Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

Mémoire doctorant 1ère année

2012 -2013

Nom - Prénom	**Ait El Faqir - Marouane**
Titre de la thèse	**Prédiction de la structure de contrôle des bactéries par optimisation robuste.**
Directeurs de thèse	**Gérard Scorletti & Vincent Fromion**
Co- encadrants	**Anne Goelzer & Julien Huillery**
Dpt. de rattachement	• **Laboratoire Ampère (CNRS UMR 5005) : Méthodes pour l'ingénierie des systèmes.** • **INRA, Jouy-en-Josas : Mathématiques, Informatique et Génome.**
Date début des travaux	**01/10/2012**
Type de financement	**CNRS : BDI**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

Rapport de première année de thèse

Marouane Ait El Faqir

marouane.ait-el-faqir@ec-lyon.fr

05 juillet 2013

Résumé

L'analyse systémique des systèmes biologiques a fait émerger un nouveau cadre de recherche en biologie des systèmes, permettant de mieux comprendre les principes généraux de fonctionnement de la cellule. Ce cadre théorique est basé sur la gestion parcimonieuse des ressources : la méthode *Resource Balance Analysis* (RBA) [26]. L'une des contributions de ce cadre théorique est d'appliquer les outils de l'optimisation convexe pour formaliser le comportement de la cellule. Globalement, l'objectif de cette thèse est de prendre en compte de l'aspect stochastique, intrinsèque aux cellules vivantes, lors du processus de la modélisation tout en gardant l'avantage de l'efficacité de résolution numérique dont dispose la méthode RBA.

L'objectif de ce document est d'offrir un aperçu aussi large que possible sur l'optimisation « moderne » [4, 9, 30] afin d'isoler et de développer les techniques nécessaires permettant à la fois de développer des méthodes d'analyse des systèmes de très grande dimension (typique aux applications biologiques), les analyser théoriquement (sur papier) sur la base de la formalisation en terme de problème d'optimisation convexe, afin de les résoudre d'une façon efficace numériquement. Au sens de Arkadi Nemirovski, l'un des contributeurs pionniers en la matière, le terme « moderne » ici, renvoie aux problèmes d'optimisation convexe bien structurés à savoir

- programmation linéaire ;
- programmation quadratique ;
- programmation semi-définie.

Ce sont les classes de problèmes les plus répandues en applications et les plus touchées par les récentes avancées en optimisation.

Je m'intéresse aux problèmes d'optimisation soumis aux incertitudes et à leur complexité algorithmique. J'explorerai les différentes approches existant dans la littérature visant à reformuler ces problèmes et à leur proposer des méthodes de résolution numérique qui sont plus ou moins efficaces.

Dans un premier temps, je propose un exemple numérique pour illustrer les points soulignés plus haut. Ensuite, je passerai à une présentation des différents aspects techniques soulevés par cet exemple.

Je renvoie le lecteur intéressé aux annexes pour plus de détails et de notes bibliographiques.

Mots-clés : Optimisation robuste, optimisation stochastique, optimisation cônique, approximation (relaxation), complexité algorithmique, simulation Monte Carlo, biologie des systèmes, dualité.

Table des matières

1 Introduction Générale

1.1 Enjeux scientifiques

Les cellules sont des systèmes organisés, composés d'un grand nombre de sous-systèmes interagissant ensemble et partageant des ressources communes : en effet, les travaux d'Anne Goelzer dans [25] ont permis l'identification de sous-systèmes (au sens de l'automatique) ou modules régis par des équations différentielles et correspondant aux différents processus cellulaires sur la base de leurs structures de contrôle permettant de répondre à des perturbations variées dans un contexte de compétition perpétuel avec les organismes vivants. Ces travaux ont permis l'émergence d'un parallèle entre biologie et automatique. L'idée ambitieuse est l'aspect d'ingénierie inverse pour dégager des principes de fonctionnement généraux de la cellule.

Au regard de la grande sophistication des systèmes biologiques, une analyse systémique de la cellule s'avère être une voie permettant un bon compromis entre modèle réaliste et modèle à complexité maîtrisée. De ce point de vue, la formulation du partage des ressources, principalement les protéines, entre les différents processus cellulaires a permis d'identifier certaines contraintes structurelles agissant sur la bactérie [26]. A travers des hypothèses réalistes, ces contraintes structurelles ont été formalisées en un problème d'optimisation quasi convexe qui peut être résolu en le ramenant de façon équivalente à la résolution d'un problème d'optimisation linéaire. Les deux propriétés de convexité et de linéarité ont permis effectivement de mettre en œuvre efficacement la méthode RBA sur des applications biologiques de grande dimension. Ceci est dû d'un côté, du fait que les problèmes d'optimisation convexes forment la classe la plus générale de problèmes d'optimisation qui admettent des algorithmes de résolution en temps polynomial en fonction de leurs tailles. D'un autre côté, les problèmes d'optimisation linéaires ont la propriété exclusive d'être efficacement résolus en très grande dimension [15, 30, 9]. Cette représentation « simplifiée » de la cellule est donc de ce point de vue, un bon choix au regard de la très grande dimension des systèmes biologiques (des milliers de variables de décision). Cet intérêt tout théorique, est renforcé par le fait que l'approche s'avère pertinente vis à vis de la biologie. Cela laisse aussi une grande marge de manœuvre pour complexifier le modèle. Ce cadre théorique est appelé méthode *Resource Balance Analysis* (*RBA*).

L'un des enjeux principaux (en commun avec la thèse d'Anne Goelzer [25]) de cette thèse est de souligner l'intérêt de développer des méthodes et des outils basés sur l'optimisation convexe pour dégager les principes de fonctionnement des « systèmes vivants ». Dans un très grand nombre des cas, et spécifiquement dans le champ de l'automatique, les méthodes d'analyse reposent sur la formulation du problème comme un problème d'optimisation bien posé. Dans ce contexte, dans la majorité des cas pratiques, le processus de la modélisation, qui coûte beaucoup de temps et d'efforts, s'effectue avant et en totale indépendance de l'étape de résolution. Le fait principal que doit savoir toute personne manipulant des modèles d'optimisation est que, en général, *les problèmes d'optimisation*

ne peuvent se résoudre [30], ce qui peut mettre en cause la validité de cette démarche. Compte tenu de ces considérations, il faut lors de la formulation et donc de la modélisation du problème, faire des compromis nécessaires afin de se ramener à des problèmes d'optimisation convexe. Cela passe par la compréhension claire du système étudié, qui va permettre, au moment de la formalisation du problème, de faire les meilleurs compromis, c'est à dire ceux qui vont permettre de le formaliser de telle sorte qu'on puisse le résoudre pratiquement.

1.2 Les questions de recherche

L'analyse des données produites dans le cadre du projet BaSynThec a permis de valider la pertinence de la méthode RBA et ses capacités de prédiction. Les principales différences entre les prédictions RBA et les données concernent principalement (a) l'apparente surdimensionnement de certaines voies métaboliques ; (b) l'absence de régulation génétique sur certaines voies métaboliques et spécifiquement sur celles comptant peu d'enzymes.

Au regard des contraintes actuellement considérées dans la méthode RBA, et de quelques travaux préliminaires, il est suspecté que certaines de ces différences soient en partie liées à la non prise en compte du caractère stochastique de la production des protéines à l'échelle de l'individu (intrinsèque au mode de production des protéines dans la bactérie).

En effet, les contraintes prises en comptent dans la méthode RBA sont actuellement « populationnelles » : les concentrations des protéines, des ribosomes, *etc*... manipulées correspondent toutes aux concentrations moyennes présentes à l'échelle de la population. Or à l'échelle de l'individu (où les contraintes que nous considérons agissent), la production des protéines est stochastique, et la concentration des protéines est plus sûrement décrite par des distributions de probabilité que par leur valeur moyenne. Au regard de la forte variabilité de la concentration des protéines, et de la nature des contraintes agissant sur la bactérie, le problème d'optimisation attaché au RBA se reformule à l'échelle de l'individu non plus comme un problème d'optimisation déterministe mais comme un problème d'optimisation stochastique.

Il semblerait donc plus judicieux de maximiser l'espérance du taux de croissance d'une population d'individus sujet à des fluctuations stochastiques. Même si les contraintes que nous considérons sont spécifiques, le problème que nous souhaitons résoudre est un problème dit à deux pas (*two stages problem* [19, 3, 14, 28, 27, 39]).

Des résultats préliminaires (obtenus sur un modèle très simplifié) indiquent au regard des contraintes attachées au RBA, que la stochasticité des concentrations des protéines tend à augmenter les valeurs moyennes des protéines impliquées dans le réseau métabolique ayant une valeur moyenne faible dans la version populationnelle.

Le premier objectif de ce travail de thèse consistera donc à intégrer ces aspects stochastiques dans le problème d'optimisation, et de traiter de façon efficace les questions portant sur le très grand nombre de variables stochastiques à intégrer. En effet, et comme nous allons le présenter dans la suite, les problèmes *Two-stage* sont difficiles à résoudre en général et la question qui nous est posée est bien celle de savoir comment intégrer dans le cadre RBA ces aspects stochastiques sans perdre pour autant les aspects touchant à leur résolution. A ce titre, on s'appuiera en particulier sur les récentes avancées faites dans ce cadre, voir [4].

1.3 Présentation et organisation du document

Durant la première année de thèse, j'ai étudié les outils de l'optimisation sous incertitude : ce cadre de recherche, actuellement en plein essor, touche plusieurs communautés scientifiques (mathématique (optimisation) et automatique). Dans ce contexte, deux axes principaux émergent clairement : l'optimisation robuste et l'optimisation stochastique. J'ai ainsi repéré leurs différentes approches ainsi que leurs contributions théoriques principales, en particulier celles ayant un grand intérêt pratique (voir annexes). Ceci a débouché sur une note de synthèse qui a pour objectif de positionner entre elles les différentes tendances existant dans la littérature pour résoudre notre problème spécifique biologique (problèmes d'optimisation convexe de grandes dimensions). On soulignera, qu'à ce jour, une telle comparaison des deux approches d'optimisation est inédite.

Ce document est organisé de la façon suivante : dans la section 2 je propose un exemple numérique type illustrant une spécificité du problème de gestion de ressources chez les bactéries. En intégrant l'aspect stochastique dans ce problème, je l'ai reformulé et résolu selon l'approche robuste (voir section 3) et selon l'approche stochastique (voir section 4). Pour les deux approches, j'ai spécifiquement souligné les aspects liés à la complexité algorithmique et aux techniques d'approximation associées aux problèmes d'optimisation. Ces aspects sont en effet critiques pour aborder les problèmes de grande dimension.

2 Exemple introductif

Je propose dans cette section, un exemple illustratif de notre problème, que je vais d'abord résoudre dans l'hypothèse de l'absence de perturbation (modèle nominal), ensuite en tenant en compte de la perturbation ; ceci passera par deux nouvelles formulations du modèle nominal. Cet exemple trouve ses origines dans l'un des aspects rencontrés dans le problème de gestion de ressources dans une bactérie. L'objectif est d'introduire, à travers cet exemple, le paradigme de l'optimisation, le concept des données affectées par des incertitudes, la modélisation des incertitudes, illustrer quelques résultats théoriques

typiques liés à la résolution des problèmes d'optimisation issus des différents modèles tenant compte de la stochasticité du problème.

2.1 Cas nominal : non prise en compte des incertitudes

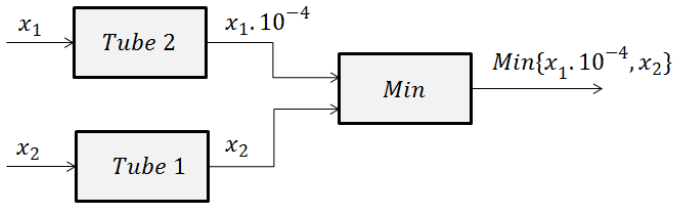


FIGURE 1 – Schéma du problème de production.

L'exemple que je vais traiter ici correspond à une spécificité du problème de gestion de ressources dans la bactérie : la bactérie doit produire des produits en abondances très différentes. Ici, on considère donc que l'un en grande quantité x_1 (acides aminés, énergie, etc ...) et l'autre en faible quantité x_2 (les vitamines, etc ...). x_1 et x_2 correspondent à des flux de matière mis en jeu par un sous-système cellulaire pour la synthèse de certaines composantes nécessaires à certains processus. Par ailleurs, on fait l'hypothèse que pour produire les flux x_1 et x_2, on doit utiliser des enzymes dont la concentration est proportionnelle aux flux : $[tube\ 1] \backsimeq x_1$,$[tube\ 2] \backsimeq x_2$. On ne dispose que d'une ressource limitée d'enzymes c'est à dire :

$$[tube\ 1] + [tube\ 2] = 10^4 + 1, \tag{1}$$

et je me propose de maximiser le produit suivant :

$$Q(x_2) := \min\{10^{-4}x_1, x_2\}. \tag{2}$$

Par abus de notation, le problème se formule de la façon suivante :

$$\max_{x_2} Q(x_2), \tag{3}$$

avec

$$\begin{cases} x_1 + x_2 = 10^4 + 1, \\ x_2 \geqslant 0, x_1 \geqslant 0. \end{cases}$$

Ce problème peut se résoudre géométriquement de la façon suivante : on trace les deux droites $y = 10^{-4}x_1 = 1 + 10^{-4} - 10^{-4}x_2$ et $y' = x_2$ par rapport à x_2, puis on détermine géométriquement la fonction $Q(x_2) = \min\{y, y'\}$ et enfin on récupère son maximum (la valeur optimale du problème 3) ainsi que la solution optimale correspondante voir figure 2. La valeur optimale Q^* du problème nominal (3) est égale à 1, atteinte en $x_2^* = 1$,

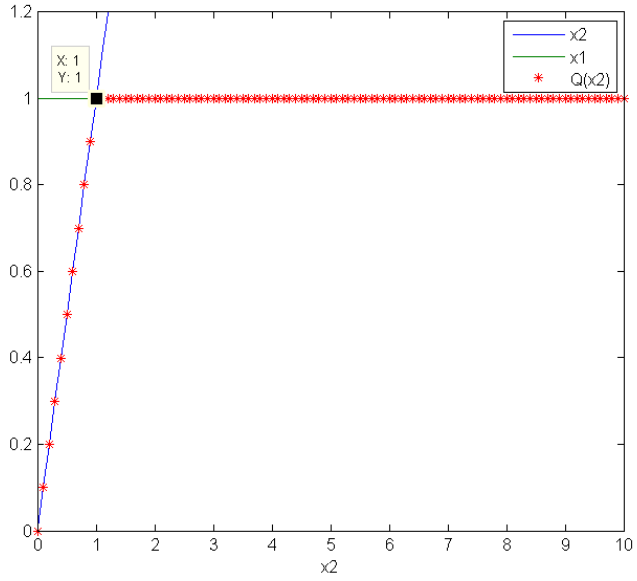


FIGURE 2 – Illustration de la résolution géométrique du problème nominal.

voir figure 2.

En réalité, dans la plupart des applications, et en particulier pour notre problème, on ne peut pas ignorer la possibilité que des perturbations (incertitudes), ce qui est intrinsèque dans la majorité des cas pratiques, affectent les données de notre problème nominal. Ceci a de graves conséquences sur la « qualité » de notre solution nominale qui peut devenir non-faisable, i.e. ne respecte plus les contraintes du problème d'optimisation perturbé, ce qui rend cette solution sans aucune utilité du point de vue pratique. Ceci peut justifier le vrai besoin de techniques capables de produire des solutions fiables : insensibles aux incertitudes en l'occurrence.

2.2 Modélisation des incertitudes

Pour illustrer la sensibilité du problème (3) aux incertitudes, on suppose, par exemple, que lors de la mise en œuvre de la solution nominale, le *tube* 2 est réalisé avec erreur, c'est à dire que la quantité x_2 est incertaine. Ce qui peut être modélisé par :

$$x_2 = \bar{x}_2 + \xi,$$

où $\bar{x}_2$ est la valeur nominale de x_2 et ξ est l'erreur d'implémentation. Cette erreur d'implémentation peut être modélisée par une variable aléatoire de loi de probabilité connue (loi normale centrée réduite par exemple). Dans le cas d'une réalisation de la variable aléatoire $\tilde{\xi}$ égale à -1 et puisque la solution nominale est égale à 1, on aura une valeur

optimale $Q^* = 0$, ce qui correspond à une production nulle.
Remarque : Deux cas peuvent se poser alors :

- on ne résout qu'une seule fois le problème d'optimisation. Dans ce cas, il est nécessaire de prendre en compte le « pire cas » pour éviter la production nulle voire négative (voir paragraphe 3.1) ;
- on résout le problème d'optimisation « plusieurs fois » : ce qui peut correspondre à réaliser des actions plusieurs fois ou en parallèle (comme dans le cas qui nous intéresse où chaque bactérie « réalise » une solution spécifique). Dans ce cas, les considérations stochastiques deviennent pertinentes, on peut souhaiter par exemple que la production globale soit bonne en moyenne et l'occurrence du pire cas dès lors qu'il est rare, n'est pas un problème (voir paragraphe 4.1).

Maintenant, et pour revenir à l'exemple, qu'en est-il pour la sensibilité de x_1 vis-à-vis des erreurs d'implémentations ? Dans le cas nominal et conformément à la contrainte structurelle [1] (1), la valeur optimale est $x_1 = 10^4$. D'une façon similaire à x_2, on suppose que x_1 est soumise à des erreurs d'implémentations qu'on modélise par $x_1 = \bar{x}_1 + \xi$. Pour le pire cas des incertitudes, i.e. $\tilde{\xi} = -1$, on a $x_1 = 10^4 - 1$ et la valeur optimale de (3) $Q(x_2) = \min\{1, 0.9999\} = 1$. On constate que x_2 est plus sensible aux incertitudes que x_1. Il faut noter au final que les erreurs d'implémentations ne sont qu'une source parmi plusieurs sources éventuelles d'incertitudes que l'on va énumérer dans la suite.

Pour résumer nos observations, les deux variables x_1 et x_2 sont liées par la contrainte (1) et on ne perd pas trop à ne supposer que x_2 perturbée. Nous considérons donc dans la suite que nous souhaitons résoudre le problème suivant :

$$\text{" } \max_{\bar{x}_2} \text{"} Q(\bar{x}_2, \xi), \tag{4}$$

avec :

$$\begin{cases} x_1 + x_2 = 10^4 + 1, \\ x_2 = \bar{x}_2 + \xi, \\ x_2 \geqslant 0, x_1 \geqslant 0. \end{cases}$$

Le programme (4) formulé sous cette forme est mal défini du fait que la maximisation de l'objectif n'a aucun sens du moment que la réalisation des incertitudes représentées par ξ est inaccessible lorsque la décision sur $\bar{x}_2$ est prise. En d'autres termes, on n'a aucune idée sur ce qu'on veut réellement maximiser. L'étape préliminaire dans la résolution de ce problème est donc de bien formaliser le problème d'optimisation.

Dans certains cas pratiques, on veut résoudre une seule fois notre problème : malgré la mal-connaissance du comportement des incertitudes (ce qui est normal), on considère que ξ est mal-inconnue mais bornées dans un intervalle bien défini ; et on exige que notre décision $\bar{x}_2$ satisfasse l'ensemble des contraintes quelles que soient les incertitudes

1. Par contrainte structurelle, on entend une contrainte intrinsèque à la nature du problème et qui doit être toujours satisfaite.

appartenant à cet intervalle. C'est l'approche d'optimisation dans le pire cas, ou robuste, selon la terminologie de Nemirovski [4]. Dans d'autres cas, on veut résoudre plusieurs fois notre problème : on assigne une nature stochastique aux incertitudes, i.e. on considère ξ une variable aléatoire avec une loi de probabilité bien définie et on maximise l'espérance de $Q(\bar{x}_2, \xi)$ (qui est une variable aléatoire) par rapport à ξ. Ce qui présente un cas particulier de l'approche d'optimisation dite *Two-stage* [14] que je vais introduire plus loin.

3 Optimisation robuste

3.1 Résolution par approche robuste

Dans le cas de l'approche robuste, le problème homologue [4] du problème (4) est défini comme suit :

$$\max_{\bar{x}_2} \hat{Q}(\bar{x}_2), \tag{5}$$

où $\Xi = [\underline{\xi}, \overline{\xi}], \underline{\xi}, \overline{\xi} \in \mathbb{R}$, $\hat{Q}(\bar{x}_2) := \inf_{\xi \in \Xi} Q(\bar{x}_2, \xi)$; on rappelle que $Q(\bar{x}_2, \xi) := \min\{10^{-4}x_1, x_2\}$ tel que :

$$\begin{cases} x_1 + x_2 = 10^4 + 1, \\ x_2 = \bar{x}_2 + \xi, \\ x_2 \geqslant 0, x_1 \geqslant 0. \end{cases}$$

On appellera ce type de problème *problème homologue robuste.* Ce problème simple peut être également résolu géométriquement : on a

$$\hat{Q}(\bar{x}_2) := \inf_{\xi \in \Xi} \min\{10^{-4}x_1, x_2\} = \min\{(10^4 + 1 - \bar{x}_2 - \overline{\xi}).10^{-4}, \bar{x}_2 + \underline{\xi}\}.$$

Pour la simulation numérique on a pris $\underline{\xi} = -3$ et $\overline{\xi} = 3$ ce qui correspond à 99% des valeurs qui peuvent être prises par la variable aléatoire ξ qu'on a supposée à loi de probabilité normale centrée réduite. Géométriquement, on trace les deux droites $y = (10^4 + 1 - \bar{x}_2 - 3).10^{-4}$ et $y' = \bar{x}_2 - 3$; on détermine la fonction $\hat{Q}(\bar{x}_2) = \min\{y, y'\}$; finalement on relève sur la courbe de $\hat{Q}(\bar{x}_2)$ sa valeur maximale ainsi que l'abscisse auquel est atteinte cette valeur. Les résultats obtenus sont illustrés sur la figure 3. La valeur optimale $\hat{Q}^*$ du problème (5) est égale à 0.9994, atteinte en $\bar{x}^*_{2r} = 4$, voir figure 3. On remarque tout d'abord que, dans le cas de l'approche robuste, la solution obtenue n'est plus la même que celle obtenue avec le modèle nominal (la solution nominale est sensible aux perturbations) ; que le domaine de faisabilité s'est rétréci, la solution $x^*_2 = 1$ du problème nominal (3) n'est plus faisable car il existe un niveau de perturbation qui conduit, si on choisit $\bar{x}_2 = 1$, à $\hat{Q} < 0$. Ceci illustre les remarques de la section

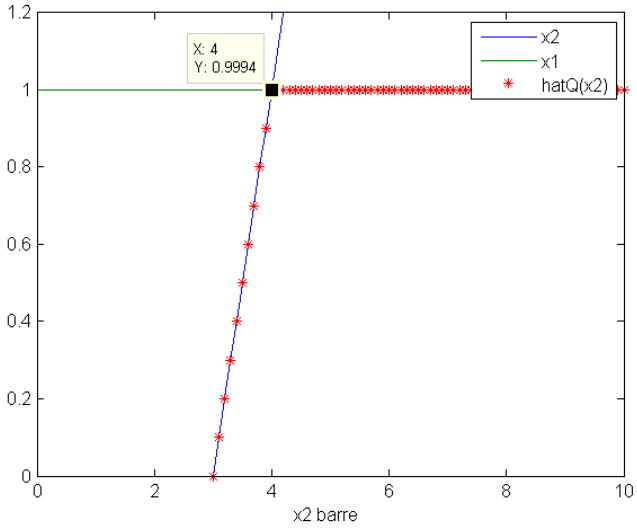


FIGURE 3 – Illustration de la résolution géométrique du problème robuste.

précédente. Cet exemple appartient à la classe de problèmes **d'optimisation linéaire robuste**. Cette classe est la classe de problèmes d'optimisation la plus riche en propriétés « sympathiques » [9] qui fait objet de la section suivante.

3.2 Optimisation linéaire robuste

La programmation linéaire [13, 20] a été introduite dans les années 40s par George B. Dantzig qui a découvert l'algorithme de résolution appelée simplexe. Ces découvertes ont permis d'étendre le cadre théorique connu maintenant sous le nom de la programmation mathématique.

Un problème d'optimisation linéaire est un problème de la forme :

$$\min_x \{c^t x | Ax \leqslant b\}, \text{ [2]} \tag{6}$$

où c, b sont des vecteurs et A est une matrice de dimensions appropriées et $\xi = (c, A, b)$ représente les données du problème.

Dans certains cas, les données du problème (6) sont incertaines (pas connues exactement). Cette incertitude peut être due à :

– Les éléments du vecteur des données ξ peuvent être non disponibles au moment où le problème doit être résolu. (Par exemple, le problème de production d'usine : on ne connaît pas, avant la fin de l'exercice la quantité de la demande, etc ...)

2. Le symbole $(.)^t$ désigne la transposée d'une matrice ou d'un vecteur.

– Les éléments du vecteur des données ξ ne peuvent pas être mesurés exactement. (Par exemple : paramètres physiques mesurés à des endroits très éloignés, etc ...)
– Certaines données comme celles des caractéristiques des appareils électriques prennent des valeurs autour de leurs valeurs nominales.
– Souvent il est impossible d'implémenter, avec grande précision, les solutions calculées : les erreurs d'implémentation.

En présence des incertitudes, le problème (6) devient une simple instance appartenant à la famille des problèmes d'optimisation suivante :

$$\left\{ \min_x \{c^t x | Ax \leqslant b\} | \quad \xi \in \mathcal{U} \right\}, \tag{7}$$

où $\mathcal{U}$ est l'ensemble des incertitudes.

L'optimisation robuste [4] prend tout son sens des deux hypothèses suivantes :

- Les données du problème sont incertaines mais bornées, i.e. sont supposées appartenir à un ensemble d'incertitudes donné $\mathcal{U}$.
- Les contraintes du problème sont dures, i.e. doivent être satisfaites quelle que soit la réalisation $\xi \in \mathcal{U}$ des données.

Sous ces deux hypothèses naissent les concepts de « solution robuste » et « problème homologue robuste » [7, 4, 8] qui définiront l'approche d'optimisation robuste.

Définition 1 (Nemirovski *et al.* [7, 8])**.** *Le problème*

$$\min_{x,\tau} \{\tau | c^t x \leqslant \tau, Ax \leqslant b \quad \forall \xi \in \mathcal{U}\} \tag{8}$$

est dit robuste homologue (ou homologue) au problème (6), sa solution est dite robuste optimale. Une solution candidate x_0 de (8) doit être robuste faisable, i.e $Ax \leqslant b$, et $c^t x \leqslant \tau \quad \forall \xi \in \mathcal{U}$. L'ensemble faisable du problème (8) est dit ensemble faisable robuste (du problème incertain (7)).

Ainsi la première étape de résolution d'un problème d'optimisation incertain est de définir, à partir du problème (6), la famille de problème (7) à laquelle il appartient. La seconde étape est de poser le problème robuste homologue (8).

La question principale associée à cette méthodologie [6] est :

> **Question (P) : Quand et comment le problème robuste homologue, à un problème d'optimisation incertain, peut-il être formulé comme un problème d'optimisation qui peut se résoudre efficacement ?**

La bonne nouvelle est que cette classe de problèmes peut être efficacement résolu [8] tant qu'on peut vérifier efficacement que l'ensemble de perturbations $\mathcal{U}$ est non vide (cf. annexe A). La réponse est résumée dans le théorème suivant :

Théorème 1 (Nemirovski et al. '09 [4]). *Supposons que l'ensemble des incertitudes est donné par un système d'inégalités linéaires $P\xi \leqslant p$, alors le problème (8) est équivalent à un problème d'optimisation linéaire de taille polynomiale en la taille de l'ensemble $\mathcal{U}$.*

Pour démonstration cf. annexe A.

3.3 Optimisation quadratique et semi-définie

Le paradigme de l'optimisation linéaire robuste se généralise intuitivement aux deux classes de problèmes d'optimisation quadratique et convexe.
En effet, un problème d'optimisation générique est un problème de la forme

$$\min_x \{f(x,\xi) | F(x,\xi) \leqslant 0\}, \tag{9}$$

où :

- x est le vecteur de décision,
- f, F sont spécifiées par la nature du problème,
- ξ est un vecteur de dimension finie représentant les données du problème.

Mis à part le cas de l'optimisation linéaire présenté dans le paragraphe précédent, les cas particuliers les plus répandus en applications sont les suivants :

- **Problème d'optimisation quadratique [16] :**

$$\min_x \{c^t x | \quad ||A_i x - b_i||_2 \leqslant c_i^t x - d_i, i = 1, \ldots, m\}, \tag{10}$$

  où d_i est un nombre réel. c, c_i, b_i sont des vecteurs A_i est une matrice de dimensions appropriées. Les données du problème sont $\xi = (c, \{A_i, b_i, c_i, d_i\}_{i=1}^m)$.
- **Problème d'optimisation semi-définie [42] :**

$$\min_x \{c^t x | \quad A_0 + \sum_{i=1}^{dimx} x_i A_i \succeq 0 \,^3\}, \tag{11}$$

  où A_i, $i = 0, \ldots, dimx$, sont des matrices symétriques ou hermitiennes. $\xi = (c, A_0, \ldots, A_{dimx})$.

Sous la présence des incertitudes, le problème (9) devient une simple instance appartenant à la famille de problèmes d'optimisation suivante :

$$\left\{ \min_x \{f(x,\xi) | \quad F(x,\xi) \leqslant 0\} | \quad \xi \in \mathcal{U} \right\}, \tag{12}$$

où $\mathcal{U}$ est l'ensemble des incertitudes.

Le problème (9) désignera l'un des deux problèmes (11), (10) et la définition 1 s'étend facilement à celle-ci :

3. La notation $A \succeq 0$, où A est une matrice symétrique, signifie que la matrice A est semi-définie positive.

Définition 2 (Problème robuste homologue (extention) : Nemirovski *et al.* [7, 8])**.** *Le problème*

$$\min_{x,\tau}\{\tau | f(x,\xi) \leqslant \tau, F(x,\xi) \leqslant 0 \quad \forall \xi \in \mathcal{U}\} \tag{13}$$

est dit robuste homologue au problème (9), sa solution est dite robuste optimale. Une solution candidate x_0 de (13) doit être robuste faisable, i.e $F(x,\xi) \leqslant 0$, et $f(x,\xi) \leqslant \tau \quad \forall \xi \in \mathcal{U}$. L'ensemble faisable du problème (13) est dit ensemble faisable robuste (du problème incertain (12)).

Pour plus de détails sur ce concept cf. annexe A.

La réponse à la question centrale **(P)** du paragraphe précédent a été explorée dans le cas des problèmes d'optimisation quadratique et semi-définie. La mauvaise nouvelle est qu'en général, ces deux classes de problèmes est NP-difficile [4] même si $\mathcal{U}$ est polyédrique[7, 4, 10, 12, 5]. Dans ce cas, on a besoin de « remplacer » les problèmes robustes homologues par leurs approximations qui soient efficacement solvables. Ceci a fait naître le concept de l'approximation sûre et efficace d'un problème d'optimisation.

Hypothèses 1. *Supposons que, et c'est typiquement le cas dans les applications, l'ensemble des incertitudes est donné par*

$$\mathcal{U} = \zeta^n + \mathcal{Z}, \tag{14}$$

- ζ^n *: les données nominales,*
- $\mathcal{Z}$ *: ensemble de perturbations compactes et convexes et* $0 \in \mathcal{Z}$.

Cette paramétrisation est cruciale pour l'obtention des approximations comme on le verra dans la suite cf. annexe A.

Définition 3 (Nemirovski *et al.* [4])**.** *Le problème d'optimisation*

$$\min_{x,u}\{c^t x : G(x,u) \geqslant 0\}, \tag{15}$$

est dit une approximation du problème (13) si : x peut être étendue en une solution faisable (x,u) du problème (15) i.e. x est faisable pour le problème (13).

Autrement dit, l'ensemble faisable du problème (13) contient la projection, sur l'espace des x, de l'ensemble faisable du problème (15). Si les deux ensembles sont égaux, alors l'approximation est dite exacte. La stratégie générale de résolution est alors la suivante :

- formuler le problème robuste homologue ;

4. Un problème est dit solvable en temps polynomial s'il est « facile à résoudre » : il admet un algorithme de résolution efficace. Un problème est dit NP-dur s'il n'admet pas d'algorithme de résolution efficace (en temps polynomial en fonction de ses données). D'amples détails sur ce sujet sont donnés dans [9, 24, 2].

– isoler l'ensemble faisable robuste, i.e. $\{x : F(x,\xi) \leqslant 0 \quad \forall \xi \in \zeta^n + \mathcal{Z}\}$;
– dériver une approximation du problème homologue.

Différents résultats, avec différents ensembles d'incertitudes, ont été obtenus en appliquant différentes techniques d'approximations. Pour une synthèse sur le sujet, voir les annexes A et B.

4 Optimisation stochastique : *Two-Stage*

4.1 Résolution par approche stochastique

Dans le cadre de la programmation stochastique le problème (4) est reformulé comme suit :

$$\max_{\bar{x}_2} \mathbb{E}_\xi Q(\bar{x}_2, \xi), \tag{16}$$

où $Q(\bar{x}_2, \xi) := \min\{10^{-4}x_1, x_2\}$ tel que :

$$\begin{cases} x_1 + x_2 = 10^4 + 1, \\ x_2 = \bar{x}_2 + \xi, \\ x_2 \geqslant 0, x_1 \geqslant 0, \end{cases}$$

où ξ est une variable aléatoire de Ω dans $\Xi \subset \mathbb{R}$ à densité de probabilité $\mathbb{P}_\xi = \mathcal{N}(0, \sigma^2)$ (par exemple), où Ξ est le support de la variable aléatoire ξ (i.e. le plus petit ensemble vérifiant $\mathbb{P}_\xi(\Xi) = 1.$), $\sigma \in \mathbb{R}^+$ est la variance, $\mathbb{E}_\xi Q(\bar{x}_2, \xi)$ désigne l'espérance mathématique de $Q(\bar{x}_2, \xi)$ par rapport à la variable aléatoire ξ.

4.1.1 Résolution analytique

Le fait que cet exemple soit simple nous donne l'avantage de calculer analytiquement la fonction objectif $\mathbb{E}_\xi Q(\bar{x}_2, \xi)$ et par conséquent la solution exacte du problème d'optimisation (16) ; ce qui est en général impossible.

En effet, tout calcul fait, on obtient :

$$\begin{aligned} \mathbb{E}_\xi Q(\bar{x}_2, \xi) = \tfrac{1}{2}\left(\bar{x}_2(1 - 10^{-4}) + (1 + 10^{-4})\right)\left(1 + erf(\tfrac{\bar{x}_2 - 1}{\sigma\sqrt{2}})\right) + ... \\ ... + \tfrac{\sigma}{\sqrt{2\pi}}(1 + 10^{-4})exp(-\tfrac{(\bar{x}_2 - 1)^2}{2\sigma^2}), \end{aligned} \tag{17}$$

où la fonction erreur est définie par $erf(x) := \frac{2}{\sqrt{\pi}} \int_0^x exp(-t^2)dt$. Pour la simulation numérique on a pris $\sigma = 1$; le tracé de la fonction (17) est donné par la figure 4. On en

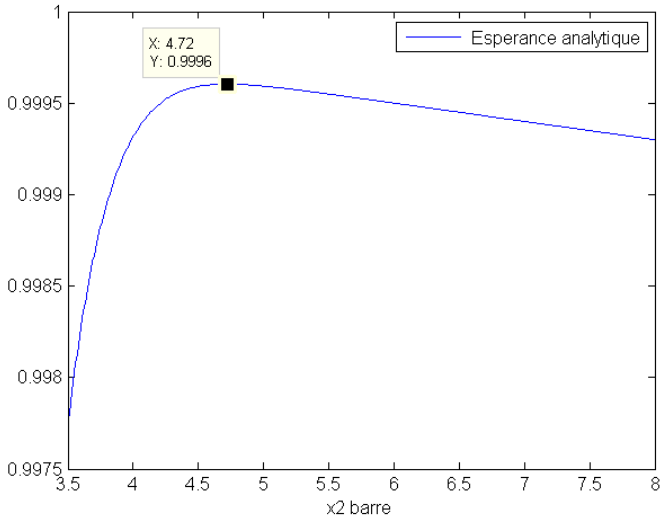


FIGURE 4 – Fonction objectif du problème (16).

déduit que la valeur optimale du problème (16) est 0.9996, atteinte en $\bar{x}_{2s}^* = 4.72$ (solution optimale).

En réalité, les problèmes d'optimisation dans les cas pratiques sont beaucoup plus compliqués que dans cet exemple (un grand nombre de contraintes et de variables de décision etc ...) ; ceci rend impossible non seulement le calcul analytique de $\mathbb{E}_\xi Q(\bar{x}_2, \xi)$ mais son évaluation numérique en un point fixé : ceci requiert un calcul d'intégrales multiples ; ce calcul avec une précision machine est en général très difficile pour $\xi \in \mathbb{R}^d, d > 4$ [32]. On trouve une discussion détaillée sur l'aspect de la complexité algorithmique de cette classe de problèmes d'optimisation plus loin dans ce document.

4.1.2 Résolution approchée basée sur la méthode Monte Carlo

Un outil classique pour l'évaluation numérique d'intégrales multiples est la méthode de Monte Carlo [32, 41] : on génère N échantillons aléatoires indépendants identiquement distribués (iid) $\xi^1, \ldots, \xi^N$ et on calcule :

$$\hat{Q}_N(\bar{x}_2) := \frac{1}{N} \sum_{i=1}^{N} Q(\bar{x}_2, \xi^i). \tag{18}$$

$\hat{Q}_N(\bar{x}_2)$ est un estimateur non biaisé de $\mathbb{E}_\xi Q(\bar{x}_2, \xi)$, i.e $\mathbb{E}_\xi \hat{Q}_N(\bar{x}_2) = \mathbb{E}_\xi Q(\bar{x}_2, \xi)$, et selon la loi forte des grands nombres [11], il est consistant i.e. converge vers $\mathbb{E}_\xi Q(\bar{x}_2, \xi)$ avec une probabilité 1 quand $N \to \infty$. De plus, la solution optimale de

$$\max_{\bar{x}_2 \geqslant 0} \hat{Q}_N(\bar{x}_2), \tag{19}$$

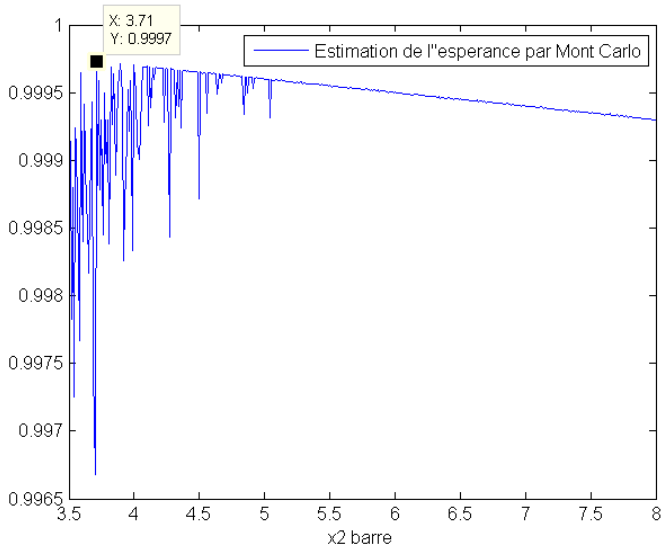


FIGURE 5 – Estimation de la fonction objectif du problème (16) par la méthode Monte Carlo : $N = 1000$.

est un estimateur de la solution de (16), vers quoi il converge en distribution normale centrée sur la solution exacte de (16) quand $N \to \infty$. On a également un résultat similaire concernant l'estimateur (18). Plusieurs propriétés détaillées sur ces deux estimateurs ont été démontrées dans [39]. On reviendra sur ces aspects plus loin dans ce document.

Nous appliquons cette méthode sur cette exemple, en supposant que $N = 1000$ et $\sigma = 1$. Le tracé de la figure 5 compare la moyenne analytique avec moyenne empirique, i.e. $\hat{Q}_N(\bar{x}_2)$. La valeur optimale du problème (19) est de 0.997 atteinte en $\bar{x}'^{*}_{2s} = 3.71$. En comparant cette solution optimale approchée et égale à 3.71 avec la solution optimale exacte égale à 4.72 du problème (16) (voir figure 4), on constate une très mauvaise précision de la solution approchée.

Conformément aux propriétés asymptotiques, mentionnées au début de ce paragraphe, de l'estimateur $\hat{Q}_N(\bar{x}_2)$ et de la solution optimale du problème (18), l'augmentation du nombre de réalisation N doit conduire à améliorer la précision de la solution approchée. Nous prenons donc $N = 100000$ et cela conduit à la courbe présentée sur la figures 6. La solution (resp. la valeur) optimale de (19), soit 4.73 (resp. 0.9996) (voir figures 6 et 7), ce qui est conforme avec les propriétés des deux estimateurs. Néanmoins, si on choisit de quantifier la complexité algorithmique de cette méthode par le nombre de réalisations N [40], on a complexifié cent fois notre problème par rapport au cas où $N = 1000$. Du point de vue temps de calcul, pour le cas $N = 1000$ on a eu un temps de $0,68$ seconde alors que pour $N = 100000$ ce temps est égal à $1,87$ seconde (à peu près trois fois $0,68$).

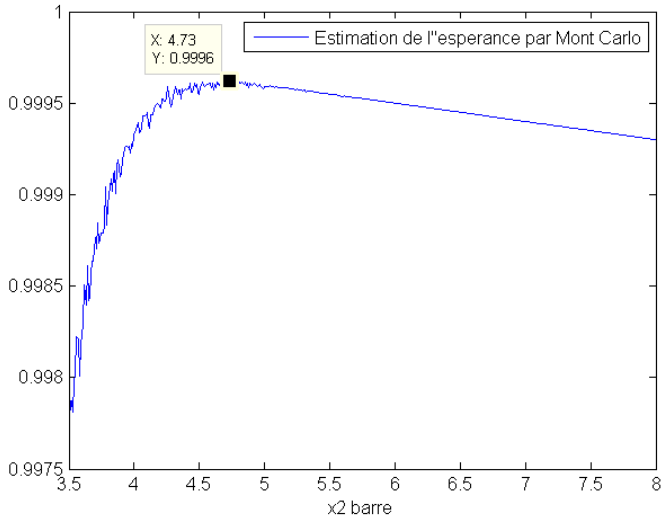


FIGURE 6 – Estimation de la fonction objectif du problème (16) par la méthode Monte Carlo : $N = 100000$.

Ce phénomène indésirable de manque de précision, n'est pas lié directement à notre méthode d'approximation, il est plutôt intrinsèque à la nature du problème. En effet, notre modèle stochastique (16) met en jeu deux variables aléatoires dont les variances sont très éloignées l'une par rapport à l'autre. La première variable aléatoire est $10^{-4}x_1 = 1 + 10^{-4} - 10^{-4}\bar{x}_2 - 10^{-4}\xi \sim \mathcal{N}(1 + 10^{-4} - 10^{-4}\bar{x}_2, 10^{-8})$, la deuxième est $x_2 = \bar{x}_2 + \xi \sim \mathcal{N}(1 + 10^{-4} - \bar{x}_2, 1)$ [5]. La variance de x_1 est 10^{-8} fois plus petite que celle de x_2. Cette situation a un impact fort sur la qualité de notre approximation comme on l'expliquera dans la suite de ce paragraphe.

