il importe de maîtriser l'impact de la quantification des coefficients, induit cette fois, et c'est la nouveauté, par la quantification des paramètres eux-mêmes. De plus, la valeur des paramètres n'étant pas connue *a priori*, c'est l'erreur sur la fonction de transfert dans le pire cas des valeurs admissibles qu'il importe peu ou prou de minimiser.
En généralisant un résultat précédent sur les systèmes LTI, cet article propose une mesure de résilience, qui tient compte de l'ensemble des valeurs de paramètres admissibles. Des réalisations optimales ou sous-optimales pourront ainsi être élaborées en regard de cette mesure.

Le plan de l'article est le suivant. Partant en section 2 de l'analyse classique quant à l'impact de la quantification des coefficients en virgule fixe, nous formaliserons à la section 3 le problème d'implantation de systèmes LTI paramétrés, et généraliserons à ce cadre une mesure basée sur la norme H_2 de l'espérance de l'erreur de la fonction de transfert. Nous envisagerons ensuite, à la section 4, la possibilité de chercher parmi les réalisations équivalentes en précision infinie, celle qui maximise la résilience au sens de cette mesure. Les résultats seront illustrés au travers du traitement d'un exemple simple mais important : celui du choix d'une implantation résiliente pour une cellule du 2ème ordre paramétrée.

2. ANALYSE CLASSIQUE DE L'IMPACT DE LA QUANTIFICATION DES COEFFICIENTS

2.1 Mesure de sensibilité

On considère un système à une entrée et une sortie (SISO) stable, commandable et observable, décrit par :

$$\begin{cases} \boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{b}u(k) \\ y(k) = \boldsymbol{c}\boldsymbol{x}(k) + du(k) \end{cases} \tag{1}$$

où $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{b} \in \mathbb{R}^{n\times 1}$, $\boldsymbol{c} \in \mathbb{R}^{1\times n}$ et $d \in \mathbb{R}$. $u(k)$ correspond à l'entrée scalaire, $y(k)$ la sortie scalaire et $\boldsymbol{x}(k) \in \mathbb{R}^{n\times 1}$ le vecteur d'état.
Sa relation entrée-sortie est donnée par la fonction de transfert (scalaire) $h : \mathbb{C} \to \mathbb{C}$ telle que :

$$h : z \mapsto \boldsymbol{c}(z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b} + d. \tag{2}$$

Remarque 1. Une forme d'état n'est pas la seule structure de calcul possible pour mettre en œuvre numériquement un système de fonction de transfert h. D'autres formes sont possibles, telles que les formes directes (I et II, transposée ou non), les réalisations avec l'opérateur δ (Middleton et Goodwin, 1986), celles avec l'opérateur ρ, telles que la forme ρ-modale (Feng et al., 2011) ou la ρDFIIt (Li et Zhao, 2004) ou encore les formes LCW et LGS (Li et al., 2007). Toutes possèdent des propriétés numériques différentes lors du passage en précision finie, et un nombre de coefficients et d'opérations différent.
Bien qu'elle ne permette pas de décrire ces autres réalisations, la forme d'état sera utilisée par la suite à des fins de simplicité. Notons malgré tout que les résultats de ce papier peuvent s'étendre à la forme implicite spécialisée proposée dans (Hilaire et al., 2007; Hilaire et Chevrel, 2011) qui permet quant à elle d'aborder une classe de réalisations plus larges, incluant toutes les réalisations à fort intérêt pratique mentionnées ci-dessus.

La quantification des nombres modifie $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$ et d en $\boldsymbol{A}+\boldsymbol{\Delta A}$, $\boldsymbol{b}+\boldsymbol{\Delta b}$, $\boldsymbol{c}+\boldsymbol{\Delta c}$ et $d+\Delta d$, respectivement. Il est classique (Gevers et Li, 1993; Yamaki et al., 2011) de considérer la sensibilité de la fonction de transfert vis-à-vis des coefficients de ces matrices comme une mesure de la dégradation due à la quantification. Pour cela, on utilisera les définitions et propositions suivantes :
Définition 1. (Sensibilité d'une fonction de transfert).
Soient $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ une matrice et $f : \mathbb{R}^{m\times n} \to \mathbb{C}$ une fonction différentiable par rapport à tous les coefficients de $\boldsymbol{X}$.
La sensibilité de f par rapport à $\boldsymbol{X}$ est définie par la matrice $\boldsymbol{S_X} \in \mathbb{R}^{m\times n}$ telle que :

$$\frac{\partial f}{\partial \boldsymbol{X}} \triangleq \boldsymbol{S_X} \quad \text{avec} \quad (\boldsymbol{S_X})_{i,j} \triangleq \frac{\partial f}{\partial \boldsymbol{X}_{i,j}}, \quad \forall i,j. \tag{3}$$

En appliquant cela à la fonction de transfert h, où $h(z)$ dépend d'une matrice donnée $\boldsymbol{X}$, alors $\frac{\partial h}{\partial \boldsymbol{X}}$ est une fonction de transfert à plusieurs entrées et sorties (MIMO) définie par :

$$\frac{\partial h}{\partial \boldsymbol{X}}(z) \triangleq \frac{\partial h(z)}{\partial \boldsymbol{X}}, \quad \forall z \in \mathbb{C}. \tag{4}$$

Définition 2. (Norme L_2). Soit $\boldsymbol{H} : \mathbb{C} \to \mathbb{C}^{k\times l}$ une fonction de transfert MIMO. $\|\boldsymbol{H}\|_2$ est la norme H_2 de $\boldsymbol{H}$, définie par :

$$\|\boldsymbol{H}\|_2 \triangleq \sqrt{\frac{1}{2\pi}\int_0^{2\pi} \left\|\boldsymbol{H}\left(e^{j\omega}\right)\right\|_F^2 \, d\omega} \tag{5}$$