Pour illustrer davantage les propriétés asymptotiques de ces deux estimateurs (on rappelle que le premier estimateur est défini dans (18) ; le deuxième estimateur est la solution du problème d'optimisation (19)) , on a répété 10000 fois la même expérience($N_{experience} = 10000$) (l'estimation de la fonction (16) par l'estimateur (18)) pour $N = 1000$, puis pour $N = 100000$, et on a tracé l'histogramme, dans les deux cas, de la solution (resp. la valeur) optimale du problème (19), estimateurs de la solution (resp. valeur) optimale du problème (16). On a obtenu les résultats illustrés sur la figure 9 et 8. On remarque que chacun des deux histogrammes de la figure 9 convergent vers une loi normale centrée sur une valeur à peu près égale à la solution (resp. valeur) optimale de (16) ; en comparant les figures 9 et 8, on remarque en plus, que cette valeur est d'autant plus proche de la solution (resp. valeur) optimale de (16) que N est grand ; ce qui est conforme avec le résultat théorique mentionné plus haut qui dit que quand N tend vers $+\infty$ les deux

5. La notation $\xi \sim \mathcal{N}(\mu, \sigma^2)$ signifie que la loi de probabilité de la variable aléatoire ξ est une gaussienne de moyenne μ et variance σ^2.

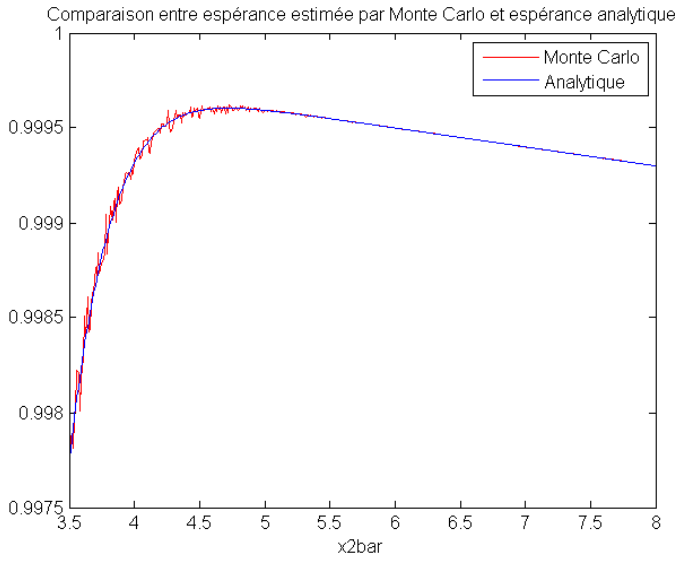


FIGURE 7 – Fonction objectif (analytique) du problème (16) et son estimation par la méthode Monte Carlo : $N = 100000$.

gaussiennes deviennent centrées sur la solution (resp. valeur) optimale exacte de (16). Néanmoins, cette convergence est très lente : le passage de $N = 1000$ à $N = 100000$ ne nous a même pas permis d'avoir des gaussiennes plus ou moins centrées sur les valeurs obtenues par la résolution analytique de notre problème : valeur et solution optimales du problème d'optimisation (16) ; les histogrammes des figures 9 et 8 nous montrent qu'aucune des 10000 expériences répétées ne nous a permis d'avoir une approximation exacte du problème (16). Autrement dit, sur les 10000 expériences répétées, l'obtention d'une approximation qui donne les mêmes solutions que les solutions analytiques du problème (16), n'est pas probable. Ce constat est en fait similaire à la remarque précédente dans le cas d'une seule expérience. Ce phénomène est illustré par la figure 10. Cette figure représente le comportement de l'estimateur (18) avec les variables aléatoires que met en jeu : $10^{-4}x_1$ et x_2 pour N fixe ainsi que le comportement asymptotique ($N = +\infty$) de cet estimateur (18) qui correspond à l'espérance analytique du problème d'optimisation (16) ; j'ai choisi de représenter 99% des valeurs qui peuvent être prises par les deux variables aléatoires $\frac{1}{N}\sum_{i=1}^{N} x_2(\xi_i)$ et $\frac{1}{N}\sum_{i=1}^{N} 10^{-4}x_1(\xi_i)$, ce qui correspond à prendre des intervalles de longueur 3 fois la variance de chaque variable aléatoire. En effet, dans la zone où x_2 est plus petit que $10^{-4}x_1$, l'estimateur $\hat{Q}_N(\bar{x}_2)$ vaut $\frac{1}{N}\sum_{i=1}^{N} x_2(\xi_i)$, cette variable aléatoire a une moyenne égale à $\bar{x}_2$ et une variance var_2 égale à $\frac{1}{N}\sigma^2$. Dans la zone où $10^{-4}x_1$ est plus petit que x_2, notre estimateur vaut $\frac{1}{N}\sum_{i=1}^{N} 10^{-4}x_1(\xi_i)$, c'est une variable aléatoire de moyenne égale à $1 + 10^{-4} - 10^{-4}\bar{x}_2$ et de variance var_1 égale à $\frac{10^{-8}}{N}$. Chaque variable

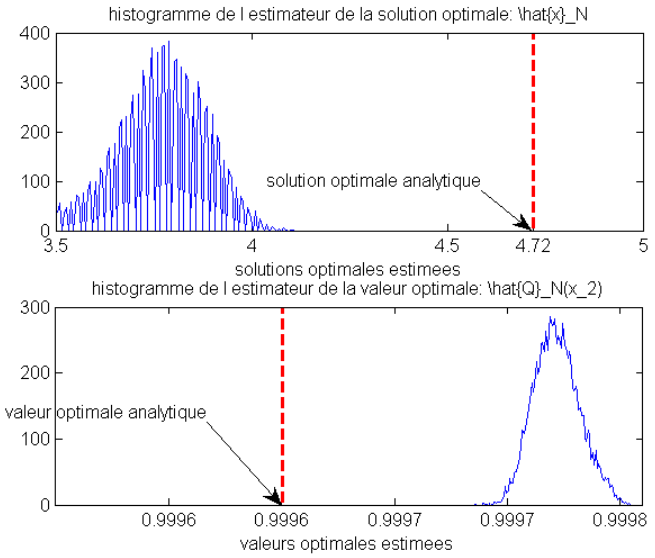


FIGURE 8 – Histogramme des solutions estimées : $N = 1000$.

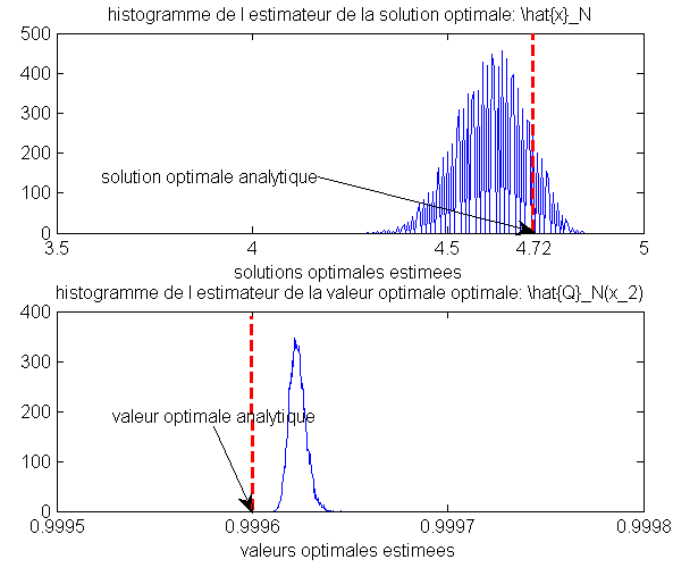


FIGURE 9 – Histogramme des solutions estimées : $N = 100000$.

aléatoire prend des valeurs qui fluctuent autour de sa moyenne dans la plage à 99% (voir figure 10). Pour un N fixe, notre simulation basée sur la méthode de Monte Carlo, donne un estimateur $\hat{Q}_N(\bar{x}_2)$ avec des oscillations fortes dans la zone où x_2 est plus petit que $10^{-4}x_1$ et des oscillations faibles dans la zone où $10^{-4}x_1$ est plus petit que x_2 (car $var_1 << var_2$). La zone transitoire est la plus problématique, car, à cause de la grande variance du x_2 et en cherchant la valeur maximale de $\hat{Q}_N(\bar{x}_2)$, on s'autorise à être décalé à gauche de la solution optimale théorique (égale à 4.72 voir figure 4), ce qui a été le cas

	$N = 1000$	$N = 100000$
$N_{experience} = 1$	0,68 secondes	1,87 secondes
$N_{experience} = 10000$	$1,42.10^2$ secondes	$1,80.10^4$ secondes

TABLE 1 – Récapitulatif des temps de calcul.

pour $N = 1000$ comme on a vu (un maximum empirique atteint en 3.71 voir figure 5). Ce phénomène de décalage par rapport au maximum théorique, dû à la variance du x_2, est d'autant moins fort que N est grand comme on a pu remarquer pour le cas $N = 100000$ (un maximum empirique égal à 4.73 voir figure 6). La variance de l'estimateur $\hat{Q}_N(\bar{x}_2)$ est d'autant plus petite que N est grand ; par conséquent, les intervalles à 99% de la figure 10 sont d'autant plus petits que N est grand. Ainsi, pour un N suffisamment grand, le phénomène de décalage par rapport au maximum théorique est faible et la solution estimée est proche de la solution exacte de telle sorte que pour $N = +\infty$ on a un décalage nul. Néanmoins, dans le cas de cet exemple numérique, un N suffisamment grand était de l'ordre de 10^5 pour un exemple de dimension 1. Les informations relatives au temps de calcul correspondant à chaque couple de nombre de répétition d'expérience $N_{experience}$ et nombre de réalisation N sont résumées dans le tableau 1. Dans le cas où $N_{experience} = 10000$ et $N = 100000$ on a eu un temps de calcul de $1,8.10^4$ secondes $=$ 5 heures, ce qui rend l'analyse de la solution obtenue très coûteuse pour un problème d'optimisation (16) de dimension 1.

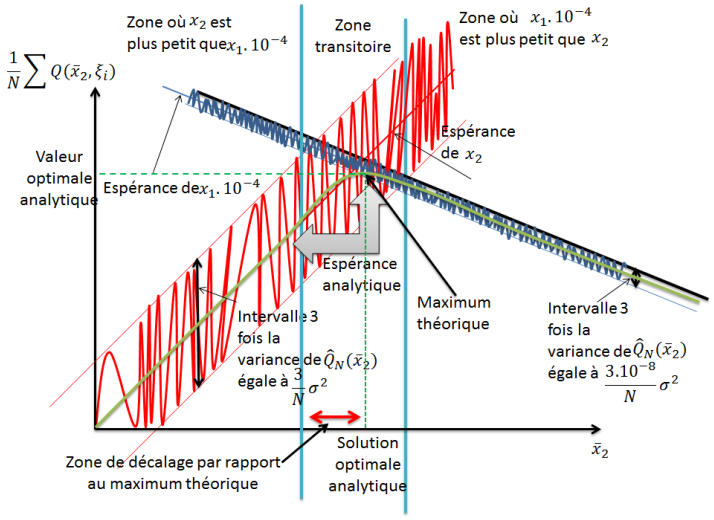


FIGURE 10 – Effet de la variance des deux variables aléatoires $10^{-4}x_1$ et x_2 sur le calcul de la moyenne empirique $\hat{Q}_N(\bar{x}_2)$.

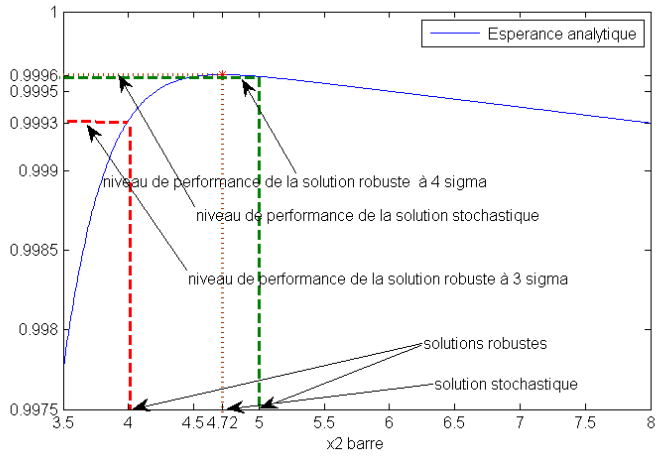


FIGURE 11 – Comparaison de performance entre solution robuste et solution stochastique.

4.1.3 Complexité de la classe de problème d'optimisation Two-stage

La résolution du problème (16), bien qu'il soit convexe, est très difficile à résoudre du fait de la difficulté pour évaluer la fonction objectif du problème d'optimisation (16). En effet on a le résultat suivant sur sa complexité :

Théorème 2 (Dyer *et al.* 2003 [23])**.** *L'évaluation de la fonction coût d'un problème d'optimisation Two-stage dans le cas de variables aléatoires à loi de probabilité continue, est $\#P$ difficile.* [6]

Cela implique que le dénombrement de solutions que peuvent admettre le problème Two-stage, ne peut pas être décidé en un temps polynomial par un algorithme déterministe. Ce résultat nous indique que, en général, les problèmes d'optimisation Two-stage ne peuvent être résolus avec une haute précision comme c'est le cas en optimisation déterministe [39]. Compte tenu de la complexité de ce problème et d'une façon similaire au cas de l'approche d'optimisation robuste, on est amené à chercher des approximations efficaces au problème Two-stage. Dans la littérature de l'optimisation stochastique, la tendance dominante a été illustrée dans le paragraphe 4.1. En effet, on approche le problème (16) par (19). En résolvant (19), on obtient une solution approchée du problème initial (16) et plus N est grand plus la précision de la solution calculée est grande. La question est comment choisir N garantissant à la solution calculée, à la fois une précision numérique et une complexité algorithmique (coût calculatoire) raisonnables. Plusieurs approches ont été développées dans ce contexte [35, 36, 40, 37, 39]. L'annexe C propose un aperçu général à ce cadre théorique appelé la méthode *Sample Average Approximation* (SAA).

6. La classe $\#P$ consiste en des problèmes de dénombrement pour lesquels le lien avec l'ensemble des objets à dénombrer peut être décidé en un temps polynomial ; pour un cadre formel cf. [2, 24]. Je suis actuellement en train d'apprendre ces différentes propriétés que je ne maîtrise pas encore suffisamment.

4.2 Discussion et comparaison entre différentes approches d'optimisation

Un tel niveau de complexité de notre problème d'optimisation Two-stage (16) peut mettre en cause l'application de l'approche d'optimisation stochastique à notre exemple numérique. Dans ce sens, on a fait appel à l'approche robuste et on a résolu le problème (5) pour différents niveaux de perturbations. Bien que la résolution du problème est beaucoup moins compliquée que celle du notre problème *Two-stage*, la question qui reste encore ouverte est comment choisir l'ensemble de perturbation de telle sorte à garantir un niveau de performance proche de celui de notre problème stochastique. En effet, on a effectué une comparaison, représentée sur figure 11, entre les performances de la solution proposée par l'approche d'optimisation robuste ($\bar{x}^*_{2r}$ voir figure 2) et celle proposée par l'approche stochastique ($\bar{x}^*_{2s} = 4.72$ voir figure 4) : on a résolu le problème d'optimisation robuste (5) dans un premier temps pour un ensemble de perturbation à 3σ, $\Xi_3 = [3\sigma, -3\sigma]$ (ce qui correspond à une plage contenant 99% des réalisations possibles de la variable aléatoire $\xi \sim \mathcal{N}(0, \sigma)$) ; dans un deuxième temps, c'était pour un intervalle à 4σ, $\Xi_4 = [4\sigma, -4\sigma]$ (ce qui correspond à une plage contenant plus de 99% des réalisations possibles de la variable aléatoire $\xi \sim \mathcal{N}(0, \sigma)$). Les solutions optimales obtenues sont respectivement $\bar{x}^*_{2r3} = 4$ et $\bar{x}^*_{2r4} = 5$. Ensuite, on a évalué la fonction coût du problème d'optimisation (16) en $\bar{x}^*_{2r3}, \bar{x}^*_{2r4}$ et on a comparé sa valeur maximale avec les deux valeurs obtenues : pour le cas 3σ on a eu une valeur optimale de 0.9993 ; pour 4σ, 0.999599 ; pour la solution analytique de notre problème *Two-stage*, 0.9996. En choisissant bien notre ensemble de perturbation, on arrive à avoir un niveau de performance satisfaisant (voir figure 11).

On remarque que cette comparaison nous a fourni un fil conducteur entre les deux approches d'optimisation : un bon choix de l'ensemble des incertitudes Ξ rend les deux problèmes d'optimisations stochastique (16) et robuste (5) équivalents dans le sens où ils ont la même solution optimale. L'importance de cette remarque réside dans le fait que les deux niveaux de performances (stochastique et robuste) sont à peu près les mêmes, ce qui nous amène, dans ce cas, à privilégier la solution robuste car la complexité algorithmique du problème d'optimisation robuste (5) est équivalente à celle d'un problème d'optimisation linéaire nominale selon le théorème 1.
Même si l'exemple traité est relativement simple, nous pouvons déjà dresser une comparaison préliminaire entre les différentes approches vues jusqu'à présent :

i) Approche nominale :
   - Simplicité de la résolution numérique.
   - Sensibilité aux incertitudes (la solution obtenue n'est plus valable en la présence des perturbations).
   - Meilleur performance (la plus grande valeur optimale obtenue parmi les trois approches).

ii) Approche robuste :
   - Incertitudes inconnues mais bornées (incertitudes intervalles par exemple) ;

- Association avec des expériences où on n'a pas un aspect répétitif (on synthétise une fois pour toutes un correcteur robuste qui respectera un cahier des charges prédéfini, etc...) ;
- Performance inférieure ou égale à celles de l'approche nominale et stochastique.
- Complexité algorithmique moins importante que celle de l'approche stochastique.

iii) Approche stochastique :

- Association avec des expériences où on a un aspect répétitif (milieux bactériens où chaque bactérie maximise son taux de croissance, etc...).
- Ne peut pas être résolu avec une grande précision en général.
- Complexité algorithmique importante.
- Difficultés de résolution intrinsèque au problème due aux propriétés numérique de l'exemple traité.

5 Conclusion et perspectives : méthodologie adoptée et développements envisagés

Comme on l'a mentionné dans l'introduction, la méthode RBA consiste en une représentation suffisamment simple mais pertinente (du point de vue biologique aussi bien que numérique) permettant d'appréhender un ensemble significatif de composantes d'un « système biologique » [25]. Le potentiel prédictif de cette méthode est basé sur la formulation et la résolution efficace (du point de vue numérique) d'un problème d'optimisation. L'un des grands avantages correspond à la classe de problème d'optimisation associé à cette méthode : les problèmes d'optimisation convexe de type Programmation Linéaire. Cette classe de problèmes d'optimisation convexe (riche en propriétés intéressantes) ayant été particulièrement étudiée [16, 30, 13, 9], des algorithmes efficaces en temps de résolution polynômial basés sur la méthode du point intérieur ont été développés [15, 30, 16, 9]. Ces algorithmes restent efficaces en particulier pour des problèmes de très grande dimension. A ce titre, ces algorithmes sont donc particulièrement appropriés pour résoudre des problèmes d'optimisation centrés sur les systèmes biologiques dont la dimension peut rapidement être très grande. Globalement, l'objectif de cette thèse est de prendre en compte de l'aspect stochastique, intrinsèque aux cellules vivantes, lors du processus de la modélisation tout en gardant l'avantage de l'efficacité de résolution numérique dont dispose la méthode RBA.

La difficulté majeure dans notre cadre est de formuler le problème de telle sorte que le problème d'optimisation associé soit de complexité algorithmique raisonnable : *(1)* il faut donc choisir une formulation adéquate du problème d'optimisation, exploiter sa structure particulière afin de choisir, parmi toutes les approches, laquelle est la mieux adaptée ; *(2)* j'explorerai alors, sur des modèles de complexité croissante si cela est nécessaire, les propriétés (mathématiques) attachées aux problèmes d'optimisation considérés ; *(3) in fine*, j'appliquerai les techniques et résultats sur le modèle de *Bacillus subtilis* développé

par A.Goelzer dans [25] et comparerai ses résultats avec les données issues de la littérature et du projet BaSynThec ; *(4)* en particulier, et après avoir étudié la classe de la programmation *Two-stage*, je dois déterminer si la méthode RBA peut être étendue à la fois du point de vue théorique et numérique dans ce contexte.

Tous ces points ont tout d'abord nécessité un travail de classification des différentes approches et modèles d'optimisation robuste et stochastique existants dans la littérature tout en insistant sur l'aspect de la complexité algorithmique induit dans chaque cas et les positionner toutes par rapport à notre problème biologique ce qui, à notre connaissance, n'existait pas auparavant. Ce rapport de première année de thèse résume les principaux éléments de ces travaux de recherche autour de la classification et exploration des propriétés et techniques de résolution des problèmes d'optimisation sous incertitudes.

Les résultats obtenus dans ce cadre me conduiront à choisir ou à défaut à développer une méthode bien adaptée pour résoudre mon problème de RBA stochastique de grande dimension.

A Annexe A : Optimisation linéaire robuste : construction d'une approximation

Dans cette annexe je vais présenter en détail le résultat central sur l'optimisation linéaire robuste mentionné dans le théorème 1 du paragraphe 3.2 : la classe des problèmes d'optimisation linéaire robuste est solvable efficacement.

Sans perte de généralité on peut supposer que le problème est à objectif certain. Le problème incertain linéaire devient alors :

$$\min_x \{c^t x : Ax \leqslant b, \quad \xi = (A, b) \in \mathcal{U}\}, \tag{20}$$

où $A \in \mathbb{R}^{m \times n}$. On considère l'inégalité linéaire incertaine correspondante :

$$\{Ax \leqslant b, \quad \xi = (A, b) \in \mathcal{U}\}. \tag{21}$$

Observations [4] :

– Pour chacune des inégalités incertaines

$$(Ax)_i \leqslant b_i \Leftrightarrow a_i^t x \leqslant b_i$$

(où a_i^t est la i^{me} colonne de A), on considère son homologue robuste, i.e.

$$a_i^t x \leqslant b_i \quad \forall (a_i, b_i) \in \mathcal{U}_i \tag{22}$$

où $\mathcal{U}_i$ est la projection de $\mathcal{U}$ sur l'espace des données de la i^{me} contrainte :

$$\mathcal{U}_i = \{(a_i; b_i) | \quad (A, b) \in \mathcal{U}\}.$$

Le problème homologue d'un problème linéaire incertain avec un objectif certain reste le même quand l'ensemble des contraintes $\mathcal{U}$ est remplacée par le produit direct

$$\hat{\mathcal{U}} = \mathcal{U}_1 \times \ldots \times \mathcal{U}_m$$

de ses projections sur l'espace des données des contraintes respectives.

– Si x est une solution faisable pour (22), alors x le reste quand on étend l'ensemble des contraintes $\mathcal{U}_i$ à leurs enveloppes convexes $Conv(\mathcal{U}_i)$.

Démonstration. Soit x tel que $a_i^t x \leqslant b_i \quad \forall (a_i, b_i) \in \mathcal{U}_i$, alors pour $(\bar{a}_i, \bar{b}_i) \in Conv(\mathcal{U}_i)$, il existe $(a_i^j, b_i^j) \in \mathcal{U}_i$, $\lambda_j \geqslant 0, \forall j$, $\sum_j \lambda_j = 1$ tels que :

$$(\bar{a}_i, \bar{b}_i) = \sum_{j=1}^{J} \lambda_j (a_i^j, b_i^j),$$

puisqu'on a

$$\bar{a}_i^t x = \sum_{j=1}^{J} \lambda_j (a_i^j)^t x \leqslant \sum_j \lambda_j b_i^j = \bar{b}_i,$$

car $(a_i^j, b_i^j) \in \mathcal{U}_i$. Alors, $\bar{a}_i^t x \leqslant \bar{b}_i$ pour tout $(\bar{a}_i, \bar{b}_i) \in Conv(\mathcal{U}_i)$. □

Avec un raisonnement similaire on arrive à démontrer que l'ensemble des solutions faisables pour (22) ne change pas si on étend l'ensemble $\mathcal{U}_i$ à sa fermeture.

En combinant toutes ces observations on arrive à la conclusion suivante :

Le problème homologue d'un problème linéaire incertain ne change pas quand on étend l'ensemble d'incertitude de chaque contrainte $\mathcal{U}_i$ à son enveloppe convexe et fermée, et quand on étend l'ensemble des incertitudes $\mathcal{U}$ au produit direct des ensembles $\mathcal{U}_i$. En d'autres termes, on peut supposer sans perte de généralité que les ensembles $\mathcal{U}_i$ des données incertains des contraintes sont convexes et fermés, et $\mathcal{U}$ est le produit direct de ces ensembles.

Ainsi le problème homologue de (20) est

$$\min_x \{c^t x : Ax \leqslant b, \quad \forall (A,b) \in \hat{\mathcal{U}} = \mathcal{U}_1 \times \ldots \times \mathcal{U}_m\}$$

ou d'une façon équivalente

$$\min_x \{c^t x : a_i x \leqslant b_i, \quad \forall (a_i, b_i) \in \mathcal{U}_i, i = 1 \ldots, m\} \tag{23}$$

Notons qu'ici, les contraintes du problème sont construites par morceaux. Cette structure présente un avantage dans la mesure où la résolution de (23) est réduite à celle des problèmes robustes homologues définis par une seule contrainte $i = 1, \ldots, m$.
Toutes ces considérations nous permettent de focaliser sur une seule inégalité linéaire incertaine, i.e. la famille

$$\{a^t x \leqslant b\}_{(a,b)\in\mathcal{U}}. \tag{24}$$

où, sous l'hypothèse 1, i.e.,

$$\mathcal{U} = \left\{ (a,b) = (a^0, b^0) + \sum_{i=1}^{L} \zeta_l (a^l, b^l) : \quad \zeta \in \mathcal{Z} \right\} \tag{25}$$

où $\mathcal{Z} \subset \mathbb{R}^L$ est l'ensemble de perturbation.

Considérons alors l'inégalité robuste homologue

$$a^t x \leqslant b, \quad \forall (a,b) \in \mathcal{U}, \tag{26}$$

On peut énoncer le résultat suivant :

Considérons le cas où l'ensemble de perturbations dans (25) est donné par *la représentation conique* (cf.[4, pages : 456-460]) :

$$\mathcal{Z} = \left\{ \zeta \in \mathbb{R}^L | \exists u \in \mathbb{R}^K | P\zeta + Qu + p \in \mathbb{K} \right\}, \tag{27}$$

où $\mathbb{K} \subset \mathbb{R}^N$ est un cône convexe fermé à intérieur non vide, P, Q sont des matrices et p un vecteur donnés de dimensions appropriées. Dans le cas où $\mathbb{K}$ n'est pas un cône polyédrique, on suppose que (27) est strictement faisable :

$$\exists(\bar{\zeta}, \bar{u}) | P\bar{\zeta} + Q\bar{u} + p \in int\mathbb{K}. \tag{28}$$

Théorème 3. *Étant donné l'ensemble de perturbation $\mathcal{Z}$ défini dans (27) et supposons de plus que dans le cas où $\mathbb{K}$ n'est pas polyédrique, (28) est vraie. Alors la contrainte incertaine (semi-infinie) (26) peut être représentée par le système d'inégalités côniques en les variables $x \in \mathbb{R}^n$, $y \in \mathbb{R}^N$:*

$$\begin{array}{l} p^t y + a^{0t} x \leqslant b^0, \\ Q^t y = 0, \\ (P^t y)_l + a^{lt} x = b^l \quad l = 1, \ldots, L, \\ y \in \mathbb{K}_*, \end{array} \tag{29}$$

où $\mathbb{K}_* = \{y | \quad y^t z \geqslant 0 \quad \forall z \in \mathbb{K}\}$ est le cône dual de $\mathbb{K}$.

Démonstration. On a,

$$\begin{array}{rl} x \text{ est faisable pour (26)} & \Leftrightarrow \sup_{\zeta \in \mathcal{Z}} \{\underbrace{a^{0t} x - b^0}_{d(x)} + \sum_{l=1}^{L} \zeta_l \underbrace{(a^{lt} x - b^l)}_{c_l(x)}\} \leqslant 0, \\ & \Leftrightarrow \sup_{\zeta \in \mathcal{Z}} \{c^t(x)\zeta + d(x)\} \leqslant 0, \\ & \Leftrightarrow \sup_{\zeta \in \mathcal{Z}} c^t(x)\zeta \leqslant -d(x), \\ & \Leftrightarrow \max_{\zeta, v} \{c^t(x)\zeta : P\zeta + Qv + p\} \leqslant -d(x), \end{array}$$

Ce qui revient à dire que x est faisable pour (26) si et seulement si la valeur optimale de :

$$(CP) : \max_{\zeta, v} \{c^t(x)\zeta : P\zeta + Qv + p\}$$

est inférieure à $-d(x)$. Supposons que (28) est vraie, alors, par le théorème de la dualité conique [9, pages 57-58] (cf. annexe B.1.5), CP est strictement faisable et sa valeur optimale est $\leqslant -d(x)$ si et seulement si son problème dual

$$(DP) : \min_{y} \{p^t y : Q^t y = 0, P^t y = -c(x), y \in \mathbb{K}_*\},$$

est solvable, sa valeur minimale atteinte et $\leqslant -d(x)$. Dans le cas où $\mathbb{K}$ a une représentation polyédrique, le théorème de la dualité classique [9, page18] donne la même condition nécessaire et suffisante (la valeur optimale de (CP) est $\leqslant -d(x)$ ssi la valeur optimale de (DP) est atteinte et $\leqslant -d(x)$) sans avoir à supposer la stricte représentabilité de $\mathbb{K}$ (28). □

Pour plus de détails sur les résultats de la dualité en optimisation [9] : théorème de dualité dans le cas de l'optimisation linéaire ou conique voir annexe B.1.5. Pour une présentation accessible sur la construction des problèmes d'optimisation duale, voir [18].

B Annexe B

Je propose dans cette annexe, un certain nombre de résultats importants sur l'optimisation robuste sous sa forme dite cônique.

B.1 Optimisation cônique

B.1.1 Classification des problèmes d'optimisation cônique

Un problème d'optimisation cônique est de la forme suivante :

$$\min_x \{c^t x + d : Ax - b \in \mathbb{K}\}, \tag{30}$$

où $x \in \mathbb{R}^n$, A est $m \times n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, $d \in \mathbb{R}$ et $\mathbb{K} \subset \mathbb{R}^m$ est un cône convexe à intérieur non vide, et l'application $x \mapsto Ax - b$ est de $\mathbb{R}^n$ dans $\mathbb{K} \subset \mathbb{R}^m$. La formulation cônique est la formulation universelle de la programmation convexe ; parmi les avantages de cette formulation spécifique il y a son pouvoir unificateur [4]. Une très large variété de problèmes d'optimisation convexe est couverte par trois types de cônes :

- i) Le produit direct des quadrants non positifs ($\mathbb{K} = \mathbb{R}^m_+$). Ces cônes donnent naissance aux problèmes d'optimisation linéaire :

$$\min_x \{c^t x + d : {a_i}^t x - b_i \geqslant 0, \quad i = 1, \ldots, m\}.$$

- ii) Le produit direct des cônes de Lorentz $\mathbb{L}^k = \{x \in \mathbb{R}^k : x_k \geqslant \sqrt{x_1^2 + \ldots + x_{k-1}^2}\}$. Ces cônes donnent naissance aux problèmes d'optimisation quadratique : La forme mathématique de ces fameux programmes est donnée comme suit :

$$\min_x \{c^t x + d : ||A_i x - b_i||_2 \leqslant c_i^t x - d_i, \quad i = 1, \ldots, m\};$$

  ici la i^{me} contrainte (dite (*inégalité cônique quadratique* :ICQ) exprime que l'application affine $x \mapsto [A_i x; c_i^t x] - [b_i; d_i]$ est de $\mathbb{R}^n$ dans un cône $\mathbb{L}^i$ de dimension appropriée, tandis que le système de toutes les contraintes signifie que l'application affine $x \mapsto [[A_1 x; c_1^t x]; \ldots; [A_m x; c_m^t x]] - [[b_1; d_1]; \ldots; [b_m; d_m]]$ est de $\mathbb{R}^n$ dans $\mathbb{L}^1 \times \ldots \times \mathbb{L}^m$.
- iii) Le produit direct de cônes semi-définis $\mathbb{S}^k_+$, où $\mathbb{S}^k_+$ est le cône des matrices semi-définies positives $k \times k$; il est contenu dans l'espace $\mathbb{S}^k$ des matrices symétriques $k \times k$. $\mathbb{S}^k$ est considérée comme l'espace Euclidien muni du produit de Frobenius interne $\langle A, B \rangle = Trace(AB)$. La famille des cônes semi-définis donne naissance aux problèmes d'optimisation semi-définie qui ont la forme suivante :

$$\min_x \{c^t x + d : \mathcal{A}_i x - B_i \succeq 0, \quad i = 1, \ldots, m\},$$

  ce qui signifie que l'application affine $x \mapsto \mathcal{A}_i x - B_i \equiv \sum_{j=1}^{n} x_j A^{ij} - B_i$ est de $\mathbb{R}^n$ dans $\mathbb{S}^{k_i}_+$ (telle que A^{ij} et B_i sont des matrices symétriques de dimension $k_i \times k_i$). La contrainte $\sum_{j=1}^{n} x_j A^{ij} - B_i \succeq 0$ est appelée *LMI* (Linear Matrix Inequality).

Remarque : Il faut noter ici que les trois types de cônes considérés admettent la propriété qui fait que le produit direct de cônes appartenant à l'un des trois types est un cône du même type. Ceci est dû à la propriété de *Richesse d'un cône* introduite par les auteurs dans ([4, page :458]), dont sont dotés ces trois types de cônes. Ce qui justifie la classification précédente basée sur la représentation de sous- produits directs de cônes de l'ensemble des contraintes.

Problèmes côniques incertains Un problème cônique incertain est un problème avec une structure fixe et un ensemble des données incertaines $\mathcal{U}_i$ paramétrisé affinement par un vecteur de perturbation $\zeta \in \mathcal{Z}$ tel que :

$$\mathcal{U}_i = \{(c, d, \{A_i, b_i\}_{i=1}^m) | \exists \zeta \in \mathcal{Z} | (c, d, \{A_i, b_i\}_{i=1}^m) = (c^0, d^0, \{A_i^0, b_i^0\}_{i=1}^m) + \sum_{l=1}^{L} \zeta_l (c^l, d^l, \{A_i^l, b_i^l\}_{i=1}^m)\}, \quad (31)$$

et la famille des problèmes côniques incertains :

$$\left\{ \min_x \left\{ c^t x + d : A_i x - b_i \in \mathcal{Q}_i, \quad i = 1, \ldots, m \right\} | (c, d, A_i, b_i) \in \mathcal{U}_i \quad i = 1, \ldots, m \right\}, \quad (32)$$

où

- i) $\mathcal{Q}_i \subset \mathbb{R}^{k_i}$ est un ensemble convexe non vide défini par un système fini de représentations côniques :

  $$\mathcal{Q}_i = \{u \in \mathbb{R}^{k_i} | \exists Q_{il}, q_{il} | Q_{il} u - q_{il} \in \mathbb{K}_{il}, \quad l = 1, \ldots, L_i\},$$

  où Q_{il}, q_{il} sont de dimensions appropriées, $\mathbb{K}_{il}$ est un cône fermé convexe (on se limite aux cas où $\mathbb{K}_{il}$ est l'un des trois cône types mentionnés dans le paragraphe précédent.) Pour une réalisation de perturbation dans (32) on a le problème d'optimisation cônique suivant :

  $$\min_x \{c^t x + d | Q_{il} A_i x - Q_{il} b_i + q_{il} \in \mathbb{K}_{il}, \quad l = 1, \ldots, L_i\},$$

- ii) $\zeta \in \mathcal{Z} \subset \mathbb{R}^L$ est un ensemble donné de perturbations.

Rappelons certains concepts nécessaires pour les développements qui vont suivre :

Problème Homologue

Définition 4. – *(i) Une solution $x \in \mathbb{R}^n$ est faisable robuste pour le problème incertain (32) si et seulement si elle reste faisable pour toutes les réalisations possibles du vecteur perturbation $\mathcal{Z}$ i.e.*

$$x \text{ est robuste faisable} \Leftrightarrow [A_i^0 + \sum_{l=1}^{L} \zeta_l A_i^l] x - [b_i^0 + \sum_{l=1}^{L} \zeta_l b_i^l] \in \mathcal{Q}_i \quad \forall (1 \leqslant i \leqslant m, \zeta \in \mathcal{Z}).$$

– *(ii) L'inégalité robuste homologue à l'inégalité incertaine*

$$[A_i^0 + \sum_{l=1}^{L} \zeta_l A_i^l]x - [b_i^0 + \sum_{l=1}^{L} \zeta_l b_i^l] \in \mathcal{Q}_i,$$

s'écrit sous la forme :

$$[A_i^0 + \sum_{l=1}^{L} \zeta_l A_i^l]x - [b_i^0 + \sum_{l=1}^{L} \zeta_l b_i^l] \in \mathcal{Q}_i \quad \forall(\zeta \in \mathcal{Z}).$$

– *(iii) Le problème robuste homologue à (32) est sous la forme :*

$$\min_{t,x} \left\{ t : \begin{array}{l} [c^{0t} + \sum_{l=1}^{L} \zeta_l c^{l^t}]x + [d^0 + \sum_{l=1}^{L} \zeta_l d^l] - t \in Q_0 \equiv \mathbb{R}_-, \\ [A_i^0 + \sum_{l=1}^{L} \zeta_l A_i^l]x - [b_i^0 + \sum_{l=1}^{L} \zeta_l b_i^l] \in Q_i, i = 1, \ldots, m \end{array} \right\} \forall \zeta \in \mathcal{Z} \left. \right\}. \quad (33)$$

Remarque : Comme dans le cas de la programmation linéaire robuste (voir annexe A), il est facile de démontrer que (33) reste le même si on remplace l'ensemble de perturbation $\mathcal{Z}$ par son enveloppe convexe. Donc on peut considérer sans perte de généralité que l'ensemble $\mathcal{Z}$ est convexe fermé.

B.1.2 Solvabilité des problèmes côniques incertains

Contrairement aux problèmes linéaires incertains, dont le problème robuste homologue peut s'avérer solvable tant que l'ensemble de perturbations l'est aussi (voir théorème 3), les problèmes côniques incertains ayant un problème homologue solvable sont très rares [4]. La stratégie adoptée par Nemirovski et.al. dans [4] est la suivante :

– Identifier la classe des problèmes d'optimisation cônique pour laquelle le problème robuste homologue est solvable ;
– Développer *une approximation sûre et efficace* sinon.

Remarque : On note que la contrainte du problème robuste homologue est construite par morceaux, ce qui réduit sa résolution à celles des problèmes robustes homologues de toutes les contraintes qui constituent le problème (Ce point a été illustré dans l'annexe A). Dans tout ce qui va suivre, on va se focaliser sur le problème homologue global qui contient toutes les contraintes du problème cônique incertain :

$$A(\zeta)x + b(\zeta) \in \mathcal{Q} \quad \forall \zeta \in \mathcal{Z}, \quad (34)$$

où $A(\zeta) \in \mathbb{R}^{k \times n}$ et $b(\zeta) \in \mathbb{R}^k$ sont affines en ζ. Rappelons ici le concept d'approximation *sûre et efficace.*

Définition 5. *Un système $\mathcal{S}$ de contraintes convexes en la variable de décision x et, éventuellement, en d'autres variables u est une approximation sûre et efficace de (34) si :*

$$\forall x \quad \textit{tel que} \quad (\exists u | (x, u) \quad \textit{satisfait} \quad \mathcal{S}) \Rightarrow x \textit{ satisfait } \quad (34).$$

B.1.3 Problèmes quadratiques incertains

Cas solvable Pour la compacité des notations, on considère l'ensemble des contraintes comme étant représenté par une seule inégalité cônique quadratique de la forme :

$$|| \underbrace{A(\zeta)y + b(\zeta)}_{\alpha(y)\zeta+\beta(y)} ||_2 \leqslant \underbrace{c^t(\zeta)y + d(\zeta)}_{\sigma^t(y)\zeta+\delta(y)}, \tag{35}$$

où $A(\zeta) \in \mathbb{R}^{k\times n}$, $b(\zeta) \in \mathbb{R}^k$, $c(\zeta) \in \mathbb{R}^n$, $d(\zeta) \in \mathbb{R}$ sont affines en ζ et $\alpha(y)$, $\beta(y)$, $\sigma(y)$, $\delta(y)$ sont de dimensions appropriées et affines en le vecteur de décision y. Le problème homologue devient :

$$\min_{x,t} \left\{ t : c^t x + d - t \leqslant 0, \quad ||A(\zeta)y + b(\zeta)||_2 \leqslant c^t(\zeta)y + d(\zeta), \quad \forall \zeta \in \mathcal{Z} \right\}. \tag{36}$$

Remarquons que ce problème dépend de la géométrie de l'ensemble de perturbation $\mathcal{Z}$ qui peut être très compliquée, ce qui se traduit en difficulté de résolution du problème homologue. L'ensemble des cas solvables est alors partitionné selon la façon dont l'ensemble des incertitudes est modélisé.

Premier cas solvable : incertitudes intervalles

Hypothèses 2. – *i) On pose $\mathcal{Z} = \mathcal{Z}^l \times \mathcal{Z}^r$. $\zeta = (\eta; \chi)$ où $\eta \in \mathcal{Z}^l$ et $\chi \in \mathcal{Z}^r$ (η et χ sont indépendants), tel que l'inégalité (35) devient :*

$$\underbrace{||A(\eta)y + b(\eta)||_2}_{\alpha(y)\eta+\beta(y)} \leqslant \underbrace{c^t(\chi)y + d(\chi)}_{\sigma^t(y)\chi+\delta(y)}, \tag{37}$$

– *ii) La perturbation à droite est décrite par la représentation cônique*

$$\mathcal{Z}^r = \{\chi : \exists u : \quad P\chi + Qu + p \in \mathbb{K}\},$$

où $\mathbb{K}$ est un cône fermé à intérieur non vide ; la représentation strictement faisable sauf dans le cas où $\mathbb{K}$ est polyédrique (défini par un ensemble fini d'inégalités linéaires).

– *iii) La perturbation à gauche est décrite par :*

$$\begin{aligned}\mathcal{Z}^l = \{\eta = [\delta A, \delta b] : |(\delta A)_{ij}| \leqslant \delta_{ij}, \quad 1 \leqslant i \leqslant k, \quad 1 \leqslant j \leqslant n,\\ |(\delta b)_i| \leqslant \delta_i, \quad 1 \leqslant i \leqslant k\}\end{aligned}$$

$$[A(\zeta), b(\zeta)] = [A^n, b^n] + [\delta A, \delta b].$$

Proposition 1. *Sous les hypothèses 2 i) − iii), le problème homologue de l'inégalité incertaine (35) est équivalent au système d'inégalités linéaires et quadratiques en les variables y, z, τ, v :*

$$\begin{array}{ll} (a) & \tau + p^t v \leqslant \delta(y), P^t v = \sigma(y), \\ & Q^t v = 0, v \in \mathbb{K}^*. \\ (b) & z_i \geqslant |(A^n y + b^n)_i| + \delta_i + \sum_{j=1}^n |\delta_{ij} y_j|, i = 1, \ldots, k \\ & ||z||_2 \leqslant \tau, \end{array} \tag{38}$$

où $\mathbb{K}_*$ est le dual de $\mathbb{K}$.

Démonstration. En effet, y est robuste faisable pour (37) ssi il existe τ tel que :

$$\begin{array}{lll} (1)\ \tau & \leqslant \min\limits_{\chi \in \mathcal{Z}^r} \{\sigma^t(y)\chi + \delta(y)\} \\ & = \min\limits_{\chi, u} \{\sigma^t(y)\chi : P\chi + Qu + p \in \mathbb{K}\} + \delta(y) \\ (2)\ \tau & \geqslant \max\limits_{\eta \in \mathcal{Z}^l} \{||A(\eta)y + b(\eta)||_2\} \\ & = \max\limits_{\delta A, \delta b} \{||[A^n + b^n] + [\delta A y + \delta b]||_2 : |\delta A|_{ij} \leqslant \delta_{ij}, |\delta b_i| \leqslant \delta_i\} \end{array}$$

Par le théorème de dualité cônique [9, pages 57-58] (voir section B.1.5.), τ satisfait (1) ssi τ peut être étendu par une variable v, bien choisie, en une solution faisable de (a) (proposition 1). Ensuite il est évident que τ satisfait (2) ssi il existe z satisfaisant (b). □

Deuxième cas solvable : Incertitudes bornées non structurées

Hypothèses 3. *On maintient les hypothèses i) − ii) de l'ensemble d'hypothèse 2 du paragraphe précédent et on définit $\mathcal{Z}^l$ de la manière suivante :*

$$\mathcal{Z}^l = \{\eta \in \mathbb{R}^{p \times q} : \quad ||\eta||_{2,2} \leqslant 1\}, \tag{39}$$

et on suppose que l'une des assertions suivantes est vraie :

- *1)*

$$A(\eta)y + b(\eta) = A^n y + b^n + L^t(y)\eta R, \tag{40}$$

  où $L(y)$ est affine en y et $R \neq 0$,

- *2)*

$$A(\eta)y + b(\eta) = A^n y + b^n + L^t \eta R(y), \tag{41}$$

  où $R(y)$ est affine en y et $L \neq 0$.

Notons que : $||\eta||_{2,2} = \max\limits_u \{||\eta u||_2 : u \in \mathbb{R}^q, ||u||_2 \leqslant 1\}$ est la norme matricielle usuelle (la valeur singulière maximale.)

Proposition 2. *Sous ces hypothèses 3, le problème homologue de l'inégalité incertaine (35) est équivalent au système d'inégalités linéaires et quadratiques en les variables* y, τ, u, λ *:*

– *i)* Le cas où (39) et (40) sont vraies

$$\begin{array}{l} (a-i)\ \tau + p^t v \leqslant \delta(y), P^t v = \sigma(y), Q^t v = 0, v \in \mathbb{K}^* \\ (b-i)\ \left[\begin{array}{c|c|c} \tau I_k & L^t(y) & A^n y + b^n \\ \hline L(y) & \lambda I_p & 0 \\ \hline [A^n y + b^n]^t & 0 & \tau - \lambda R^t R \end{array}\right] \succeq 0 \end{array} \tag{42}$$

– *ii)* Le cas où (39) et (41) sont vraies

$$\begin{array}{l} (a-ii)\ \tau + p^t v \leqslant \delta(y), P^t v = \sigma(y), Q^t v = 0, v \in \mathbb{K}^* \\ (b-ii)\ \left[\begin{array}{c|c|c} \tau I_k - \lambda L^t L & L^t(y) & A^n y + b^n \\ \hline 0 & \lambda I_q & R(y) \\ \hline [A^n y + b^n]^t & R^t(y) & \tau \end{array}\right] \succeq 0 \end{array} \tag{43}$$

Démonstration. L'idée de la démonstration est basée sur le théorème 3, sur le théorème de la dualité cônique [9, pages 57-58], le lemme du complément de Schur [16, pages :650-651] et la $\mathcal{S}$-procédure [16, pages :655][33] (voir section B.1.5). Le même raisonnement que la proposition précédente est utilisé ici. En effet, y est robuste faisable pour (37) ssi il existe τ tel que :

$$\begin{array}{rl} (cond-i)\ \tau & \leqslant \min\limits_{\chi \in \mathcal{Z}^r} \{\sigma^t(y)\chi + \delta(y)\} \\ & = \min\limits_{\chi, u} \{\sigma^t(y)\chi : P\chi + Qu + p \in \mathbb{K}\} \\ (cond-ii)\ \tau & \geqslant \max\limits_{\eta \in \mathcal{Z}^l} \{||A(\eta)y + b(\eta)||_2\} \\ & = \max\limits_{\delta A, \delta b} \{||[A^n + b^n] + [\delta A y + \delta b]||_2 : |\delta A|_{ij} \leqslant \delta_{ij}, |\delta b_i| \leqslant \delta_i\}. \end{array}$$

Un tel τ satisfait $(cond-i)$ ssi il peut être étendu, par une variable v bien choisie, en une solution de $(a-i) \Leftrightarrow (a-ii)$ (de la proposition 2). Reste à comprendre quand τ satisfait $(b-i), (b-ii)$. Ceci est lié à la $\mathcal{S}-$procédure et au lemme de base suivant :

Lemme 1 (Représentation semi-définie du cône de Lorentz [4])**.** *Le vecteur* $(y,t)^t \in \mathbb{R}^k \times \mathbb{R}$ *appartient au cône* $\mathbb{L}^{k+1} = \{(y,t)^t \in \mathbb{R}^{k+1} : ||y||_2 \leqslant t\}$ *ssi la matrice*

$$Arrow(y,t) = \left[\begin{array}{c|c} t & y^t \\ \hline y & tI_k \end{array}\right],$$

est semi-définie positive.

Le résultat découle immédiatement du lemme du complément de Schur.
Commençons par le cas de l'équation(40). On a (y, τ) satisfait $(cond-ii)$ si et seulement si,

$$(\underbrace{[A^n y + b^n]}_{\hat{y}} + L^t(y)\eta R; \tau) \in \mathbb{L}^{k+1} \quad (\forall(\eta : ||\eta||_{2,2} \leqslant 1)), \quad (\text{par l'équation} \quad 40),$$

si et seulement si,

$$\left[\begin{array}{c|c} \tau & \hat{y}^t + R^t\eta^t L(y) \\ \hline \hat{y} + L^t(y)\eta R & \tau I_k \end{array}\right] \succeq 0 \quad (\forall \eta : ||\eta||_{2,2} \leqslant 1)(\text{par le lemme} \quad 1),$$

si et seulement,

$$\tau s^2 + 2sr^t[\hat{y} + L^t(y)\eta R] + \tau r^t r \geqslant 0 \quad \forall(s; r)\forall(\eta : ||\eta||_{2,2} \leqslant 1)$$

si seulement si,

$$(*) \quad \tau s^2 + 2s\hat{y}^t r + 2 \min_{||\eta||_{2,2}\leqslant 1} [s(\eta^t L(y)r)^t R] + \tau r^t r \geqslant 0 \quad \forall(s; r).$$

Or on a le lemme suivant :

Lemme 2.

$$\min_{\eta:||\eta||_{2,2}\leqslant 1} [(\eta^t L(y)r)^t (sR)] = -||L(y)r||_2||sR||_2.$$

Démonstration. En effet, par l'inégalité de Cauchy-Schwarz [16] on a : $|(\eta^t L(y)r)^t(sR)| \leqslant ||\eta^t L(y)r||_2||sR||_2$, or η est telle que $||\eta||_{2,2} \leqslant 1$; de plus, pour tout y,r,s,R,$\eta : ||\eta||_{2,2} \leqslant 1$, on a $||\eta^t L(y)r||_2 \leqslant ||\eta||_{2,2}||L(y)r||_2$ (car $||.||_{2,2}$ est norme matricielle induite par la norme $||.||_2$). On en déduit : $-||L(y)r||_2||sR||_2 \leqslant (\eta^t L(y)r)^t(sR) \leqslant ||L(y)r||_2||sR||_2$ pour tout $\eta : ||\eta||_{2,2} \leqslant 1$. Il s'ensuit que $\inf_{\eta:||\eta||_{2,2}\leqslant 1} [(\eta^t L(y)r)^t(sR)] \geqslant -||L(y)r||_2||sR||_2$. D'autre part, cette borne inférieure est atteinte ; en effet, prenons la matrice $\eta_0 = -\frac{(L(y)r)(sR)}{||L(y)r||_2||sR||_2}$. On a $\eta_0^t\eta_0 = 1$ donc $\eta_0 \in \{\eta : ||\eta||_{2,2} \leqslant 1\}$. Par suite, $\inf_{\eta:||\eta||_{2,2}\leqslant 1} [(\eta^t L(y)r)^t(sR)] = \min_{\eta:||\eta||_{2,2}\leqslant 1} [(\eta^t L(y)r)^t(sR)] \leqslant (L(y)r)\eta_0(sR) = -||L(y)r||_2||sR||_2$.

□

Par conséquent, $(*)$ est vraie si et seulement si :

$$(\star) \quad \tau s^2 + 2s\hat{y}^t r - 2||L(y)r||_2||sR||_2 + \tau r^t r \geqslant 0 \quad \forall(s; r).$$

Or on a le lemme suivant :

Lemme 3.

$$\min_{\xi:\xi^t\xi\leqslant s^2 R^t R} (L(y)r)^t\xi = -||L(y)r||_2||sR||_2$$

Démonstration. Par Cauchy-Schwarz on a $|(L(y)r)^t\xi| \leqslant ||L(y)r||_2||\xi||_2$, or $\xi^t\xi \leqslant s^2R^tR$, donc $|(L(y)r)^t\xi| \leqslant ||L(y)r||_2||sR||_2$ pour tout $\xi : \xi^t\xi \leqslant s^2R^tR$. On en déduit que $\inf_{\xi:\xi^t\xi\leqslant s^2R^tR}(L(y)r)^t\xi \geqslant -||L(y)r||_2||sR||_2$.
Reste à démontrer que cette borne inférieure est atteinte en $-||L(y)r||_2||sR||_2$. En effet, il suffit de prendre $\xi_0 = sR$ et de vérifier que $sR \in \{\xi : \xi^t\xi \leqslant s^2R^tR\}$. □

Par conséquent, $(\star)$ est vraie si et seulement si :

$$\left[\begin{array}{c|c|c} \tau I_k & L(y)^t & \hat{y} \\ \hline L(y) & \lambda I_p & 0 \\ \hline \hat{y}^t & 0 & \tau - \lambda R^tR \end{array}\right] \succeq 0 \quad \text{(par la } \mathcal{S}-\text{procédure).}$$

Passons maintenant au cas (41). On a (y, τ) satisfait $(cond - ii)$

si seulement si $\quad (\underbrace{[A^ny + b^n]}_{\hat{y}} + L^t\eta R(y); \tau) \in \mathbb{L}^{k+1} \quad (\forall(\eta : ||\eta||_{2,2} \leqslant 1))$ (par 41),

si seulement si $\quad \left[\begin{array}{c|c} \tau & \hat{y}^t + R(y)^t\eta^tL \\ \hline \hat{y} + L^t\eta R(y) & \tau I_k \end{array}\right] \succeq 0 \quad (\forall\eta : ||\eta||_{2,2} \leqslant 1)$ (par le lemme 1),

si seulement si $\quad \tau s^2 + 2sr^t[\hat{y} + L^t\eta R(y)] + \tau^t\tau r \geqslant 0 \quad \forall(s; r)\forall(\eta : ||\eta||_{2,2} \leqslant 1),$

si seulement si $\quad \tau s^2 + 2s\hat{y}^tr + 2\min_{||\eta||_{2,2}\leqslant 1}[s(\eta^tLr)^tR(y)] + \tau r^tr \geqslant 0 \quad \forall(s; r),$

si seulement si $\quad \tau s^2 + 2s\hat{y}^tr - 2||Lr||_2||sR(y)||_2 + \tau r^tr \geqslant 0 \quad \forall(s; r),$

si seulement si $\quad \tau r^tr + 2sR(y)^t\xi + 2sr^t\hat{y} + \tau s^2 \geqslant 0 \quad \forall(s; r), \xi : \xi^t\xi \leqslant r^tL^tLr,$

si seulement si $\quad \left[\begin{array}{c|c|c} \tau I_k - \lambda L^tL & 0 & \hat{y} \\ \hline 0 & \lambda I_q & R(y) \\ \hline \hat{y}^t & R(y)^t & \tau \end{array}\right] \succeq 0$ (par la $\mathcal{S}$-procédure).

□

Cas non solvables efficacement Dans la majorité des cas, les problèmes d'optimisation cônique robuste ne sont pas solvables efficacement. Ceci nous amène à développer une approximation efficacement calculable de ce genre de problèmes homologues. Dans ce qui va suivre, je présenterai quelques outils pour approximer efficacement quelques classes de problèmes robustes homologues non solvables efficacement. Comme dans le

paragraphe précédent, cet ensemble d'outils s'articule sur les différents types de perturbation à savoir la classe des *perturbations bornées structurées* et celle dite *perturbations* $\bigcap$*-ellipsoïdales.*

Premier cas : perturbations bornées structurées

Hypothèses 4. – *i)On considère l'inégalité cônique quadratique (35) et on définit l'ensemble des perturbations à gauche de la façon suivante :*

$$\mathcal{Z}_\rho^l = \left\{ \begin{array}{c} \eta^\nu \in \mathbb{R}^{p_\nu \times q_\nu} \ \forall \nu \leqslant N \\ \eta = (\eta^1, \ldots, \eta^N) : \ ||\eta^\nu||_{2,2} \leqslant \rho \ \forall \nu \leqslant N \\ \eta^\nu = \theta_\nu I_{p_\nu}, \ \theta_\nu \in \mathbb{R}, \ \nu \in \mathcal{I}_S \end{array} \right\} \tag{44}$$

où $\mathcal{I}_S$ *est un sous- ensemble de* $\{1, \ldots, N\}$ *tel que* $p_\nu = q_\nu = 1 \ \forall \nu \in \mathcal{I}_S$.

– *ii)On a :*

$$A(\eta)y + b(\eta) = A^n y + b^n + \sum_{\nu=1}^{N} L_\nu^t(y)\eta^\nu R_\nu(y), \tag{45}$$

où $R_\nu(y) \neq 0, \ L_\nu(y) \neq 0$ *sont affines en* y *pour tout* ν.

– *iii)*

$$\mathcal{Z}_\rho^r = \rho\mathcal{Z}_1^r, \mathcal{Z}_1^r = \{\chi : \exists u : \quad P\chi + Qu + p \in \mathbb{K}\} \tag{46}$$

où $\mathcal{Z}_1^r$ *désigne la famille des ensembles des perturbations à droite par rapport à l'inégalité (37),* $0 \in \mathcal{Z}_1^r$ $\mathbb{K}$ *est un cône convexe fermé à intérieur non vide (on suppose de plus que cette représentation cônique est stricte dans le cas où* $\mathbb{K}$ *n'est pas polyédrique, (pourquoi ? cf.annexe A)).*

Discussion de la méthode d'approximation

– Compte tenu des hypothèses du paragraphe précédent, la déduction d'une approximation efficace pour cette classe de problèmes se fait en deux étapes majeures après avoir éclaté l'inégalité (37)en deux inégalités ayant un membre constant chacune. En effet, y est faisable pour (37) *ssi* il existe τ tel que :

$$\begin{array}{ll} (a) & ||A(\eta)y + b(\eta)||_2 \leqslant \tau \quad \forall \eta \in \mathcal{Z}_\rho^l \\ (b) & c^t(\chi)y + d(\chi) \geqslant \tau \quad \forall \chi \in \mathcal{Z}_\rho^r \end{array} \tag{47}$$

– Dans un premier temps on développe une approximation pour ($(a)47$). Ceci s'effectue en deux étapes. D'abord la reformulation de ($(a)47$) sous une LMI en introduisant éventuellement des nouvelles variables de décision. La reformulation est obtenue en combinant le lemme de Schur et la $\mathcal{S}$-procédure. Ensuite une approximation de cette reformulation au sens de la définition donnée plus haut est fournie en appliquant à nouveau la $\mathcal{S}$-procédure.

– Ensuite, à l'aide de la proposition 1 on calcule une transformation équivalente de ($(b)47$) efficacement calculable.

En combinant les résultats obtenus on arrive à énoncer le théorème suivant :

Théorème 4. *Compte tenu des hypothèses précédentes 4, le système explicite des LMI suivant :*

$$\begin{array}{ll}
(a) & \tau + \rho p^t v \leqslant \delta(y), P^t v = \sigma(y), Q^t v = 0, \quad v \in \mathbb{K}^* \\
(b.1) & Y_\nu \succeq \pm \left(\hat{L}_\nu^t(y) \hat{R}_\nu(y) + \hat{R}_\nu^t(y) \hat{L}_\nu(y) \right), \quad \nu \in \mathcal{I}_S \\
(b.2) & \left[\begin{array}{c|c} Y_\nu - \lambda_\nu \hat{L}_\nu^t \hat{L}_\nu & \hat{R}_\nu^t(y) \\ \hline \hat{R}_\nu(y) & \lambda_\nu I_{k\nu} \end{array} \right] \succeq 0, \quad \nu \notin \mathcal{I}_S \\
(b.3) & Arrow(A^n y + b^n, \tau) - \rho \sum_{\nu=1}^N Y_\nu \ \succeq 0.
\end{array} \tag{48}$$

en les variables $Y_1, \ldots Y_N$, λ_ν, y, τ, v *est une approximation efficacement calculable du problème cônique quadratique (37). L'approximation est exacte pour* $N = 1$.
Où

$$Arrow(u,t) = \left[\begin{array}{c|c} \tau & u^t \\ \hline u & \tau I_k \end{array} \right],$$

où u *est un vecteur de dimension* k, $\hat{L}_\nu = [0_{k\times 1}, I_k] L_\nu^t \eta^\nu$, $\hat{R}_\nu(y) = R_\nu(y)\mathcal{R}$, $\mathcal{R} = [1, 0, \ldots, 0]$ *est de taille* $1 \times (k+1)$.

Démonstration. Le calcul de la transformation équivalente de $((b)47)$ étant évident, on va détailler juste le calcul de l'approximation efficace de la contrainte moindre carrée incertaine $((a)47)$.

Stratégie d'approximation On se propose de déduire une approximation sûre et efficacement calculable de l'inégalité

$$||A(\eta)y + b(\eta)||_2 \leqslant \tau \quad \forall \eta \in \mathcal{Z}_\rho^l \tag{49}$$

– **1. Reformulation du problème robuste homologue (49),(44),(45)** Par le lemme 1, l'inégalité moindre carrée semi-infinie (49) est équivalente à :

$$Arrow(A(\eta)y + b(\eta), \tau) \succeq 0 \quad \forall \eta \in \mathcal{Z}_\rho^l. \tag{50}$$

En introduisant les deux matrices $\mathcal{L} = [0_{k\times 1}, I_k] \in \mathbb{R}^{k\times(k+1)}$ et la matrice $\mathcal{R} = [1, 0, \ldots, 0] \in \mathbb{R}^{1\times(k+1)}$, on obtient :

$$\begin{aligned}
Arrow(A(\eta)y + b(\eta), \tau) = Arrow(A^n y + b^n, \tau) \\
+ \sum_{\nu=1}^N \left(\mathcal{L}^t L_\nu^t(y) \eta^\nu R_\nu(y) \mathcal{R} + \mathcal{R}^t R_\nu(y)^t \eta^{\nu t} L_\nu(y) \mathcal{L} \right).
\end{aligned} \tag{51}$$

Puisque pour tout ν, soit $L_\nu(y)$ soit $R_\nu(y)$, ou bien tous les deux sont indépendants de y, en remplaçant, si c'est nécessaire, $\eta^{\nu t}$ par η^ν, et en permutant $L_\nu(y)\mathcal{L}$ et

$R_\nu(y)\mathcal{R}$, on peut supposer sans perte de généralité que les facteurs $L_\nu(y)$ sont indépendants de y. L'équation (51) devient,

$$Arrow(A(\eta)y+b(\eta),\tau) = Arrow(A^n y+b^n,\tau) + \sum_{\nu=1}^{N}\left(\underbrace{\mathcal{L}^t L_\nu^t}_{\hat{L}_\nu^t}\eta^\nu \underbrace{R_\nu(y)\mathcal{R}}_{\hat{R}_\nu(y)} + \hat{R}_\nu^t(y)\eta^{\nu t}\hat{L}_\nu\right),$$

où les $\hat{R}_\nu(y)$ sont affines en y et $\hat{L}_\nu \neq 0$. De plus les matrices

$$B_\nu(y,\eta^\nu) = \hat{L}_\nu^t\eta^\nu\hat{R}_\nu(y) + \hat{R}_\nu^t(y)\eta^{\nu t}\hat{L}_\nu$$

sont de rang inférieur ou égal à 2.
Un premier résumé de cette recherche est le suivant :

Le problème homologue de (49),(44),(45) est équivalent à la LMI semi-infinie suivante :

$$\underbrace{Arrow(A^n y+b^n,\tau)}_{B_0(y,\tau)} + \sum_{\nu=1}^{N} B_\nu(y,\eta^\nu) \succeq 0 \quad \forall\eta^\nu \in \{\eta^\nu | \eta^\nu \in \mathbb{R}^{p_\nu\times q_\nu}, ||\eta^\nu||_{2,2} \leqslant \rho \quad \forall\nu \leqslant N\} \tag{52}$$

– **Approximation de (52)** Une condition suffisante sur la validité de (52) pour un y donné est l'existence de matrices symétriques Y_ν, $\nu = 1,\ldots,N$, tel que

$$Y_\nu \succeq B_\nu(y,\eta^\nu) \quad \forall(\eta^\nu \in \{\eta^\nu|\eta^\nu \in \mathbb{R}^{p_\nu\times q_\nu}, ||\eta^\nu||_{2,2} \leqslant 1; \nu \in \mathcal{I}_S \Rightarrow \eta^\nu \in \mathbb{R}\}) \tag{53}$$

et

$$B_0(y,\tau) - \rho\sum_{\nu=1}^{N} Y_\nu \succeq 0. \tag{54}$$

L'idée est de démontrer que la LMI semi-infinie (53) en les variables Y_ν, y, τ peut être représentée par un système fini et explicite de LMIs, dans la mesure où le système $\mathcal{S}_0$ des contraintes semi-infinies (53),(54) en les variables $Y_1,\ldots,Y_N,y,\tau$ soit équivalent à un système $\mathcal{S}$ des LMIs. Puisque $\mathcal{S}_0$ est une approximation sûre de (52), alors $\mathcal{S}$ le sera aussi (de plus, elle est efficacement calculable.)
– 1°) Commençons par le cas où $\nu \in \mathcal{I}_S$ ce qui revient à dire que $-1 \leqslant \eta^\nu \leqslant 1$. et $p_\nu = q_\nu = 1$. En utilisant le lemme suivant :
Lemme 4 (Lemme polytopique)**.**

$$\forall\lambda_i \geqslant 0, \quad \sum_i \lambda_i = 1, \Omega_0 + \sum_i^L \lambda_i\Omega_i \succeq 0 \Leftrightarrow \forall i \in \{1,\ldots,L\}, \quad \Omega_0 + \Omega_i \succeq 0,$$

on obtient alors que (53) est équivalent à :

$$Y_\nu \succeq B_\nu(y) = \hat{L}_\nu^t \hat{R}_\nu(y) + \hat{R}_\nu^t(y)\hat{L}_\nu \quad \& \quad Y_\nu \succeq -B_\nu(y), \tag{55}$$

Ce qui est évident, car dans notre cas η^ν appartient au polytope $[-1, 1]$. Il est alors (par ce lemme) nécessaire et suffisant de vérifier (53) pour $\eta^\nu = \pm 1$.

- 2°) Considérons maintenant le cas où $\nu \notin \mathcal{I}_S$. Le raisonnement est similaire à celui de la proposition 2 et peut être facilement refait pour aboutir au trois dernières LMIs énoncées dans le théorème.

- Reste à trouver une approximation équivalente (comme c'est le cas) pour l'inégalité linéaire semi-infinie :

$$c^t(\chi)y + d(\chi) \geqslant \tau \quad \forall \chi \in \mathcal{Z}_\rho^r. \tag{56}$$

Remarquons que pour $\rho > 0$, on a $\rho\mathcal{Z}_1^r = \rho.\{\chi : \exists u : \quad P(\chi/\rho) + Qu + p \in \mathbb{K}\} = \{\chi : \exists u' : \quad P\chi + Qu' + \rho p \in \mathbb{K}\}$. Ainsi, en appliquant la proposition 1 à (56), on arrive à retrouver la première inégalité du théorème.

□

Deuxième cas : perturbations$\bigcap$-ellipsoïdales.

Hypothèses 5. *La démarche est exactement similaire à celle de la section précédente.*
Hypothèses :

- *i) L'ensemble des perturbations à gauche est défini par :*

$$\mathcal{Z}_\rho^l = \{\eta \ : \ \eta^t Q_j \eta \leqslant \rho^2 \ , \ j = 1, \ldots, J\}, \tag{57}$$

où $Q_j \succeq 0$ et $\sum_{j=1}^J Q_j \succ 0$ pour tout j.

- *ii)*

$$\mathcal{Z}_\rho^r = \rho\mathcal{Z}_1^r, \mathcal{Z}_1^r = \{\chi : \exists u : \quad P\chi + Qu + p \in \mathbb{K}\}$$

- *iii) La structure des incertitudes est donnée par :*

$$A(\zeta)y + b(\zeta) = \underbrace{[A^n y + b^n]}_{\beta(y)} + \underbrace{\sum_{l=1}^{L} \eta_l [A^l y + b^l]}_{\alpha(y)\eta} \tag{58}$$

où $L = dim(\eta)$.

Alors en suivant exactement la même démarche que dans la section précédente, on aboutit à une approximation efficacement calculable du problème cônique quadratique (37). Les résultats sont résumés dans le théorème suivant :

Théorème 5. *Compte tenu des hypothèses ci-dessus, le système explicite des LMI suivants :*

$$
\begin{array}{ll}
(a) & \tau + \rho p^t v \leqslant \delta(y), P^t v = \sigma(y), Q^t v = 0, \quad v \in \mathbb{K}^* \\
(b.1) & \left[\begin{array}{c|c|c} \mu & 0 & \beta^t(y) \\ \hline 0 & \sum_{l=1}^{J} \lambda_j Q_j & \alpha^t(y) \\ \hline \beta(y) & \alpha(y) & I \end{array}\right] \succeq 0, \\
(b.2) & \mu + \rho^2 \sum_{j=1}^{J} \lambda_j \leqslant \tau, \lambda_j \ \succeq 0
\end{array}
\tag{59}
$$

en les variables y,v,μ,λ_j,τ est une approximation efficacement calculable pour le problème cônique quadratique (37). où $\beta(y) = [A^n y + b^n]$ *et* $\alpha(y)\eta = \sum_{l=1}^{L} \eta_l [A^l y + b^l]$.
Pour $J = 1$ *l'approximation est exacte.*

Démonstration. La démonstration ici suit le même enchaînement que celle du paragraphe précédent. En effet, y est faisable pour (37) *ssi* il existe τ tel que :

$$||A(\eta)y + b(\eta)||_2 \leqslant \tau \quad \forall \eta \in \mathcal{Z}_\rho^l \tag{60}$$

$$c^t(\chi)y + d(\chi) \geqslant \tau \quad \forall \chi \in \mathcal{Z}_\rho^r. \tag{61}$$

Seul le calcul de l'approximation sûre et efficace pour l'inégalité moindres carrés (60) sera détaillé. Le calcul d'une représentation équivalente de (61) est obtenu en appliquant la proposition 1 à l'inégalité (61). Le problème robuste homologue de (60),(57) est équivalent au système de contraintes suivant :

$$\tau \geqslant 0 \quad \& \quad ||\beta(y) + \alpha(y)\eta||_2^{\,2} \leqslant \tau^2 \quad \forall (\eta : \eta^t Q_j \eta \leqslant \rho^2, j = 1, \ldots, J),$$

ce qui équivaut à

$$
\begin{array}{ll}
(a) & \mathcal{A}_\rho \equiv \max_{\eta,t} \{\eta^t \alpha^t(y)\alpha(y)\eta + 2t\beta^t(y)\alpha(y)\eta : \eta^t Q_j \eta \leqslant \rho^2 \ \forall j, t^2 \leqslant 1\} \leqslant \tau^2 - \beta^t(y)\beta(y), \\
(b) & \tau \geqslant 0
\end{array}
\tag{62}
$$

Vers une condition suffisante pour (62) En supposant qu'il existe des réels positifs γ, γ_j, $j = 1, \ldots, J$ tels que la forme quadratique homogène suivante

$$\gamma t^2 + \sum_{j=1}^{J} \gamma_j \eta^t Q_j \eta - (\eta^t \alpha^t(y)\alpha(y)\eta + 2t\beta^t(y)\alpha(y)\eta) \tag{63}$$

soit positive. Alors on obtient que

$$\mathcal{A}_\rho \equiv \max_{\eta,t} \{\eta^t \alpha^t(y)\alpha(y)\eta + 2t\beta^t(y)\alpha(y)\eta : \eta^t Q_j \eta \leqslant \rho^2 \ \forall j, t^2 \leqslant 1\} \leqslant \gamma + \rho^2 \sum_{j=1}^{J} \gamma_j. \tag{64}$$

En effet, si on pose $F = \{(\eta, t) : \eta^t Q_j \eta \leqslant \rho^2 \ j = 1, \ldots, J; t^2 \leqslant 1\}$, on a

$$\begin{aligned}\mathcal{A}_\rho &= \max_{(\eta,t)\in F}\{\eta^t\alpha^t(y)\alpha(y)\eta + 2t\beta^t(y)\alpha(y)\eta : \eta^t Q_j\eta \leqslant \rho^2\ \forall j, t^2 \leqslant 1\}\\ &\leqslant \max_{(\eta,t)\in F}\{\gamma t^2 + \textstyle\sum_{j=1}^J \gamma_j\eta^t Q_j\eta\} \text{ (par la positivité de (63))}\\ &\leqslant \gamma + \rho^2\textstyle\sum_{j=1}^J\gamma_j \text{ (par la définition de } F \text{ et du fait que } \gamma \geqslant 0, \gamma_j \geqslant 0.)\end{aligned}$$

On peut résumer nos remarques précédentes de la façon suivante : S'ils existent γ, γ_j, $j = 1, \ldots, J$ tels que (63) est positive ou bien, ce qui est identiquement équivalent, tels que :

$$\left[\begin{array}{c|c} \gamma & -\beta^t(y)\alpha(y) \\ \hline -\alpha^t(y)\beta(y) & \displaystyle\sum_{j=1}^J \gamma_j Q_j - \alpha^t(y)\alpha(y)\end{array}\right] \succeq 0,$$

et

$$\gamma + \rho^2\sum_{j=1}^J\gamma_j \leqslant \tau^2 - \beta^t(y)\beta(y),$$

alors (y, τ) est faisable pour (62.a). En posant $\nu = \gamma + \beta^t(y)\beta(y)$, on peut récrire cette conclusion sous la forme suivante : S'il existe ν et $\gamma_j \geqslant 0$ tels que

$$\left[\begin{array}{c|c} \nu - \beta^t(y)\beta(y) & -\beta^t(y)\alpha(y) \\ \hline -\alpha^t(y)\beta(y) & \displaystyle\sum_{j=1}^J \gamma_j Q_j - \alpha^t(y)\alpha(y)\end{array}\right] \succeq 0,$$

et

$$\nu + \rho^2\sum_{j=1}^J\gamma_j \leqslant \tau^2,$$

alors (y, τ) est faisable pour (62.a). Si on suppose que τ est positif et on pose $\lambda_j = \gamma_j/\tau$, $\mu = \nu/\tau$, la condition précédente peut être reformulée comme suit : S'ils existent μ et $\lambda_j \geqslant 0, j = 1, \ldots, J$

$$\left[\begin{array}{c|c} \mu - \tau^{-1}\beta^t(y)\beta(y) & -\tau^{-1}\beta^t(y)\alpha(y) \\ \hline -\tau^{-1}\alpha^t(y)\beta(y) & \displaystyle\sum_{j=1}^J \lambda_j Q_j - \tau^{-1}\alpha^t(y)\alpha(y)\end{array}\right] \succeq 0,$$

et

$$\mu + \rho^2\sum_{j=1}^J\lambda_j \leqslant \tau,$$

alors (y, τ) est faisable pour (62.a).

$$\left[\begin{array}{c|c} \mu - \tau^{-1}\beta^t(y)\beta(y) & -\tau^{-1}\beta^t(y)\alpha(y) \\ \hline -\tau^{-1}\alpha^t(y)\beta(y) & \sum_{j=1}^{J}\lambda_j Q_j - \tau^{-1}\alpha^t(y)\alpha(y) \end{array}\right] = \left[\begin{array}{c|c} \mu & 0 \\ \hline 0 & \sum_{j=1}^{J}\lambda_j Q_j \end{array}\right] - \underbrace{\left[\begin{array}{c|c} \tau^{-1}\beta^t(y)\beta(y) & \tau^{-1}\beta^t(y)\alpha(y) \\ \hline \tau^{-1}\alpha^t(y)\beta(y) & \tau^{-1}\alpha^t(y)\alpha(y) \end{array}\right]}_{\mathcal{M}(\alpha(y),\beta(y))}, \tag{65}$$

où

$$\mathcal{M}(\alpha(y), \beta(y)) = \left[\frac{\beta^t(y)}{\alpha^t(y)}\right] \tau I \left[\begin{array}{c|c} \beta(y) & \alpha(y) \end{array}\right],$$

par le lemme de complément de Schur appliqué à (65), on obtient le résultat énoncé dans le théorème. En appliquant la proposition 1 à (61), on finit par retrouver le système de contraintes (80) définissant l'approximation sûre, efficacement calculable de l'inégalité incertaine (37) dans le cas des perturbations ellipsoïdales (57),(58). □

B.1.4 Problème semi-définis incertains

Rappelons qu'un problème d'optimisation semi-définie (notons le SDP) est défini sous la forme suivante :

$$\min_x \left\{ c^t x + d : \mathcal{A}_i(x) = \sum_{j=1}^{n} x_j A^{ij} - B_i \succeq 0 \right\} \tag{66}$$

où $A^{ij} : k_i \times k_i$ et $B_i : k_i \times k_i$ sont des matrices symétriques, $x = (x_1, \ldots, x_n)^t : n \times 1$ est le vecteur des variables décision. Une contrainte de la forme

$$\mathcal{A}x - B \equiv \sum_{i=1}^{n} x_i A^i - B_i \succeq 0,$$

avec A^i et B_i sont symétriques, est dite *LMI* (Linear Matrix Inequality). Un problème semi-défini est, sans perte de généralité, un problème de minimisation d'une fonction linéaire(objectif) sous un nombre fini de contraintes LMI. Une notation plus compacte [4, page 203] sur laquelle cette présentation sera basée est la suivante :

$$\min_x \left\{ c^t x + d : A_i x - b_i \in \mathcal{Q}_i, i = 1, \ldots, L_i \right\}, \tag{67}$$

où $\mathcal{Q}_i$ est donné par :

$$\mathcal{Q}_i = \left\{ u \in \mathbb{R}^{p_i} : \sum_{s=1}^{p_i} u_s Q^{sil} - Q^{il} \succeq 0, l = 1, \ldots, L_i \right\}.$$

Les données du SDP peuvent être regroupées sous la forme suivante :

$$(c, \{A_i, b_i\}_{i=1}^m).$$

Notons que, dans ce cas, l'ensemble $\mathcal{Q}_i$ fixe la structure du problème d'optimisation. Un SDP incertain est une famille de problèmes (67) avec une structure commune (les ensembles $\mathcal{Q}_i$). Rappelons que les données du problème varient dans un ensemble d'incertitudes ; on suppose toujours que les données sont affinement paramétrées par le vecteur de perturbation $\zeta \in \mathbb{R}^L$ qui, sans perte de généralité, varie dans un ensemble de perturbation fermé et convexe (ce point a été discuté en détail dans l'annexe A) $\mathcal{Z}$ tel que $0 \in \mathcal{Z}$:

$$\begin{aligned} (c; d) &= (c^n; d^n) + \sum_{l=1}^{L} \zeta_l (c^l; d^l), \\ [A_i; b_i] &= [A_i^n; b_i^n] + \sum_{l=1}^{L} \zeta_l [A_i^l, b_i^l], i = 1, \ldots, m \end{aligned} \tag{68}$$

Le problème robuste homologue de l'incertain SDP (67), (68) à un niveau de perturbation $\rho > 0$ est le problème d'optimisation semi-infini :

$$\min_{y=(x,t)} \left\{ t : \begin{array}{l} [c^{nt}x + d^n] + \sum_{l=1}^{L} \zeta_l [c^{lt}x] \leqslant t \\ [A_i^n x + b_i^n] + \sum_{l=1}^{L} \zeta_l [A_i^l x + b_i^l] \in \mathcal{Q}_i, i = 1, \ldots, m \end{array} \right\} \forall \zeta \in \rho\mathcal{Z} \right\}. \tag{69}$$

Remarque : Rappelons qu'une approximation efficace du problème semi-infini (67), (68) est un système fini $\mathcal{S}$ de contraintes en (x, t), et éventuellement en d'autres variables supplémentaires u, convexes et efficacement calculables dont la projection sur l'espace des t est incluse dans l'ensemble des solutions faisables du problème (69).

Cas solvables La construction du problème robuste homologue d'un SDP incertain se ramène à la construction des inégalités robustes homologues (voir définition 4 dans les annexes) de toutes les contraintes incertaines du problème SDP. Les problèmes de solvabilité efficace dans le cadre de l'optimisation semi-définie robuste se ramènent alors à ceux de l'inégalité robuste homologue suivants :

$$\mathcal{A}_\zeta(y) \equiv \mathcal{A}^n(y) + \sum_{l=1}^{L} \zeta_l \mathcal{A}_l(y) \succeq 0 \quad \forall \zeta \in \rho\mathcal{Z} \tag{70}$$

de l'inégalité incertaine :

$$\mathcal{A}_\zeta(y) \equiv \mathcal{A}^n(y) + \sum_{l=1}^{L} \zeta_l \mathcal{A}_l(y) \succeq 0 \tag{71}$$

où $\mathcal{A}^n(y)$ et $\mathcal{A}_l(y)$ sont symétriques dépendants affinement du vecteur de décision y.

Remarque : Comme on a vu dans les chapitres précédents, les problèmes robustes homologues des problèmes côniques incertains sont souvent insolvables numériquement ce qui se confirme en particulier pour les SDP incertains (voir [4]). En effet, on a vu que dans le cas des problèmes côniques quadratiques, seuls certains cas particuliers peuvent être résolus ou approximés efficacement (cas scénario des perturbations , incertitudes bornées non-structurées, incertitudes ellipsoïdales, etc ... [4]). Dans le cas des SDP incertains seule la classe des problèmes non-structurés reste solvable [4, 5, 7, 10].

Perturbations non-structurées à norme bornée

Définition 6. *La LMI (71) est avec perturbation à norme bornée non structurée si,*

- *i) L'ensemble de perturbations $\mathcal{Z}$ est l'ensemble des matrices ζ de taille $p \times q$ dont la norme matricielle usuelle est inférieure à ρ ($||\zeta||_{2,2} \leqslant \rho$).*
- *ii) La LMI incertaine (71) est définie dans ce cas par :*

$$\mathcal{A}_\zeta(y) \equiv \mathcal{A}^n(y) + [L^t(y)\zeta R(y) + R^t(y)\zeta^t L(y)] \succeq 0, \tag{72}$$

où $L(.)$ et $R(.)$ sont affines et au moins une de ces deux fonctions ne dépend pas de y. (On suppose sans perte de généralité que R ne dépend pas de y : on peut toujours permuter ξ^t et ξ ainsi que L et R simultanément).

Ainsi on a défini le problème homologue de cette classe de SDP incertain. Le calcul de l'approximation efficace du problème homologue est résumé dans le théorème suivant :

Théorème 6. *le problème homologue de la LMI incertaine (72) dans le cas des perturbations bornées non-structurées peut être représenté d'une façon équivalente par la LMI suivante :*

$$\left[\begin{array}{c|c} \lambda I_p & \rho L(y) \\ \hline \rho L(y)^t & \mathcal{A}^n(y) - \lambda R^t R \end{array}\right] \succeq 0, \tag{73}$$

en les variables y, λ

Démonstration. La démonstration est similaire à celle de la proposition 2. En effet y est robuste faisable pour (72), si et seulement si,

$$\xi^t \left(\mathcal{A}^n(y) + [L^t(y)\zeta R + R^t\zeta^t L(y)]\right) \xi \geqslant 0 \quad \forall(\xi : ||\xi||_{2,2} \leqslant \rho),$$

si et seulement si,

$$\xi^t\mathcal{A}^n(y)\xi + 2\xi^t L^t(y)\zeta R\xi \geqslant 0 \quad \forall(\xi : ||\xi||_{2,2} \leqslant \rho),$$

si et seulement si,

$$\xi^t\mathcal{A}^n(y)\xi + 2\underbrace{\min_{||\xi||_{2,2}\leqslant\rho} \xi^t L^t(y)\zeta R\xi}_{-\rho||L(y)\xi||_2||R\xi||_2} \geqslant 0 \quad \forall\xi$$

(pour détails voir démonstration de la proposition 2 lemme 2),

si et seulement si,

$$\clubsuit \quad \xi^t\mathcal{A}^n(y)\xi - 2\rho||L(y)\xi||_2||R\xi||_2 \geqslant 0 \quad \forall\xi,$$

or on a le lemme suivant

Lemme 5.

$$\min_{\eta\in\Theta}\{\rho\eta^t L(y)\xi\} = -\rho||L(y)\xi||_2||R\xi||_2,$$

où $\Theta = \{\eta : \eta^t\eta \leqslant \xi^t R^t R\xi\}$

Démonstration. par Cauchy-Schwarz on a $|\eta^t L(y)\xi\| \leqslant ||\eta||_2||L(y)\xi||_2, \forall\eta$ or $\eta^t\eta \leqslant \xi^t R^t R\xi$, donc on a $|\eta^t L(y)\xi\| \leqslant ||R\xi||_2||L(y)\xi||_2 \forall\eta \in \Theta$, ainsi $\inf_{\eta\in\Theta}\{\rho\eta^t L(y)\xi\} \geqslant -\rho||L(y)\xi||_2||R\xi||_2$. Reste à démontrer que cette borne inférieure est atteinte en $-\rho||L(y)\xi||_2||R\xi||_2$. En effet, il suffit de prendre $\eta_0 = R\xi$ et de vérifier que $\eta_0 \in \Theta$, ce qui est le cas. □

On en déduit alors que ♣ est vraie si et seulement si,

$$\xi^t\mathcal{A}^n(y)\xi + 2\rho\eta^t L(y)\xi \geqslant 0 \quad \forall\xi \quad \forall(\eta \in \Theta = \{\eta : \eta^t\eta \leqslant \xi^t R^t R\xi\}),$$

si et seulement si

$$\exists\lambda \geqslant 0 : \left[\begin{array}{c|c} 0 & \rho L(y) \\ \hline \rho L^t(y) & \mathcal{A}^n(y) \end{array}\right] \succeq \lambda \left[\begin{array}{c|c} -I_p & 0 \\ \hline 0 & \lambda R^t R \end{array}\right] \text{(par la } \mathcal{S} - \text{procédure)},$$

si et seulement si,

$$\exists\lambda \geqslant 0 : \left[\begin{array}{c|c} \lambda I_p & \rho L(y) \\ \hline \rho L^t(y) & \mathcal{A}^n(y) - \lambda R^t R \end{array}\right] \succeq 0$$

□

Cas non solvables efficacement On a vu que dans le cadre de la programmation cônique incertaine et en particulier celle de la programmation semi-définie, la possibilité de reformuler un problème robuste homologue qui soit efficacement solvable est largement rare. D'où l'introduction de l'approximation efficace (voir [4]) dans le cas où le problème homologue admet une telle approximation. Selon Nemirovski (voir [4]) seules les incertitudes structurées à norme bornées sont sous ce cas de figure.