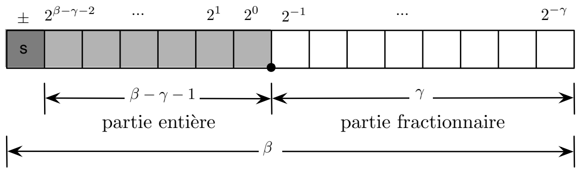

FIGURE 1. Représentation en virgule fixe

où $\|\boldsymbol{.}\|_F$ est la norme de Froebenius.
Proposition 1. Si $\boldsymbol{H}$ est un système MIMO défini par ses matrices de la représentation d'état $(\boldsymbol{K}, \boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$, alors sa norme H_2 peut être calculée par :

$$\|\boldsymbol{H}\|_2^2 = tr(\boldsymbol{N}\boldsymbol{N}^\top + \boldsymbol{M}\boldsymbol{W}_c\boldsymbol{M}^\top) \tag{6}$$

$$= tr(\boldsymbol{N}^\top\boldsymbol{N} + \boldsymbol{L}^\top\boldsymbol{W}_o\boldsymbol{L}) \tag{7}$$

où $\boldsymbol{W}_c$ et $\boldsymbol{W}_o$ sont les Gramiens de commandabilité et d'observabilité, respectivement. Ils sont solutions des équations de Lyapunov :

$$\boldsymbol{W}_c = \boldsymbol{K}\boldsymbol{W}_c\boldsymbol{K}^\top + \boldsymbol{L}\boldsymbol{L}^\top \text{ et } \boldsymbol{W}_o = \boldsymbol{K}^\top\boldsymbol{W}_o\boldsymbol{K} + \boldsymbol{M}^\top\boldsymbol{M}. \tag{8}$$

Gevers et Li (1993) ont proposé une mesure de sensibilité L_2 pour *evaluer* l'impact sur la fonction de transfert de la quantification des coefficients. Cette mesure est définie par

$$M_{L_2} \triangleq \left\|\frac{\partial h}{\partial \boldsymbol{A}}\right\|_2^2 + \left\|\frac{\partial h}{\partial \boldsymbol{b}}\right\|_2^2 + \left\|\frac{\partial h}{\partial \boldsymbol{c}}\right\|_2^2 + \left\|\frac{\partial h}{\partial d}\right\|_2^2 \tag{9}$$

et peut être calculée par $\frac{\partial h}{\partial \boldsymbol{A}}(z) = \boldsymbol{G}^\top(z)\boldsymbol{F}^\top(z)$, $\frac{\partial h}{\partial \boldsymbol{b}}(z) = \boldsymbol{G}^\top(z)$, $\frac{\partial h}{\partial \boldsymbol{c}}(z) = \boldsymbol{F}(z)$ et $\frac{\partial h}{\partial d}(z) = 1$, avec

$$\boldsymbol{F}(z) \triangleq (z\boldsymbol{I}_n - \boldsymbol{A})^{-1}\boldsymbol{b}, \text{ et } \boldsymbol{G}(z) \triangleq \boldsymbol{c}(zI_n - \boldsymbol{A})^{-1}. \tag{10}$$

Remarque 2. Pour simplifier les expressions, il est préférable de regrouper tous les coefficients en une unique matrice $\boldsymbol{Z}$:

$$\boldsymbol{Z} \triangleq \begin{pmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{c} & d \end{pmatrix}. \tag{11}$$

Alors, grâce à une propriété de la norme H_2, il est possible d'écrire

$$M_{L_2} = \left\|\frac{\partial h}{\partial \boldsymbol{Z}}\right\|_2^2. \tag{12}$$

En utilisant l'équation (10) , il est possible d'écrire la sensibilité de la fonction de transfert $\frac{\partial h}{\partial \boldsymbol{Z}}$ sous la forme d'un système MIMO $(\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}, \tilde{\boldsymbol{D}})$ avec :

$$\begin{gathered} \tilde{\boldsymbol{A}} \triangleq \begin{pmatrix} \boldsymbol{A} & \boldsymbol{bc} \\ \boldsymbol{0} & \boldsymbol{A} \end{pmatrix}, \tilde{\boldsymbol{B}} \triangleq \begin{pmatrix} \boldsymbol{0} & \boldsymbol{b} \\ \boldsymbol{I}_n & \boldsymbol{0} \end{pmatrix}, \\ \tilde{\boldsymbol{C}} \triangleq \begin{pmatrix} \boldsymbol{I}_n & 0 \\ \boldsymbol{0} & \boldsymbol{c} \end{pmatrix}, \tilde{\boldsymbol{D}} \triangleq \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{pmatrix}. \end{gathered} \tag{13}$$

La proposition 1 permet ensuite de calculer M_{L_2}. On se reportera à (Gevers et Li, 1993) pour plus de détails.

2.2 Espérance de l'erreur de fonction de transfert

Malheureusement, cette mesure n'est pas complètement représentative de la dégradation réelle de la fonction de transfert h lors de la quantification des coefficients $\boldsymbol{Z}$ car ceux-ci n'ont pas tous la même magnitude et donc ne seront pas modifiés avec une quantité du même ordre de grandeur.

On considère dans cet article la représentation des nombres en **virgule fixe**. Ainsi, pour tout réel x, on notera β_x le

nombre total de bits pour le représenter et γ_x le nombre de bits de sa partie fractionnaire (la position de la virgule), comme présenté à la figure 1. Contrairement à la virgule flottante, le format virgule fixe est fixé (constant) pour chaque variable (entrée, sortie, états) et chaque coefficient, et la position de la virgule (γ) est implicite et non codée dans le nombre lui-même.
Dans la notation utilisée, β et γ pourront être scalaires, vectoriels ou matriciels, et on les indicera par les noms de variables auxquels ils se réfèrent (par exemple, $\gamma_{\boldsymbol{Z}}$ sera la matrice des positions des virgules de $\boldsymbol{Z}$).

Ainsi, si l'on suppose connu le nombre de bits β_x utilisés pour représenter un coefficient x, la position de la virgule de x se calcule par

$$\gamma_x = \beta_x - 2 - \lfloor \log_2 |x| \rfloor \tag{14}$$

où $\lfloor . \rfloor$ représente l'opération d'arrondi vers $-\infty$.

Remarque 3. La position de la virgule ne peut être définie pour les coefficients nuls. Cependant, ce n'est pas un problème car ces coefficients ne seront pas utilisés dans l'algorithme final.
De plus, pour considérer les coefficients qui ne seront pas modifiés par la quantification, on introduit la matrice de *pondération* $\boldsymbol{\delta_Z}$ (de même taille que $\boldsymbol{Z}$) définie par

$$(\delta_{\boldsymbol{Z}})_{ij} \triangleq \begin{cases} 0 & \text{si } \boldsymbol{Z}_{ij} \text{ est exactement implanté} \\ 1 & \text{sinon.} \end{cases} \tag{15}$$

Les coefficients exactement implantés sont 0, ± 1 et les puissances (négatives et positives) de 2.

Durant la quantification, les coefficients $\boldsymbol{Z}$ sont changés en $\boldsymbol{Z}^\dagger \triangleq \boldsymbol{Z} + \boldsymbol{\Delta Z}$. Pour un arrondi au plus près, on considère $\{\boldsymbol{\Delta Z}_{ij}\}$ comme des variables aléatoires indépendantes uniformément distribuées (Sripad et Snyder, 1977) dans l'intervalle $-2^{-\gamma_{\boldsymbol{Z}_{ij}}-1} \leqslant \boldsymbol{\Delta Z}_{i,j} < 2^{-\gamma_{\boldsymbol{Z}_{ij}}-1}$. Si on note $\sigma^2_{\boldsymbol{Z}_{ij}}$ leur variance (moment d'ordre deux), on a

$$\begin{aligned} \sigma^2_{\boldsymbol{Z}_{ij}} &\triangleq E\left\{(\boldsymbol{\Delta Z}_{ij})^2\right\} && (16)\\ &= \frac{2^{-2\gamma_{\boldsymbol{Z}_{ij}}}}{12}\delta_{\boldsymbol{Z}_{ij}}, && (17) \end{aligned}$$

où $E\{.\}$ est l'espérance mathématique d'une variable aléatoire (v.a.).

De plus, de par la quantification des coefficients, la fonction de transfert h du système considéré est transformée en $h^\dagger \triangleq h + \Delta h$. Il est alors possible de voir Δh comme une fonction de transfert dont les coefficients sont les variables aléatoires $\boldsymbol{\Delta Z}_{i,j}$, et l'on peut évaluer *statistiquement* cette dégradation par les définitions et propositions suivantes :

Définition 3. Une mesure de l'erreur de la fonction de transfert peut être définie par (Hinamoto et al., 2002; Hilaire et Chevrel, 2011) :

$$\sigma^2_{\Delta h} \triangleq \frac{1}{2\pi}\int_0^{2\pi} E\left\{\left|\Delta h\left(e^{j\omega}\right)\right|^2\right\} d\omega. \tag{18}$$

On appellera cette mesure *Norme de l'espérance de l'erreur de fonction de transfert* (NEEFT).

Proposition 2. La mesure NEEFT se calcule par (Hilaire, 2009) :

$$\sigma^2_{\Delta h} = \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \boldsymbol{\Xi_Z} \right\|_2^2 \tag{19}$$

où $\times$ est le produit direct (Schur), $\boldsymbol{\Xi_Z} \in \mathbb{R}^{(n+1)\times(n+1)}$ défini par :

$$(\boldsymbol{\Xi_Z})_{ij} \triangleq \begin{cases} \dfrac{2^{-\boldsymbol{\beta}_{\boldsymbol{Z}_{ij}}+1}}{\sqrt{3}} \lfloor \boldsymbol{Z}_{ij} \rfloor_2 (\boldsymbol{\delta_Z})_{ij} & \text{si } \boldsymbol{Z}_{ij} \neq 0 \\ 0 & \text{si } \boldsymbol{Z}_{ij} = 0 \end{cases} \tag{20}$$

et $\lfloor x \rfloor_2$ la puissance de 2 immédiatement inférieure à $|x|$:

$$\lfloor x \rfloor_2 \triangleq 2^{\lfloor \log_2 |x| \rfloor}. \tag{21}$$

Preuve 1. La preuve complète se trouve dans (Hilaire, 2009) et repose sur une approximation au 1$^{\text{er}}$ ordre de Δh :

$$\Delta h(z) = \sum_{i,j} \frac{\partial h}{\partial \boldsymbol{Z}_{ij}}(z)\boldsymbol{\Delta Z}_{ij}, \qquad \forall z \in \mathbb{C}. \tag{22}$$

et sur l'indépendance des v.a. $\boldsymbol{\Delta Z}_{i,j}$. ■

Remarque 4. Il est intéressant d'analyser le lien avec la mesure originelle M_{L_2} de l'équation (9). Dans Hinamoto et al. (2002) où la norme de l'espérance de l'erreur de la fonction de transfert est définie pour la 1$^{\text{ère}}$ fois, les coefficients sont tous supposés avoir la même représentation virgule fixe, et donc leurs variances sont toutes identiques et égales à ce que nous noterons σ_0^2. On a alors dans ce cas :

$$M_{L_2} = \frac{\sigma^2_{\Delta h}}{\sigma^2_0}. \tag{23}$$

La mesure NEEFT (19) peut être vue comme une extension de cette mesure originelle en considérant la représentation des nombres en virgule fixe. Ceci conduit à pondérer la sensibilité relativement à la variance de l'erreur due à la quantification de chaque coefficient.