Définition 7. *On dit que le LMI incertaine (71) est affectée par des incertitudes structurées à normes bornées si,*

- *1. L'ensemble de perturbation $\mathcal{Z}_\rho$ (voir (68))est sous la forme suivante :*

$$\mathcal{Z}_\rho = \left\{ \zeta = (\zeta^1, \ldots, \zeta^L) : \begin{array}{l} \zeta^l \in \mathbb{R}, |\zeta^l| \leqslant \rho, l \in \mathcal{I}_S \\ \zeta^l \in \mathbb{R}^{p_l \times q_l} : ||\zeta^l||_{2,2} \leqslant \rho, l \notin \mathcal{I}_S \end{array} \right\} \tag{74}$$

- *2. La LMI (71) peut être écrite :*

$$\mathcal{A}_\zeta(y) = \mathcal{A}^n(y) + \sum_{l \in \mathcal{I}_S} \zeta^l \mathcal{A}_l(y) + \sum_{l \notin \mathcal{I}_S} [L_l^t(y) \zeta^l R_l + R_l^t [\zeta^l]^t L_l(y)], \tag{75}$$

où $\mathcal{A}_l$, $l \in \mathcal{I}_S$ et les $L_l(y)$ sont affines en y, R_l est non nulle. Ici on a supposé, comme dans le paragraphe précédent, que les R_l sont indépendants de y ; ceci est sans perte de généralité.

Le résultat permettant de calculer une approximation efficace de cette classe de problèmes est résumé dans le théorème suivant :

Théorème 7. *Étant donnée une incertaine LMI (71) avec des incertitudes structurées à norme bornées (74),(75) ; on lui associe le système des LMI explicites suivants en les variables Y_l, $l = 1, \ldots, L$, λ, $l \notin \mathcal{I}_S$, y :*

$$\begin{array}{ll} (a) & Y_l \succeq \pm \mathcal{A}_l(y), l \in \mathcal{I}_S \\ (b) & \left[\begin{array}{c|c} \lambda_l I_{p_l} & L_l(y) \\ \hline L_l^t(y) & Y_l - \lambda_l R_l^t R_l \end{array} \right] \succeq 0, l \notin \mathcal{I}_S \\ (c) & \mathcal{A}^n(y) - \rho \sum_{l=1}^L Y_l \succeq 0 \end{array} \tag{76}$$

Alors le système (76) est approximation efficace du problème robuste homologue (70) (74),(75). L'approximation est exacte pour $L = 1$ où, quand toutes les perturbations sont des scalaires (i.e $\mathcal{I}_S = \{1, \ldots, L\}$) et que toutes les matrices $\mathcal{A}_l(y)$ sont de rang inférieur ou égal à 1.

Démonstration. Il est clair que y est faisable pour (71),(74),(75) si et seulement si y peut être étendu en les variables $Y_1, \ldots, Y_L$ telles que :

$$\forall \zeta \in \mathcal{Z} : \left\{ \begin{array}{l} (a) - \rho Y_l \preceq \zeta^l \mathcal{A}_l(y), l \in \mathcal{I}_S, \\ (b) - \rho Y_l \preceq [L_l^t(y) \zeta^l R_l + R_l^t [\zeta^l]^t L_l(y)], l \notin \mathcal{I}_S \\ (c) \mathcal{A}^n(y) - \rho \sum_{l=1}^L Y_l \succeq 0. \end{array} \right. \tag{77}$$

A partir de cette condition, la démonstration de ce théorème se basera sur les techniques que j'ai détaillées dans les démonstrations de la section précédente. En effet, et en ce qui concerne (a), si on divise cette inégalité par ρ on a $||\zeta_0 = \zeta/\rho||_{2,2} \leqslant 1$ et puisque $\in \mathcal{I}_S$ alors sans perte de généralité on peut considérer que $\zeta_0 \in \mathbb{R}$ et selon le lemme polytopique évoqué dans la section précédente, une condition nécessaire et suffisante est de vérifier l'inégalité (a) aux sommets du polytope $[-1, 1]$, ainsi on retrouve la première inégalité du théorème. Maintenant pour le cas où $l \notin \mathcal{I}_S$, on peut utiliser le résultat du théorème 6 pour remonter à la LMI (b) du présent théorème. □

B.1.5 Récapitulatif des techniques d'optimisation de base

Dans cette section, je vais isoler les techniques vues jusqu'à présent, de résolution de problèmes d'optimisation robuste. En plus du théorème de la dualité conique (cf. [9, page 57]). Ces techniques sont les suivantes :

Théorème 8 (Théorème de dualité cônique : [9])**.** *Considérons le problème*

$$(CP) \quad c^* = \min_x \{c^t x | Ax \geqslant_{\mathbb{K}} b\}\ ^7,$$

son dual s'écrit sous la forme

$$(D) \quad b^* = \max_x \{b^t \lambda | A\lambda = c, \lambda \in \mathbb{K}_*\},$$

où $\mathbb{K}_* = \{y| \quad y^t z \geqslant 0 \quad \forall z \in \mathbb{K}\}$ *est le cône dual de* $\mathbb{K}$*.*

- *1. La dualité est symétrique : le problème dual du dual est le primal.*
- *2. La valeur de l'objectif du dual en toute solution faisable* λ *est* $\leqslant$ *à la valeur de l'objectif du primal en toute solution* x *faisable, par conséquent le gap de dualité*

$$c^t x - b^t \lambda$$

  est positif pour tout pair faisable (x, λ)
- *3.a. Si le primal (CP) est borné inférieurement et strictement faisable, i.e.* $Ax >_{\mathbb{K}} b$ *pour un certain* x*, alors le dual (D) est solvable (i.e borné supérieurement faisable et sa valeur optimale est atteinte) et les valeurs optimales dans les deux problèmes sont égales :* $c^* = b^*$*.*
- *3.b. Si le dual (D) est borné supérieurement et strictement faisable, i.e.* $\lambda >_{\mathbb{K}_*} 0$*, alors le primal (CP) est solvable (i.e borné inférieurement faisable et sa valeur optimale est atteinte) et* $c^* = b^*$*.*

Lemme 6 (Lemme du complément de Schur [16])**.** *Pour toute matrice symétrique* M *de la forme suivante :*

$$\left[\begin{array}{c|c} A & B \\ \hline B^t & C \end{array}\right],$$

7. $a \geqslant_{\mathbb{K}} b$ si est seulement si $a - b \in \mathbb{K}$ où $\mathbb{K}$ est un cône donné.

si C est inversible, alors on a les deux résultats suivants :

- *(1) $M \succ 0$ ssi $C \succ 0$ et $A - BC^{-1}B^t \succ 0$.*
- *(2) Si $C \succ 0$, alors $M \succeq 0$ ssi $A - BC^{-1}B^t \succeq 0$.*

Lemme 7 (Représentation semi-définie du cône de Lorentz[4])**.** *Le vecteur $(y,t)^t \in \mathbb{R}^k \times \mathbb{R}$ appartient au cône $\mathbb{L}^{k+1} = \{(y,t)^t \in \mathbb{R}^{k+1} : ||y||_2 \leqslant t\}$ ssi la matrice*

$$Arrow(u,t) = \left[\begin{array}{c|c} t & y^t \\ \hline y & tI_k \end{array}\right],$$

est semi-définie positive.

Lemme 8 ($\mathcal{S}$-procédure : version sans perte [9, 33])**.** *(i) [version homogène] Soient A et B deux matrices symétriques de la même taille telles que $\bar{x}^t A\bar{x} > 0$ pour un certain $\bar{x}$, alors l'implication suivante :*

$$x^t Ax \geqslant 0 \Rightarrow x^t Bx \geqslant 0,$$

est vraie ssi

$$\exists \lambda \geqslant 0 : B \succeq \lambda A.$$

(ii)[version non homogène] Soient A et B deux matrices symétriques de la même taille, supposons aussi que la forme quadratique $x^t Ax + 2a^t x + \alpha$ est strictement positive en un certain point. Alors l'implication suivante :

$$x^t Ax + 2a^t x + \alpha \geqslant 0 \Rightarrow x^t Bx + 2b^t x + \beta \geqslant 0$$

est vraie ssi

$$\exists \lambda \geqslant 0 : \left[\begin{array}{c|c} B - \lambda A & b^t - \lambda a^t \\ \hline b - \lambda a & \beta - \lambda\alpha \end{array}\right] \succeq 0,$$

.

Lemme 9 ($\mathcal{S}$-procédure : version avec perte[9, 33])**.** *Soient A et B_i, $i = 1, \ldots, p$, C_j, $j = 1, \ldots, q$ des matrices symétriques de la même taille $(n \times n)$et soit*

$$\Gamma = \left\{x \in \mathbb{R}^n | x^t B_1 x > 0, \ldots, x^t B_p x > 0; x^t C_1 x = 0, \ldots, x^t C_q x = 0\right\}.$$

S'il existe $\tau_i \in \mathbb{R}^+$, $i = 1, \ldots, p$, $\nu_j \in \mathbb{R}$, $j = 1, \ldots, q$ tels que :

$$\forall x \in \mathbb{R}^n \quad x^t \left(A - \sum_{i=1}^{p} \tau_i B_i - \sum_{i=1}^{q} \nu_i C_i\right) \succ 0,$$

alors :

$$\forall x \in \Gamma, \quad x^t Ax > 0.$$

Lemme 10 (Lemme polytopique)**.**

$$\forall \lambda_i \geqslant 0, \sum_i \lambda_i = 1, \Omega_0 + \sum_i^L \lambda_i \Omega_i \succeq 0 \Leftrightarrow \forall i \in \{1, \ldots, L\}, \Omega_0 + \Omega_i \succeq 0,$$

Lemme 11 ([9])**.** *Considérons l'inégalité matricielle suivante :*

$$Y - Q^t \Delta^t P^t Z^t R - R^t Z P \Delta Q \succeq 0, \tag{78}$$

où Y est symétrique de taille $n \times n$, Δ est $k \times l$, P, Q, Z, R sont rectangulaires de dimensions appropriées. Etant données les matrices Y, P, Q, Z, R avec $Q \neq 0$ (c'est le seul cas non trivial), cette LMI est satisfaite pour tout Δ telle $|\Delta| \leqslant \rho$ [8] si et seulement s'il existe λ tel que

$$\left[\begin{array}{c|c} Y - \lambda Q^t Q & -\rho R^t Z P \\ \hline -\rho P^t Z^t R & \lambda I_k \end{array}\right] \succeq 0.$$

B.2 Quelques techniques supplémentaires

Dans cette section je vais revisiter certains résultats prouvés plus haut, notamment la classe des problèmes d'optimisation quadratique et semi-défini e soumis à des perturbations $\bigcap$-ellipsoïdales, afin de mettre en oeuvre certaines techniques clés permettant d'aboutir aux mêmes résultats précédents. Le grand intérêt de ces techniques est qu'elles nous font gagner un grand degré de généricité dans la façon d'aborder les problèmes d'optimisation comme on va le voir dans la suite de ce document ; ceci n'était pas le cas pour les techniques que j'ai présentées jusqu'à maintenant. En effet, on a utilisé une relaxation lagrangienne particulièrement dans le cas du problème quadratique soumis à des perturbations $\bigcap$-ellipsoïdales (voir théorème 5). Un autre avantage majeur de ces techniques c'est qu'elles permettent en général de trouver systématiquement des conditions suffisantes aux problèmes de faisabilité robustes.

Cas d'inégalité moindre carré robuste avec perturbations $\bigcap$-ellipsoïdales On considère le problème posé dans la section B.1.3. D'après la formulation (58), l'inégalité moindre carré (60) devient :

$$|| \underbrace{[A^n y + b^n]}_{\beta(y)} + \underbrace{\sum_{l=1}^{L} \eta_l [A^l y + b^l]}_{\alpha(y)\eta} ||_2 \leqslant \tau.$$

ce qui est équivalent à

$$\tau \geqslant 0 \quad \& \quad ||\beta(y) + \alpha(y)\eta||_2^{\,2} \leqslant \tau^2 \quad \forall (\eta : \eta^t Q_j \eta \leqslant \rho^2, j = 1, \ldots, J). \tag{79}$$

Ici, je donnerai une nouvelle démonstration du théorème 5.

8. cette notation $|.|$ désigne la norme spectrale de Δ définie comme étant la racine carré de la plus grande valeur propre de la matrice carrée $A^t A$.

Théorème 9. *Compte tenu des hypothèses 5 de la section B.1.3, le système explicite des LMI suivant :*

$$
\begin{array}{ll}
(c.1) & \left[\begin{array}{c|c|c} \mu & 0 & \beta^t(y) \\ \hline 0 & \sum_{l=1}^{J} \lambda_j Q_j & \alpha^t(y) \\ \hline \beta(y) & \alpha(y) & I \end{array}\right] \succeq 0, \\
(c.2) & \mu + \rho^2 \sum_{j=1}^{J} \lambda_j \leqslant \tau^2, \lambda_j \ \geqslant 0
\end{array}
\tag{80}
$$

en les variables y,λ_j, μ est une approximation efficacement calculable pour le problème cônique quadratique (79). Pour $J = 1$, l'approximation est exacte.

Démonstration. Dans la suite on va omettre la dépendance de y dans α et β. La démonstration se fait en 3 étapes :

- i) Vers la $\mathcal{S}$-procédure :
  L'inégalité robuste (79) est équivalente à :

$$
\left[\begin{array}{c} 1 \\ \eta \end{array}\right]^t \left[\begin{array}{c|c} \beta^t\beta - \tau & \beta^t\alpha \\ \hline \alpha^t\beta & \alpha^t\alpha \end{array}\right] \left[\begin{array}{c} 1 \\ \eta \end{array}\right] \leqslant 0 \tag{81}
$$

  pour tout η tel que :

$$
\left[\begin{array}{c} 1 \\ \eta \end{array}\right]^t \left[\begin{array}{c|c} -\rho^2 & 0 \\ \hline 0 & Q_j \end{array}\right] \left[\begin{array}{c} 1 \\ \eta \end{array}\right] \leqslant 0 \quad j = 1 \ldots L \tag{82}
$$

- ii) Application de la $\mathcal{S}$-procédure :
  Une condition suffisante pour la validité de (81), (82) et par suite de (79) est donnée par la $\mathcal{S}$-procédure sous la forme suivante : $\exists \lambda_j \geqslant 0, \quad j = 1 \ldots, L$ tels que :

$$
\left[\begin{array}{c|c} \tau^2 - \sum_{j=1}^{L} \lambda_j \rho^2 & 0 \\ \hline 0 & \sum_{j=1}^{L} \lambda_j Q_j \end{array}\right] - \left[\begin{array}{c|c} \beta^t\beta & \beta^t\alpha \\ \hline \alpha^t\beta & \alpha^t\alpha \end{array}\right] \succeq 0, \tag{83}
$$

  or

$$
\left[\begin{array}{c|c} \beta^t\beta & \beta^t\alpha \\ \hline \alpha^t\beta & \alpha^t\alpha \end{array}\right] = \left[\begin{array}{c} \beta^t \\ \hline \alpha^t \end{array}\right] I \left[\begin{array}{c|c} \alpha & \beta \end{array}\right].
$$

  Par suite (83) devient :

$$
\left[\begin{array}{c|c} \tau^2 - \sum_{j=1}^{L} \lambda_j \rho^2 & 0 \\ \hline 0 & \sum_{j=1}^{L} \lambda_j Q_j \end{array}\right] - \left[\begin{array}{c} \beta^t \\ \hline \alpha^t \end{array}\right] I \left[\begin{array}{c|c} \alpha & \beta \end{array}\right] \succeq 0. \tag{84}
$$

– iii) Application du lemme de Schur : Puisqu'on a I une matrice définie positive alors par le lemme du complément de Schur (voir lemme 6), (84) est vraie si et seulement si les deux inégalités ((c.1),(c.2) de 80) sont vraies. Ainsi on obtient le même résultat que celui énoncé dans le théorème 5.

□

Avec cette technique on voit bien qu'on a abouti systématiquement au résultat qui est une approximation (condition suffisante) du problème de faisabilité robuste (79).

Problèmes d'optimisation semi-définie : cas des perturbations $\bigcap$-ellipsoïdales

Position du problème[5] Considérons l'inégalité matricielle suivante :

$$F(y,\delta) = F^0 + \sum_{i=1}^{l} \delta^i F^i(y) \succeq 0, \tag{85}$$

où $y \in \mathbb{R}^m$ est le vecteur de décision, δ est le vecteur perturbation appartenant à un certain ensemble donné $\mathcal{D} \subset \mathbb{R}^l$ défini de la manière suivante :

$$\mathcal{D} = \left\{ \delta \in \mathbb{R}^l | \delta = (\delta^1, \ldots, \delta^N)^t | \quad \delta^k \in \mathbb{R}^{n_k}, ||\delta^k||_2 \leqslant \rho, k = 1, \ldots, N \right\}, \tag{86}$$

où ρ est un paramètre positif. Les entiers n_k représentent la taille de chaque sous-bloc δ^k de δ (on a évidemment $n_1 + \ldots + n_N = l$). F est une application de $\mathbb{R}^m \times \mathcal{D}$ dans $\mathcal{S}^n$. On considère le problème de faisabilité robuste suivant :

$$F(y,\delta) \succeq 0, \forall \delta \in \mathcal{D} \tag{87}$$

Les auteurs dans [5] proposent une approximation efficacement calculable pour ce problème. Le problème présente plusieurs motivations du fait de la structure des incertitudes (86) qui est assez générale. En effet, la configuration fixée dans (86) peut être exploitée dans plusieurs cas de figure : on peut citer par exemple le cas où le vecteur de perturbation a chaque composante bornée en module ; on aura $n_1 = \ldots = n_N = 1$, il y a aussi le cas où la norme euclidienne du vecteur perturbation est bornée... Il s'agit d'une classe de problèmes qui est déjà NP-difficile [5, 4]. Dans ce paragraphe, je vais en proposer une approximation efficacement calculable (pour (87)).

Théorème 10. *Considérons le problème semi-défini incertain (85) et l'ensemble d'incertitudes (86), et soit* $\nu_0 = 0$, $\nu_k = \sum_{s=1}^{k} n_s$. *Alors le système explicite de LMI certaines*

est le suivant :

$$(d.1)\quad \left[\begin{array}{c|cccc} T_k-\rho^2 D_k & F_{\nu_{k-1}+1} & F_{\nu_{k-1}+2} & \dots & F_{\nu_k} \\ \hline F_{\nu_{k-1}+1} & D_k & & & \\ F_{\nu_{k-1}+2} & & D_k & & \\ \vdots & & & \ddots & \vdots \\ F_{\nu_k} & & & & D_k \end{array}\right] \succeq 0, k=1,\dots,N; \tag{88}$$

$$(d.2)\quad \sum_{k=1}^{N} T_k \preceq 2F_0,$$

en les variables y, $T_1,\dots,T_N$, $D_1,\dots,D_N$ est une approximation efficace du problème homologue (87) (i.e. la projection de l'ensemble faisable de (88) sur l'espace des y est incluse dans l'ensemble des solutions faisables de (87)).

Démonstration. La démonstration s'effectue systématiquement en trois temps principaux :

– i) Isolation et reformulation adéquate avec la structure des incertitudes :

On commence par reformuler l'inégalité incertaine (85) d'une façon adéquate avec la structure des incertitudes (86). En effet, on doit démontrer que pour tout $\delta = (\delta_1,\dots,\delta_l)$ tels que :

$$\sum_{i=\nu_{k-1}+1}^{\nu_k} \delta_i^2 \leqslant \rho^2, \quad k=1,\dots,N, \tag{89}$$

on a

$$F_0 + \sum_i \delta F_i \succeq 0.$$

Ce qui est équivalent à

$$F^0 + \sum_{k=1}^{N} \left[F_{\nu_{k-1}-1},\dots F_{\nu_k}\right] \left[\begin{array}{c} \delta_{\nu_{k-1}+1} I_n \\ \vdots \\ \delta_{\nu_k} I_n \end{array}\right] \succeq 0. \tag{90}$$

Il est clair qu'il existe toujours une matrice de permutation $Q = Q^t = Q^{-1} \in \mathbb{R}^{(n_k\times n)\times(n_k\times n)}$ telle que :

$$Q \left[\begin{array}{c} \delta_{\nu_{k-1}+1} I_n \\ \vdots \\ \delta_{\nu_k} I_n \end{array}\right] = [I_n \otimes \delta^k];$$

l'inégalité (90) devient alors :

$$F^0 + \sum_{k=1}^{N} \left[F_{\nu_{k-1}-1},\dots F_{\nu_k}\right] Q[I_n \otimes \delta^k] \succeq 0. \tag{91}$$

Cette formulation s'articule bien sur la partition du vecteur perturbation (voir (86)).

– ii) Vers une paramétrisation des incertitudes :

Une condition suffisante (et nécessaire) pour la validité du problème homologue (91),(86) est :
$\exists T_k \in \mathcal{S}_n$ $k = 1, \ldots, N$ telles que :

$$\begin{aligned} (a) \quad & T_k \succeq - \sum_{i=\nu_{k-1}+1}^{\nu_k} \delta_i F_i \\ (b) \quad & F_0 \succeq \sum_{k=1}^{N} Tk. \end{aligned} \tag{92}$$

On se fixe un $k \in \{1, \ldots, N\}$ et on considère alors l'inégalité suivante :

$$T_k + \left[F_{\nu_{k-1}-1}, \ldots F_{\nu_k}\right] Q[I_n \otimes \delta^k] \succeq 0. \tag{93}$$

$$\Updownarrow$$

$$x^t T_k x + x^t \left[F_{\nu_{k-1}-1}, \ldots F_{\nu_k}\right] Q[I_n \otimes \delta^k] x \geqslant 0 \quad \forall x \in \mathbb{R}.$$

L'idée est d'isoler la partie « connue » de la partie « inconnue » de cette inégalité. Pour ce faire, on peut voir la partie incertaine $Q[I_n \otimes \delta^k]$ comme système dont l'entrée est x et la sortie peut être notée $p = Q[I_n \otimes \delta^k]x$. Le couple (x, p) est dit graphe du système $Q[I_n \otimes \delta^k]$ (cf. [22]). En effet, on va chercher un ensemble de contraintes quadratiques vérifiées par le graphe (x, p). La suffisance de notre approximation énoncée dans le théorème est liée alors au fait qu'on a remplacé le graphe par cet ensemble de contraintes quadratiques. Cette technique s'appelle *procédé d'immersion* (cf. [22]). On pose $p = Q[I_n \otimes \delta^k]x$. Une première contrainte peut être déjà énoncée :

$$p^t p = x^t [I_n \otimes \delta^k]^t Q^t Q [I_n \otimes \delta^k] x$$

$$\Updownarrow$$

$$p^t p = x^t [I_n \otimes \delta^{k^t} \delta^k] x,$$

or on a : $\delta^{k^t} \delta^k = ||\delta^k||_2^2 \leqslant \rho^2$, donc

$$p^t p \leqslant \rho^2 x^t x.$$

Bien que cette contrainte est vérifiée par le graphe (x, p), elle n'est pas suffisante car on doit chercher un ensemble de contraintes qui contient ce graphe. D'après l'observation suivante :

$$[Q \otimes \delta^k] R_k = diag(R_k, \ldots, R_k) \begin{bmatrix} \delta_{\nu_{k-1}+1} I_n \\ \vdots \\ \delta_{\nu_k} I_n \end{bmatrix} \quad \forall R_k \in \mathbb{R}^{n \times n}, \tag{94}$$

on a la paramétrisation suivante :

$$x^t \left(Q[I_n \otimes \delta^k]R_k\right)^t \left(Q[I_n \otimes \delta^k]R_k\right) x \leqslant \rho^2 x^t R_k^t R_k x, \quad \forall x \in \mathbb{R}^n,$$

en remplaçant $R_k^t R_k$ par D_k matrice symétrique quelconque dans $\mathbb{R}^{n\times n}$, et par $\tilde{D}_k$, $diag(R_k^t R_k, \ldots, R_k^t R_k) = [I_n \otimes R_k^t R_k]$, on obtient finalement :

$$p^t (I \otimes D_k) p \leqslant \rho^2 x^t D_k x, \quad \forall D_k \in \mathcal{S}^n; \tag{95}$$

d'après l'observation (94) ci-dessus et en appliquant les règles du produit de Kronecker (cf. [17]), ce résultat est immédiat. Reste à prouver qu'il s'agit d'une opération d'immersion. En effet, on a l'inclusion suivante :

$$\left\{(x,p)|p = Q[I_n \otimes \delta^k]x\right\} \subset \left\{(x,p)|p^t(I \otimes D_k)p \leqslant \rho^2 x^t D_k x, \quad \forall D_k \in \mathcal{S}^n\right\}.$$

D'où

$$\begin{cases} x^t T_k x + x^t \left[F_{\nu_{k-1}-1}, \ldots F_{\nu_k}\right] p \geqslant 0 \quad \forall x \in \mathbb{R}, \\ \text{avec :} \quad (p,x) \in \{(p,x)|\exists \delta^k : ||\delta^k||_2 \leqslant 1, p = Q[I_n \otimes \delta^k]x\}. \end{cases}$$

Ce qui est impliqué par :

$$\begin{cases} x^t T_k x + x^t \left[F_{\nu_{k-1}-1}, \ldots F_{\nu_k}\right] p \geqslant 0 \quad \forall x \in \mathbb{R}, \\ \text{avec :} \quad (p,x) \in \{(p,x)|p^t(I \otimes D_k)p \leqslant \rho^2 x^t D_k x, \quad \forall D_k \in \mathcal{S}^n\}. \end{cases}$$

$$\Updownarrow$$

$$\begin{bmatrix} x \\ p \end{bmatrix}^t \left[\begin{array}{c|ccc} T_k & F_{\nu_{k-1}+1} & F_{\nu_{k-1}+2} & \ldots & F_{\nu_k} \\ \hline F_{\nu_{k-1}+1} & 0 & \ldots & & 0 \\ F_{\nu_{k-1}+2} & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ F_{\nu_k} & 0 & \ldots & & 0 \end{array}\right] \begin{bmatrix} x \\ p \end{bmatrix} \geqslant 0 \tag{96}$$

$\forall (x,p)$ tels que :

$$\begin{bmatrix} x \\ p \end{bmatrix}^t \left[\begin{array}{c|c} \rho^2 D_k & 0 \\ \hline 0 & -[I_n \otimes D_k] \end{array}\right] \begin{bmatrix} x \\ p \end{bmatrix} \geqslant 0. \tag{97}$$

– iii) Application de la $\mathcal{S}$-procédure :
En appliquant la $\mathcal{S}$-procédure aux deux formes quadratiques de (96) et (97), on obtient ((d.1).88) ; et lui rajoutant la condition ((b).92) on obtient le résultat du théorème.

□

Matrix Cube theorem [4, page 489] Le problème *Matrix cube* a été introduit par les auteurs dans [4, page 489]. Dans ce paragraphe, je propose une nouvelle démonstration en m'appuyant sur les techniques que j'ai déjà introduites dans cette section. Je vais mettre également en oeuvre la technique *(D,G)-scaling* [34] sur les contraintes quadratiques de ce problème.

Je présenterai seulement le cas complexe : le cas réel est similaire au cas complexe à quelques simplifications près.

Position du problème 1. *(Cas Complexe) Soit m, p_1, q_1, $\ldots$, p_l, q_l des entiers positifs, $A \in \mathcal{H}_+^m$ [9], $L_j \in \mathbb{C}^{p_j \times m}$, $R_j \in \mathbb{C}^{q_j \times m}$ deux matrices données, $L_j \neq 0$. On considère la partition $\{1, \ldots, L\} = I_s^r \cup I_s^c \cup I_f^c$ telle que $p_j = q_j$ pour $j \in I_s^r \cup I_s^c$. On associe à ces données la famille de « boîtes de matrices » définies par* [10] *:*

$$\mathcal{U}_\rho = \left\{ A + \rho \sum_{j=1}^{L} \left[L_j^* \Theta_j R_j + R_j^* \Theta_j^* L_j \right] : \begin{array}{c} \Theta_j \in \mathcal{Z}_j, \\ 1 \leqslant j \leqslant L \end{array} \right\} \tag{98}$$

où est $\rho \geqslant 0$ est un paramètre et

$$\mathcal{Z}_j = \left\{ \begin{array}{c} \{\Theta_j = \theta I_{p_j} : \theta \in \mathbb{R}, |\theta| \leqslant 1\}, j \in I_s^r \\ (\text{« Bloc perturbation : scalaire réel répété »}) \\ \{\Theta_j = \theta I_{p_j} : \theta \in \mathbb{C}, |\theta| \leqslant 1\}, j \in I_s^c \\ (\text{« Bloc perturbation : scalaire complexe répété »}) \\ \{\Theta_j \in \mathbb{C}^{p_j \times q_j} : ||\Theta||_{2,2} \leqslant 1\}, j \in I_f^c. \\ (\text{« Bloc perturbation : Bloque complexe plein »}) \end{array} \right. \tag{99}$$

Le problème s'énonce alors comme suit : étant donné $\rho \geqslant 0$, vérifier si

$$\mathcal{U}_\rho \subset \mathcal{H}_+^m. \tag{100}$$

Théorème 11. *S'il existe $Y_j \in \mathcal{H}_m$, $j = 1, \ldots, L$ telle que*

$$\begin{array}{ll} (a) & Y_j \succeq L_j^* \Theta_j R_j + R_j^* \Theta_j^* L_j \quad \forall (\Theta_j \in \mathcal{Z}_j, 1 \leqslant j \leqslant L) \\ (b) & A - \rho \sum_{j=1}^{L} Y_j \succeq 0, \end{array} \tag{101}$$

alors la proposition (100) est vraie. De plus, le système de contraintes LMIs (101) admet

9. Cette notation désigne l'ensemble de matrices hermitiennes semi-définies positives.
10. Le symbole $(.)^*$ désigne la matrice transconjuguée d'une matrice.

l'approximation efficace :

$$
\begin{aligned}
&(i) \quad Y_j \pm [L_j^* R_j + R_j^* L_j] \succeq 0, \quad j \in I_s^r, \\
&(ii) \quad \left[\begin{array}{cc} Y_j - V_j & L_j^* R_j \\ R_j^* L_j & V_j \end{array} \right] \succeq 0, j \in I_s^c, \\
&(iii) \quad \left[\begin{array}{cc} Y_j - \lambda_j L_j^* L_j & R_j^* \\ R_j & \lambda_j I_{p_j} \end{array} \right] \succeq 0, j \in I_f^c, \\
&(vi) \quad A - \rho \sum_{j=1}^{L} Y_j \succeq 0.
\end{aligned}
\tag{102}
$$

en les variables matricielles $Y_j \in \mathcal{H}^m$, $j = 1, \ldots, k$, $V_j \in \mathcal{H}^m$, $j \in I_s^c$ *et les variables réelles* λ_j, $j \in I_f^c$.

Démonstration. Les auteurs dans [4] ont démontré que le système de contraintes (102) est une transformation équivalente pour (101). Ici, je vais démontrer juste la suffisance. Ceci dit, l'un des avantages des techniques introduites dans cette section en outre de permettre de trouver des conditions suffisantes systématiquement, c'est qu'elles représentent un outil systématique pour trouver des pistes vers les conditions nécessaires, ce qui est typiquement le cas de ce problème.

– i) Considérons le cas où $j \in I_s^r$:
soit $j \in I_s^r$, on a $\Theta_j = \theta I_{p_j}$ avec $\theta \in [-1, 1]$, alors la condition ((a) 101) est vraie si et seulement si elle vérifiée sur les sommets du polytope $[-1, 1]$, ce qui est équivalent à la condition ((i) 102). On peut déduire une autre condition pour ce cas ($j \in I_s^r$) en appliquant la technique (D,G)-scaling [34]. En effet on a le lemme suivant :

Lemme 12. *Pour tout* $\Theta_j \in \mathcal{Z}_j$ *telle que* $j \in I_s^r$, *et pour tout* $D_j = D_j^* \succ 0 \in \mathbb{C}^{p_j \times p_j}$ *et* $G_j = -G_j^* \in \mathbb{C}^{p_j \times p_j}$, *on a :*

$$
\left[\begin{array}{c} I_{p_j} \\ \hline I_{p_j} \otimes \theta \end{array} \right]^* \left[\begin{array}{c|c} D_j & G_j \\ \hline G_j^* & -D_j \end{array} \right] \left[\begin{array}{c} I_{p_j} \\ \hline I_{p_j} \otimes \theta \end{array} \right] \succeq 0. \tag{103}
$$

Le résultat est immédiat (cf.[17] pour les règles du produit de Kronecker) compte tenu des observations évidentes suivantes :
– 1)

$$(I_{p_j} \otimes \theta)^* D_j (I_{p_j} \otimes \theta) \preceq D_j;$$

– 2)

$$G_j (I_{p_j} \otimes \theta) + (I_{p_j} \otimes \theta)^* G_j^* = 0.$$

D'une façon similaire à la démonstration du théorème10, je vais isoler la partie incertaine de la partie certaine dans la contrainte ((a)101) et procéder à une immersion. En effet ((a)101) est équivalente à

$$x^*Y_jx + x^*\theta L_j^* I_{p_j} R_j x + x^* R_j^* I_{p_j} L_j \theta^* x \geqslant 0, \quad \forall x \in \mathbb{C}^m.$$

en posant $p = I_{p_j}\theta^* x$ on obtient de façon équivalente :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & L_j^* R_j \\ \hline R_j^* L_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0. \tag{104}$$

pour (x, p) tel que $p = I_{p_j}\theta^* x$.

Vers la synthèse de l'immersion : L'immersion est formulée sous la forme de l'inclusion suivante :

$$\{(x,p)|\exists j \in I_s^r | p = \Theta_j^* x\} \subset \left\{(x,p)| \left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} D_j & G_j \\ \hline G_j^* & -D_j \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \succeq 0 \right\}. \tag{105}$$

Ce résultat est immédiat compte tenu du lemme 12. A partir de (105), on en déduit alors ce qui suit :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & L_j^* R_j \\ \hline R_j^* L_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \quad \forall (x,p) : p = \Theta_j^* x. \tag{106}$$

$$\tag{107}$$

Ce qui est impliqué par :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & L_j^* R_j \\ \hline R_j^* L_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \tag{108}$$

pour tout (x, p) tel que :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} D_j & G_j \\ \hline G_j^* & -D_j \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0. \tag{109}$$

En appliquant la $\mathcal{S}$-procédure à (108) et (109) on obtient l'approximation de ((a) 101), dans le cas où $j \in I_s^r$, suivante :

$$\left[\begin{array}{c|c} Y_j - D_j & L_j^* R_j - G_j \\ \hline R_j^* L_j - G_j^* & D_j \end{array}\right] \succeq 0, \tag{110}$$

en la variable Y_j et les variables supplémentaires D_j, G_j.

– ii) Considérons ensuite le cas où $j \in I_s^c$: le raisonnement est exactement similaire au précédent, ceci est dû au fait que la matrice $\Theta_j = \theta I_{p_j}$, avec $\theta \in \mathbb{C}$, commute avec n'importe quelle matrice. En effet ((a)101) est équivalente à :

$$x^* Y_j x + x^* \theta L_j^* I_{p_j} R_j x + x^* R_j^* I_{p_j} L_j \theta^* x \geqslant 0, \quad \forall x \in \mathbb{C}^m.$$

L'immersion suivante est obtenue de manière similaire à la précédente :

$$\{(x,p) | \exists j \in I_s^r | p = \Theta_j^* x, \theta \in \mathbb{C}\} \subset \left\{ (x,p) | \left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} V_j & 0 \\ \hline 0 & -V_j \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \quad \forall V_j \in \mathbb{C}^{m \times m} \right\}.$$

A partir de cette observation, on en déduit immédiatement le résultat suivant :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & L_j^* R_j \\ \hline R_j^* L_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0 \quad \forall (x,p) : p = \Theta_j^* x, \tag{111}$$

ce qui est impliqué par :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & L_j^* R_j \\ \hline R_j^* L_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \tag{112}$$

pour tout (x,p) tel que :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} V_j & 0 \\ \hline 0 & -V_j \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0. \tag{113}$$

En appliquant la $\mathcal{S}$-procédure à (112) et (113), on obtient l'approximation de ((a) 101), dans le cas où $j \in I_s^c$, suivante :

$$\left[\begin{array}{c|c} Y_j - V_j & L_j^* R_j \\ \hline R_j^* L_j & V_j \end{array}\right] \succeq 0, \tag{114}$$

en la variable Y_j et la variable supplémentaire V_j.

– iii) Considérons le cas où $j \in I_f^c$: En effet ((a)101) est équivalente à :

$$x^* Y_j x + x^* L_j^* \Theta_j R_j x + x^* R_j^* \Theta_j^* L_j x \geqslant 0, \quad \forall x \in \mathbb{C}^m.$$

Déduction de l'immersion : Si on pose :

$$\left\{ \begin{array}{l} p = \Theta_j^* q \\ q = L_j x, \end{array} \right. \tag{115}$$

on obtient $p^*p = q^*\Theta_j^*\Theta_j q \leqslant q^*q$ du fait que $||\Theta_j||_{2,2} \leqslant 1$, ce qui est équivalent à

$$\left[\begin{array}{c} q \\ p \end{array}\right]^* \left[\begin{array}{c|c} I_{p_j} & 0 \\ \hline 0 & -I_{p_j} \end{array}\right] \left[\begin{array}{c} q \\ p \end{array}\right] \geqslant 0,$$

ce qui implique

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} L^*L & 0 \\ \hline 0 & -I_{p_j} \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0,$$

du fait que $q = L_j x$. A partir de ces observations, on peut énoncer le résultat suivant :

$$\{(x,p)|\exists j \in I_c^f | p = \Theta_j^* x, \Theta \in \mathcal{Z}_j\} \subset \left\{(x,p)| \left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} L_j^*L_j & 0 \\ \hline 0 & -I_{p_j} \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \right\}$$

par conséquent, on a :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & R_j^* \\ \hline R_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \forall (x,p) : p = \Theta_j^* x, \tag{116}$$

ce qui est impliqué par :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} Y_j & R_j^* \\ \hline R_j & 0 \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0, \tag{117}$$

pour tout (x,p) tel que :

$$\left[\begin{array}{c} x \\ p \end{array}\right]^* \left[\begin{array}{c|c} L_j^*L_j & 0 \\ \hline 0 & -I_{p_j} \end{array}\right] \left[\begin{array}{c} x \\ p \end{array}\right] \geqslant 0. \tag{118}$$

En appliquant la $\mathcal{S}$-procédure à (117) et (118), on obtient une approximation, dans le cas où $j \in I_f^c$, pour ((a) 101) :

$$\left[\begin{array}{cc} Y_j - \lambda L_j^* L_j & R_j^* \\ R_j & \lambda_j I_{p_j} \end{array}\right] \succeq 0,$$

en la variable Y_j et les variables supplémentaires positives λ_j, $j \in I_f^c$.

□

B.3 Conclusion

Durant cette présentation, on a présenté quelques méthodes pour transformer d'une façon équivalente ou approchée un problème homologue. La classification de tous les problèmes présentés jusqu'à maintenant s'articule sur les trois critères suivants :

- Représentation de l'ensemble des perturbations : supposée toujours cônique.
- Représentation de l'ensemble des données incertaines :
  - Sans perte de généralité, cet ensemble est convexe.
- structure de la perturbation :
  - Structurées,
  - Non structurées,
  - $\bigcap$-Ellipsoïdale,

On a pu constater que dans le cadre du paradigme de la robustesse, ces critères jouent un rôle central dans la résolution des problèmes d'optimisation. Ensuite, viennent les techniques d'optimisation que j'ai présentées dans la dernière section de ce document, permettant de déduire, systématiquement, des conditions suffisantes pour chaque classe de problèmes et éventuellement offrir des pistes vers les conditions nécessaires dans certains cas. Ces techniques se divisent en trois catégories principales :

- i) Procédés d'immersion [22],
- ii) (D,G)-scaling [34],
- iii) $\mathcal{S}$- procédure [4, 16, 9] (cf. annexe B, section B.1.5).

C Annexe C : Sur la complexité du problème d'optimisation stochastique *Two-stage*

C.1 Introduction et position du problème

On considère la classe de problèmes d'optimisation stochastiques dite problème d'optimisation *two-stage* linéaire avec recours. Le concept de programmation *two-stage* a été introduit dans les années 50s dans [3] et [19] et discuté dans de nombreuses publications ; on cite par exemple [14, 27, 39] ; une introduction facile et accessible est donnée dans [28]. Tels problèmes peuvent être écrits sous la forme suivante :

$$\min_{x\in X}\{f(x) := \mathbb{E}_\xi[F(x,\xi)]\}, \tag{119}$$

où

$$\begin{array}{l} X := \{x \in \mathbb{R}^n : Ax = b, x \geqslant 0\}, \\ F(x,\xi) := c^t x + Q(x,\xi), \\ Q(x,\xi) := \min\limits_{y\geqslant 0}\{q^t y | Wy = h + Tx\}. \end{array} \tag{120}$$

Ici $\xi := (q, h, T)$ sont les données du problème second-stage qui sont considérées comme des v.a. [11] (i.e. ξ est un vecteur de $\mathbb{R}^d$ dont les éléments sont ceux des vecteurs q et h et matrice T) ξ est alors une v.a. de Ω dans $\Xi \subset \mathbb{R}^d$ à densité de probabilité $\mathbb{P}_\xi$, où Ξ est le support de la v.a. ξ (i.e. le plus petit ensemble vérifiant $\mathbb{P}_\xi(\Xi) = 1$). On considère également que toutes les variables x, y, vecteurs c, b, q, h et matrices A, T, W sont de dimensions appropriées. $\mathbb{E}_\xi(.)$ dénote l'espérance mathématique par rapport à la v.a. ξ de loi de probabilité $\mathbb{P}_\xi$.
Cette classe de problème d'optimisation est un cas particulier d'une classe plus générale qui peut être écrite sous la même forme :

$$\min_{x\in X}\{g(x) := \mathbb{E}_\xi[G(x,\xi)]\}, \tag{121}$$

où G est une fonction de $\mathbb{R}^n \times \Xi$ dans $\mathbb{R}$. Il faut noter que la différence entre les deux classes de problèmes (119) et (121) est que la fonction $F(x,\xi)$, dans le cas du problème two-stage, n'est pas données explicitement, ce qui introduit certaines difficultés par rapport à la formulation générale (121). En effet, $F(x,\xi)$ peut être, pour x et ξ données, non bornée inférieurement, par conséquant $F(x,\xi) = -\infty$, i.e. pour une certaine solution faisable $x \in X$ et une possible réalisation de ξ, on peut améliorer la valeur de la fonction objectif du second stage $F(x,\xi)$ indéfiniment. On suppose alors que ce cas dégénéré a été évité lors de l'étape de modélisation, i.e. $F(x,\xi) > -\infty$ pour tout $(x,\xi) \in X \times \Xi$. On suppose également que le problème du second-stage est à recours relativement complet, i.e. pour tout $x \in X$, l'ensemble de contraintes du problème second stage est non vide avec une probabilité 1. Par conséquent $F(x,\xi) < +\infty$.