Enfin, on notera qu'un changement de base, défini par

$$\tilde{\boldsymbol{x}}(k) \triangleq \boldsymbol{\mathcal{U}}^{-1}\boldsymbol{x}(k) \tag{24}$$

sur le système d'état $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}, d)$ amène au nouveau système $(\boldsymbol{\mathcal{U}}^{-1}\boldsymbol{A}\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{U}}^{-1}\boldsymbol{b}, \boldsymbol{c}\boldsymbol{\mathcal{U}}, d)$. Ces deux systèmes, équivalents en précision infinie, ne le sont plus nécessairement en précision finie (quantification des coefficients). Les mesures de précision finie (sensibilité L_2, NEEFT, etc.) dépendent donc du choix de la base.

Ainsi, le problème de réalisation optimale consiste à trouver la réalisation (le changement de base à partir d'un système initial) qui minimisera la dégradation en précision finie :

$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ inversible}}{\arg\min}\ \mathcal{J}(\boldsymbol{\mathcal{U}}), \tag{25}$$

où $\mathcal{J}$ est un critère de dégradation.
Pour $\mathcal{J} = M_{L_2}$, Gevers et Li (1993) ont montré que ce problème était convexe en $\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{U}}\boldsymbol{\mathcal{U}}^\top$ et possédait une unique solution $\boldsymbol{\mathcal{P}}^{opt}$ que l'on peut obtenir par un algorithme de descente de gradient. L'ensemble des solutions s'obtient donc par multiplication d'une solution $\boldsymbol{\mathcal{U}}$ par n'importe quelle matrice orthogonale.
Pour $\mathcal{J} = \sigma^2_{\Delta h}$, Hilaire (2009) a montré que $\sigma^2_{\Delta h}$ était invariant par un changement d'échelle en puissance de 2 ($\boldsymbol{\mathcal{U}}$ diagonal avec des termes diagonaux en puissance de 2) et que l'on pouvait utiliser une étape de *normalisation* du système en imposant une contrainte de mise à l'échelle (L_2-*scaling*), ce qui revient à une contrainte sur les termes diagonaux du Gramien de commandabilité du système, suite à quoi il était possible de trouver une réalisation optimale (*normalisée*).

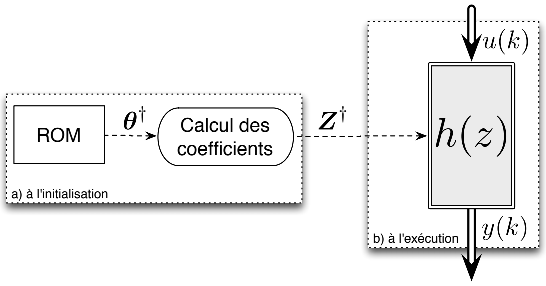

FIGURE 2. Mode de fonctionnement du régulateur

3. SYSTÈMES LTI PARAMÉTRÉS

Cette approche n'est cependant utilisable que pour les systèmes LTI pour lesquels les coefficients (constants) sont connus au moment de l'implantation. Parfois, particulièrement dans le domaine automobile, ces coefficients ne sont définis que plus tard dans le cycle de développement, lors de la phase ultime de réglage, et donc bien après la phase d'implantation en virgule fixe. Au moment du passage en virgule fixe et de l'écriture du code, seule une plage de valeur est identifiée.

De plus, ces coefficients sont parfois calculés *in-situ* au moment de l'initialisation à partir de paramètres *externes*, indépendants du régulateur/filtre. On notera $\boldsymbol{\theta}$ ces paramètres. On supposera donc que le mode de fonctionnement est le suivant (voir figure 2) :

a) à l'initialisation du régulateur, les paramètres $\boldsymbol{\theta}$ sont connus (lus depuis une mémoire). Ces valeurs sont souvent des paramètres de réglages susceptibles d'être modifiés par une mise-à-jour du système ;
b) ensuite les coefficients $\boldsymbol{Z}$ sont calculés en interne ;
c) puis ces coefficients sont utilisés pour toute la durée de la régulation.

Il nous faut donc prendre en considération, désormais, l'effet de la quantification de ces paramètres, le calcul (approché) des coefficients qui en découlent et leur quantification. Ceci afin de pouvoir répondre aux questions suivantes : que se passe-t-il si les paramètres ne sont connus qu'avec une certaine précision (par exemple si les paramètres ne sont pas exactement représentables, comme par exemple π ou même 0.1 qui ne possèdent pas un développement binaire fini) ? Quel en est l'impact sur la fonction de transfert ? Comment considérer cela pour la recherche de la réalisation *optimale* ?

3.1 Formalisation des LTI paramétrés

Ainsi, nous considèrerons un vecteur $\boldsymbol{\theta}$ de a paramètres réels qui sera quantifié en $\boldsymbol{\theta}^\dagger \triangleq \boldsymbol{\theta} + \boldsymbol{\Delta\theta}$. Ces paramètres étant inconnus pour le moment, il est juste possible de dire que les $\boldsymbol{\Delta\theta}_k$ sont des variables aléatoires, indépendantes, centrées et uniformément réparties dans l'intervalle $-2^{-\gamma_{\boldsymbol{\theta}_k}-1} \leqslant \boldsymbol{\Delta\theta}_k < 2^{-\gamma_{\boldsymbol{\theta}_k}-1}$, où $\gamma_{\boldsymbol{\theta}_k}$ se calcule selon l'équation (14) à partir de la représentation virgule fixe.

Puis $\boldsymbol{Z}$ sera calculé à partir de $\boldsymbol{\theta}^\dagger$, et ensuite quantifié pour être représenté en virgule fixe. On supposera que le calcul $\boldsymbol{Z}(\boldsymbol{\theta}^\dagger)$ se fera avec un arrondi exact pour ne considérer que la quantification finale du résultat (par exemple en utilisant une précision suffisante pour que le résultat du calcul suivi de la quantification aboutisse au même résultat que le résultat exact suivi de la quantification). Ainsi, $\boldsymbol{Z}^\dagger \triangleq \boldsymbol{Z}(\boldsymbol{\theta}^\dagger) + \boldsymbol{\Delta Z}$, et la quantification sur $\boldsymbol{\Delta Z}$ vérifie les mêmes propriétés que celles énoncées pour l'équation (17).

3.2 Nouvelle expression de la mesure NEEFT

Proposition 3. Dans le contexte de systèmes LTI paramètrés, la NEEFT, définie à l'équation (18) est donnée par

$$\sigma^2_{\Delta h} = \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \boldsymbol{\Xi}_{\boldsymbol{Z}} \right\|_2^2 + \sum_{k=1}^{a} \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \boldsymbol{\Theta}_k \right\|_2^2 \tag{26}$$

où $\boldsymbol{\Xi}_{\boldsymbol{Z}}$ a été défini à l'équation (20) et $\boldsymbol{\Theta}_k \in \mathbb{R}^{(n+1)\times(n+1)}$ est donné par

$$\boldsymbol{\Theta}_k \triangleq \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{\theta}_k} \frac{2^{-\beta_{\boldsymbol{\theta}_k}+1}}{\sqrt{3}} \left\lfloor \boldsymbol{\theta}_k \right\rceil_2 (\boldsymbol{\delta_\theta})_k \,. \tag{27}$$

Le 1^{er} terme de l'équation (26) correspond à la quantification des coefficients $\boldsymbol{Z}$ dûe au calcul avec arrondi exact de $\boldsymbol{Z}(\boldsymbol{\theta})$, tandis que le 2nd terme correspond à la quantification de chaque paramètre $\boldsymbol{\theta}_k$.

Preuve 2. Une approximation au 1^{er} ordre nous donne

$$\boldsymbol{Z}^\dagger = \boldsymbol{Z} + \sum_k \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{\theta}_k} \boldsymbol{\Delta\theta}_k + \boldsymbol{\Delta Z}, \tag{28}$$

et puisque h est modifié en $h^\dagger = h + \Delta h$, on a (au 1^{er} ordre) :

$$\Delta h(z) = \sum_{i,j} \frac{\partial h}{\partial \boldsymbol{Z}_{i,j}}(z) \left(\boldsymbol{\Delta Z}_{i,j} + \sum_k \frac{\partial \boldsymbol{Z}_{i,j}}{\partial \boldsymbol{\theta}_k} \boldsymbol{\Delta\theta}_k \right) \tag{29}$$

Dans le cas où les arrondis sur $\boldsymbol{\theta}$ et sur le calcul de $\boldsymbol{Z}$ sont des arrondis au plus proche, les $\boldsymbol{\Delta Z}_{i,j}$ et $\boldsymbol{\Delta\theta}_k$ sont des variables aléatoires indépendantes, centrées [1] et uniformément distribuées sur leur intervalle.
En calculant $E\{|\Delta h(e^{j\omega})|^2\}$, on peut écrire :

$$\sigma^2_{\Delta h} = \sum_{i,j} \left\| \frac{\partial h}{\partial \boldsymbol{Z}_{i,j}} \right\|_2^2 \left(\sigma^2_{\boldsymbol{Z}_{i,j}} + \sum_k \left(\frac{\partial \boldsymbol{Z}_{i,j}}{\partial \boldsymbol{\theta}_k} \right) \sigma^2_{\boldsymbol{\Delta\theta}_k} \right) \tag{30}$$

Avec $\sigma^2_{\boldsymbol{\Delta\theta}_k} = \frac{2^{-2\gamma_{\boldsymbol{\theta}_k}}}{12} \delta_{\boldsymbol{\theta}_k}$, où $\delta_{\boldsymbol{\theta}_k}$ est défini de manière similaire à l'équation (15), et $\boldsymbol{\Theta}_k$ défini selon (27), on retrouve l'équation (26). ■

Remarque 5. Dans le cas où les largeurs de bits des coefficients et des paramètres sont toutes identiques (cas classique d'une implantation logicielle), il est possible de normaliser la mesure en supprimant les termes $\frac{2^{-\beta+1}}{\sqrt{3}}$, où β est la largeur commune. On pourra donc utiliser :

$$\tilde{\sigma}^2_{\Delta h} = \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \lfloor \boldsymbol{Z} \rceil_2 \times \delta_{\boldsymbol{Z}} \right\|_2^2 + \sum_{k=1}^{a} \left\| \frac{\partial h}{\partial \boldsymbol{Z}} \times \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{\theta}_k} \lfloor \boldsymbol{\theta}_k \rceil_2 \, \delta_{\boldsymbol{\theta}_k} \right\|_2^2 . \tag{31}$$

Bien que non détaillé ici, cette mesure peut s'étendre au cas où h est MIMO (voir Hilaire et al. (2007) pour la mesure M_{L_2} en MIMO).