11. Dans cette annexe, l'abréviation v.a. désigne variable aléatoire ou vecteur aléatoire selon le contexte.

Quand on a modélisé notre problème d'optimisation sous la forme (119) on doit répondre aux deux questions de bases suivantes :

– (i) Est-ce que le problème (119) est bien défini ?

– (ii) Est-ce qu'on peut le résoudre efficacement (numériquement) ?

Ces deux questions doivent être traitées de façon inséparable : le résultat de modélisation, même s'il est satisfaisant, peut amener à un problème d'optimisation qu'on ne peut pas résoudre dans un temps raisonnable et/ou avec une précision raisonnable ; l'utilité d'un tel modèle doit être mise en cause.

La réponse à ces deux questions n'est pas évidente et dépend en général de la classe de problème considérée. Par ailleurs, et dans le même esprit que la question (i), deux autres questions viennent à l'esprit à savoir :

– (i')Qu'est- ce qu'on sait de la loi de probabilité $\mathbb{P}_\xi$: sans spécifier la loi de probabilité, on ne peut même pas formuler mathématiquement le problème. Dans certains cas cette loi peut être estimée avec une précision raisonnable si un historique des données du problème est disponible. Toutefois, dans certains cas cette loi peut changer en fonction du temps ou bien être assignée par un jugement subjectif...

– (ii') Pourquoi on optimise sur la moyenne : si la procédure d'optimisation est répétée plusieurs fois avec la même loi de probabilité $\mathbb{P}_\xi$, alors avec la loi des grands nombres, on peut justifier que cette répétition donne une solution optimale en moyenne. Par contre, si $\mathbb{P}_\xi$ varie, une décision sur la moyenne devient inutile.

Par rapport à la question (ii), la résolution du problème (119) dans le cas continu est délicate du fait que ceci requiert une intégration multi-dimensionnelle qui, du point de vue numérique rend l'évaluation, avec une grande précision, de la fonction objectif dans (119) impossible pour $d > 4$ [32]. Le cas discret est aussi difficile à résoudre avec une grande précision, du fait du nombre de scénarios à prendre en compte, qui explose exponentiellement. En effet, si la v.a. ξ a d éléments dont chacun peut prendre 3 valeurs possibles, alors le nombre de scénarios à prendre en compte est égal à 3^d ! ! Dans la section suivante on va exposer des résultats, plus explicites, de la littérature sur la complexité de la programmation two-stage.

C.2 Aperçu sur la complexité algorithmique de la programmation two-stage

La complexité du problème (119) se ramène à celle de la minimisation d'une fonction objective déterministe $f(x)$ donnée implicitement. Ce problème est au moins aussi difficile que celui de la minimisation d'une fonction objective explicite sur un domaine de faisabilité X. Le cas solvable dans ce cas est connu, c'est le cas de la programmation convexe, i.e. X est fermé et convexe et la fonction objective $f : X \longrightarrow \mathbb{R}$ est une fonction convexe. En effet, on sait que la classe des problèmes d'optimisation convexe peut être résolue en temps polynomial en fonction de la taille du problème, contrairement à la classe des problèmes non convexes qui est *NP-difficile*. Il faut noter que la propriété *NP-dur* ne concerne que les classes de problèmes d'optimisation (problème générique) ce qui veut

bien dire que la classe concernée peut contenir des problèmes qui peuvent être résolus particulièrement d'une façon efficace et c'est le cas de la classe d'optimisation two-stage (on montrera ça dans la suite du document). Cela étant, pour étudier la complexité de la classe d'optimisation two-stage, la première étape est l'étude de sa convexité. La bonne nouvelle est que cette classe de problèmes est en effet convexe. Cette propriété a été démontrée dans [28] et [39]. On a la proposition suivante :

Proposition 3. *Le problème d'optimisation (119) est convexe.*

Démonstration. On a pour tout $\xi \in \Xi$, la fonction $Q(.,\xi)$ est convexe [39, page28]. Il s'ensuit que la fonction $F(.,\xi)$ (120) est convexe pour tout $\xi \in \Xi$. Par conséquent la fonction $f(.) := \mathbb{E}_\xi[F(.,\xi)]$ (119) est convexe [28, page 31]. De plus on a l'ensemble de faisabilité first-stage X (120) est polytopique donc convexe. On en déduit que le programme (119) est convexe. □

Cette propriété est perdue si des contraintes d'intégralité (on se limite aux solutions faisables entières) ont été imposées [14]. Maintenant que la propriété de convexité a été établie, on passe à l'étude de la complexité algorithmique de (119). Plusieurs méthodes de résolution des problèmes d'optimisation two-stage sont basées sur le concept du problème équivalent déterministe. La première étape de telles méthodes consiste à formuler ce problème. Ceci est basé sur l'hypothèse qui fait que dans (119), les réalisations de la v.a. ξ sont spécifiées sous forme de scénarios et chaque scénario contient une description complète de la v.a. $\xi := (q, h, T)$ comme étant les valeurs que ces objets prennent en une seule réalisation. Ensuite les scénario sont énumérés $\xi^1, \ldots, \xi^K$ avec leurs probabilités respectives p_k où K le nombre possible de réalisations de ξ. Ainsi, l'équivalent déterministe à (119) s'écrit sous la forme :

$$\min c^t x + \sum_k^K p^k (q^k)^t y^k \tag{122}$$

$$\text{tel que : } \begin{cases} Ax = b, \\ Wy^k = T^k x + h^k, j = 1, \ldots, K \\ x \geqslant 0, y^k \geqslant 0, j = 1, \ldots, K. \end{cases}$$

Le problème (122) est $\#P$ difficile.

En effet, le résultat suivant a été démontré dans [23] :

Théorème 12 (Dyer *et al.* 2003 [23])**.** *La classe de problème d'optimisation two-stage avec une loi de distribution discrète est $\#P$ difficile.*

Ceci est dû au fait que le nombre de scénarios à prendre en compte pour évaluer, en un x donné, la valeur de $Q(x,\xi)$ (120) augmente exponentiellement en fonction de la taille da la v.a. ξ. Si on discrétise la loi de probabilité de ξ en seulement 4 points, déjà avec une taille du vecteur ξ égale à 20, on aura à faire à $4^{20} = 10^{12}$. Dans cette situation, il faut oublier d'évaluer l'espérance $\mathbb{E}_\xi[Q(x,\xi)]$.

Concernant le cas où la loi de distribution de ξ est continu, on a le résultat suivant :

Théorème 13 (Dyer *et al.* 2003 [23]). *L'évaluation* $\mathbb{E}_\xi[Q(x,\xi)]$ *(120) du problème two-stage (119) est* $\#P$ *difficile, même si* ξ *a une loi de probabilité uniforme* $[0,1]$.

La nature de la difficulté est la même que dans le cas discret, dans ce cas, celle-ci se ramène à calculer numériquement une intégrale multiple, ce qui est difficile. Pour plus de détails sur ce point voir [32]. Ces résultats sur la complexité rendent la tâche de dénombrement des solutions, que peut admettre un problème d'optimisation Two-stage, indécidable en temps polynomiale en fonction de la taille de ses données. Ceci indique qu'en général, les problèmes d'optimisation Two-stage ne peuvent être résolus avec une haute précision comme c'est le cas en optimisation déterministe[39]. Toutefois, dans certains cas, le problème (119) peut être résolu numériquement avec une précision machine. En effet, si on considère par exemple le cas de problèmes two-stage dits avec *recours simple* (ce cas particulier est traité dans [14, page 114]). Ce cas particulier correspond à $W = [I - I]$; on peut alors partitionner la variable de décision du second-stage y et q de la forme suivante : $y = (y+ \quad y-)^t$ et $q = (q+ \quad q-)^t$. On a :

$$Q_i(x,\xi) = q_i^+(h_i - T_i x)^+ + q_i^-(-h_i + T_i x)^+.$$

Si on suppose que seulement $h \in \mathbb{R}^n$ est aléatoire, alors on obtient :

$$Q_i(x,h_i) = q_i^+(h_i - T_i x)^+ + q_i^-(-h_i + T_i x)^+.$$

De plus $Q(x,\xi)$ sera décomposable en la somme suivante :

$$Q(x,\xi) = \sum_{i=1}^{n} Q_i(x,h_i),$$

avec $(.)_+ = \max\{.,0\}$ et $q^+ + q^- \geqslant 0$ pour tout $i = 1,\ldots,n$. et par suite

$$\mathbb{E}_h[Q(x,\xi)] = \sum_{i=1}^{n} \mathbb{E}_{h_i}[Q_i(x,h_i)],$$

le calcul de l'estimation multi-dimensionnelle de $\mathbb{E}_h[Q(x,\xi)]$ se réduit au calcul des estimations mono-dimensionnelles $\mathbb{E}_{h_i}[Q_i(x,h_i)]$ (largement suffisant de détails dans [27, 226]). De ce point de vue, il est clair que la façon dont les paramètres stochastiques du problème d'optimisation two-stage, et en général les problèmes d'optimisation stochastique, agissent considérablement sur la complexité de la situation. Ce cas reste très particulier et, en général, il n'y a pas de chance de résoudre les problèmes (119) et (122) avec une précision machine à cause des considérations que je viens de présenter. Toujours du point de vue précision des solutions obtenues et pour revenir sur la question (ii'), comme la loi de probabilité peut varier à cause de variabilité des données du problème d'optimisation, il peut s'avérer raisonnable d'optimiser dans notre problème initial (119), une sorte de somme pondérée de l'espérance avec un indice de dispersion statistique représentant cette variabilité, soit :

$$f(x) := \mathbb{E}_\xi[F(x,\xi)] + \lambda Var_\xi[F(x,\xi)],$$

où λ est une constante positive. Cette variante dite programmation stochastique *mean-risk* a été introduite dans les années 50s par Markowitz [29]. La mauvaise nouvelle est que cette sous-classe de problèmes est NP-difficile. En effet, on trouve le résultat suivant dans [1] :

Théorème 14 (Ahmed 2006 [1]). *Le problème d'optimisation stochastique mean-risk*

$$\min\{\mathbb{E}_\xi[F(x,\xi)] + \lambda Var_\xi[F(x,\xi)], x \in X\}$$

correspondant au problème (119) avec recours simple est NP-dur pour tout $\lambda > 0$.

La difficulté ici, vient du fait que l'utilisation d'un critère de variance fait perdre la propriété de la convexité au problème (119).

Compte tenu de la complexité de la classe de problèmes two-stage, la résolution efficace de tels problèmes est en général très difficile, ce qui influence considérablement la qualité des solutions calculées par les différentes méthodes de résolution. De plus, et du point de vue pratique, les erreurs dues à la modélisation, à l'estimation de lois de probabilité, etc, sont plus importantes que celles dues à l'optimisation. Par conséquent, chercher à résoudre un problème d'optimisation stochastique avec une grande précision n'a pas de sens. Pour plus de discussions voir Nesterov et al. [31], Shapiro [37], Nemirovski et al. [40]. Il faut alors chercher une approximation à notre problème (119) qui permet d'avoir des solutions avec une précision raisonnable. Dans la section suivante je vais présenter une approche d'approximation basée sur la technique de simulation de Monte Carlo dite méthode *Sample Average Approximation.*

C.3 La méthode d'approximation *Sample Average Approximation.*

C.3.1 Introduction et propriétés de convergence

Supposons qu'on peut générer des échantillons aléatoires $\xi^1, \ldots, \xi^N$ de N réalisations du vecteurs aléatoires ξ. Supposons aussi que ξ^j, $j = 1, \ldots, N$ sont iid (indépendants identiquement distribués), i.e. ont la même loi de probabilité et indépendants, et considérons la fonctions *sample average*

$$\hat{f}_N := \frac{1}{N}\sum_{i=1}^{n} F(x,\xi^j). \tag{123}$$

La fonction $\hat{f}_N$ dépend des échantillons générés ξ^j et est, par conséquent ,aléatoire. Pour tout $x \in X$ cette fonction est un estimateur non biaisé de $f(x)$, i.e, $\mathbb{E}[\hat{f}_N(x)] = f(x)$ et par la loi des grands nombres on démontre que $\hat{f}_N(x)$ tend vers $f(x)$ avec une probabilité un quand $N \to \infty$. (En fait, cette convergence est uniforme sur tout sous- ensemble compact C de X. Voir conditions [39, chapitre 5]). De plus, par le théorème de la limite

centrale on démontre que pour tout $x \in X$, $N^{1/2}(\hat{f}_N(x) - f(x))$ converge en distribution de probabilité vers la loi normale $\mathcal{N}(0, \sigma^2(x))$, avec $\sigma^2(x) = Var[F(x, \xi)]$ (par contre la vitesse de convergence de $\hat{f}_N(x)$ vers $f(x)$ est notoirement lente : de l'ordre de $O(N^{-1/2})$ [37]). Compte tenu de toutes ces observations, pour tout $x \in X$, on peut estimer la valeur de $f(x)$ en calculant la moyenne des valeurs $F(x, \xi^j)$, $j = 1, \ldots, N$. Ceci nous amène à approximer notre problème initial (119) par le problème dit *Sample Average Approximation* :

$$\min_{x \in X} \left\{ \hat{f}_N(x) := \frac{1}{N} \sum_{i=1}^{n} F(x, \xi^j) \right\}, \tag{124}$$

d'autant plus que la valeur optimale $\hat{v}_N$ et l'ensemble des solutions optimales $\hat{S}_N$ de (124) sont des estimateurs consistants [12] de leurs homologues v et S dans le problème (119). En effet, les propriétés de convergence des deux estimateurs $\hat{v}_N$ et $\hat{S}_N$ ont été étudiées dans [37] et on a en particulier le résultat suivant :

Théorème 15 (Sharpio *et al.* 2007 [37])**.** *Supposans qu'il existe un compacte compact $C \in \mathbb{R}^n$ tel que :*

- *(i) L'ensemble S des solutions optimales de (119) est non vide et inclus dans C,*
- *(ii) la fonction $f(x)$ est continue et à valeurs finies sur C,*
- *(iii) $\hat{f}_N(x)$ converge, uniformément en $x \in C$, vers $f(x)$ quand $N \to \infty$,*
- *(iv) avec probabilité 1 et pour N suffisamment grand $\hat{S}_N$ est non vide et $\hat{S}_N \subset C$.*

Alors, $\hat{v}_N \to v$ et $\mathbb{D}(\hat{S}_N, S) \to 0$ [13] avec probabilité 1 quand $N \to \infty$.

L'assertion $\mathbb{D}(\hat{S}_N, S) \to 0$ avec probabilité 1 veut dire que pour toute sélection $\hat{x}_N \in \hat{S}_N$, on a $dist(\hat{x}_N, S) \to 0$ avec probabilité 1. Si de plus on a $S = \{\bar{x}\}$ un singleton, i.e, le problème (119) a une solution unique $\bar{x}$, alors cela veut dire $\hat{x}_N \to \bar{x}$ avec probabilité 1 ; la vitesse de cette convergence est de l'ordre de $O(N^{-1/2})$ [36]. De plus, on a [35] :

$$\hat{v}_N = \min_{x \in S} \hat{f}_N(x) + o(N^{-1/2})$$

avec probabilité 1. Ce qui veut dire, dans le cas où $S = \{\bar{x}\}$ est un singleton, $\hat{v}_N$ converge vers v avec la même vitesse de convergence de $\hat{f}_N(x)$ vers $f(x)$.

D'après la discussion ci-dessus, il est clair que la méthode SAA (*Sample Average Approximation*) a des propriétés de convergence lentes : par exemple la vitesse de convergence de $\hat{f}_N(x)$ vers $f(x)$ est de l'ordre de $O(N^{-1/2})$. Par conséquent, pour améliorer la précision de l'estimateur $\hat{f}_N(x)$ par un digit, on a besoin d'une taille d'échantillonnage N 100 fois plus grande. Cette convergence lente a été héritée par l'estimateur $\hat{v}_N$ aussi. Pour pallier ce problème, Shapiro *et al.* propose une amélioration nettement plus efficace (vitesse de convergence exponentielle) basée sur la théorie des larges déviations (pour discussion générale sur cette théorie voir [21].) Voir également [27, page351] pour une accessible introduction sur la méthode SAA.

12. $\hat{\theta}_N$ est un estimateur consistant de θ si $\hat{\theta}_N$ converge vers θ avec probabilité 1 quand $N \to \infty$.
13. $\mathbb{D}(A, B) := \sup_{x \in A} dist(x, B)$ est appelé déviation de A dans B.

C.3.2 Vitesse de convergence exponentielle des estimateurs de la méthode SAA.

Comme la résolution exacte du problème (119) est en général impossible, on doit le remplacer par le problème suivant :

$$\text{Trouver} \quad x \in X \quad \text{t.q} : f(x) - v \leqslant \epsilon, \tag{125}$$

avec $v = f(x^*)$ est la valeur optimale du problème initial (119), et x^* est sa solution.

Une telle solution x^* est dite ϵ optimale pour le problème (119). Les résultats présentés dans cette section sont basés sur la notion des ensembles ϵ optimaux.
On définit les deux ensembles S^ϵ et $\hat{S}_N^\delta$:

$$S^\epsilon := \{x \in X : f(x) \leqslant v + \epsilon\}, \quad \hat{S}_N^\delta := \{x \in X : \hat{f}_N(x) \leqslant \hat{v}_N + \delta\},$$

où S^ϵ et $\hat{S}_N^\delta$ est l'ensemble des solutions ϵ (resp. δ) de (119) (resp. (124)).

L'objectif est d'étudier l'événement aléatoire $\hat{S}_N^\delta \subset S^\epsilon$ et de calculer une borne sur la probabilité de cet événement $Pr(\hat{S}_N^\delta \subset S^\epsilon)$ qui va dépendre à priori de N, ϵ, δ. On fait les hypothèses suivantes :

Théorème 16 (Shapiro *et al.* 2009 [39])**.** *On suppose que l'ensemble faisable X du problème (119) est fini et note $|X|$ son cardinal. Soit ϵ et δ deux nombres positifs. Alors*

$$1 - Pr(\hat{S}_N^\delta \subset S^\epsilon) \leqslant |X| e^{-N\eta(\epsilon,\delta)},$$

où

$$\eta(\epsilon, \delta) := \min_{x \in X \setminus S^\epsilon} I_x(-\delta). \tag{126}$$

Démonstration. Du moment que X est fini, les ensembles S^ϵ et $\hat{S}_N^\delta$ sont non vides et finis. Soit $\epsilon \geqslant 0$ et $\delta \in [0, \epsilon]$, considérons l'événement $\{\hat{S}_N^\delta \subset S^\epsilon\}$. Cet événement signifie que toute solution δ optimale de (124) est ϵ optimale pour (119). Estimant maintenant la probabilité de cet événement. On peut écrire

$$\begin{aligned}
\{\hat{S}_N^\delta \not\subset S^\epsilon\} &= \cup_{x \in X \setminus S^\epsilon} \left\{ \hat{f}_N(x) \leqslant \inf_{y \in X} f_N(y) + \delta \right\}, \\
&= \cup_{x \in X \setminus S^\epsilon} \left\{ \hat{f}_N(x) \leqslant f_N(y) + \delta, \forall y \in X \right\}, \\
&= \cup_{x \in X \setminus S^\epsilon} \cap_{y \in X} \left\{ \hat{f}_N(x) \leqslant f_N(y) + \delta \right\},
\end{aligned}$$

et par conséquent,

$$Pr(\hat{S}_N^\delta \not\subset S^\epsilon) \leqslant \sum_{x \in X \setminus S^\epsilon} Pr\left(\cap_{y \in X} \left\{ \hat{f}_N(x) \leqslant f_N(y) + \delta \right\} \right).$$

Considérons maintenant une application $u : X \setminus S^\epsilon \to X$. Si $X \setminus S^\epsilon$ est vide, alors toute solution faisable $x \in X$ pour (119) est ϵ optimale pour (119). On suppose alors que $X \setminus S^\epsilon$ est non vide. On a alors, en particulier,

$$Pr(\hat{S}_N^\delta \not\subset S^\epsilon) \leqslant \sum_{x \in X \setminus S^\epsilon} Pr\left\{\hat{f}_N(x) \leqslant f_N(u(x)) + \delta\right\}.$$

On choisit l'application u telle que

$$f(u(x)) \leqslant f(x) - \epsilon^* \quad \forall x \in X \setminus S^\epsilon,$$

où $\epsilon^* \geqslant \epsilon$. Cette application existe toujours. En effet, soit $x \in X \setminus S^\epsilon$, on a $\exists y \in X : f(x) > f(y) + \epsilon$ et si on choisit $\epsilon^* > \epsilon$, on aura $f(x) \geqslant f(y) + \epsilon^*$, par suite $\epsilon^* \leqslant f(x) - f(y) \leqslant f(x) - \inf_{y \in X} f(y), \forall x \in X \setminus S^\epsilon$; on choisit alors $\epsilon^* = \min_{x \in X \setminus S^\epsilon} f(x) - v$. Par ailleurs, s'il existe $y_1, \ldots, y_k \in X$ tq $f(x) \geqslant f(y_i) + \epsilon^*, i = 1, \ldots, k$ on pend par exemple $y_0 = \max\{y_1, \ldots, y_k\}$. Dans ce cas, on peut définir u comme étant l'application qui à x associe y_0 et cette application est bien définie. On constate que le choix de l'application u n'est pas unique. Posons

$$Y(x, \xi) := F(u(x), \xi) - F(x, \xi). \tag{127}$$

Notons que $\mathbb{E}_\xi[Y(x, \xi)] = f(u(x)) - f(x)$, et par suite $\mathbb{E}_\xi[Y(x, \xi)] \leqslant -\epsilon^*$ pour tout $x \in x \in X \setminus S^\epsilon$. La moyenne empirique correspondante est

$$\hat{Y}_N(x) := \frac{1}{N} \sum_{j=1}^{N} Y(x, \xi^j) = \hat{f}_N(u(x)) - \hat{f}_N(x).$$

Par conséquent on a

$$Pr(\hat{S}_N^\delta \not\subset S^\epsilon) \leqslant \sum_{x \in X \setminus S^\epsilon} Pr\left(\hat{Y}_N(x) \geqslant -\delta\right).$$

On définit la fonction $I_x(.)$ par

$$I_x(z) := \sup_{t \in \mathbb{R}} \{tz - \log\left(\mathbb{E}_\xi[e^{tY(x,\xi)}]\right)\}.$$

On a $Pr\left(\hat{Y}_N(x) \geqslant -\delta\right) = Pr\left(e^{t\hat{Y}_N(x)} \geqslant e^{-t\delta}\right)$, par conséquent en appliquant l'inégalité de Chebyshev au second membre de cette égalité on obtient $Pr\left(\hat{Y}_N(x) \geqslant -\delta\right) \leqslant e^{t\delta}\mathbb{E}_\xi[e^{t\hat{Y}_N(x)}]$. Or $\mathbb{E}_\xi[e^{t\hat{Y}_N(x)}] = [\mathbb{E}_\xi[e^{t/NY(x,\xi)}]]^N$. En passant au logarithme dans les deux membres de l'inégalité et en faisant le changement de variable $t' = t/N$ et en minimisant sur t'> 0, on obtient

$$\frac{1}{N} \ln\left(Pr\left(\hat{Y}_N(x) \geqslant -\delta\right)\right) \leqslant -I_x(-\delta).$$

Par suite

$$Pr(\hat{S}_N^\delta \not\subset S^\epsilon) \leqslant \sum_{x \in X \setminus S^\epsilon} e^{-NI_x(-\delta)} \leqslant |X| e^{-N\eta(\delta,\epsilon)},$$

où $\eta(\epsilon,\delta) := \min_{x \in X \setminus S^\epsilon} I_x(-\delta)$. □

Corollaire 1. *Dans le cas où, $\delta < \epsilon^*$ et si $\mathbb{E}_\xi[e^{t/NY(x,\xi)}]$ est à valeurs finies au voisinage de $t = 0$, alors $\eta(\epsilon,\delta) > 0$.*

Démonstration. Si $\mathbb{E}_\xi[e^{t/NY(x,\xi)}]$ est à valeurs finies au voisinage de $t = 0$ alors $\mathbb{E}_\xi[Y(x,\xi)]$ est finie, et si de plus $\delta < \epsilon^*$ alors $-\delta > \mathbb{E}_\xi[Y(x,\xi)]$ (car $\mathbb{E}_\xi[Y(x,\xi)] \leqslant -\epsilon^*$) et puisque $I_x(.)$ est strictement croissante ([39, 388]), par conséquent $I_x(-\delta) > I_x(\mathbb{E}_\xi[Y(x,\xi)])$. Or $I_x(\mathbb{E}_\xi[Y(x,\xi)]) = 0$ (en effet, par l'inégalité de Jensen on a $\ln\left(\mathbb{E}_\xi[e^{tY(x,\xi)}]\right) \geqslant \ln\left(e^{t\mathbb{E}_\xi[Y(x,\xi)]}\right) = t\mathbb{E}_\xi[Y(x,\xi)]$. Par conséquent $I_x(\mathbb{E}_\xi[Y(x,\xi)]) \leqslant 0$. Or $I_x(.)$ est positive sur $\mathbb{R}$ (remplacer par $t = 0$ dans la définition de $I_x(.)$.)) Ainsi $I_x(-\delta) > 0$ et par suite $\eta(\epsilon,\delta) > 0$. □

Remarque :sous les mêmes hypothèse du corollaire 1, on a

$$I_x(\mathbb{E}_\xi[Y(x,\xi)]) = I'_x(\mathbb{E}_\xi[Y(x,\xi)]) = 0, I''_x(\mathbb{E}_\xi[Y(x,\xi)]) = 1/\sigma_x^2$$

où $\sigma_x^2 = Var[Y(x,\xi)]$ [38] (le résultat est obtenu par une simple application du théorème des fonction implicite). Par conséquent, on peut calculer une approximation de $I_x(\mathbb{E}_\xi[Y(x,\xi)])$ par un développement en série de Taylor d'ordre 2 au voisinage $\mathbb{E}_\xi[Y(x,\xi)]$ comme suit : $I_x(\mathbb{E}_\xi[Y(x,\xi)]) = \frac{(-\delta-\mathbb{E}_\xi[Y(x,\xi)])^2}{2\sigma_x^2} + o((-\delta-\mathbb{E}_\xi[Y(x,\xi)])^2) \geqslant \frac{(\epsilon-\delta)^2}{2\sigma_x^2}$. On obtient finalement une borne inférieure sur $\eta(\epsilon,\delta) \geqslant \frac{(\epsilon-\delta)^2}{2\sigma_x^2}$.

Corollaire 2. *S'il existe $\sigma > 0$ telle que pour tout $x', x \in X$, la fonction génératrice des moments*[14] *$M^*(t)$ de $F(x',\xi) - F(x,\xi) - \mathbb{E}_\xi[F(x',\xi) - F(x,\xi)]$ satisfait :*

$$M^*(t) \leqslant exp(\frac{\sigma^2 t^2}{2}), \quad \forall t \in \mathbb{R}, \tag{128}$$

alors

$$\eta(\delta,\epsilon) \geqslant \frac{(-\delta-\mu_x)^2}{2\sigma^2} \geqslant \frac{(\epsilon-\delta)^2}{2\sigma^2},$$

où $\mu_x = \mathbb{E}_\xi[Y(x,\xi)]$.

Démonstration. Sous l'hypothèse du corollaire et pour $x' = u(x)$, la v.a. $F(x',\xi) - F(x,\xi)$ coïncide avec $Y(x,\xi)$ (voir (127)). (128) implique alors : $M_x(t) \geqslant exp(\mu_x t + \frac{\sigma^2 t^2}{2})$. Par suite on a : $I_x(z) \geqslant \sup_{t \in \mathbb{R}}\{zt - \mu_x t - \frac{\sigma^2 t^2}{2}\} = \frac{(z-\mu_x)^2}{2\sigma^2}$ et par conséquent, par définition de $\eta(\epsilon,\delta)$ (126) on a pour tout $\epsilon > 0$ et $\delta \in [0,\epsilon)$, le résultat du corollaire est vrai. □

14. La fonction génératrice de la v.a. Y est définie par $M(t) := \mathbb{E}_\xi[e^{tY}]$.

Remarque : :

D'après le corollaire 2 une estimation du nombre d'échantillons N, peut être écrit comme suit :

$$N \geqslant \frac{2\sigma^2}{(\epsilon - \delta)^2} \ln(\frac{|X|}{\alpha}),$$

où $\alpha = Pr(\hat{S}_N^\delta \not\subset S^\epsilon)$. Ce qui veut dire que ce N garantit que la probabilité de l'événement $\{\hat{S}_N^\delta \not\subset S^\epsilon\}$ soit égale à $1 - \alpha$ au moins.

Ces résultats peuvent être étendus au cas où X est continu et borné, en rajoutant certaines conditions techniques et calculatoires [36, 40, 37].

Références

[1] S. Ahmed. Convexity and decomposition of mean-risk stochastic programs. *Math. Program.*, 106(3) :433–446, May 2006.

[2] S. Arora and B. Barak. *Computational Complexity : A Modern Approach*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[3] E. M. L. Beale. On minimizing a convex function subject to linear inequalities. *J Royal Statistical Society*, 17(2) :173–184, 1955.

[4] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[5] A. Ben-tal, L. El Ghaoui, and A. Nemirovski. Robust semidefinite programming. In *Handbook on Semidefinite Programming, Kluwer Academis Publishers*, pages 139–162, 1998.

[6] A. Ben-Tal and A. Nemirovski. Robust optimization - methodology and applications. *Math. Program.*, (3) :453–480.

[7] A. Ben-tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23 :769–805, 1998.

[8] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Oper. Res. Lett.*, 25(1) :1–13, August 1999.

[9] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization : analysis, algorithms, and engineering applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.

[10] A. Ben-Tal, A. Nemirovski, and C. Roos. Robust solutions of uncertain quadratic and conic-quadratic problems. *SIAM Journal on Optimization*, (2) :535–560.

[11] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to probability*. Athena Scientific, Belmont (Mass.).

[12] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Rev.*, 53(3) :464–501, August 2011.

[13] D. Bertsimas and J. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

[14] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*, volume 49. Springer, 1997.

[15] F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization : Theoretical and practical aspects*. Universitext. Springer, Berlin.

[16] S. Boyd and L. Vandenberghe. *Convex Optimization*, volume 25. Cambridge University Press, 2010.

[17] J. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, 25(9) :772–781, 1978.

[18] D. Chaerani. Recipes for building the dual of conic optimization problem. *Journal of indonesian mathematical society*, 16(1) :9–23, 2010.

[19] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3-4) :197–206, 1955.

[20] G. B. Dantzig and M. N. Thapa. *Linear Programming. 2. , Theory and extensions.* Springer series in operations research. Springer, New York.

[21] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Applications of mathematics. Springer, New York, Berlin, Heidelberg, 1998.

[22] M. Dinh. *Synthèse dependant de paramètres par optimisation LMI de dimension finie : application a la synthèse de correcteurs re réglables.* PhD thesis, Thèse de doctorat de l'Université de Caen Basse-Normandie, UFR de Sciences, Ecole doctorale SIMEM, 2005.

[23] M. Dyer and L. Stougie. Computational complexity of stochastic programming problems. *Math. Program.*, 106(3) :423–432, May 2006.

[24] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., New York, NY, USA, 1990.

[25] A. Goelzer. *Emergence de structures modulaires dans les régulations des systèmes biologiques : théorie et applications à Bacillus subtilis.* These, Ecole Centrale de Lyon, November 2010.

[26] A. Goelzer, V. Fromion, and G. Scorletti. Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6) :1210–1218, 2011.

[27] P. Kall and J. Mayer. *Stochastic Linear Programming : Models, Theory, and Computation.* International series in operations research & management science. Springer, 2011.

[28] P. Kall and S.W. Wallace. *Stochastic programming.* Wiley-Interscience series in systems and optimization. Wiley, 1994.

[29] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1) :77–91, 1952.

[30] Y. Nesterov. *Introductory lectures on convex optimization : a basic course.* Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.

[31] Y. Nesterov and J.P. Vial. Confidence level solutions for stochastic programming. *CORE Discussion Papers*, 2000 :00–2000, 2000.

[32] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[33] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Rev.*, 49(3) :371–418, July 2007.

[34] G. Scorletti, X. Bombois, M. Barenthin, and V. Fromion. Improved efficient analysis for systems with uncertain parameters, 2007.

[35] A. Shapiro. Asymptotic analysis of stochastic programs. *Ann. Oper. Res.*, 30(1-4) :169–186, June 1991.

[36] A. Shapiro. Monte carlo simulation approach to stochastic programming. In *Proceedings of the 33nd conference on Winter simulation*, WSC '01, pages 428–431, Washington, DC, USA, 2001. IEEE Computer Society.

[37] A. Shapiro. Stochastic programming approach to optimization under uncertainty. *Math. Program.*, 112(1) :183–220, July 2007.

[38] A. Shapiro, T. Homem de Mello, and J. Kim. Conditioning of convex piecewise linear stochastic programs. *Math. Program.*, 94(1) :1–19, 2002.

[39] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming : modeling and theory.* MPS-SIAM series on optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[40] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems, 2004.

[41] R. Tempo, G. Calafiore, and F Dabbene. *Randomized algorithms for analysis and control of uncertain systems.* Springer, Berlin, 2004.

[42] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38 :49–95, 1994.

Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

Mémoire doctorant 1ère année

2012 -2013

Nom - Prénom	**AMEUR - Omar**
Titre de la thèse	**Commande et stabilité des systèmes commutés: Applications Génie Electrique et Fluid Power**
Directeur de thèse	**G. SCORLETTI**
Co- encadrants	**X. BRUN, M. SMAOUI, P. MASSIONI, A. HIDJAZI**
Dpt. de rattachement	**Méthodes pour l'Ingénierie des Systèmes**
Date début des travaux	**01 Octobre 2012**
Type de financement	**Bourse ministérielle**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

Table des matières

Notations

a	Accélération linéaire $[m.s^{-2}]$
A	Matrice d'état
b_v	coefficient de frottement visqueux [N/(m/s)]
I_0	L'ensemble des indices de cellules qui incluent l'origine
I_1	L'ensemble des indices de cellules qui n'incluent pas l'origine
X_i	La i$^{\text{ème}}$ cellule
F_{ext}	Force extérieure $[N]$
F_f	Force des frottements $[N]$
M	Masse $[Kg]$
k	Coefficient polytropique
P_P	Pression dans la chambre P $[Pa]$
P_N	Pression dans la chambre N $[Pa]$
q_m	Débit massique $[kg.s^{-1}]$
r	Constante des gaz parfaits relative à l'unité de masse $[J.kg^{-1}.K^{-1}]$
$\mathbb{R}^n$	L'ensemble des vecteurs réels de dimension n
S_P	Section du piston dans la chambre P $[m^2]$
S_N	Section du piston dans la chambre N $[m^2]$
T	Température absolue $[K]$
u	Tension $[V]$
v	Vitesse $[m.s^{-1}]$
v_d	Vitesse désirée $[m.s^{-1}]$
V_p	Volume dans la chambre P $[V^3]$
V_N	Volume dans la chambre N $[V^3]$
y	Position $[m]$
y_d	Position désirée $[m]$
y_{stop}	Position à l'équilibre $[m]$
τ	Constante de temps $[s]$

Introduction générale

Les systèmes électropneumatiques sont très utilisés dans l'industrie. Leur bonne utilisation dépend de la maîtrise de leur commande, ce qui constitue un axe de recherche au laboratoire Ampère. Différentes lois de commandes ont été synthétisées afin de répondre aux différents cahiers des charges testés sur les "benchmark"'du laboratoire.

En effet, le vérin éléctropneumatique est parmi les "benchmark" les plus utilisés au laboratoire, pour lequel des lois de commandes en position et en pression ont été synthétisées. Cette application pneumatique souffre d'un problème majeur dit "phénomène de redécollage" lors de l'application de ces lois de commandes. Il se traduit par un mouvement de secousses saccadées quand le système est à l'arrêt et constitue un problème crucial pour la mise en œuvre industrielle de cette technologie. Afin d'éviter ce phénomène, des lois de commandes commutées ont été synthétisées et proposées par [Tur10]. Cette approche a été implémentée sur le vérin électropneumatique, les résultats expérimentaux ont été très satisfaisants et le phénomène de redécollage a été évité. La stabilité pour le vérin électropneumatique avec des lois de commandes commutées a été vérifiée en se basant sur la simulation de ce dernier. Cependant, il n'y a pas de démonstration mathématique qui a été fait pour la démontrer. Il ne faut pas perdre de vue que le fait de simuler un système à partir d'un état initial n'est pas suffisant pour évaluer le comportement globale du système car à chaque point initial correspond une trajectoire différente et il faudrait simuler à partir d'une infinité de points initiaux pour étudier correctement le comportement du système en générale, ce qui est impossible à réaliser. C'est en ce sens que ce travail de recherche a été mené, pour répondre à la problématique suivante : prenant des systèmes à commutation, quels sont les verrous scientifiques qui empêchent l'étude de la stabilité ? Comment alors mettre en œuvre des méthodes ou des outils efficaces aptes à les débloquer ?

Afin d'analyser la stabilité, il faut disposer d'un modèle suffisamment détaillé, tel qu'il soit capable de prendre en compte à la foi l'effet non linéaire du vérin éléctropneumatique à cause des frottements et l'effet de la commutation entre les lois de commandes. Les systèmes linéaires par morceaux sont une classe des systèmes non linéaires qui permettent de répondre à ces contraintes. Dans la littérature, plusieurs travaux on été développés sur cette classe de systèmes. Parmi ces travaux, il y a [JR98] qui permet d'étudier le problème de la stabilité en le transformant en un problème d'optimisation sous contraintes LMI (Linear Matrix Inequality). L'optimisation convexe sous contraintes LMI apparaît actuellement comme une des plus larges classes d'optimisation convexe pour laquelle on dispose d'algorithmes de résolution efficace proposés dans les logiciels de calcul scientifique généraux comme Matlab ou Scilab et qui a eu d'importantes applications en Sciences de l'ingénieur [BGFV94]. Les problèmes d'optimisation convexe apparaissent comme une sous classe de problèmes d'optimisation "faciles", c'est-à-dire dotée d'algorithmes de résolution en temps polynomial.

L'objectif des travaux de [JR98] est de développer des méthodes pour la construction de fonctions de Lyapunov qui démontrent la stabilité pour une classe des systèmes non linéaires qui sont les systèmes dynamiques affines par morceaux. D'après [JR98], c'est naturel de prendre

les fonctions de Lyapunov quadratiques par morceaux qui semblent être une puissante extension de la stabilité quadratique pour démontrer la stabilité de cette classe de systèmes. Le calcul de ces fonctions de Lyapunov quadratiques par morceaux peut être traduit en utilisant le principe de la S-procédure à la résolution d'un problème d'optimisation sous contraintes LMI.

Cependant, nos premiers travaux de thèse montre que l'application de la méthode de [JR98] pour démontrer la stabilité du vérin éléctropneumatique avec des lois de commande commutées ne donne pas des résultats satisfaisants et cela pour la raison suivante :

- Les critères de la solution présentée dans [JR98] sont suffisantes et ne sont pas nécessaires : il existe des systèmes stables pour lesquels la méthode de [JR98] ne marche pas (problème de conservatisme).

L'objectif de ce travail de thèse est donc de travailler à améliorer le résultat de [JR98] en essayant d'enlever le conservatisme, de manière à trouver des outils systématiques de validations des lois des commandes et éventuellement des méthodes de synthèse.

Pour avoir donc une meilleur compréhension du problème et de la méthode appliquée, notre travail est subdivisé de la façon suivante :

- Dans le Chapitre 1, une description du système électropneumatique du laboratoire AMPERE sera développée en décrivant la problématique de la commutation au niveau des commandes synthétisées, les hypothèses et ses différentes caractéristiques.
- Dans le Chapitre 2, nous développerons la méthode d'analyse choisie pour analyser la stabilité et qui se base sur les idées développées dans [JR98]. Dans cette partie, nous allons voir que le résultat de [JR98] ne donne pas des résultats satisfaisants.
- Dans le Chapitre 3 , nous discuterons les étapes suivis pour améliorer le résultat de [JR98] et les résultats trouvés au cours de nos travaux.

Chapitre 1

Le domaine “Fluid Power” : Actionneur Electropneumatique

1.1 Introduction

L’objectif de ce chapitre est double, une introduction à l’application pneumatique et la présentation de la problématique.
Dans un premier temps, il s’agit de faire une petite description du procédé électropneumatique, sa modélisation et ses caractéristiques. Puis nous décrirons la problématique rencontrée sur ce procédé liée au phénomène de redécollage en proposant ensuite une solution issue des travaux de [Tur10]. La dernière partie sera consacrée à l’étude de la stabilité de la solution proposée en éclaircissant la difficulté et les obstacles qui empêchent cette étude.

1.2 Description du procédé

Nous pouvons représenter les différents éléments et les étages de puissance qui constituent un système électropneumatique ou tout système “Fluid Power” d’une façon générale donnée dans la figure suivante (Fig 1.1) [Bru99]

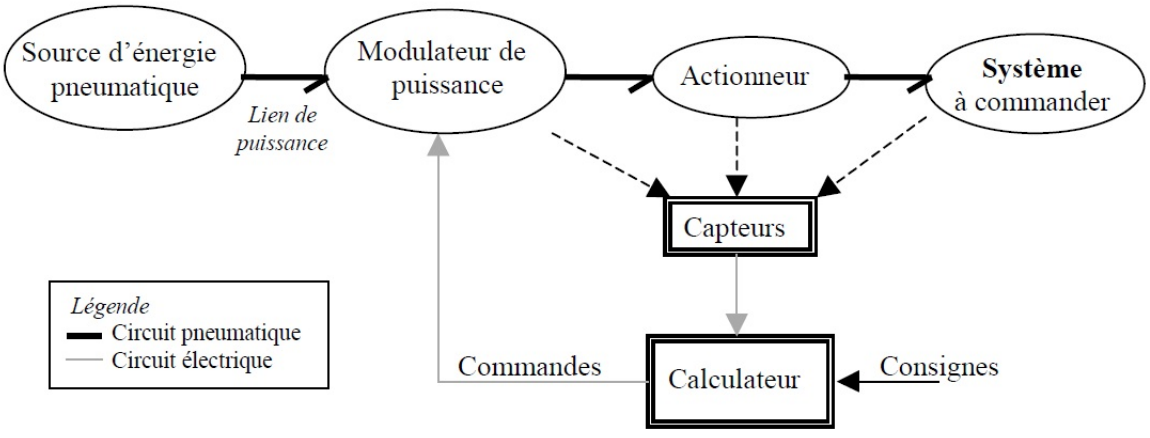


FIGURE 1.1 – Système Electropneumatique : Principe de base

D’après cette figure et dans le but de répondre au cahier des charges imposé par l’utilisateur que ce soit en temporelle ou fréquentielle, l’asservissement électropneumatique de position est réalisé via un actionneur pneumatique qui entraine une charge et qui utilise l’énergie délivrée

via le modulateur de puissance. Des capteurs sont mis en place pour délivrer des informations sur la position, l'accélération, les pressions ...).

La figure (1.2) présente le banc d'essai du Laboratoire Ampère à l'INSA qui est destiné à des applications en mouvements rectilignes comme le positionnement d'une charge à masse variable.

Le banc d'essai peut être alimenté de par l'air comprimé et contient deux servodistributeurs pour réguler le débit fourni aux deux chambres du vérin, il présente un actionneur sous la forme d'un vérin linéaire pneumatique double effet avec une tige qui relie un chariot guidé sur rail.