1. Cela n'est plus vrai si l'arrondi effectué sur $\boldsymbol{Z}$ est une troncature. Dans ce cas, les v.a. $\boldsymbol{\Delta Z}_{i,j}$ ne sont plus centrés et l'équation (30) n'est plus valide car les termes croisés $E\{\boldsymbol{\Delta Z}_{ij} \boldsymbol{\Delta\theta}_k\}$ ne sont plus nuls.

3.3 Exemple de régulateur/filtre paramétré

On considère la fonction de transfert continue suivante
$$h(s) = \frac{g}{s^2 + 2\xi\omega_n s + \omega_n^2} \tag{32}$$
qui dépend de 3 paramètres : un gain g, le facteur d'amortissement ξ et la pulsation naturelle ω_n (le gain statique est g/ω_n^2).

La transformée bilatérale conduit à la fonction de transfert discrète
$$h(z) = \frac{b_0 z^2 + b_1 z + b_2}{a_0 z^2 + a_1 z + a_2}$$
avec $\mathbf{b_0} = gT^2$, $\mathbf{b_1} = 2gT^2$, $\mathbf{b_2} = gT^2$, $\mathbf{a_0} = 4\xi\omega_n T + \omega_n^2 T^2 + 4$, $\mathbf{a_1} = 2\omega_n^2 T^2 - 8$ et $\mathbf{a_2} = \omega_n^2 T^2 - 4\xi\omega_n T + 4$.

On remarque donc que chaque coefficient est fonction des quatre paramètres g, ξ, ω_n et T.
Ensuite, pour l'implantation, il est possible par exemple d'utiliser la forme canonique commandable :
$$\begin{cases} \boldsymbol{x}(k+1) = \begin{pmatrix} -\frac{a_1}{a_0} & -\frac{a_2}{a_0} \\ 1 & 0 \end{pmatrix} \boldsymbol{x}(k) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(k) \\ y(k) = \begin{pmatrix} \frac{b_1 a_0 - b_0 a_1}{a_0^2} & \frac{b_2 a_0 - b_0 a_2}{a_0^2} \end{pmatrix} \boldsymbol{x}(k) + \frac{b_0}{a_0} u(k) \end{cases} \tag{33}$$

On a donc ici $\boldsymbol{\theta} = \begin{pmatrix} g \\ \xi \\ \omega_n \\ T \end{pmatrix}$ et $\boldsymbol{Z}$ est donné par l'équation (34). On peut donc, par le biais d'outils de calcul formel, en déduire $\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{\theta}_k}$ pour $k \in \{1,..,4\}$ et ainsi calculer $\sigma^2_{\Delta h}$. Par manque de place, seul $\frac{\partial \boldsymbol{Z}}{\partial g}$ est donné par l'équation (35).

4. PROBLÈME DE RÉALISATIONS OPTIMALES

4.1 Cas où une seule réalisation est considérée

Lorsqu'une seule réalisation est considérée, c'est-à-dire lorsqu'une seule valeur du vecteur de paramètres θ est considérée, il est possible de rechercher quelle réalisation présente la meilleure *résilience*, c'est-à-dire celle pour laquelle l'impact de la quantification des coefficients et paramètres sera le plus faible.

Pour cela, on considère de nouveau le changement de base décrit à l'équation (24). Partant de la réalisation $\boldsymbol{Z}_0$ on obtient une réalisation équivalente $\boldsymbol{Z}_1$ avec une matrice inversible $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n\times n}$:
$$\boldsymbol{Z}_1 = \boldsymbol{\mathcal{T}}^{-1} \boldsymbol{Z}_0 \boldsymbol{\mathcal{T}} \quad \text{avec} \quad \boldsymbol{\mathcal{T}} = \begin{pmatrix} \boldsymbol{\mathcal{U}} & \\ & 1 \end{pmatrix} \tag{36}$$

On peut remarquer que $\frac{\partial h}{\partial \boldsymbol{Z}}$, $\boldsymbol{\Xi_Z}$ et $\boldsymbol{\Theta}_k$ sont dépendants de $\boldsymbol{\mathcal{U}}$, avec
$$\left.\frac{\partial h}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_1} = \boldsymbol{\mathcal{T}}^\top \left.\frac{\partial h}{\partial \boldsymbol{Z}}\right|_{\boldsymbol{Z}_0} \boldsymbol{\mathcal{T}}^{-1}, \tag{37}$$
et
$$\left.\boldsymbol{\Theta}_k\right|_{\boldsymbol{Z}_1} = \boldsymbol{\mathcal{T}}^{-1} \left.\boldsymbol{\Theta}_k\right|_{\boldsymbol{Z}_0} \boldsymbol{\mathcal{T}}. \tag{38}$$

Le problème de réalisation optimale consiste donc à trouver $\boldsymbol{\mathcal{U}}$ qui minimise l'erreur de la fonction de transfert
$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ inversible}}{\arg\min}\ \sigma^2_{\Delta h}(\boldsymbol{\mathcal{U}}). \tag{39}$$

Proposition 4. Tout comme l'espérance de l'erreur de fonction de transfert pour les systèmes LTI, la mesure NEEFT pour les systèmes LTI paramétrés est invariant par un changement en puissance de 2, c'est-à-dire avec $\boldsymbol{\mathcal{U}}$ tel que $\boldsymbol{\mathcal{U}}_{ii} = 2^{p_i}$ et $p_i \in \mathbb{Z}$.

Preuve 3. On trouvera dans (Hilaire, 2009) la démonstration de
$$\left.\left(\frac{\partial h}{\partial \boldsymbol{Z}_{ij}} \Xi_{ij}\right)\right|_{Z_1} = \left.\left(\frac{\partial h}{\partial \boldsymbol{Z}_{ij}} \Xi_{ij}\right)\right|_{Z_0} \tag{40}$$
pour un tel $\boldsymbol{\mathcal{U}}$, basée sur les propriétés de l'opérateur $\lfloor . \rceil_2$. De plus, l'équation (38) permet de prouver l'invariance de l'erreur due à la quantification des $\boldsymbol{\theta}_k$ pour un changement d'échelle en puissance de 2.

Cela nous permet de *normaliser* les changements de base possibles, en ne considérant que ceux qui amènent dans un état particulier. La mise à l'échelle L_2 (L_2-scaling), qui consiste à imposer la norme L_2 de la fonction de transfert de l'entrée à chaque état a été considérée dans (Hilaire, 2009). Elle consiste à faire combiner tout changement de base $\boldsymbol{\mathcal{U}}$ par un autre $\boldsymbol{\mathcal{V}}$ qui ramène les termes diagonaux du Gramien de commandabillité dans l'intervalle $[1, 4[$.

Ainsi, le problème de réalisation optimale peut-il être résolu en considérant le problème d'optimisation non contraint suivant :
$$\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}} \text{ inversible}}{\arg\min}\ \sigma^2_{\Delta h}(\boldsymbol{\mathcal{U}\mathcal{V}}), \tag{41}$$
où $\boldsymbol{\mathcal{V}}$ est une matrice diagonale telle que
$$\boldsymbol{\mathcal{V}}_{ii} = \left\lfloor \sqrt{(\boldsymbol{\mathcal{U}}^{-1} \boldsymbol{W}_c \boldsymbol{\mathcal{U}}^{-\top})_{ii}} \right\rceil_2 \tag{42}$$
et $\boldsymbol{W}_c$ est le Gramien de commandabilité du système, défini en (8).

Enfin, ce problème d'optimisation peut être abordé par des algorithmes d'optimisation globale, telle qu'un recuit simulé adaptatif (Ingber, 1996; Chen et Luk, 1999). Une méthode basée sur les gradients, comme l'algorithme de Newton amène vers un minimal local, qui, en pratique, s'avère proche de celui trouvé avec les méthodes d'optimisation globale.

Cas où une famille de réalisations est considérée

Il est tout de même plus fréquent que les paramètres $\boldsymbol{\theta}$ ne soient pas connus exactement au moment de l'implantation du régulateur/filtre, et que les concepteurs se laissent une marge de manœuvre pour régler et retoucher les paramètres du régulateur après-coup. Pour notre exemple, on pourrait n'avoir comme information que $\xi \in [0.1, 0.2]$ et non une valeur fixée.

On suppose alors connu l'ensemble $\boldsymbol{\Omega_\theta}$ des valeurs possibles pour $\boldsymbol{\theta}$: ce sera dans la plupart des cas une boîte de $\mathbb{R}^a$ car les paramètres seront définis par un intervalle $[\underline{\theta_k}; \overline{\theta_k}]$, et sinon ce sera un sous-ensemble compact de $\mathbb{R}^a$.

Une mesure intéressante concerne alors la pire mesure NEEFT lors de l'implantation en virgule fixe pour l'ensemble des paramètres possibles :

Définition 4. (Erreur de pire cas). On note $\sigma^2_{\Delta h, \boldsymbol{\Omega_\theta}}$ la norme de l'espérance de l'erreur de la fonction de transfert maximum pour l'ensemble des valeurs possibles des paramètres :

$$
\boldsymbol{Z} = \begin{pmatrix} -\dfrac{2\left(T^2\omega_n^2-4\right)}{T^2\omega_n^2+4T\omega_n\xi+4} & -\dfrac{T^2\omega_n^2-4T\omega_n\xi+4}{T^2\omega_n^2+4T\omega_n\xi+4} & 1 \\ 1 & 0 & 0 \\ 2T^2g-\dfrac{2\left(T^2\omega_n^2-4\right)T^2g}{\left(T^2\omega_n^2+4\,T\omega_n\xi+4\right)^2} & T^2g-\dfrac{2\left(T^2\omega_n^2-4\right)T^2g}{\left(T^2\omega_n^2+4\,T\omega_n\xi+4\right)^2} & \dfrac{T^2g}{T^2\omega_n^2+4T\omega_n\xi+4} \end{pmatrix} \tag{34}
$$

$$
\frac{\partial \boldsymbol{Z}}{\partial g} = \begin{pmatrix} \dfrac{8\left(T\omega_n-2\right)\left(T\omega_n+2\right)T\omega_n}{\left(T^2\omega_n^2+4T\omega_n\xi+4\right)^2} & \dfrac{8\left(T^2\omega_n^2+4\right)T\omega_n}{\left(T^2\omega_n^2+4T\omega_n\xi+4\right)^2} & 0 \\ 0 & 0 & 0 \\ \dfrac{16\left(T\omega_n-2\right)\left(T\omega_n+2\right)T^3g\omega_n}{\left(T^2\omega_n^2+4T\omega_n\xi+4\right)^3} & \dfrac{16\left(T\omega_n-2\right)\left(T\omega_n+2\right)T^3g\omega_n}{\left(T^2\omega_n^2+4T\omega_n\xi+4\right)^3} & \dfrac{-4T^3g\omega_n}{\left(T^2\omega_n^2+4T\omega_n\xi+4\right)^2} \end{pmatrix} \tag{35}
$$

$$
\sigma^2_{\Delta h,\boldsymbol{\Omega_\theta}}(\boldsymbol{\mathcal{U}}) \triangleq \max_{\boldsymbol{\theta}\in\boldsymbol{\Omega_\theta}} \sigma^2_{\Delta h}(\boldsymbol{\theta},\boldsymbol{\mathcal{U}}) \tag{43}
$$