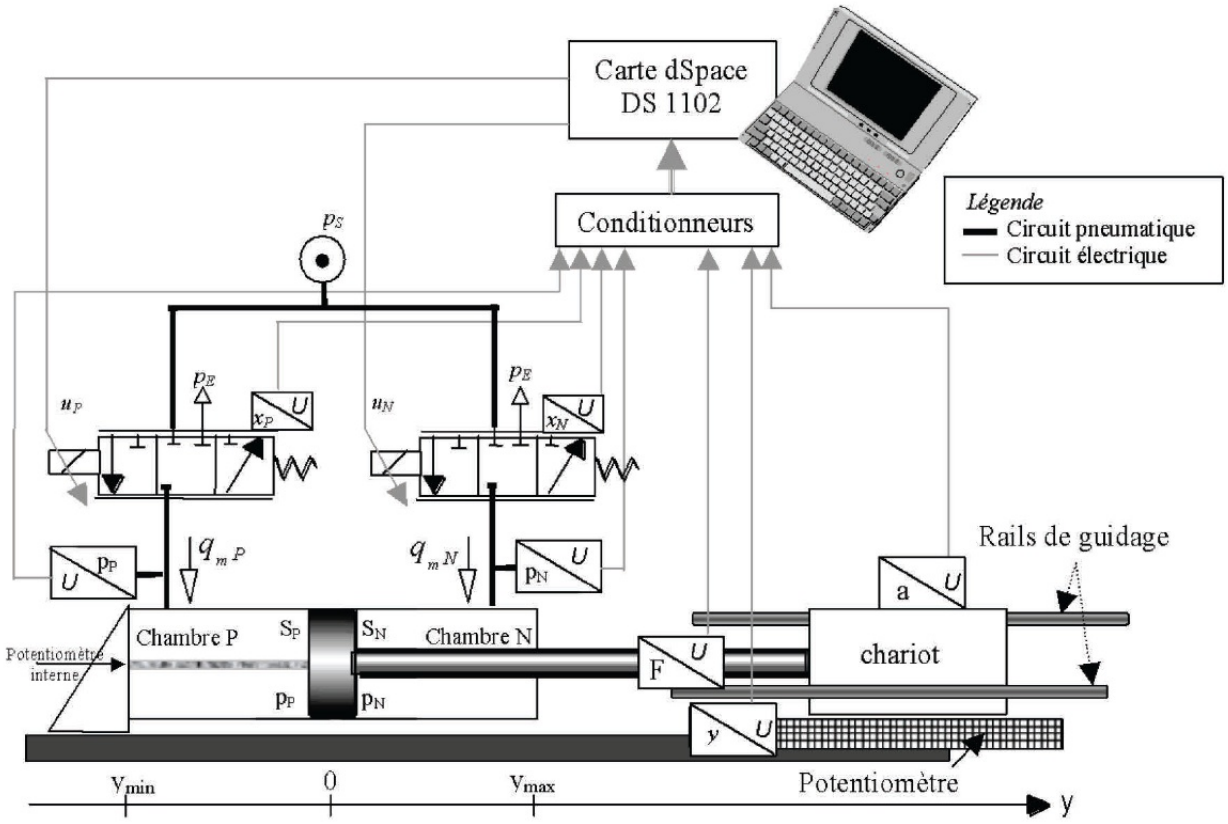


FIGURE 1.2 – Système Electropneumatique : Banc d'essai

1.3 Modélisation

Le modèle non linéaire du système électropneumatique après les hypothèses simplificatrices est donné comme suit [Bru99] [Tur10] :

$$\begin{cases} \dot{y} = v \\ \dot{v} = \frac{1}{M}(S_p P_p - S_N P_N - b_v v - F_f(v) - F_{sec}) \\ \dot{P}_N = \frac{krT}{V_N(y)}[\phi(P_N) + \frac{S_N}{rT} P_N v] + \frac{krT}{V_N(y)}\psi(P_N, sgn(u_N))u_N \\ \dot{P}_p = \frac{krT}{V_P(y)}[\phi(P_P) + \frac{S_P}{rT} P_P v] + \frac{krT}{V_P(y)}\psi(P_p, sgn(u_P))u_P \end{cases} \tag{1.1}$$

Le domaine physique des grandeurs :

$$D \subset \mathbb{R}^4 = \begin{bmatrix} -0.25 \leq y \leq 0.25 \\ -2.7 \leq v \leq 2.7 \\ 1 \leq P_P \leq 7 \\ 1 \leq P_N \leq 7 \end{bmatrix} \begin{bmatrix} m \\ m/s \\ bar \\ bar \end{bmatrix}$$

Nous pouvons exprimer ce modèle non linéaire dans une nouvelle base $\begin{bmatrix} y & v & a & P_p \end{bmatrix}^T$. Le modèle peut s'écrire comme suit :

$$\begin{cases} \dot{y} = v \\ \dot{v} = a \\ \\ \dot{a} = \frac{S_P krT}{MV_P(y)}[\phi(P_P) + \frac{S_P}{rT}P_P v] - \frac{S_N krT}{MV_N(y)}[\phi(P_N) + \frac{S_N}{rT}P_N v] - \frac{b_v}{M}a + \frac{S_P krT}{MV_P(y)}\psi_N u_P - \frac{S_N krT}{MV_N(y)}\psi_N u_N \\ \\ \dot{P}_p = \frac{krT}{V_P(y)}[\phi(P_P) + \frac{S_P}{rT}P_P v] + \frac{krT}{V_P(y)}\psi(P_p, sgn(u_P))u_P \end{cases} \tag{1.2}$$

avec :

$$P_N = \frac{1}{S_N}(S_p P_p - b_v v - F_f(v) - F_{sec})$$

1.4 Problématique de Redécollage

1.4.1 Phénomène du Rédécollage

Les systèmes pneumatiques et plus particulièrement les actionneurs électropneumatiques souffrent d'un problème majeur qui rend plus délicat leur développement et qui favorise la concurrence des autres domaines (les procédés électriques...). Ce phénomène appelé phénomène du redécollage de l'actionneur électropneumatique et qui se présente sous la forme de mouvements ou secousses spontanées et saccadées.

Nous allons expliquer brièvement ce phénomène du redécollage sans entrer dans les détails. Il est similaire au phénomène de "stick-slip" (ou collé-glissé) connu dans le domaine mécanique et il peut se présenter ou apparaitre à n'importe quel moment et quelque soit la commande appliquée. Nous pouvons envisager ce phénomène au moment ou nous faisons un asservissement de position, c'est à dire avoir un équilibre ($y = y_{stop}$, $v = 0$ et $a = 0$) qui est un équilibre partiel et non pas total car il y aura toujours une évolution des pressions dans les chambres pneumatiques. Il n'y aura pas d'équilibre pneumatique, l'effort moteur qui est donné comme étant $S_p P_P - S_N P_N$ peut devenir donc supérieur aux forces de frottement sec et le vérin va s'écarter de sa position en influençant sur l'erreur en position qui aura tendance à augmenter. La commande va de nouveau agir pour ramener le vérin à sa position d'où le phénomène saccadé. Résoudre ce problème était pas assez évident à réaliser et il n'existait pas dans la littérature des travaux qui généralisent une solution donnée pour tous les systèmes pneumatiques. [HVBB96] ont essayer de réduire l'effet des frottements sec de façon progressive qui est la cause de ce phénomène et la même chose a été fait soit en agissant sur les lois de commande [CSI95], soit en agissant sur les vannes [Hä02].

1.4.2 Solution proposée

Dernièrement, une solution a été proposée dans les travaux de [Tur10] dont le principe est le suivant.

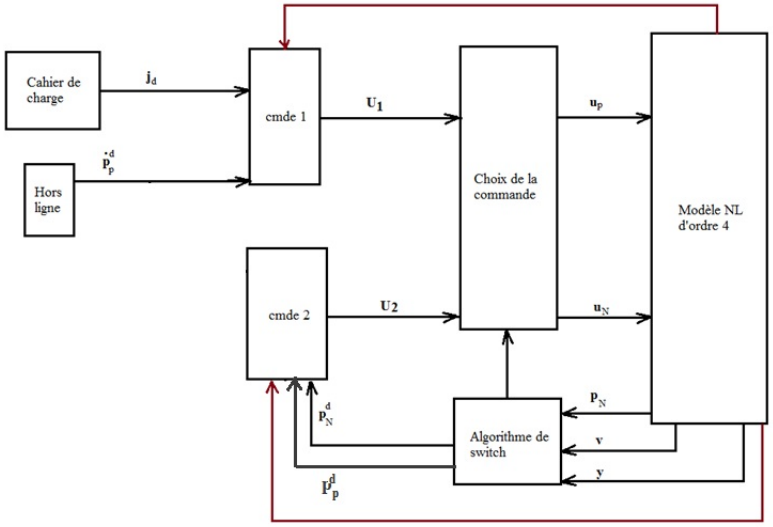


FIGURE 1.3 – Schéma de commutation au niveau de la commande

Cette solution consiste à maintenir les pressions constantes dans les chambres pneumatiques en régime statique. En d'autres termes, l'asservissement de trajectoire en position est fait en régime dynamique et durant le régime statique le système est commandé en régulation des pressions avec pressions désirées choisies comme étant les pressions au début du régime statique, c'est à dire au moment de la commutation vers la régulation des pressions. La commutation est assurée par un algorithme de "switch" selon la configuration donnée dans la figure 1.3
Pour assurer un bon fonctionnement, la commutation s'effectue selon des critères sur la position et la vitesse de la façon suivante :

Critère 01 : vérifier que le régime statique de la trajectoire désirée est atteint c'est à dire $y_d = y_{stop}$ en vérifiant le critère :

$$v_d = 0$$

Critère 02 : vérifier que l'erreur en position soit faible et inférieure à une certaine valeur fixée ε_1 très faible en vérifiant le critère :

$$|y - y_{stop}| \leq \varepsilon_1$$

Critère 03 : vérifier que le système soit à l'arrêt, c'est à dire que la vitesse soit faible à une certaine valeur très faible ε_2 en vérifiant le critère :

$$|v| \leq \varepsilon_2$$

1.4.3 Problème de la stabilité

Ce phénomène de redécollage a été un des problèmes majeurs rencontré dans les travaux précédents : la commande linéarisante et par platitude [Bru99], la synthèse par la technique de backstepping et mode glissant [SBT06][SBT08] [BT01]. En appliquant la solution proposée dans le paragraphe précédent qui se base sur la commutation entre deux lois de commandes, poursuite de la position en régime dynamique et régulation des pressions en régime statique, le phénomène de redécollage a été évité et aucun mouvement saccadé n'a été observé expérimentalement.

Par simulation, la stabilité de ce système a été vérifiée au moment de la commutation entre les deux lois de commande. Cependant théoriquement il n'existe pas de travaux qui

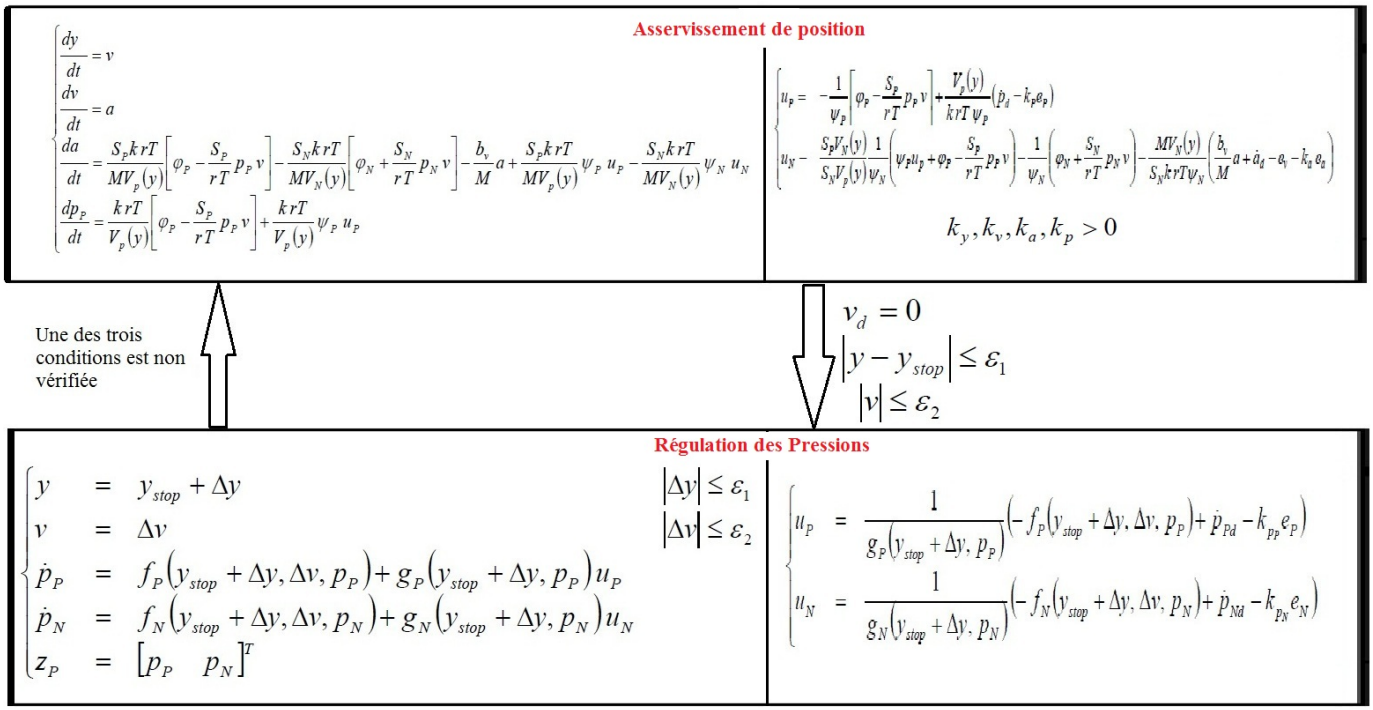


FIGURE 1.4 – aspect hybride de la problématique

ont démontré la stabilité en prenant en compte la commutation entre les deux modèles et en appliquant les deux lois de commandes citées précédemment. Cette problématique, nous pouvons l'envisager sur la figure 1.4 où il est évident que la distinction entre les conditions de commutations causées par les frottements secs et celles causées en changeant la loi de commande n'est pas assez claire. En plus l'influence des frottements n'apparait pas dans le deuxième modèle ce qui empêche l'étude du phénomène de redécollage. Ces dernières raisons nous ont forcé à attaquer le problème autrement et à suivre une autre démarche pour pouvoir l'analyser et le résoudre.

1.4.4 La démarche à suivre

Vues la complexité et l'ambiguïté d'étudier la commutation par rapport à la commande et l'influence des frottements, il a fallu procéder autrement en prenant un modèle simplifié et étudier ses caractéristiques pas à pas en prenant comme cause de la commutation dans ce cas là uniquement les frottements. Nous prenons alors le modèle linéarisé tangent du modèle non linéaire précédent autour d'un ensemble d'équilibre suivant :

$$\begin{cases} y^e \text{quelconque} \\ v^e = 0 \\ q_m(u^e, P_P^e) = 0 \\ q_m(-u^e, P_N^e) = 0 \end{cases}$$

Le modèle linéarisé tangent avec les frottements $F_f(v)$ est donc donné comme suit :

$$\begin{cases} \dot{y} = v \\ \dot{v} = \frac{S_p}{M} P_p - \frac{S_N}{M} P_N - \frac{1}{M} F_f(v) \\ \\ \dot{P}_p = -\frac{1}{\tau_p^e} P_p - \frac{kP_p^e S_p}{V_p(y^e)} v + \frac{krT_s G_{up}^e}{V_p(y^e)} u_1 \\ \\ \dot{P}_N = -\frac{1}{\tau_N^e} P_N + \frac{kP_N^e S_N}{V_N(y^e)} v + \frac{krT_s G_{uN}^e}{V_N(y^e)} u_2 \end{cases} \tag{1.3}$$

avec :

$$F_f(v) = \begin{cases} F_s : & e < v \\ \\ \frac{F_s}{e} v : & -e < v < e \\ \\ -F_s \end{cases}$$

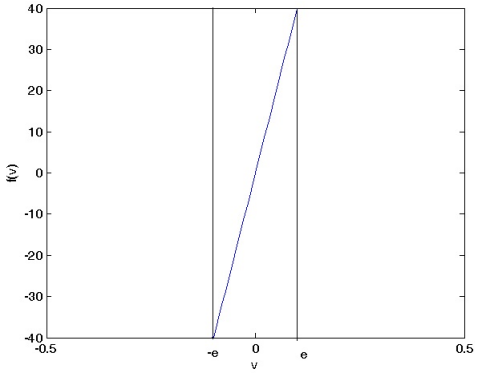


FIGURE 1.5 – modèle des frottements

synthèse des commandes

Dans cette partie,le même type de commandes synthétisées sur le modèle non linéaire dans [Tur10] a été choisi sur le modèle linéarisé tangent. Ce choix a été fait pour ne pas s'écarter du problème de départ car l'objectif principal est de démontrer la stabilité pour le modèle non linéaire commuté. Une loi de commande qui transforme notre modèle en une chaine d'intégrateurs a été synthétisée donc en appliquant un premier bouclage, puis un deuxième bouclage pour faire le retour de position, vitesse, accélération et pression. Pour cela, nous prenons le modèle de commande donné comme suis :

$$\begin{cases} \dot{y} = v \\ \dot{v} = \frac{S_p}{M} P_p - \frac{S_N}{M} P_N \\ \\ \dot{P}_p = -\frac{1}{\tau_p^e} P_p - \frac{kP_p^e S_p}{V_p(y^e)} v + \frac{krT_s G_{up}^e}{V_p(y^e)} u_1 \\ \\ \dot{P}_N = -\frac{1}{\tau_N^e} P_N + \frac{kP_N^e S_N}{V_N(y^e)} v + \frac{krT_s G_{uN}^e}{V_N(y^e)} u_2 \end{cases}$$

Ce modèle peut s'écrire dans une nouvelle base $\begin{bmatrix} y & v & a & P_P \end{bmatrix}^T$ sous la forme suivante :

$$\begin{cases} \dot{y}_1 = v \\ \dot{v}_2 = a \\ \dot{a}_3 = \frac{S_p}{M}(\frac{1}{\tau_N^e} - \frac{1}{\tau_p^e})P_P - [\frac{k}{M}(\frac{P_p^e S_p^2}{V_p^e} + \frac{P_N^e S_N^2}{V_N^e})]v - \frac{1}{\tau_N^e}a + \frac{krT_s S_p G_{up}^e}{MV_p}u_1 - \frac{krT_s S_N G_{uN}^e}{MV_N}u_2 \\ \dot{P}_{P4} = -\frac{1}{\tau_p^e}P_P - \frac{kP_p^e S_p}{V_p(y^e)}v + \frac{krT_s G_{up}^e}{V_p(y^e)}u_1 \end{cases} \tag{1.4}$$

les commandes sont calculées et données comme suit.

$$\begin{cases} u_1 = \frac{V_p(y^e)}{krT_s G_{up}^e}[kP_p^e \frac{S_p}{V_p(y^e)}v + \frac{1}{\tau_p^e}P_P + \dot{p}_p^d - k_p e_p] \\ u_2 = \frac{V_N(y^e)}{krT_s G_{uN}^e}[\frac{krT_s S_p G_{up}^e}{MV_p^e}u_1 - \frac{1}{\tau_N^e}a - [\frac{k}{M}(\frac{P_p^e S_p^2}{V_p^e} + \frac{P_N^e S_N^2}{V_N^e})]v + \frac{S_p}{M}(\frac{1}{\tau_N^e} - \frac{1}{\tau_p^e})P_P - \dot{a}^d + k_a e_a + k_v e_v + k_y e_y] \end{cases}$$

En appliquant ces commandes, le modèle précédent se transforme de la manière suivante :

$$\begin{cases} \dot{y}_1 = v \\ \dot{v}_2 = a \\ \dot{a}_3 = \dot{a}^d - k_a e_a - k_v e_v - k_y e_y \\ \dot{P}_P = \dot{p}_p^d - k_p e_p \end{cases} \tag{1.5}$$

Étude de la stabilité

Dans cette section, la stabilité a été vérifiée pour le modèle précédent et cela premièrement par simulation en simulant le modèle sans frottements et en regardant la stabilité de ce dernier, puis en ajoutant les frottements et voir si la stabilité reste toujours vérifiée. Deuxièmement, par une étude théorique en se basant sur le principe de Lyapunov [JR98].

Simulation

Les simulations ont été faites sur "Simulink" en étudiant les deux cas : modèle sans frottements et modèle avec frottements pour une réponse forcée, les trajectoires d'état sont convergentes et rejoignent les trajectoires désirées. Les réponses sont dans l'annexe "B" (page 41).

Méthode appliquée

Afin d'étudier la stabilité pour ce système qui est une classe des systèmes commutés présentés sous la forme $\dot{x} = A_i x + a_i$, nous allons exploiter un résultat de la littérature basé sur le travail développé dans l'article [JR98] où l'auteur a proposé une méthode pour étudier le problème de la stabilité en le transformant en un problème d'optimisation sous contraintes LMI via l'utilisation du principe de la S-procédure.

Chapitre 2

Les fonctions de Lyapunov quadratiques par morceaux

2.1 Introduction

En automatique, les systèmes sont soit représentés par un modèle dynamique continu soit par un modèle à événements discrets. Cependant dans la plupart des cas, la majorité des systèmes complexes mélangeant les deux dynamiques continue et discrète ne peuvent pas être classés dans les deux catégories précédentes, d'où l'intérêt d'étudier les modèles hybrides permettant la prise en compte simultanément les variables continues et discrètes [SS99][Bra95][Bra96]. Ce travail porte donc sur l'étude d'une classe de systèmes dynamiques hybrides présentant un formalisme englobant de nombreuses classes de modèles. Par celles-ci, nous choisissons alors celles qui décrivent le mieux la problématique de commutation tout en étant de complexité limitée et ce qu'on appelle les systèmes dynamiques à commutations [Lib03][BJ12].

De multiples pistes ont été déjà explorées en profondeur pour analyser ce type de systèmes complexes qui est considérée comme une étape indispensable pour répondre aux différents cahiers des charges.

Les systèmes à commutation sont caractérisés par des commutations entre plusieurs modes de fonctionnement où chaque mode est régi par ses propres lois dynamiques continues. Les transitions entre les modes peuvent alors être déclenchées selon un modèle événementiel : événements d'état, événements de temps, événements d'entrée. Ces transitions ou ces commutations dans un système commuté peuvent influencer sur la stabilité de ce dernier comme dans le cas suivant : les systèmes commutées sont constitués en sous systèmes, il est possible d'avoir un système instable pour une commutation entre deux sous systèmes stables ou au contraire, pour deux systèmes instables qui peuvent donner par une loi de commutations adéquat un système global stable [LM99]. Il est alors intéressant d'étudier la stabilité asymptotique uniforme des systèmes à commutation.

Plusieurs travaux ont été développés en se basant sur le principe de Lyapunov pour l'étude de la stabilité et des résultats plus particulièrement pour les systèmes hybrides commutés : l'approche basée sur les fonctions de Lyapunov communes[VL07][NB94][SN98][AOFY11], d'autres basées sur les fonctions de Lyapunov multiples [Bra98][DBPL00][SWM+07], des résultats basés sur l'algèbre et l'inclusion différentielles [RSD10][LHM99] et des travaux sur les critères de stabilité d'une manière générale pour les systèmes hybrides commutés. Pour ce type de systèmes hybrides, nous trouverons plusieurs façons d'étudier la stabilité avec de degré de complexité différent : la stabilité sous une commutation arbitraire et inconnue, ou vérifiant la stabilité

pour une séquence donnée [PL96] [JR98], ou même l'existence et la réalisation d'un signal de commutation qui garantie la stabilité. Dans notre cas, la commutation pour le système électro-pneumatique se fait par rapport à des critères d'état. Pour ce problème, nous allons nous baser pour l'analyse de la stabilité par les fonctions de Lyapunov quadratiques. Elles constituent un outil très efficace et très utilisé et beaucoup de travaux qui ont étendu ce concept pour le cas des systèmes commutés. On a aussi abouti au concept des fonctions de Lyapunov quadratiques par morceaux qui sera notre méthode que nous avons développé et qui est présentée dans ce document pour analyser la stabilité [TT09][AGAA12][MTM00][Joh02].

Cette partie est consacrée à l'étude d'une méthode d'analyse de la stabilité pour une classe particulière de systèmes dynamiques qui sont les systèmes dynamiques linéaires par morceaux. Une condition suffisante pour prouver la stabilité de ces systèmes est de pouvoir trouver une fonction de Lyapunov commune, cependant, soit il est difficile de trouver cette fonction qui assure la stabilité dans toutes les parties linéaires du système, soit on peut avoir des systèmes linéaires par morceaux stables et qui ne possèdent pas une telle fonction.

L'objectif est d'analyser la stabilité en considérant des fonctions de Lyapunov dites quadratiques **par morceaux**. Cependant, le problème du calcul des matrices de Lyapunov ne se présente pas sous un problème où on a des méthodes capables à le résoudre notamment le problème d'optimisation convexe. Ce qui nous a obligé d'essayer de trouver une série de transformations et d'opérations qui le ramène vers un problème où on sait efficacement le programmer et c'est le cas du problème d'optimisation sous contraintes LMI (Inégalités Matricielles Linéaires). Parmi ces opérations, nous avons une méthode efficace pour ce type de systèmes appelée la S-procédure [JR98] [Joh02] [PP09] [AMWZ12].

2.2 La S-procédure

2.2.1 Objectif

La S-procédure sert à démontrer la non négativité d'une forme quadratique sous la condition de la non-négativité d'une autre.

2.2.2 Principe de la S-procédure pour deux formes quadratiques

Soit $F_0 = F_0^T$, $F_1 = F_1^T \in \mathbb{R}^{n\times n}$.

On veut montrer la non négativité de la forme quadratique $z^T F_0 z$ pour les z tels que la forme quadratique $z^T F_1 z$ est positive.

Lemme 1. *[Jö04] L'implication :*

$$z^T F_1 z \geq 0 \Longrightarrow z^T F_0 z \geq 0 \tag{2.1}$$

qui est équivalente à :

$$z^T F_0 z \geq 0 \quad \forall z : z^T F_1 z \geq 0$$

est vraie si et seulement si la condition suivante est vérifiée :

$$\exists \lambda \in \mathbb{R}, \lambda \geq 0 : F_0 - \lambda F_1 \succeq 0. \tag{2.2}$$

2.2.3 Principe de la S-procédure pour des formes quadratiques multiples

Le principe est le même, selon le lemme suivant.

Lemme 2. *[BGFV94] Soient $F_0 = F_0^T, ..., F_k = F_k^T \in \mathbb{R}^{n\times n}$. L'implication :*

$$z^T F_1 z \geq 0, ..., z^T F_k z \geq 0 \Longrightarrow z^T F_0 z \geq 0 \tag{2.3}$$

qui est équivalente à :

$$z^T F_0 z \geq 0 \quad \forall z : z^T F_1 z \geq 0, ..., z^T F_k z \geq 0$$

est vraie si la condition suivante est vérifiée :

$$\exists \lambda_1, ..., \lambda_k \in \mathbb{R}, \lambda_1, ..., \lambda_k \geq 0 : F_0 - \sum_{i=1}^{k} \lambda_i F_i \succeq 0. \tag{2.4}$$

Remarque : Dans le lemme 1, on a une seule contrainte donc le passage entre (2.1) et (2.2) est non conservatif (ou sans perte) et la S-procédure dans ce cas est dite "sans perte". Cependant, dans le lemme 2, on a plusieurs contraintes donc le passage entre (2.3) et (2.4) est conservatif (ou avec perte) et la S-procédure dans ce cas est dite "avec perte".

2.3 Systèmes linéaires par morceaux

Soit $x \in \mathbb{R}^n$ le vecteur d'état d'un système dynamique. On considère une partition de $\mathbb{R}^n$ de plusieurs cellules X_i, pour $i \in I$. On définit I_1 comme l'ensemble des indices de cellules qui n'incluent pas l'origine, c'est-à-dire $I_1 = \{i : 0 \notin X_i\}$. De la même façon, on définit $I_0 = \{i : 0 \in X_i\}$. Les systèmes linéaires par morceaux sont représentés de la façon suivante :

$$\begin{cases} \dot{x}(t) = A_i x(t) \quad \text{pour} \quad x \in X_i,\ i \in I_0 \\ \\ \dot{x}(t) = A_i x(t) + a_i \quad \text{pour} \quad x \in X_i,\ i \in I_1 \end{cases} \tag{2.5}$$

Le domaine de chaque cellule X_i est défini par :

$$\begin{cases} E_i x \geq 0 \quad \text{pour} \quad x \in X_i,\ i \in I_0 \\ \\ \overline{E}_i \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_i,\ i \in I_1 \end{cases} \tag{2.6}$$

Une condition suffisante pour démontrer la stabilité de ce type de système est de trouver une fonction de Lyapunov $V(x) \in \mathbb{R}$, définie par :

$$V(x) = V_i(x) \quad \text{pour} \quad x \in X_i \quad \forall i \in I$$

telle qu'elle vérifie les deux conditions suivantes :

1.

$$\begin{cases} V(0) = 0 \\ \\ V_i(x) > 0 \quad \text{pour} \quad x \in X_i \setminus \{0\} \quad \forall i \in I, \\ \\ \dot{V}_i(x) < 0 \quad \forall i \in I, \end{cases}$$

2. Pour $x \in X_i \cap X_j$: $V_i(x) = V_j(x)$ (condition de la continuité).

Pour la 1ère condition, il suffit de trouver une fonction de Lyapunov quadratique de la forme :

$$\begin{cases} V_i(x) = x^T P_i x > 0 \quad \text{pour} \quad x \in X_i \setminus \{0\},\ i \in I_0 \\ \\ V_i(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}^T P_i \begin{bmatrix} x \\ 1 \end{bmatrix} > 0 \quad \text{pour} \quad x \in X_i \setminus \{0\},\ i \in I_1 \end{cases}$$

telle que

$$\begin{cases} \dot{V}_i = x^T (A_i^T P_i + P_i A_i) x < 0 \quad \text{pour} \quad x \in X_i \setminus \{0\},\ i \in I_0 \\ \\ \dot{V}_i = \begin{bmatrix} x \\ 1 \end{bmatrix}^T (A_i^T P_i + P_i A_i) \begin{bmatrix} x \\ 1 \end{bmatrix} < 0 \quad \text{pour} \quad x \in X_i \setminus \{0\},\ i \in I_1 \end{cases} \tag{2.7}$$

Donc, l'objectif est de calculer les matrices de Lyapunov P_i à travers les LMI formulées en partant des conditions (2.7).

Si on est dans le cas suivant :

$$\dot{x}(t) = Ax(t)$$

alors, dans ce cas $I_1 = \{\emptyset\}$ et $I_0 = \{1\}$ avec $X_1 = \mathbb{R}^n$. Calculer la matrice de Lyapunov P avec $x^T P x > 0$ pour $x \in \mathbb{R}^n \setminus \{0\}$ qui démontre la stabilité pour ce système telle que $x^T(A^T P + PA)x < 0$ pour $x \in \mathbb{R}^n \setminus \{0\}$ est équivalent à résoudre une inégalité matricielle linéaire, c'est-à-dire trouver $P \succ 0$, tel que :

$$A^T P + PA \prec 0$$

Pour les systèmes linéaires par morceaux représentés par (2.5), dire que $V_i(x) = x^T P_i x > 0$ (resp. $\dot{V}_i(x) = x^T(A_i^T P_i + P_i A_i)x < 0$) n'implique pas que $P_i \succ 0$ (resp. $A_i^T P_i + P_i A_i \prec 0$) car x appartient à un sous-espace $X_i \subset \mathbb{R}^n$ et pas à $\mathbb{R}^n$.

Pour pouvoir trouver les matrices de Lyapunov P_i et résoudre la problématique précédente, nous allons appliquer la méthode présentée dans la section 2.2 appelée la S-procédure qui donne à partir des formes quadratiques les inégalités matricielles linéaires associées.
Pour pouvoir appliquer cette méthode, les ensembles X_i doivent être présentés sous formes quadratiques. Pour cela, nous devons donc réécrire l'inégalité exprimant chaque domaine d'une cellule pour aboutir à cette forme quadratique.

Objectif

Trouver l'ensemble Ω_i défini par des contraintes quadratiques du type :

$$\Omega_i = \left\{ \ x \quad : \quad x^T Q_i x \geq 0 \ \right\}$$

tel que :

$$X_i \subset \Omega_i$$

Justification

Du fait que $X_i \subset \Omega_i$, on a :

$$x^T P_i x \succ 0 \quad \forall x \in \Omega_i \Longrightarrow x^T P_i x \succ 0 \quad \forall x \in X_i \tag{2.8}$$

En appliquant la S-procédure, d'après Lemme 1 (page 11), l'implication :

$$x^T Q_i x \geq 0 \Longrightarrow x^T P_i x > 0$$

qui est équivalente à :

$$x^T P_i x > 0 \quad \forall x : x^T Q_i x \geq 0$$

est vraie si et seulement si la condition suivante est vérifiée :

$$\exists \lambda \in \mathbb{R}, \lambda \geq 0 : P_i - \lambda Q_i \succ 0.$$

ce qui est bien une inégalité matricielle linéaire qui a comme variable de décision P_i et λ.
Le même principe est appliqué pour l'inégalité $x^T(A_i P_i + P_i A_i)x < 0$.

Le choix de Q_i

Dans l'article [JR98], on trouve un choix particulier de la forme quadratique $x^T Q_i x$.
Soit E_i qui définit la cellule X_i selon (2.6) pour le cas $i \in I_0$. En choisissant des matrices U_i à coefficients positifs, on obtient :

$$x^T E_i^T U_i E_i x \geq 0 \quad \text{pour} \quad x \in X_i, i \in I_0$$

L'idée principale est de trouver un ensemble décrit par une inégalité sous forme quadratique (dans ce cas $x^T E_i^T U_i E_i x \geq 0$) qui englobe ou qui inclut l'ensemble de départ décrit par l'inégalité $E_i x \geq 0$ tel que la forme quadratique "$x^T(P_i)x$" soit positive.
Dans ce cas :

$$Q_i = E_i^T U_i E_i \quad \text{avec} \quad U_i = [u_{ij}], u_{jk} \geq 0$$

Alors :

$$X_i \subset \Omega_i.$$

2.3.1 Illustration

Dans cette section, nous allons illustrer la raison du choix des coefficients positifs dans la matrice U_i. Pour cela, on prend un exemple simple.

2.3.1.1 Exemple :

Prenons :

$$U = \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, E = \begin{bmatrix} e_{11} & e_{12} \\ e_{12} & e_{22} \end{bmatrix}.$$

On cherche à déterminer le signe des coefficients U_i de telle sorte que l'inégalité suivante soit vérifiée :

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix}^T \begin{bmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geq 0. \tag{2.9}$$

Avec :

$$Ex \geq 0. \tag{2.10}$$

Développant la relation (2.10) :

$$Ex \geq 0 \Leftrightarrow \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} x \geq 0.$$

Telle que :

$$e_1^T = \begin{bmatrix} e_{11} & e_{12} \end{bmatrix}, e_2^T = \begin{bmatrix} e_{21} & e_{22} \end{bmatrix}$$

D'après (2.9), on aura :

$$u_{11}\, x^T e_1 e_1^T x + 2\, u_{12}\, x^T e_1 e_2^T x + u_{22}\, x^T e_2 e_2^T x \geq 0 \tag{2.11}$$

Avec :

$$\begin{cases} x^T e_1 e_1^T x \geq 0 \\ x^T e_1 e_2^T x \geq 0 \\ x^T e_2 e_2^T x \geq 0 \end{cases} \tag{2.12}$$

Une solution suffisante pour (2.11) est de prendre tous les coefficients u_i positifs ou nuls.

2.3.2 Le passage vers des inégalités matricielles linéaires

Nous allons maintenant exploiter les formes quadratiques en appliquant la S-procédure pour aboutir à des LMI. Pour cela, nous allons procéder de la façon suivante.

Nous allons faire le passage vers les LMI en considérant que la forme quadratique $x^T P_i x$ est positive sous condition que la forme quadratique $x^T E_i^T U_i E_i x$ soit non négative et cela en utilisant la S-procédure présentée dans le lemme 1. L'implication :

$$x^T E_i^T U_i E_i x \geq 0 \Longrightarrow x^T P_i x > 0 \tag{2.13}$$

qui est équivalente à :

$$x^T P_i x > 0 \quad \forall x : x^T E_i^T U_i E_i x \geq 0$$

est vraie si et seulement si la condition suivante est vérifiée :

$$\exists \lambda \in \mathbb{R}, \lambda \geq 0 : P_i - \lambda E_i^T U_i E_i \succ 0. \tag{2.14}$$

En faisant le changement de variable $\overline{U}_i = \lambda U_i$, on aura l'inégalité suivante :

$$P_i - E_i^T \overline{U}_i E_i \succ 0 \tag{2.15}$$

On est dans le cas d'une seule inégalité, donc la S-procédure est sans perte car d'après le lemme1, si on a une seule contrainte alors dans ce cas, le passage entre (2.13) et (2.14) sera sans perte ou, en d'autres termes, c'est une équivalence.

2.3.3 Résultat

En se basant sur le résultat cité dans l'article [JR98], l'analyse de stabilité des systèmes commutés a été faite de la manière suivante.
Soit le système commuté :

$$\begin{cases} \dot{x}(t) = A_i x(t) \quad \text{pour} \quad x \in X_i,\, i \in I_0 \\ \dot{x}(t) = A_i x(t) + a_i \quad \text{pour} \quad x \in X_i,\, i \in I_1 \end{cases}$$

On peut construire les matrices E_i qui définissent chacune des cellules X_i selon la relation :

$$\begin{cases} E_i x \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_0 \\ \overline{E}_i \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_1 \end{cases}$$

et les matrices qui définissent les frontières entre chaque deux cellules X_i et X_j et vérifient la relation :

$$\overline{F}_i \begin{bmatrix} x \\ 1 \end{bmatrix} = \overline{F}_j \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{pour} \quad x \in X_i \cap X_j,\ i, j \in I \tag{2.16}$$

Les cellules sont de formes polyédriques, nous pouvons prendre donc $\overline{F}_i = \begin{bmatrix} F_i & f_i \end{bmatrix}$ et $\overline{E}_i = \begin{bmatrix} E_i & e_i \end{bmatrix}$ avec $e_i = 0$ et $f_i = 0$ pour $i \in I_0$.
Choisissons U_i des matrices à coefficients positifs, on obtient donc :

$$\begin{cases} x^T E_i^T U_i E_i x \geq 0 \quad \text{pour} \quad x \in X_i,\ i \in I_0 \\ \\ \begin{bmatrix} x \\ 1 \end{bmatrix}^T \overline{E}_i^T U_i \overline{E} \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_i,\ i \in I_1 \end{cases}$$

Les fonctions de Lyapunov quadratiques par morceaux permettent de démontrer la stabilité de ce système. Elles sont définies par :

$$V_i(x) = \begin{cases} x^T P_i x \quad \text{pour} \quad x \in X_i,\ i \in I_0 \\ \\ \begin{bmatrix} x \\ 1 \end{bmatrix}^T \overline{P}_i \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{pour} \quad x \in X_i,\ i \in I_1 \end{cases}$$

continues sur les frontières, telles que :

$$\begin{cases} P_i = F_i^T T F_i \quad \text{pour} \quad i \in I_0 \\ \\ \overline{P}_i = \overline{F_i}^T T \overline{F_i} \quad \text{pour} \quad i \in I_1 \end{cases} \tag{2.17}$$

avec F_i qui satisfait (2.16).

Théorème 1. *[JR98] S'il existe les matrices T, U_i et W_i, avec U_i et W_i à coefficients non négatifs telles que les inégalités matricielles linéaires suivantes sont satisfaites :*

$$\begin{cases} \begin{cases} A_i^T P_i + P_i A_i + E_i^T U_i E_i \prec 0 \\ \\ P_i - E_i^T W_i E_i \succ 0 \end{cases} & i \in I_0 \\ \\ \begin{cases} \overline{A}_i^T \overline{P}_i + \overline{P}_i \overline{A}_i + \overline{E}_i^T U_i \overline{E}_i \prec 0 \\ \\ \overline{P}_i - \overline{E}_i^T W_i \overline{E}_i \succ 0 \end{cases} & i \in I_1 \end{cases} \tag{2.18}$$

où P_i et $\overline{P}_i$ sont définies dans (2.17).
Alors toute trajectoire continue par morceaux $x(t) \in \cup_{i \in I} X_i$ tend vers zéro exponentiellement.

2.3.4 Discussion

1. Dans cette partie, nous allons discuter de la notion du conservatisme. D'où vient le conservatisme ?

   L'idée principale est de trouver un ensemble décrit par une inégalité sous forme quadratique qui englobe l'ensemble de départ tel que la forme quadratique $x^T P_i x$ soit toujours positive. Comme $X_i \subset \Omega_i$ et $\forall x \in \Omega_i$ l'inégalité $x^T P_i x > 0$ est vérifié, alors le conservatisme vient du passage de l'ensemble X_i à Ω_i.

2. Dans (2.8), on a l'implication dans un sens mais pas dans l'autre, c'est-à-dire l'implication suivante est fausse : $x^T P_i x > 0 \quad \forall x \in X_i \Longrightarrow x^T P_i x > 0 \quad \forall x \in \Omega_i$
3. Nous avons vu précédemment que le choix de l'ensemble décrit par une forme quadratique n'est pas unique (le choix des U_i positifs est une condition suffisante) et qu'il dépend du problème traité. Assurer donc le bon fonctionnement de la méthode revient à trouver le bon Ω_i qui dépend du problème considéré.

2.3.4.1 Interprétation de la condition (2.15)

Nous allons maintenant appliquer la S-procédure pour faire le passage vers les LMI en sachant que $(x^T P_i x)$ est positive, quand x vérifie les inégalités :

$$\begin{cases} x^T e_1 e_1^T x \geq 0 \\ x^T e_1 e_2^T x \geq 0 \\ x^T e_2 e_2^T x \geq 0 \end{cases} \tag{2.19}$$

En utilisant le lemme 2 :

$$\begin{cases} x^T e_1 e_1^T x \geq 0 \\ x^T e_1 e_2^T x \geq 0 \\ x^T e_2 e_2^T x \geq 0 \end{cases} \Longrightarrow x^T P_i x > 0 \tag{2.20}$$

l'implication est vraie, si $\exists \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0$ tels que :

$$P_i - \lambda_1 e_1 e_1^T - \lambda_3 e_1 e_2^T - \lambda_2 e_2 e_2^T \succ 0 \tag{2.21}$$

ou :

$$P_i - \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix}^T \begin{bmatrix} \lambda_1 & \frac{1}{2}\lambda_3 \\ \frac{1}{2}\lambda_3 & \lambda_2 \end{bmatrix} \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} \succ 0 \tag{2.22}$$

prenons $U_i = \begin{bmatrix} \lambda_1 & \frac{1}{2}\lambda_3 \\ \frac{1}{2}\lambda_3 & \lambda_2 \end{bmatrix}$:

$$P_i - E_i^T U_i E_i \succ 0 \tag{2.23}$$

avec $E_i = \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix}$.
Nous avons trois contraintes, dans ce cas la S-procédure est dite avec perte [BGFV94].

2.3.5 Ω_i dans des cas particulier de X_i

Nous avons vu que l'ensemble Ω_i englobe l'ensemble X_i. Nous allons le vérifier sur la figure 2.1.