Et le problème de réalisation optimale consiste donc à trouver

$$
\boldsymbol{\mathcal{U}}_{opt} = \underset{\boldsymbol{\mathcal{U}}\text{ inversible}}{\arg\min}\ \sigma^2_{\Delta h,\boldsymbol{\Omega_\theta}}(\boldsymbol{\mathcal{U}}) \tag{44}
$$

Un tel problème d'optimisation n'est pas évident à résoudre, et une 1$^{\text{ère}}$ solution est de discrétiser l'ensemble $\boldsymbol{\Omega_\theta}$ en un ensemble discret $\boldsymbol{\Psi_\theta}$ de N points. Dans le cas où chaque $\boldsymbol{\theta}$ est donné par un intervalle, on pourra discrétiser chaque intervalle en r points pour obtenir $N = r^a$ points de $\boldsymbol{\Psi_\theta}$. Cela n'est bien sûr réalisable que si N est faible.

Remarque 6. Si l'intervalle $[\underline{\theta_k};\overline{\theta_k}]$ dans lequel peut évoluer $\boldsymbol{\theta}_k$ est trop grand (c'est-à-dire si toutes les valeurs n'ont pas la même représentation virgule fixe), alors seule la position de la virgule de la plus grande valeur prise (en valeur absolue) par $\boldsymbol{\theta}_k$ doit être considérée pour le calcul de $\boldsymbol{\Theta}_k$:

$$
\boldsymbol{\Theta}_k \triangleq \frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{\theta}_k} \frac{2^{-\beta_{\boldsymbol{\theta}_k}+1}}{\sqrt{3}} \left\lfloor \overset{\max}{\boldsymbol{\theta}_k} \right\rfloor_2 (\boldsymbol{\delta_\theta})_k \,. \tag{45}
$$

5. CONCLUSION

Une formalisation du problème de recherche de réalisations de systèmes LTI paramétrés, résilientes vis-à-vis d'une implantation en virgule fixe, a été proposée. Une première voie pour résoudre ce problème a été dessinée.

Nous avons considéré comme mesure de résilience, la norme H_2 de l'espérance de l'écart, entre la fonction de transfert nominale d'une part et la fonction de transfert résultant du choix de la réalisation et du niveau de précision retenus pour l'implantation d'autre part. Ceci a pu être formalisé en définissant précisément l'impact de la quantification des paramètres externes sur les coefficients qui en dépendent. Une voie pour obtenir une réalisation paramétrée, qui soit résiliente quelque soit la valeur des paramètres dans leur intervalle d'admissibilité, a été tracée. Elle mérite d'être approfondie, soit en proposant des réalisations paramétrées intrinsèquement résilientes (quoique possiblement sous optimales en regard de la mesure proposée), soit en proposant une heuristique permettant de traiter efficacement le problème de minimisation de l'erreur du pire cas paramétrique.

REMERCIEMENTS

Ce travail a été financé en partie par le CNRS à travers le projet PEPS *ReSyst* de l'INSIS.

RÉFÉRENCES

Chen, S. et Luk, B. (1999). Adaptive Simulated Annealing for optimization in signal processing applications. *Signal Processing*, 79, 117–128.

Feng, Y., Chevrel, P., et Hilaire, T. (2011). Generalised modal realisation as a practical and efficient tool for fwl implementation. *International Journal of Control*, 84(1), 66–77.

Gevers, M. et Li, G. (1993). *Parametrizations in Control, Estimation and Filtering Probems*. Springer-Verlag.

Hilaire, T. (2009). On the transfer function error of state-space filters in fixed-point context. *IEEE Trans. on Circuits & Systems II*, 56(12), 936–940.

Hilaire, T. et Chevrel, P. (2011). Sensitivity-based pole and input-output errors of linear filters as indicators of the implementation deterioration in fixed-point context. *EURASIP Journal on Advances in Signal Processing*, special issue on Quantization of VLSI Digital Signal Processing Systems.

Hilaire, T., Chevrel, P., et Whidborne, J. (2007). A unifying framework for finite wordlength realizations. *IEEE Trans. on Circuits and Systems*, 8(54), 1765–1774.

Hinamoto, T., Yokoyama, S., Inoue, T., Zeng, W., et Lu, W. (2002). Analysis and minimization of L_2-sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability gramians. In *IEEE Transactions on Circuits and Systems, Fundamental Theory and Applications*, volume 49.

Ingber, L. (1996). Adaptive simulated annealing (ASA) : Lessons learned. *Control and Cybernetics*, 25(1), 33–54.

Li, G., Chu, J., et Wu, J. (2007). A matrix factorization-based structure for digital filters. *Signal Processing, IEEE Transactions on*, 55(10), 5108–5112.

Li, G. et Zhao, Z. (2004). On the generalized DFIIt structure and its state-space realization in digital filter implementation. *IEEE Trans. on Circuits and Systems*, 51(4), 769–778.

Middleton, R. et Goodwin, G. (1986). Improved finite word length characteristics in digital control using delta operators. *IEEE Transactions on Automatic Control*, 31(11), 1015–1021.

Sripad, A. et Snyder, D. (1977). A necessary and sufficient condition for quantization error to be uniform and white. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, volume 5, 442–448.

Yamaki, S., Abe, M., et Kawamata, M. (2011). Derivation of the class of digital filters with all second-order modes equal. *IEEE Transactions on Signal Processing*, 59(11), 5236–5242.