Dans cette figure, prenons comme un exemple l'ensemble X_1 le quadrant positif (x_1, x_2) (hachurée en bleu) qui est inclus dans la région hachurée coloré en marron qui constitue l'ensemble Ω_1. L'ensemble Ω_1 englobe donc X_1 et le symétrique de X_1 par rapport à l'origine. Elle est déterminée par l'union des régions délimitées par les droites de pentes qui dépendent des valeurs des coefficients de la matrice U_1 dans la relation $x^T E_1^T U_1 E_1 x \geq 0$, donc en jouant sur la valeur de la pente de chaque droite on va modifier en fait la forme de l'ensemble Ω_1 (voir Annexe A).
Le fait que l'ensemble Ω_1 englobe le symétrique de X_1 est un résultat très intéressant car il va réduire le nombre de contraintes dans le programme d'optimisation LMI. C'est-à-dire qu'au lieu de prendre chaque région séparément et d'écrire ses contraintes, on prend un seul ensemble

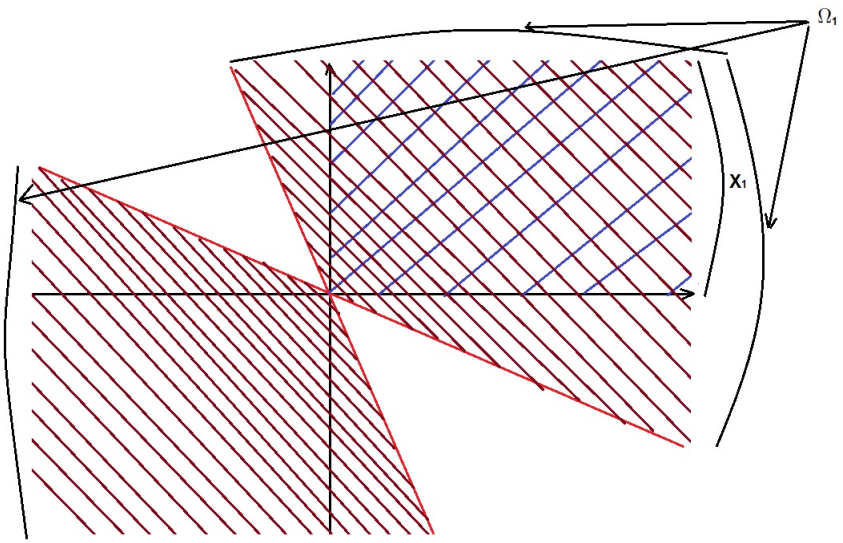


FIGURE 2.1 – Ω_i dans un cas particulier de X_i

avec des contraintes valables pour les deux régions.
Pour envisager cette propriété, nous pouvons prendre l'exemple suivant :

$$\begin{cases} \dot{x} = A_1 x + a_1 & \text{pour} \quad x > 0 \\ \dot{x} = A_2 x + a_2 & \text{pour} \quad x < 0 \end{cases}$$

avec : $A_2 = A_1$ et $a_2 = -a_1$ X_1 dans ce cas est la partie pour laquelle $x > 0$, si $x \in X_1$ alors $-x \in X_2$ qui est le symétrie de X_1 par rapport à 0 ou c'est le quart de plan négatif.
D'après la figure 2.1, on voit bien que l'ensemble Ω_1 contient les deux régions X_1 et son symétrique X_2, ou bien :

$$X_1 \cup X_2 \subset \Omega_1$$

2.3.6 Exemples

L'idée de la S-procédure dans le paragraphe précédent est de transformer une relation (inégalité) vers une forme quadratique, puis d'exploiter ces formes quadratiques pour pouvoir obtenir des contraintes LMI.

Cependant, cette méthode n'est toujours valable dans tous les cas de figure, et nous allons voir effectivement à travers des exemples simples que le terme ajouté dans les inégalités pour avoir des LMI n'apporte en fait rien et n'a aucun influence sur le signe de l'inégalité.

Exemple 1

Concernant l'exemple, nous avons le système commuté suivant :

$$\begin{cases} \dot{x} = A_1 x & x \in X_1 \\ \dot{x} = A_2 x & x \in X_2 \end{cases} \tag{2.24}$$

où X_1 et X_2 sont respectivement le demi plan à gauche et à droite (voir Figure 2.2).

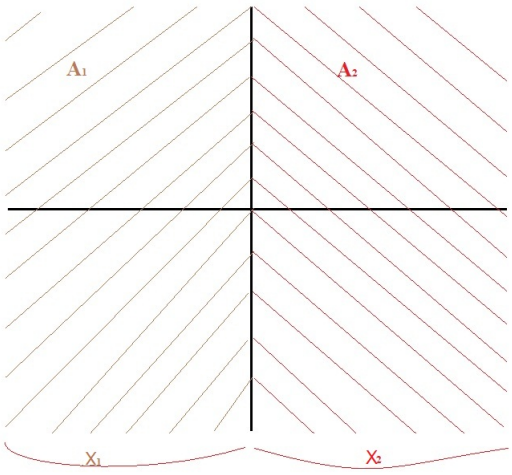


FIGURE 2.2 – les sous-espaces X_i

L'objectif est de prouver la stabilité pour le système lorsqu'il commute entre les deux régions X_1 et X_2. c'est-à-dire chercher dans chaque région une fonction de Lyapunov d'une forme quadratique $V_i(x) = x^T P_i x > 0, \quad i = 1, 2$ pour $x \in X_i$ telle que :

$$\dot{V}_i(x) = x^T (A_i^T P_i + P_i A_i) x < 0 \quad i = 1, 2, \quad x \in X_i \tag{2.25}$$

Le seul inconvénient dans ce cas, c'est qu'on n'a pas tout l'espace d'état, en d'autres termes : $x \in X_i$ (tel que X_i est un sous-espace de l'espace d'état $\mathbb{R}^2$) et donc :

$$x^T (A_i^T P_i + P_i A_i) x < 0 \quad \text{n'implique pas forcément} \quad A_i^T P_i + P_i A_i \prec 0 \tag{2.26}$$

C'est là où intervient l'intérêt de la S-procédure. En appliquant le théorème 1,
l'inégalité $E_i x \geq 0$ représente le domaine X_i, on veut exprimer cette inégalité sous forme quadratique, et pour cela, on prend une matrice U à coefficients positifs telle que :

$$x^T E_i^T U E_i x \geq 0 \tag{2.27}$$

Alors, on a :

$$x^T E_i^T U E_i x \geq 0 \quad \Longrightarrow \quad x^T (A_i^T P_i + P_i A_i) x < 0 \tag{2.28}$$

$$\Longleftrightarrow A_i^T P_i + P_i A_i + E_i^T U E_i \prec 0 \tag{2.29}$$

Cependant, dans ce cas de figure et dans ce qui suit, nous allons voir que l'inégalité $A_i^T P_i + P_i A_i \prec 0$ est toujours vraie et que le terme $E_i^T U E_i$ servira à rien.

Pour la région X_1 : on a : $E_1 x \geq 0$ cette inégalité correspond au demi-plan gauche. On va l'exprimer sous forme quadratique : choisissant U telle que :

$$E_1 x \geq 0 \Longrightarrow x^T E_1^T U E_1 x \geq 0 \tag{2.30}$$

$x^T E_1^T U E_1 x \geq 0$: la région est plus grande que $E_1 x \geq 0$ et on peut définir les ensembles suivants :

$$D_1 = \left\{ \; x \in X_1 \quad E_1 x \geq 0 \; \right\} \tag{2.31}$$

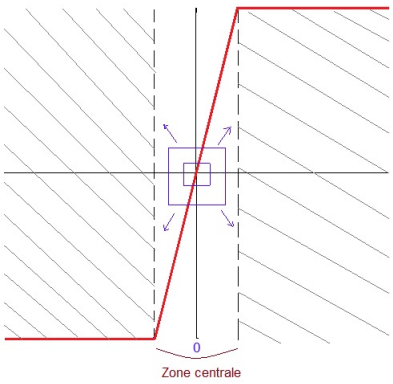


FIGURE 2.3 – la zone centrale d'une saturation

$$\Omega_1 = \left\{ \; x \quad | \quad x^T E_1^T U E_1 x \geq 0 \; \right\} \tag{2.32}$$

$E_1 x \geq 0 \Longrightarrow x^T E_1^T U E_1 x \geq 0 \Longleftrightarrow D_1 \subset \Omega_1$

Pour $x \in D_1$: La fonction de Lyapunov :$V_1(x) = x^T P_1 x > 0 \quad \forall x \neq 0$ telle que :
$\dot{V}_1(x) = x^T (A_1^T P_1 + P_1 A_1) x < 0$

D'après le théorème 1 :

$A_1^T P_1 + P_1 A_1 + E_1^T U E_1 \prec 0$

Pour $x \in D_1 \subset \Omega_1$: $x^T E_1^T U E_1 x \geq 0$, donc pour avoir la négativité de ce terme $(A_1^T P_1 + P_1 A_1 + E_1^T U E_1)$, $x^T (A_1^T P_1 + P_1 A_1) x$ doit forcément être négatif (c'est ce qui est demandé).

Aussi, $x \in X_1$ donc $(-x) \in X_2$ (par symétrie) et la relation suivante est toujours vraie :
$x^T E_1^T U E_1 x \geq 0$ avec $\mathbb{R}^2 = X_1 \cup X_2$
Et dans ce cas aussi, pour assurer la négativité du terme $(A_1^T P_1 + P_1 A_1 + E_1^T U E_1)$,il faut que le terme $A_1^T P_1 + P_1 A_1$ le soit forcément car le terme $x^T E_1^T U E_1 x$ est toujours non négatif $\forall x \in \mathbb{R}^2$.
En d'autres termes, $\forall x \in \mathbb{R}^2 \quad A_1^T P_1 + P_1 A_1 \prec 0$, ce qui veut dire que le terme $(E_1^T U E_1)$ servira à rien.

Exemple 2 :

Nous allons illustrer un autre exemple où le terme $E_1^T U E_1$ de la S-procédure n'a aucun influence sur la négativité de l'inégalité $A_1^T P_1 + P_1 A_1 + E_1^T U E_1 \prec 0$
Pour cela ,prenons l'exemple de la figure(2.3) où l'espace d'état est subdivisé en trois régions,une région centrale et deux autres régions symétriques. Pour la zone centrale, l'inégalité qui représente le domaine de cette zone est donnée comme suit :

$$-\varepsilon \leq x_2 \leq \varepsilon \tag{2.33}$$

où x_2 représente la deuxième variable d'état qui est la vitesse dans notre cas(voir chapitre précédent). On peut écrire cette inégalité d'une autre façon :

$$E_0 \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{avec} \quad E_0 = \begin{bmatrix} 0 & -1 & 0 & \varepsilon \\ 0 & 1 & 0 & \varepsilon \end{bmatrix} \tag{2.34}$$

Prouver la stabilité dans cette région revient à trouver une fonction de Lyapunov $V_0 = x^T P_0 x > 0$ telle que :

$$\dot{V}_0 = x^T (A_0^T P_0 + P_0 A_0) x < 0 \tag{2.35}$$

Et comme $x \in X_0 \subset \mathbb{R}^n$,alors dans ce cas on ne peut pas écrire : $A_0^T P_0 + P_0 A_0 \prec 0$.
En Multipliant l'inégalité (2.35) par "α^2" tel que : $\alpha \in \mathbb{R}$:

$$(\alpha x)^T (A_0^T P_0 + P_0 A_0)(\alpha x) < 0 \tag{2.36}$$

Prenons maintenant un sous espace dans la région centrale représenté par un carré (voir la figure 2.3). La multiplication par α cause donc une dilatation ou contraction du carré dans cette région, et pour α assez grand, la région centrale englobe tout l'espace d'état, et dans ce cas là :
Si on prend $\tilde{x} = (\alpha x)$ tel que :

$$\tilde{x}^T (A_0^T P_0 + P_0 A_0) \tilde{x} < 0 \tag{2.37}$$

On pourra écrire :$A_0^T P_0 + P_0 A_0 \prec 0$ car $\tilde{x} \in \mathbb{R}^n$.

C'est-à-dire que le terme ($E_0^T U E_0$ de la S-procédure n'a aucun influence et il n'y a aucun intérêt de l'ajouter dans l'inégalité.

Chapitre 3

Extension de la méthode de [JR98] pour l'étude de la stabilité de l'application "Fluid Power"

Nous présentons dans cette partie l'étude de la stabilité du système présenté dans le chapitre 1. Premièrement, la méthode présentée dans [JR98] a été appliquée en donnant des résultats non satisfaisants. Ensuite, des améliorations et des modifications ont été réalisées afin de résoudre les problématiques rencontrées en appliquant la méthode de [JR98]. finalement, une méthode générale à été proposée pour démontrer la stabilité de cette classe de systèmes commutés et qui a été appliquée sur l'application pneumatique en donnant de meilleurs résultats.

3.1 Rappel sur le modèle de l'application pneumatique

Le modèle du système électropneumatique présenté au premier chapitre est rappelé par la relation :

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_3 \\ \dot{x}_3 = \frac{S_p}{M}(\frac{1}{\tau_N^e} - \frac{1}{\tau_p^e})x_4 - [\frac{k}{M}(\frac{P_p^e S_p^2}{V_p^e} + \frac{P_N^e S_N^2}{V_N^e})]x_2 - \frac{1}{\tau_N^e}x_3 + \frac{krT_s S_p G_{up}^e}{MV_p}u_1 - \frac{krT_s S_N G_{uN}^e}{MV_N}u_2 - \frac{1}{M\tau_N}F_s ec - \frac{1}{M}\dot{F}_s ec \\ \dot{x}_4 = -\frac{1}{\tau_p^e}x_4 - \frac{kP_p^e S_p}{V_p(y^e)}x_2 + \frac{krT_s G_{up}^e}{V_p(y^e)}u_1 \end{cases} \tag{3.1}$$

avec :

$$F_{sec} = \begin{cases} F_s : & v > e \\ \frac{F_s}{e}v : & -e \leq v \leq e \\ -F_s : & v < -e \end{cases}$$

et

$$\dot{F}_{sec} = \begin{cases} 0 : & v > e \\ \frac{F_s}{e}a : & -e \leq v \leq e \\ 0 : & v < -e \end{cases}$$

Le vecteur d'état : $x = \begin{pmatrix} y & v & a & P_p \end{pmatrix}^T$

En appliquant les commandes développées au chapitre 1 :

$$\begin{cases} u_1 = \frac{V_p(y^e)}{krT_sG^e_{up}}[kP^e_p\frac{S_p}{V_p(y^e)}x_2 + \frac{1}{\tau^e_p}x_4 + \dot{p}^d_p - k_pe_p] \\ \\ u_2 = \frac{V_N(y^e)}{krT_sG^e_{uN}}[\frac{krT_sS_pG^e_{up}}{MV^e_p}u_1 - \frac{1}{\tau^e_N}x_3 - [\frac{k}{M}(\frac{P^e_pS^2_p}{V^e_p} + \frac{P^e_NS^2_N}{V^e_N})]x_2 + \frac{S_p}{M}(\frac{1}{\tau^e_N} - \frac{1}{\tau^e_p})x_4 - \dot{a}^d + k_ae_a + k_ve_v + k_ye_y] \end{cases}$$

le modèle en boucle fermée :

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_3 \\ \\ \dot{x}_3 = -\dot{a}^d + k_ae_a + k_ve_v + k_ye_y - \frac{1}{M\tau_N}F_sec - \frac{1}{M}\dot{F}_sec \\ \\ \dot{x}_4 = \dot{p}^d_p - k_pe_p \end{cases} \tag{3.2}$$

3.2 Analyse de la stabilité par optimisation sous contraintes LMI

3.2.1 Utilisation de la méthode présentée dans [JR98]

Nous allons utiliser pour faire l'étude de la stabilité seulement le vecteur d'état réduit $[y \quad v \quad a]^T$ car pendant la commutation, l'état P_p et la commande appliquée là-dessus seront les mêmes dans les deux modèles.

En utilisant le modèle de frottement retenu et en considérant que les trois premières équations d'états, on aura un modèle commuté de la forme $\dot{x} = Ax + a_0$. La commutation est suivant la valeur de x_2 :
pour la zone centrale :$-e < x_2 < e$

$$A_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -k_y & -k_v - \frac{F_s}{M\tau_N e} & -k_a - \frac{F_s}{Me} \end{bmatrix}, \quad x = \begin{bmatrix} y & v & a \end{bmatrix}^T$$

pour la zone $x_2 > +e$:

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -k_y & -k_v & -k_a & -\frac{F_s}{M\tau_N} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad x = \begin{bmatrix} y & v & a & 1 \end{bmatrix}^T$$

pour la zone $x_2 < -e$:

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -k_y & -k_v & -k_a & \frac{F_s}{M\tau_N} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad x = \begin{bmatrix} y & v & a & 1 \end{bmatrix}^T$$

Il s'agit d'un modèle linéaire par morceaux dont nous étudions la stabilité en utilisant la méthode présentée précédemment.

Le système est de la forme :

$$\begin{cases} \dot{x}(t) = A_i x(t) \quad \text{pour} \quad x \in X_i,\, i \in I_0 \\ \\ \dot{x}(t) = A_i x(t) + a_i \quad \text{pour} \quad x \in X_i,\, i \in I_1 \end{cases}$$

dans notre cas : $I_0 = \{0\}$ et $I_1 = \{1, 2\}$

On peut construire les matrices E_i qui définissent chacune des trois cellules X_i et qui vérifient les relations :

$$\begin{cases} E_i x \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_0 \\ \\ \overline{E}_i \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_1 \end{cases} \tag{3.3}$$

pour la région centrale :$-e < x_2 < e$

$$E_0 x \geq 0 \quad \text{avec} - e < x_2 < e \Longleftrightarrow E_0 = \begin{bmatrix} 0 & -1 & 0 & e \\ 0 & 1 & 0 & e \end{bmatrix}$$

pour les régions extérieures :
la région : $x_2 > +e$

$$E_1 x \geq 0 \quad \text{avec} + e < x_2 \Longleftrightarrow E_1 = \begin{bmatrix} 0 & 1 & 0 & -e \end{bmatrix} \tag{3.4}$$

la région : $x_2 < -e$

$$E_2 x \geq 0 \quad \text{avec} \quad x_2 < -e \Longleftrightarrow E_2 = \begin{bmatrix} 0 & -1 & 0 & -e \end{bmatrix}$$

Afin d'analyser la stabilité de ce système en utilisant le résultat de l'article [JR98], nous devons spécifier deux contraintes :

1. déterminer la forme quadratique représentant chaque cellule pour pouvoir écrire les LMI et calculer les matrices de Lyapunov. Elle est de la forme suivante :

$$\begin{cases} x^T E_i^T U_i E_i x \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_0 \\ \\ \begin{bmatrix} x \\ 1 \end{bmatrix}^T \overline{E}_i^T U_i \overline{E} \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_i,\, i \in I_1 \end{cases}$$

2. définir la continuité à la frontière :

$$\overline{F}_i \begin{bmatrix} x \\ 1 \end{bmatrix} = \overline{F}_j \begin{bmatrix} x \\ 1 \end{bmatrix} \quad \text{pour} \quad x \in X_i \cap X_j,\, i, j \in I \tag{3.5}$$

en déterminant les matrices $\overline{F}_i$ et $\overline{F}_j$ telles que :

$$\begin{cases} P_i = F_i^T T F_i \quad \text{pour} \quad i \in I_0 \\ \\ \overline{P}_i = \overline{F_i}^T T \overline{F_i} \quad \text{pour} \quad i \in I_1 \end{cases}$$

Cependant, en prenant en compte ces contraintes et les spécifications indiquées dans notre problème de départ (le système commuté à trois régions), le programme d'optimisation sous contraintes LMI est infaisable pour les deux raisons suivante :

– raison 1 : le problème se pose au premier lieu au nivaux du choix de la forme quadratique pour pouvoir utiliser le principe de la s-procédure et trouver les LMI. Dans l'article [JR98], le choix $x^T E_i^T U_i E_i x$ comme forme quadratique dans notre problème ne va pas apporter des solutions pour la simple raison suivante :
  – U_i dans notre cas est un scalaire, ce qui veux dire :

$$x^T E_i^T U_i E_i x \geq 0 \Longleftrightarrow U_i (E_i x)^2 \geq 0$$

  cette relation est valable quelque soit le terme $E_i x$ (positif ou négatif), et donc cette forme quadratique ne reflète pas la cellule considérée ce qui engendre dans la programmation des problème de faisabilité et l'inexistence de la solution.

– raison 2 : Des difficultés ont été rencontrées par rapport ç la condition de la continuité de la frontière. En effet, il n'y a pas une méthode générale pour choisir les matrices $\overline{F}_i$ et $\overline{F}_j$:
  1. En choisissant la matrice $\overline{F}_i$, la matrice $\overline{F}_j$ sera déterminée en fonction de cette matrice à partir de la condition (3.5). En imposant plus de conraintes, il n'est pas évident de déterminer les matrices de Lyapunov qui satisfaites ces contraintes.
  2. si en exprimant de manière générale les termes de $\overline{F}_j$ en fonction de ceux de $\overline{F}_i$, on obtient un problème bilinéaire non convexe dont la solution n'est pas évidente.

Pour ces raisons, il a été proposé d'exprimer les deux contraintes précédentes autrement.

3.2.2 Proposition d'une nouvelle méthode pour analyser la stabilité

Dans cette partie, nous présentons les améliorations qui ont été faites sur la méthode de [JR98] pour démontrer la stabilité du modèle précédent.

3.2.2.1 Choix de la contrainte quadratique

Notre objectif est de définir l'ensemble Ω_i défini par des contraintes quadratiques pour pouvoir appliquer le principe de la s-procédure, autrement dit trouver :

$$\Omega_i = \left\{ \; x \quad : \quad x^T Q_i x \geq 0 \; \right\}$$

tel que :

$$X_i \subseteq \Omega_i$$

Et dans ce cas, en appliquant la S-procédure et d'après Lemme 1 (page 11), l'implication suivante

$$x^T Q_i x \geq 0 \Longrightarrow x^T P_i x > 0$$

qu'est équivalente à :

$$x^T P_i x > 0 \quad \forall x : x^T Q_i x \geq 0$$

est vraie si et seulement si la condition suivante est vérifiée :

$$\exists \lambda \in \mathbb{R}, \lambda \geq 0 : P_i - \lambda Q_i \succ 0.$$

cette dernière inégalité matricielle linéaire est utilisée pour calculer la matrice de Lyapunov P_i qui est considérée comme variable de décision avec le paramètre λ.

Nous avons vu que l'application de la méthode cité dans [JR98] pour déterminer l'ensemble Ω_i n'a pas pu donné de résultat et le problème était infaisable. Dans ce qui suit, nous allons choisir Ω_i de deux façons différentes.

3.2.2.2 Première méthode : utilisation de la forme quadratique

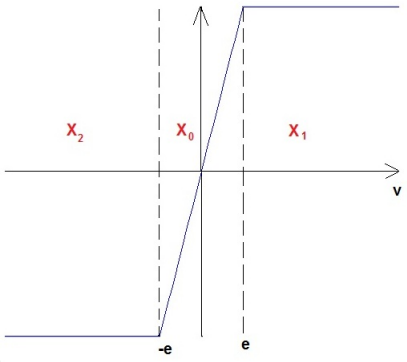


FIGURE 3.1 – les cellules X_i

La figure 3.1 présente les cellules X_i. Les deux régions X_1 et X_2 sont symétrique par rapport à 0, nous pouvons donc exploiter cette propriété pour décrire les deux régions par une seule contrainte inégalité :

la première région X_1 est décrite par l'inégalité $x_2 > +e$ et la deuxième région X_2 est décrite par $x_2 < -e$. Alors, l'inégalité qui décrit les deux régions à la fois est donnée comme suivant :

$$|x_2| > e \Longleftrightarrow x_2^2 > e^2$$

À partir de cette inégalité, nous allons essayer de formuler la forme quadratique qui décrit l'ensemble Ω_i.

On a $x_2^2 > e^2$, on peut l'exprimer de la façon suivante :

$$\begin{bmatrix} x_1 & x_2 & x_3 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -e^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{bmatrix} > 0$$

$$\Longleftrightarrow \begin{bmatrix} x \\ 1 \end{bmatrix}^T Q \begin{bmatrix} x \\ 1 \end{bmatrix} > 0$$

avec $Q = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -e^2 \end{bmatrix}$.

En appliquant le principe de la S-procédure (voir page 26), s'il existe $\tau_1, \tau_2, \tau_3, \tau_4$ des coefficients non négatifs telles que les inégalités matricielles linéaires suivantes sont satisfaites :

$$\begin{cases} \begin{cases} A_i^T P_i + P_i A_i + \tau_1 Q \prec 0 \\ P_i - \tau_2 Q \succ 0 \end{cases} & i \in I_0 \\ \begin{cases} \overline{A}_i^T \overline{P}_i + \overline{P}_i \overline{A}_i + \tau_3 Q \prec 0 \\ \overline{P}_i - \tau_4 Q \succ 0 \end{cases} & i \in I_1 \end{cases} \tag{3.6}$$

Alors toute trajectoire continue par morceaux $x(t) \in \cup_{i \in I} X_i$ tend vers zéro exponentiellement.

Discussion

- En fait, dans notre cas on a pas besoin d'ajouter le terme τQ à toutes les inégalités grâce aux propriétés du ce système. Pour la région centrale, d'après le raisonnement qui a été fait précédemment, l'inégalité $A_i^T P_i + P_i A_i \prec 0$ si elle est valable dans cette région, alors le terme de la S-procédure n'a aucun influence et il n'y a aucun intérêt de l'ajouter.
- pour les régions extérieures, nous avons vu qu'elles sont symétriques par rapport à 0 et nous avons choisi une forme quadratique $\begin{bmatrix} x \\ 1 \end{bmatrix}^T Q \begin{bmatrix} x \\ 1 \end{bmatrix} > 0$ qui décrit les deux régions au même temps, donc il n'est pas obligatoire d'imposer les contraintes des deux régions et il suffit d'écrire celles d'une région et elles serons valables pour l'autre grâce à la propriété de la symétrie. Nous verrons par la suite aussi que la solution trouvée pour une région sera valable pour l'autre région et cela sera évident dans la condition de la continuité.

Nous pouvons donc réécrire les inégalités si nous prenons que la région centrale nommée région 0 et la région $x_2 > e$ appelée région 1 :

$$\begin{cases} \begin{cases} A_0^T P_0 + P_0 A_0 \prec 0 \\ P_0 \succ 0 \end{cases} & i \in I_0 \\ \begin{cases} \overline{A}_1^T \overline{P}_1 + \overline{P}_1 \overline{A}_1 + \tau_3 Q \prec 0 \\ \overline{P}_1 - \tau_4 Q \succ 0 \end{cases} & i \in I_1 \end{cases}$$

Ces inégalités vont être utilisées pour calculer les matrices de Lyapunov.

3.2.2.3 Deuxième méthode : utilisation de la forme linéaire

Dans cette partie, nous allons exploiter l'inégalité linéaire qui décrit la cellule pour aboutir aux LMI. Pour la première région où $x_2 > +e$, la matrice E_1 qui définit cette cellule X_1 et qui vérifie la relation :

$$\overline{E}_1 \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \quad \text{pour} \quad x \in X_1$$

est donnée comme suivant :

$$\overline{E}_1 = \begin{bmatrix} 0 & 1 & 0 & -e \end{bmatrix}$$

Cherchons maintenant la matrice Q_1 telle que :

$$\begin{bmatrix} x \\ 1 \end{bmatrix}^T Q_1 \begin{bmatrix} x \\ 1 \end{bmatrix} > 0$$

Nous pouvons prendre un choix à partir de l'inégalité linéaire décrivant la région X_1 et sa symétrie. La matrice Q_1 est donc de la forme suivante :

$$Q_1 = \begin{bmatrix} 0 \\ \overline{E}_1 \end{bmatrix} + \begin{bmatrix} 0 & \overline{E}_1^T \end{bmatrix} \tag{3.7}$$

$$\Longleftrightarrow Q_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2e \end{bmatrix}$$

En appliquant le même principe que la première méthode et en prenant aussi les mêmes contraintes considérées (région centrale, symétrie), nous pouvons dire :

s'il existe τ_1, τ_2 non négatifs telles que les inégalités matricielles linéaires suivantes sont satisfaites :

$$\begin{cases} \begin{cases} A_0^T P_0 + P_0 A_0 \prec 0 \\ P_0 \succ 0 \end{cases} & i \in I_0 \\ \begin{cases} \overline{A}_1^T \overline{P}_1 + \overline{P}_1 \overline{A}_1 + \tau_1 Q_1 \prec 0 \\ \overline{P}_1 - \tau_2 Q_1 \succ 0 \end{cases} & i \in I_1 \end{cases}$$

alors toute trajectoire continue par morceaux $x(t) \in \cup_{i \in I} X_i$ tend vers zéro exponentiellement.

Remarque

Nous démontrons par la suite que la solution pour ce problème d'optimisation sous contraintes LMI en considérant que les contraintes de la première région est aussi une solution pour celles de la régions symétrique à la première. C'est la raison de prendre dans la formulation des LMI que les contraintes décrivant la première région extérieure.

3.2.2.4 Continuité sur la frontière

Nous avons vu que l'application de la méthode citée dans [JR98] pour imposer la continuité de la fonction de Lyapunov sur la frontière n'est pas évidente.

Pour remédier à ces problèmes, nous proposons d'exprimer la condition de la continuité de la fonction de Lyapunov de façon directe sans passer par les matrices F_i et F_j.

La condition de continuité de la fonctions de Lyapunov sur la frontière entre la région centrale et la région extérieure $x_2 > +e$ impose :

$$\begin{bmatrix} x \\ 1 \end{bmatrix}^T P_1 \begin{bmatrix} x \\ 1 \end{bmatrix} = x^T P_0 x, \quad x = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$$

pour $x_2 = e$.

En raison de la présence de la contrainte d'égalité, il se peut que des problèmes de conditionnement numériques se résultent au moment de la programmation du problème d'optimisation sous contraintes LMI. Pour éviter ce problème, nous proposons une paramétrisation de la matrice P_1 en fonction de la matrice P_0. Cette paramétrisation dans les LMI nous permet de vérifier de façon implicite la continuité de la fonction de Lyapunov.

Les matrices P_0 et P_1 dans le programme d'optimisation LMI sont des variables de décision, elles sont données comme suit :

$$P_0 = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix}, \quad P_1 = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{12} & q_{22} & q_{23} & q_{24} \\ q_{13} & q_{23} & q_{33} & q_{34} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{bmatrix}$$

d'après la condition de la continuité pour $x_2 = e$, nous pouvons écrire :

$$\begin{bmatrix} x_1 \\ e \\ x_3 \\ 1 \end{bmatrix}^T P_1 \begin{bmatrix} x_1 \\ e \\ x_3 \\ 1 \end{bmatrix} = \begin{bmatrix} x_1 \\ e \\ x_3 \end{bmatrix}^T P_0 \begin{bmatrix} x_1 \\ e \\ x_3 \end{bmatrix}$$

$$\Longleftrightarrow (p_{11}-q_{11})x_1^2+2(p_{12}e-(q_{12}e+q_{14}))x_1+2(p_{13}-q_{13})x_1x_3+2(p_{23}e-(q_{23}e+q_{34}))x_3+(p_{33}-q_{33})x_3^2$$
$$+(p_{22}e^2-2q_{24}e-q_{22}e^2-q_44)=0$$

nous pouvons donc exprimer les variables de la matrices P_1 en fonction de celles de la matrice P_0 :

$$\begin{cases} q_{11}=p_{11} \\ q_{14}=(p_{12}-q_{12})e \\ q_{13}=p_{13} \\ q_{34}=(p_{23}-q_{23})e \\ q_{33}=p_{33} \\ q_44=(p_{22}-q_{22})e^2-2q_{24}e \end{cases} \tag{3.8}$$

En remplaçant ces variables dans les LMI décrites précédemment et en prenant le reste comme variables de décision (p_{11}, p_{12}, p_{13}, p_{22}, p_{23}, p_{33}, q_{12}, q_{22}, q_{23} et q_{24}), des résultats intéressants ont été trouvés.

Cependant, cette méthode reste compliquée surtout dans le cas des matrices de grandes dimensions. pour cette raison, nous proposons d'exprimer cette condition de continuité de manière plus simple.

Nous avons paramétré la matrice P_1 en fonction de la matrice P_0 et la caractéristique de la frontière qui délimite la cellule $x_2 = e$, nous pouvons donc écrire :

$$P_1=\begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix}+G$$

comme P_1 est symétrique, G doit l'être aussi, donc elle s'écrire comme suit :

$$G=G_1^T+G_1$$

G_1 dépend de la caractéristique de la frontière qui délimite la cellule extérieure qui correspond à $x_2 > e$, elle peut donc se mettre de la façon suivante :

$$G_1=L\overline{E}_1$$

avec : $\overline{E_1}$ est le vecteur qui défini la cellule extérieure et L la partie restante. La matrice P_1 peut se paramétrer de façon générale comme suit :

$$P_1=\begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix}+(L\overline{E}_1)^T+L\overline{E}_1$$

P_0 et L dans ce cas sont les variables de décision.
Nous proposons de vérifier cette condition sur notre exemple de trois cellules.
D'après (3.8) :

$$\begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{12} & q_{22} & q_{23} & q_{24} \\ q_{13} & q_{23} & q_{33} & q_{34} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{bmatrix}=\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 \\ p_{12} & p_{22} & p_{23} & 0 \\ p_{13} & p_{23} & p_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}+\begin{bmatrix} 0 & q_{12}-p_{12} & 0 & (p_{12}-q_{12})e \\ p_{12}-q_{14}) & q_{22}-p_{22} & q_{23}-p_{23} & q_{24} \\ 0 & q_{23}-p_{23} & 0 & (p_{23}-q_{23})e \\ (p_{12}-q_{12})e & q_{24} & (p_{23}-q_{23})e & (p_{22}-q_{22})e^2-2q_{24}e \end{bmatrix} \tag{3.9}$$

$$(L\overline{E}_1)^T+L\overline{E}_1=\begin{bmatrix} 0 & l_1 & 0 & -el_1 \\ l_1 & 2l_2 & l_3 & l_4-el_2 \\ 0 & l_3 & 0 & -el_3 \\ -el_1 & l_4-el_2 & -el_3 & -2el_4 \end{bmatrix}$$

avec :

$$L = \begin{bmatrix} l_1 & l_2 & l_3 & l_4 \end{bmatrix},\ \overline{E}_1 = \begin{bmatrix} 0 & 1 & 0 & -e \end{bmatrix}$$

On a donc :

$$\begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{12} & q_{22} & q_{23} & q_{24} \\ q_{13} & q_{23} & q_{33} & q_{34} \\ q_{14} & q_{24} & q_{34} & q_{44} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 \\ p_{12} & p_{22} & p_{23} & 0 \\ p_{13} & p_{23} & p_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & l_1 & 0 & -el_1 \\ l_1 & 2l_2 & l_3 & l_4 - el_2 \\ 0 & l_3 & 0 & -el_3 \\ -el_1 & l_4 - el_2 & -el_3 & -2el_4 \end{bmatrix}$$

par comparaison avec (3.9), on aura donc :

$$\begin{cases} l_1 = q_{12} - p_{12} \\ l_2 = \frac{1}{2}(q_{22} - p_{22}) \\ l_3 = q_{23} - p_{23} \\ l_4 = q_{24} + el_2 = q_{24} + \frac{e}{2}(q_{22} - p_{22}) \end{cases}$$

Nous avons donc réussi à paramétrer les variables de la matrice P_1 sous la forme :

$$P_1 = \begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix} + (L\overline{E}_1)^T + L\overline{E}_1$$

Démonstration

Dans ce paragraphe, nous démontrons que la paramétrisation qui a été fait sur la matrice P_1 soit une relation suffisante et valable de façon générale.

On a :

$$P_1 = \begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix} + (L\overline{E}_1)^T + L\overline{E}_1$$

$$\Longleftrightarrow \begin{bmatrix} x \\ 1 \end{bmatrix} (P_1 - \begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix}) \begin{bmatrix} x \\ 1 \end{bmatrix} + \begin{bmatrix} x \\ 1 \end{bmatrix} (L\overline{E}_1)^T \begin{bmatrix} x \\ 1 \end{bmatrix} + \begin{bmatrix} x \\ 1 \end{bmatrix} L\overline{E}_1 \begin{bmatrix} x \\ 1 \end{bmatrix} = 0 \tag{3.10}$$

sachant que sur la frontière $x_2 = e$:

$$\text{pour } \begin{bmatrix} x \\ 1 \end{bmatrix} / \ \overline{E}_1 \begin{bmatrix} x \\ 1 \end{bmatrix} = 0 \tag{3.11}$$

en remplaçant cette condition dans 3.10, on aura :

$$\begin{bmatrix} x \\ 1 \end{bmatrix} (P_1 - \begin{bmatrix} P_0 & 0 \\ 0 & 0 \end{bmatrix}) \begin{bmatrix} x \\ 1 \end{bmatrix} = 0 \quad \text{car} \quad \begin{bmatrix} x \\ 1 \end{bmatrix} (L\overline{E}_1)^T \begin{bmatrix} x \\ 1 \end{bmatrix} + \begin{bmatrix} x \\ 1 \end{bmatrix} L\overline{E}_1 \begin{bmatrix} x \\ 1 \end{bmatrix} = 0 \tag{3.12}$$

Donc, cette paramétrisation linéaire est une condition suffisante qui nous a permet de préserver la propriété de la continuité de la fonction de Lyapunov sur la frontière.

3.2.2.5 Théorème proposé

nous pouvons exprimer les résultats suivants.

Théorème 2. *S'il existe $\tau_j,\ \tau'_j$ des coefficients non négatifs, $\overline{E}_i$ qui satisfait 3.3, pendant que :*

$$\overline{P}_j = P_i + \overline{E}_j^T L^T + L\overline{E}_j,\ i \in I_0,\ j \in I_1$$

$$Q_j = \begin{bmatrix} 0 \\ \overline{E}_j \end{bmatrix} + \begin{bmatrix} 0 & \overline{E}_j^T \end{bmatrix}$$

telles que les inégalités matricielles linéaires suivantes sont satisfaites :

$$\begin{cases} \begin{cases} A_i^T P_i + P_i A_i \prec 0 \\ P_i \succ 0 \end{cases} & i \in I_0 \\ \begin{cases} \overline{A}_i^T \overline{P}_i + \overline{P}_i \overline{A}_i + \tau_j Q_j \prec 0 \\ \overline{P}_i - \tau_j' Q_j \succ 0 \end{cases} & i \in I_1 \end{cases} \tag{3.13}$$

Alors toute trajectoire continue par morceaux $x(t) \in \cup_{i \in I} X_i$ tend vers zéro exponentiellement.

3.2.3 Étude du taux de décroissance

Le taux de décroissance représente le plus grand α positif tel que pour toute condition initiale x_0 [Sco12] :

$$\lim_{t \to \infty} e^{\alpha t} \parallel x(t) \parallel = 0$$

Il a été démontré une condition nécessaire et suffisante : le taux de décroissance est (strictement) supérieur à α s'il existe une fonction $V(x)$ telle que pour $x \neq 0$:

$$\begin{cases} V(x) > 0 \\ \dot{V}(x) < -2\alpha V(x) \end{cases} \tag{3.14}$$

Si on prend $V(x) = x^T P x$, P est une matrice symétrique à calculer.

$$(3.14) \Longleftrightarrow \begin{cases} P > 0 \\ A^T P + PA + 2\alpha P < 0 \end{cases} \Longleftrightarrow \begin{cases} -\alpha(2P) - (A^T P + PA) > 0 \\ 2P > 0 \end{cases}$$

En posant $\lambda = -\alpha$, ce problème ce transforme en un problème de minimisation de la valeur propre généralisée maximale donné comme suit :

$$\begin{aligned} \text{minimiser} \qquad & \lambda, \quad \text{pour } \lambda \in \mathbb{R}, \xi \in \mathbb{R}^{\geqslant} \\ \text{contraint par} \qquad & \lambda F(\xi) - G(\xi) > 0 \\ & F(\xi) > 0 \end{aligned}$$

avec : $F(\xi) = 2P$, $G(\xi) = A^T P + PA$

3.3 Application numérique

La programmation du problème d'optimisation sous contraintes LMI pour démontrer la stabilité de l'application pneumatique peut être effectué sous Matlab en utilisant sa propre bibliothèque ou d'autres bibliothèques. Dans notre cas, nous avons utilisé la bibliothèque Yalmip et Sedumi.

3.3.1 Le problème d'optimisation sous contraintes LMI

En appliquant le théorème 2, le problème d'optimisation sous contraintes LMI pour calculer les matrices de Lyapunov qui démontrent la stabilité du modèle pneumatique s'écrit comme suit.
S'il existe $\tau_1,\ \tau_2$ des coefficients non négatifs, soit E_1 donné par 3.4, pendant que :

$$P_1 = P_0 + E_1^T L^T + L E_1$$

$$Q_1 = \begin{bmatrix} 0 \\ E_1 \end{bmatrix} + \begin{bmatrix} 0 & E_1^T \end{bmatrix}$$

telles que les inégalités matricielles linéaires suivantes sont satisfaites :

$$\begin{cases} \begin{cases} A_0^T P_0 + P_0 A_0 + 2\alpha P_0 \prec 0 \\ P_0 \succ 0 \end{cases} \\ \begin{cases} A_1^T P_1 + P_1 A_1 + \tau_1 Q_1 + 2\alpha P_1 \prec 0 \\ P_1 - \tau_2 Q_1 \succ 0 \end{cases} \end{cases} \tag{3.15}$$

Alors toute trajectoire continue par morceaux $x(t) \in \cup_{i\in I} X_i$ tend vers zéro exponentiellement.

3.3.2 Résultats

Nous avons programmé l'exemple traité précédemment en utilisant la bibliothèque Yalmip. La programmation a été faite par dichotomie pour résoudre le problème de minimisation de la valeur propre généralisée maximale.
La valeur du taux de décroissance α trouvé est donné comme suit.

$$\alpha = 7.9041$$

Les autres résultats numériques sont présentés dans l'annexe A page 39.

Nous allons maintenant comparer la valeur de α aux valeurs propres des matrices d'état de chaque région du système commuté. Les valeurs propres des deux matrices d'état sont données comme suit.

$$VP(A_0) = \begin{bmatrix} -124.8 \\ -7.9041 + 7.54i \\ -7.9041 - 7.54i \end{bmatrix} \tag{3.16}$$

$$VP(A_1) = \begin{bmatrix} -100 \\ -8.54 + 8.71i \\ -8.54 - 8.71i \\ 0 \end{bmatrix} \tag{3.17}$$

Le taux de décroissance est inférieur ou égale à la valeur propre minimale de la matrice d'état de la région qui englobe l'origine. Il est calculé quelque soit la condition initiale. Pour une condition initiale proche de 0, la trajectoire d'état converge rapidement vers l'origine et le taux de décroissance dans ce cas donc est inférieur ou égale à la valeur 7.9041 qui est exactement la même valeur trouvée par le programme d'optimisation LMI. Alors, le taux de décroissance calculé colle sur la borne supérieure ce qui valorise ces résultats et l'efficacité de la méthode appliquée.

3.3.3 Conclusion

La méthode proposée a donné de bons résultats et nous a permis de démontrer la stabilité pour l'application pneumatique ce qui est impossible à réaliser en appliquant la méthode de [JR98].

Les résultats de la programmation nous ont montré la différence entre l'utilisation de la forme quadratique et la forme linéaire pour le choix du terme de la S-procédure. En effet, le choix de la forme quadratique va créer un certain conservatisme ce qui donne une valeur du taux de décroissance inférieur à celle trouvé en choisissant la forme linéaire.

Le résultat trouvé nous a permis de dire que la méthode proposé a apporté de contributions pour l'étude de la stabilité de systèmes commutés.

Conclusion et Perspectives

Durant ce travail, nous avons essayé de traiter une problématique de la stabilité sur une classe particulière de systèmes dynamiques hybrides en considérant une application connue dans le monde industriel qui est le domaine pneumatique. Nous avons vu que la solution pour le problème de départ n'était pas évidente à trouver vu la complexité de ce problème. Il fallait donc penser autrement et simplifier ce problème de façon à pouvoir appliquer les méthodes développées dans la littérature et les améliorées si nécessaire.

Nous avons pris comme méthode d'analyse celle développée dans l'article [JR98] qui transforme l'étude de la stabilité en un problème d'optimisation sous contraintes LMI et de calculer numériquement les matrices de la fonction de Lyapunov quadratique par morceaux. L'application de cette méthode sur notre problème simplifié n'a pas pu apporter de résultats pour plusieurs raisons, ce qui nous a obligé à améliorer cette méthode en considérant les points suivants :

- nous avons commencé par changer le terme de la S-procédure en faisant la liaison avec les relations qui délimités les cellules, c'est à dire la forme quadratique considérée pour pouvoir appliquer la S-procédure est la superposition des deux formes linéaires qui présentent la cellule et sa symétrique. Prendre la forme quadratique sous une telle forme a l'avantage de réduire le conservatisme et de chercher l'ensemble de solutions uniquement pour le problème considéré.
- le problème persistait toujours, cette fois-ci le choix des matrices de Lyapunov et la façon d'écrire la condition de continuité de la fonction de Lyapunov sur la frontière présentaient plus de conservatisme et le problème était infaisable. Il a fallu écrire cette condition de continuité de manière plus simple et générale en faisant une paramétrisation des matrices de Lyapunov. Ce choix a cassé le conservatisme et nous avons trouvé de bons résultats.

Ces résultats trouvées nous ont donné la volonté de continuer et d'essayer de réaliser les perspectives suivantes :

- généraliser les deux points améliorés pour d'autre types de problèmes ;
- revenir au problème de départ en essayant de trouver un modèle qui permet de générer le phénomène de redécollage ;
- appliquer la méthode proposée sur ce modèle et essayer de résoudre le problème.

Appendices

Annexe A

Formes quadratiques et résultats de programmation

A.1 La forme de Ω_i

1) Cas $u_{11} \neq 0$:

On a :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^T \begin{bmatrix} 1 & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \geq 0$$

avec :

$$y_1 = e_1^T x \, , y_2 = e_2^T x$$

$$\Longleftrightarrow y_1^2 + 2\, u_{12}\, y_1 y_2 + u_{22}\, y_2^2 \geq 0$$

prenons $t = (\frac{y_1}{y_2})$:

$$t^2 + 2\, u_{12}\, t + u_{22} \geq 0$$

$$\begin{aligned} &\Longleftrightarrow (t + u_{12})^2 + u_{22} - u_{12}^2 \geq 0 \\ &\Longleftrightarrow \qquad (t + u_{12})^2 - \Delta' \geq 0 \end{aligned} \tag{A.1}$$

avec $\Delta' = u_{12}^2 - u_{22}$
Deux cas à envisager selon le discriminant réduit $\acute{\Delta}$:

1) $\Delta' \leq 0$: l'inégalité $(t + u_{12})^2 \geq \Delta'$ est toujours vérifiée $\forall t \in \mathbb{R}$, et donc la solution de (A.1) dans la base (x_1, x_2) est $\mathbb{R}^2$.

3) $\Delta' > 0$: l'inégalité (A.1) a deux solutions :

$$\begin{cases} t > -u_{12} + \sqrt{\Delta'} \\ t < -u_{12} - \sqrt{\Delta'} \end{cases}$$

$t_1 = -u_{12} + \sqrt{\Delta'}$ et $t_2 = -u_{12} - \sqrt{\Delta'}$. Dans ce cas (A.1) est vraie dans l'intervalle $]-\infty, t_2] \cup [t_1, +\infty[$

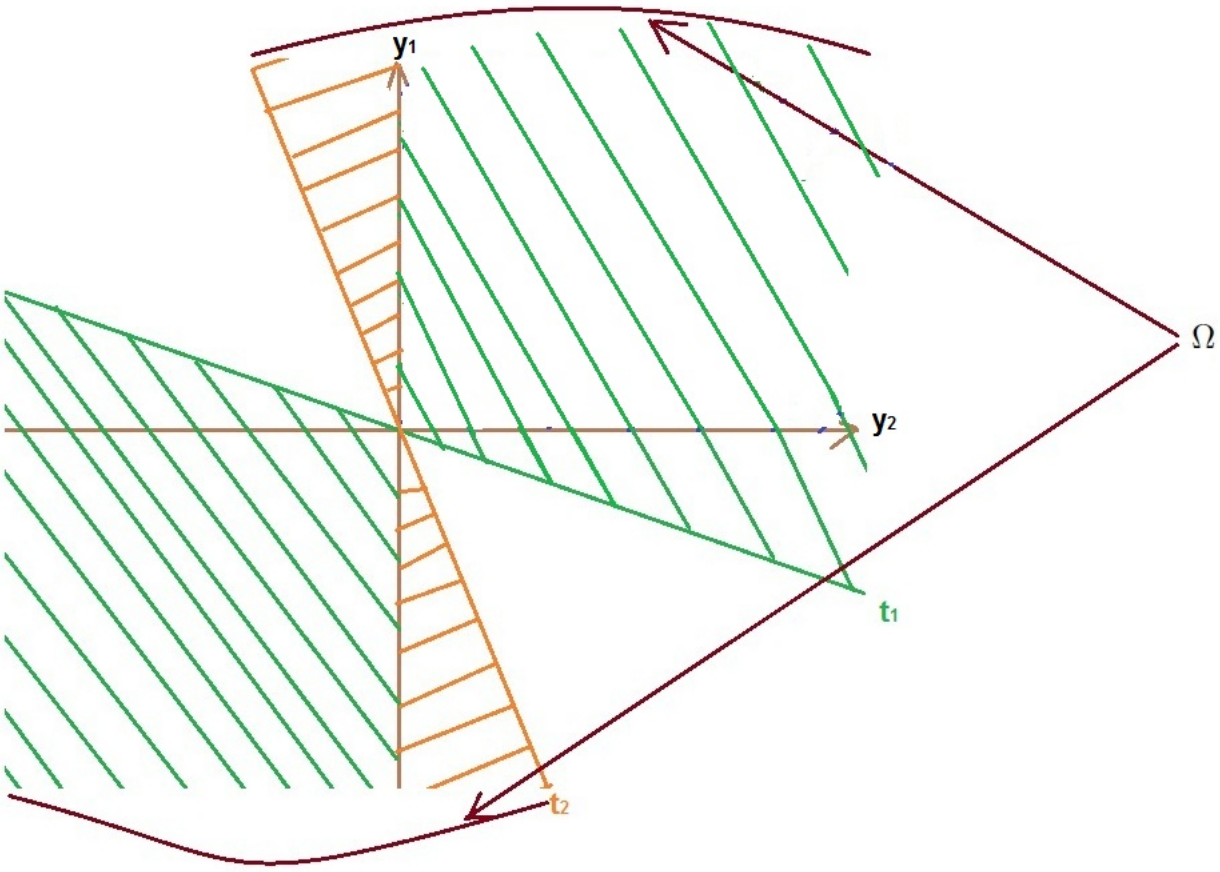


FIGURE A.1 – La forme de Ω_i

Pour $t > t_1 \Longleftrightarrow \frac{y_1}{y_2} > t_1$:

$$\begin{cases} y_2 > 0 \Longrightarrow y_1 > t_1 y_2 & \text{la partie hachurée en vert dans le demi plan } y_2 > 0 \quad (figure\,(A.1)) \\ y_2 < 0 \Longrightarrow y_1 < t_1 y_2 & \text{la partie hachurée en vert dans le demi plan } y_2 < 0 \quad (figure\,(A.1)) \end{cases}$$

Le même raisonnement pour $t < t_2 \Longleftrightarrow \frac{y_1}{y_2} < t_2$:

$$\begin{cases} y_2 > 0 \Longrightarrow y_1 < t_2 y_2 & \text{la partie hachurée en orange dans le demi plan } y_2 > 0 \quad (figure\,(A.1)) \\ y_2 < 0 \Longrightarrow y_1 > t_2 y_2 & \text{la partie hachurée en orange dans le demi plan } y_2 > 0 \quad (figure\,(A.1)) \end{cases}$$

Exemple

Prenons comme un exemple :

$$E_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

le sous espace X_1 décrit par $E_1 x \geq 0$ correspond au quart de plan positif (figure (2.1)). Dans ce cas :

$$\begin{cases} y_1 = e_{11}x_1 + e_{12}x_2 = x_1 \\ y_2 = e_{21}x_1 + e_{22}x_2 = x_2 \end{cases}$$

choisissons maintenant des coefficients positifs pour la matrice U_1 :

$$\begin{cases} u_{12} = 2 \\ u_{22} = 1 \end{cases}$$

En remplaçant tout dans l'équation A.1, on obtient alors :

$$(t+2)^2 - 3 \geq 0 \tag{A.2}$$

$\Delta' = 3 \Longrightarrow t_{1,2} = -2 \pm \sqrt{3}$
la solution de l'inégalité A.2 est donc l'intervalle $]-\infty, t2] \cup [t_1, +\infty[$
Pour $t > t_1 \Longleftrightarrow \frac{x_1}{x_2} > t_1$:

$$\begin{cases} x_2 > 0 \Longrightarrow x_1 > t_1 x_2 & \text{la partie hachurée en vert dans le demi plan } x_2 > 0 \quad (figure\ 2.1) \\ x_2 < 0 \Longrightarrow x_1 < t_1 x_2 & \text{la partie hachurée en vert dans le demi plan } x_2 < 0 \quad (figure\ 2.1) \end{cases}$$

Le même raisonnement pour $t < t_2 \Longleftrightarrow \frac{x_1}{x_2} < t_2$:

$$\begin{cases} x_2 > 0 \Longrightarrow x_1 < t_2 x_2 & \text{la partie hachurée en orange dans le demi plan } x_2 > 0 \quad (figure\ 2.1) \\ x_2 < 0 \Longrightarrow x_1 > t_2 x_2 & \text{la partie hachurée en orange dans le demi plan } x_2 > 0 \quad (figure\ 2.1) \end{cases}$$

2) Cas $u_{11} = 0$:

On a :

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}^T \begin{bmatrix} 0 & u_{12} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \geq 0$$

avec :

$$y_1 = e_1^T x \, , y_2 = e_2^T x$$

est équivalent à

$$2\, u_{12}\, y_1 y_2 + u_{22}\, y_2^2 \geq 0.$$

Prenons $t = (\frac{y_1}{y_2})$:

$$2\, u_{12}\, t + u_{22} \ \geq 0 \tag{A.3}$$

Si $u_{12} = 0$: l'inégalité (A.3) est toujours vérifiée, donc on a tout le plan $\mathbb{R}^2$.
Si $u_{12} \neq 0$:

$$t \geq -\frac{u_{22}}{2u_{12}}$$

$t \geq -\frac{u_{22}}{2u_{12}} \Longleftrightarrow \frac{y_1}{y_2} \geq -\frac{u_{22}}{2u_{12}}$:

$$\begin{cases} y_2 > 0 \Longrightarrow y_1 \geq -\frac{u_{22}}{2u_{12}} y_2 & \text{la partie hachurée en bleue } (figure\,(A.2)) \\ y_2 < 0 \Longrightarrow y_1 \leq -\frac{u_{22}}{2u_{12}} y_2 & \text{la partie hachurée en vert } (figure\,(A.2)) \end{cases}$$

A.2 Les formes quadratiques affines

L'idée présentée dans le paragraphe précédent sert à réécrire les inégalités exprimant chaque domaine sous une forme quadratique standard "$x^T Q x$" pour pouvoir manipuler à la fin des LMI à l'aide de la méthode S-procédure. Généralement, il est toujours possible de se ramener à partir d'une forme quadratique présentée de la manière suivante :

$$x^T P x + x^T q + q^T x + r \geq 0 \tag{A.4}$$

vers cette forme quadratique standard.

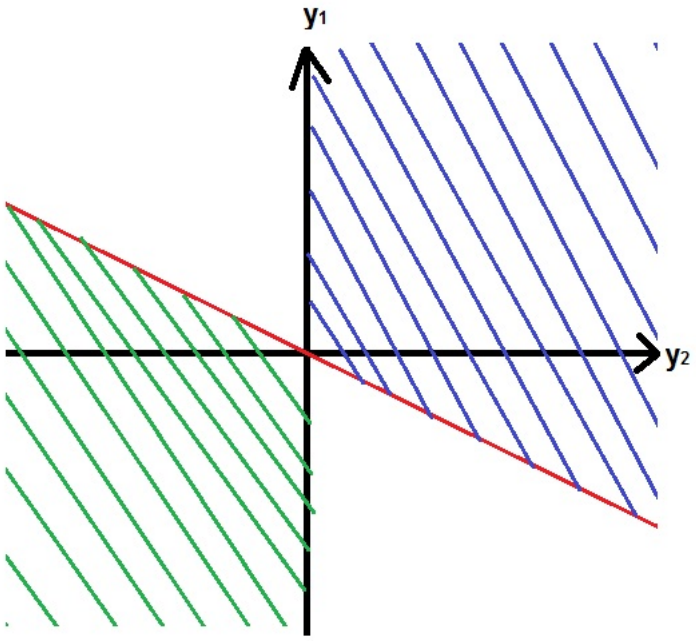


FIGURE A.2 – la forme de Ω_i dans un cas particulier de la matrice U_i

l'idée est simple, procédons comme suit :
On a :

$$x^T P x + x^T q + q^T x + r = \begin{bmatrix} x \\ 1 \end{bmatrix}^T \begin{bmatrix} P & q \\ q^T & r \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \tag{A.5}$$

En multipliant par un coefficient $\alpha \in \mathbb{R}^+$:

$$\begin{bmatrix} \alpha x \\ \alpha \end{bmatrix}^T \begin{bmatrix} P & q \\ q^T & r \end{bmatrix} \begin{bmatrix} \alpha x \\ \alpha \end{bmatrix} \geq 0 \tag{A.6}$$

Prenons : $\tilde{x} = \begin{bmatrix} \alpha x \\ \alpha \end{bmatrix}$, donc on aura la forme quadratique standard :

$$\tilde{x}^T \begin{bmatrix} P & q \\ q^T & r \end{bmatrix} \tilde{x} \geq 0, \quad \forall \tilde{x} \in \mathbb{R}^{n+1} \tag{A.7}$$

A.3 Les résultats de la programmation LMI

La matrice de Lyapunov pour la région centrale :

$$P_0 = \begin{bmatrix} 23.6724 & 16.5873 & 4.7421 \\ 16.5873 & 22.6412 & 3.1647 \\ 4.7421 & 3.1647 & 1.2196 \end{bmatrix}$$

La matrice L issue de la paramétrisation de P_1

$$L = \begin{bmatrix} 1.9540 \\ 1.0399 \\ 0.3581 \\ 0.1077 \end{bmatrix}$$

La matrice de Lyapunov pour la région extérieur :

$$P_1 = \begin{bmatrix} 23.6724 & 18.5413 & 4.7421 & -0.1954 \\ 18.5413 & 24.7210 & 3.5228 & 0.0037 \\ 4.7421 & 3.5228 & 1.2196 & -0.0358 \\ -0.1954 & 0.0037 & -0.0358 & -0.0215 \end{bmatrix}$$

les valeurs propres de P_0 :

$$\lambda_0 = \begin{bmatrix} 0.2557 \\ 6.7252 \\ 40.5522 \end{bmatrix}$$

les constante t1 et t2 :

$$t1 = 5.3667$$

$$t2 = 0.7477$$

les valeurs propres de $P_1 - t_2 Q_1$:

$$\lambda_1 = \begin{bmatrix} 0.0153 \\ 0.2829 \\ 5.8903 \\ 43.5524 \end{bmatrix}$$

les valeurs propres de $A_0^T P_0 + P_0 A_0 + 2\alpha P_0$:

$$\lambda_2 = \begin{bmatrix} -71.0991 \\ -0.0002 \\ -0.0000 \end{bmatrix}$$

les valeurs propres de $A_1^T P_1 + P_1 A_1$:

$$\lambda_3 = \begin{bmatrix} -711.2865 \\ -94.9381 \\ -51.1618 \\ 0.1732 \end{bmatrix}$$

les valeurs propres de $A_1^T P_1 + P_1 A_1 + t_1 Q_1 + 2\alpha P_1$

$$\lambda_4 = \begin{bmatrix} -47.5029 \\ -23.3626 \\ -2.9587 \\ -0.5163 \end{bmatrix}$$

La condition de la continuité : vérifiée!!!

le coefficient de décroissance α trouvé :

$$\alpha = 7.9041$$

Annexe B

Simulation

B.1 Modèle de commande “avec frottements”

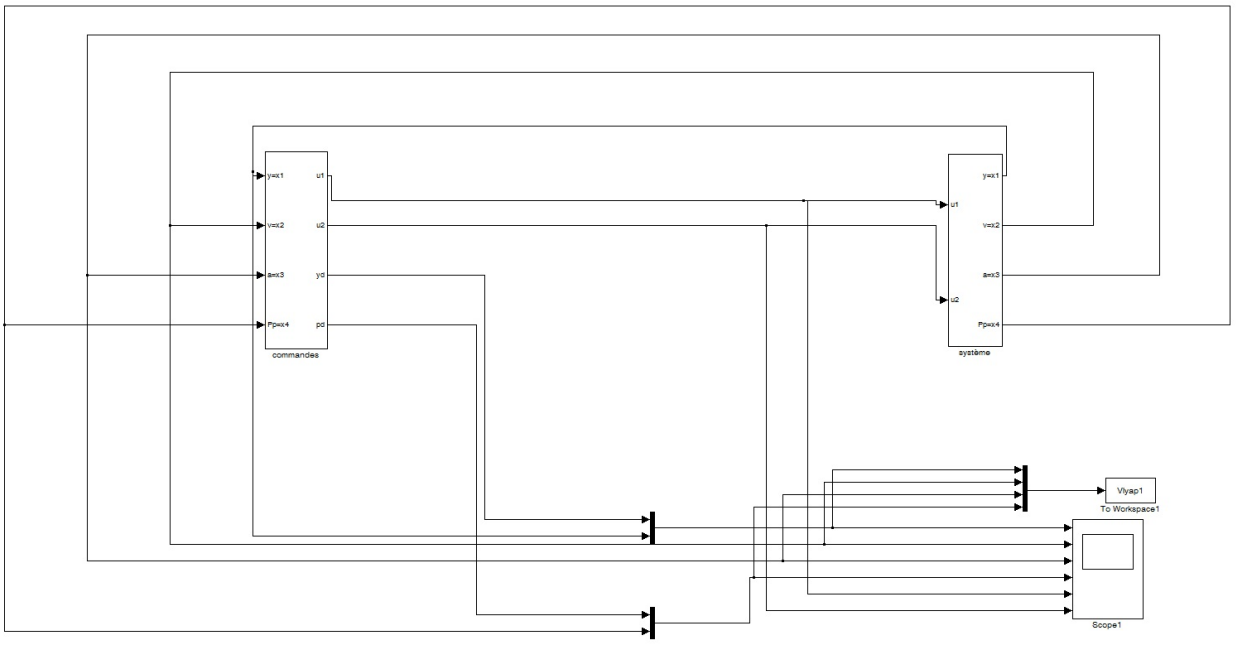


Figure B.1 – Modèle simulink : modèle avec frottements

Modèle d’analyse “avec frottements”

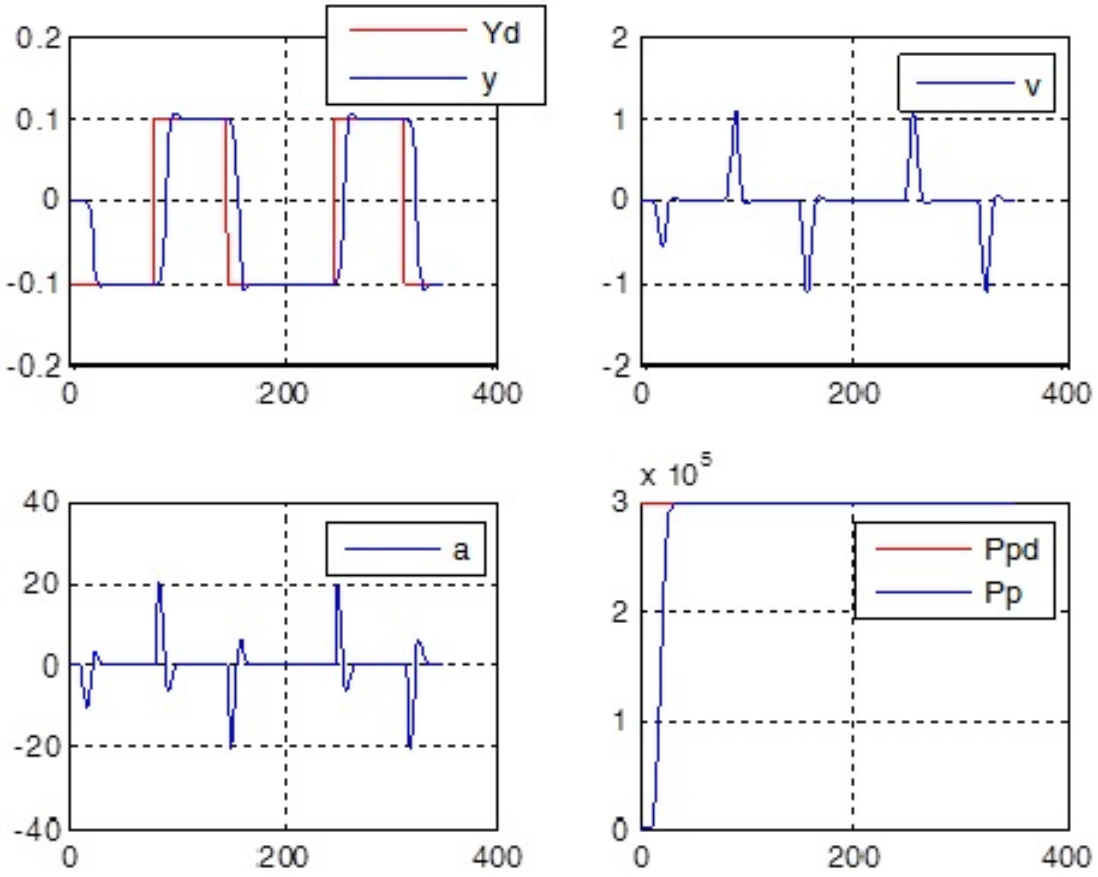


FIGURE B.2 – les trajectoires d’état : modèle sans frottements

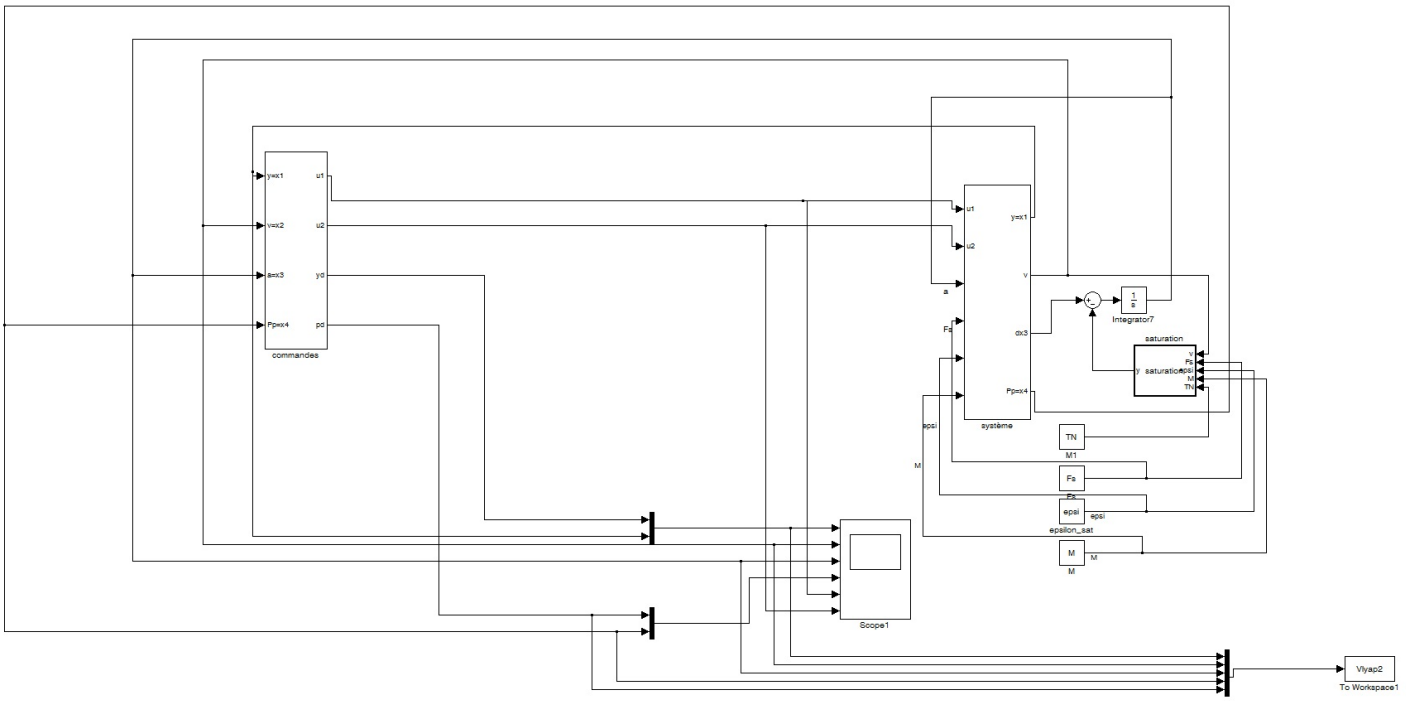


FIGURE B.3 – Modèle simulink : modèle avec frottements

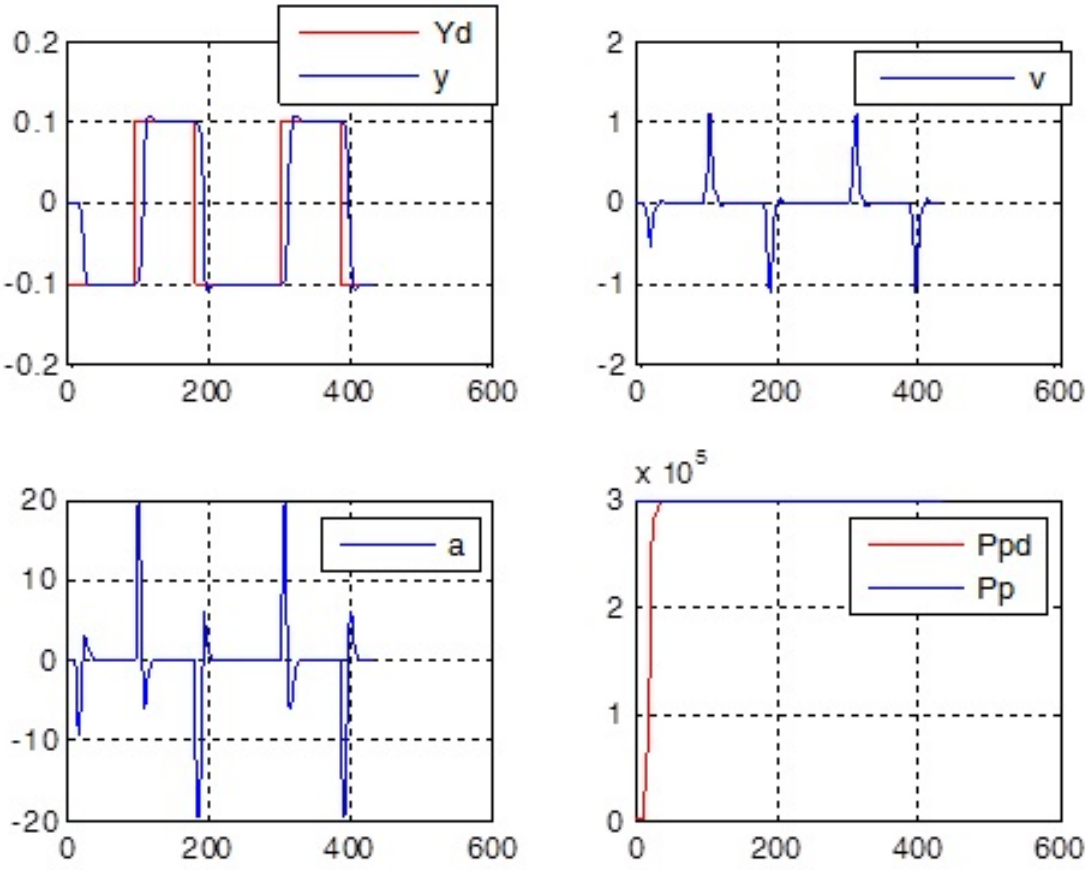


FIGURE B.4 – les trajectoires d’état : modèle avec frottements

Bibliographie

[AGAA12] R. Ambrosino, E. Garone, M. Ariola, and F. Amato. Piecewise quadratic functions for finite-time stability analysis. *51st IEEE Conference on Decision and Control,kh Maui, Hawaii, USA*, 2012.

[AMWZ12] M. Ahmadi, H. Mojallali, R. Wisniewski, and R. Izadi Zamanabadi. Robust stability and h control of uncertain piecewise linear switched systems with filippov solutions. *IEEE Multi-Conference on Systems and Control*, 2012.

[AOFY11] O. M. Abou Al-Ola, K. Fujimoto, and T. Yoshinaga. Common lyapunov function based on kullback–leibler divergence for a switched nonlinear system. *Hindawi Publishing Corporation, Mathematical Problems in Engineering*, 10.1155/2011/723509, 2011.

[BGFV94] S. Boyd, L. El Ghaoui, E. Feron, and V.Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15. Society for Industrial and Applied Mathematics, 1994.

[BJ12] M. BALDE and P. JOUAN. Stability of linear switched systems with quadratic bounds and observability of bilinear systems. 2012.

[Bra95] M. S. Branicky. *Studies in Hybrid Systems : Modeling, Analysis, and Control*. Department of Electrical Engineering and Computer Science, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 1995.

[Bra96] M. S. Branicky. *GeneralHybrid Dynamical Systems : Modeling, Analysis, and Control*. Laboratory for Informationand Decision Systems, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 1996.

[Bra98] M. S. Branicky. Multiple lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE Trans. Autom. Control*, 43(4) :475–482, 1998.

[Bru99] X. Brun. *Commandes linéaires et non linéaires en électropneumatique. Méthodologies et Applications*. PhD thesis, l'INSA de Lyon, 1999.

[BT01] M. Bouri and D. Thomasset. Sliding control of an electropneumatic actuator using an integral switching surfacer. *IEEE Transactions on control systems technology*, Vol. 9. n :2, pp. 368-375, 2001.

[CSI95] M. Chang and S. Shy-I. Identification and position control of a servo pneumatic cylinder. *Control Engineering Practice*, Vol. 9. pp. 1285–1290, 1995.

[DBPL00] R. A. DeCarlo, M. S. Branicky, S. Petterson, and B. Lennartson. Perspectives and results on the stability and stabilizability of hybrid systems. *In Proceedings of the IEEE. Special issue of hybrid systems*, volume 88, pages 1069–1082,, 2000.

[Hä02] T. Hägglund. A friction compensator for pneumatic control valves. *Journal of Process Control*, Vol. 12, N :8, pp. 897-904, 2002.

[HVBB96] K. Hamiti, A. Voda-Besancon, and R. Buisson. Position control of a pneumatic actuator under the influence of stiction. *Control Engineering Practice*, Vol. 4, N :8, pp. 1079-1088, 1996.

[Jö04] U. Jönsson. *The S-Procedure and its applications in IQC Anaysis*. Optimisation and Systems Theory, Departement of Mathematics,Royal Institute of Technology (KTH),Stockholm, Sweden, Feb 2004.

[Joh02] M. Johansson. Picewise quadratic estimates of domains of attraction for linear systems with saturation. *15th Triennial World Congress, Barcelona, Spain*, 2002.

[JR98] M. Johansson and A. Rantzer. Computaion of piecewise quadratique Lyapunov functions for hybrid systems. vol.43(.4), April 1998.

[LHM99] D. Liberzon, J. P. Hespanha, and A. S. Morse. Stability of switched linear systems : a lie-algebraic condition. *Systems and Control Letters*, :117–122,37(3), 1999.

[Lib03] D. Liberzon. *Switching in systems and control*. Systems and Control : Foundations and Applications. Birkh auser Boston. Inc., Boston, MA, 2003.

[LM99] D. Liberzon and A. S. Morse. Basic problems in stability and design of switched system. *IEEE Control Systems*, 19(5) :59–70, 1999.

[MTM00] D. Mignone, G. F. Trecate, and M. Morari. Stability and stabilization of piecewise affine and hybrid systems : An lmi approach. 2000.

[NB94] K. S. Narendra and J. Balakrishnan. A common lyapunov function for stable lti system with commuting a-matrices. *IEEE Trans. Automat. Control*, 39(12) :2469–2471, 1994.

[PL96] S. Pettersson and B. Lennartson. Stability and robustness for hybrid systems. *In 35th Conf. Decision and Contr*, 1996.

[PP09] A. Papachristodoulou and S. Prajna. Robust stability analysis of nonlinear hybrid systems. 2009.

[RSD10] P. Riedinger, M. Sigalotti, and J. Daafouz. On the algebraic characterization of invariant sets of switchedlinear systems. *Automatica*, 46, 6 (2010) 1047-1052, 2010.

[SBT06] M. Smaoui, X. Brun, and D. Thomasset. A study on tracking position control of an electropneumatic system using backstepping design. *Control Engineering Practice*, Vol. 14, n :8, pp. 923-933, 2006.

[SBT08] M. Smaoui, X. Brun, and D. Thomasset. High order sliding mode for an electropneumatic system : A differentiator-controllers design. *International Journal of Robust and Nonlinear Control*, Vol. 18, pp. 481-501, 2008.

[Sco12] G. Scorletti. Optimisation et sciences de l'ingénieur : une approche basée sur l'automatique. Technical report, Master Recherche ESCI 2A Université de Caen Basse Normandie, 2012.

[SN98] R. N. Shorten and K. S. Narendra. On the stability and existence of common lyapunov functions for stable linear switching systems. *IEEE, Conference on Decision and Control Tampa, Florida USA December 1998*, 39(12) :2469–2471, 1998.

[SS99] A. V. Dar Schaft and H. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Lecture Notes in Control and Information Sciences, LNCIS 251, Springer-Verlag, London, 1999.

[SWM+07] R. Shorten, F. Wirth, O. Mason, K. Wulff, and C. King. Stability criteria for switched and hybrid systems. *SIAM Rev*, 49(4) :545–592, 2007.

[TT09] T. Hu T. Thibodeau and A. R. Teel. Analysis of oscillation and stability for systems with piecewise linear components via saturation functions. *2009 American Control Conference. Hyatt Regency Riverfront, St. Louis, MO, USA*, 2009.

[Tur10] K. Turki. *Nouvelles Approches Pour La Synthèse De Loi de Commande Non Linéaires Robustes. Application À Un Actionneur Electropneumatique Et Proposition D'une Solution Au Problème De Redécollage*. PhD thesis, l'INSA de Lyon, 2010.

[VL07] L. Vu and D. Liberzon. *Common Lyapunov functions for families of commuting nonlinear systems*. Coordinated Science Laboratory Univ. of Illinois at Urbana-Champaign Urbana, IL 61801, U.S.A, 2007.

Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

---

Mémoire doctorant 1ère année 2012 -2013

Nom - Prénom	**AZHARI Soufiane**
Titre de la thèse	**Contribution à l'étude d'un houlogénérateur**
Directeur de thèse	**Guy CLERC**
Co- encadrants	**Eric BLANCO**
Dpt. de rattachement	**EEA**
Date début des travaux	**02/10/2012**
Type de financement	**FUI**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

Table des matières

Liste des figures

Liste des tableaux

Abréviations et Notations

u_{dc_0} : Valeur limite de la tension seuil du bus continu du back-to-back

$u_{dc_{min}}$: Valeur seuil de la tension de du bus continu du back-to-back

V_{1eff} : Tension efficace simple de la machine

u_{dc} : Tension du bus continu du back-to-back

Δu_{dc} : Ondulation de tension du bus continu du back-to-back

$\Delta u_{\alpha\beta}$: Tensions de compensation dans le référentiel (α, β)

e_{dc} : Tension continue équivalent de la tension réseau pour le back-to-back

t_d : Temps mort

t_{on} : Temps de mise en conduction d'un IGBT

V_{dc} : Tension d'alimentation de l'onduleur triphasé de tension

f_s : Fréquence de découpage de la MLI

A^+ : Ordre de commande de l'interrupteur haut du bras A de l'onduleur

A^- : Ordre de commande de l'interrupteur haut du bras A de l'onduleur

T_a: Durée de conduction réelle de l'IGBT haut du bras de l'onduleur

T_a^* : Durée de conduction idéale de l'IGBT haut du bras de l'onduleur

u_{ao} : Tension réelle de sortie du bras de l'onduleur

i_a : Courant du bras A de l'onduleur

u_{n0} : Tension du neutre

T_{dead_i} : Erreur en temps d'application de la tension de sortie du bras i de l'onduleur ($i \in \{a, b, c\}$)

V_{in} : Tension d'entrée du bloc de commutation

V_{out} : Tension de sortie du bloc de commutation

I_{in} : Courant d'entrée du bloc de commutation

I_{out} : Courant de sortie du bloc de commutation

ρ_1 : Rapport cyclique de conduction de l'IGBT 1

ρ_2 : Rapport cyclique de conduction de l'IGBT 2

ρ_d : Rapport cyclique relatif au temps mort

T_{32} : Transformation de Concordia

u_{err} : Distorsion de tension pour la méthode 2/catégorie 1

E : Tension d'alimentation de l'onduleur triphasé

MLI : Modulation de largeur d'impulsions

$SVPWM$: Space Vector Width Modulation

V_{dqref_ond}: Tensions de modulation de l'onduleur du convertisseur back-to-back dans le référentiel (d,q)

V_{dqref_red} Tensions de modulation du redresseur du convertisseur back-to-back dans le référentiel (d,q)

T_{1hon} : Instant d'ordre de fermeture de l'IGBT (interrupteur haut du bras)

T_{1hoff} : Instant d'ordre d'ouverture de l'IGBT (interrupteur haut du bras)

T_{1hon*} : Instant corrigé d'ordre de fermeture de l'IGBT (interrupteur haut du bras)

T_{1hoff*} Instant corrigé d'ordre d'ouverture de l'IGBT (interrupteur haut du bras)

T_{1lon} : Instant d'ordre de fermeture de l'IGBT (interrupteur bas du bras)

T_{1lon*} : Instant corrigé d'ordre de fermeture de l'IGBT (interrupteur bas du bras)

T_{1loff} : Instant d'ordre d'ouverture de l'IGBT (interrupteur bas du bras)

T_{1loff*} : Instant corrigé d'ordre de fermeture de l'IGBT (interrupteur bas du bras)

C_{em} : Couple électromagnétique de la machine

ϕ : Flux total de la machine synchrone

P : Puissance active

Q : Puissance réactive

V_d : Tension visualisée, sur l'axe (d), de la charge sans compensation des temps morts

V_{dcomp} : Tension visualisée de la charge avec compensation des temps morts

V_q : Tension visualisée sur, sur l'axe (q), de la charge sans compensation

V_{qcomp} : Tension de la charge sur l'axe q avec compensation

1. Introduction

Les ressources énergétiques renouvelables sont, à notre échelle de temps, les seules ressources dispensées continûment par la nature. Sur la Terre, elles ont pour origine le rayonnement solaire, la chaleur du noyau terrestre et les interactions gravitationnelles de la lune et du soleil avec les océans.

L'humanité consomme annuellement, en ce début de troisième millénaire, très approximativement 12 Gtep[1] d'énergie primaire, soit une quantité correspondant à 80. 10^{-12} de l'énergie solaire qui arrive à la surface de la terre [1][2]. La production d'électricité mondiale quant à elle représente environ 17.1012 kWh/an.

Parmi ces ressources, notre intérêt porte essentiellement sur l'énergie des vagues de la mer notamment la houle dont la quantité disponible est évaluée de 140 à 170 TWh[2] par an d'après le World Energy Council (WEC)[3], soit 1 à 5% de la demande annuelle mondiale en électricité. La puissance moyenne par mètre de front de vague possède des valeurs comprises entre 10 et 100 kW/m.

Bien entendu cette ressource totale est très supérieure à la ressource effectivement accessible en tenant compte des limitations techniques et des limitations naturelles ou légales qui réduisent le domaine utilisable, sans parler de l'acceptabilité sociale qui peut interdire encore certains sites.

Bien que fluctuante, elle répond à la fois aux problèmes économiques et environnementaux. Afin de ne pas l'exploiter dans les lieux les plus défavorables et de lisser ses fluctuations, il s'avère indispensable de diversifier les solutions. A cet effet, de nombreuses recherches dans ce domaine sont réalisées depuis quelques dizaines d'années avec une certaine accélération récente.

Le présent rapport traite non seulement la mise en contexte de développement d'un nouveau système de récupération de l'énergie de la houle, mais aussi l'étude de la partie électrique de sa chaine de conversion d'énergie notamment le niveau de la commande rapprochée des convertisseurs statiques mis en jeu pour assurer le transfert de la puissance extraite et le modèle moyen qui sera adopté pour décrire le comportement dynamique de la puissance active et/ou réactive transférés au réseau.

Aussi nous allons présenter le développement des méthodes de compensation des non linéarités inhérentes aux composants à semi-conducteurs (IGB/Diode) d'un onduleur triphasé de tension après avoir exposé un état de l'art sur ces méthodes, à la lumière de l'étude bibliographique que nous avons faite. La mise au point des méthodes de compensation nécessite un modèle très fin des convertisseurs exigeant des pas de

[1] Gtep= 41,86 GJ
[2] 1TWh=3.6 x 10^{15} Joules
[3] WEC : Fondé en 1923, le Conseil Mondial de l'Energie est la principale organisation multi-énergétique mondiale. Son objectif est de « promouvoir la fourniture et l'utilisation durables de l'énergie pour le plus grand bien de tous » en mettant en avant les questions d'accessibilité, de disponibilité et d'acceptabilité énergétiques.

simulation très fins. Cependant, ces temps de simulation incompatibles avec la mise au point des commandes haut niveau pour le pilotage du Bilboquet. Il est donc nécessaire d'élaborer un modèle moyen dont le développement est présenté dans ce document. Des résultats de simulations seront présentés pour valider les méthodes de compensation et le modèle moyen.

2. Contexte général de l'étude

2.1. Présentation du contexte de la thèse

Pendant ces deux dernières décennies, de nombreuses recherches ont été menées sur la conception des systèmes de récupération d'énergie des vagues. Aussi trois catégories technologiques ont été développées : les systèmes à rampe de déferlement, systèmes à colonne d'eau oscillants et système à corps flottants [2] [3].

La troisième catégorie consiste en des systèmes à corps mus par la houle .Ils sont composés souvent d'une partie qui oscille avec les mouvements de la houle et d'une partie fixe permettant d'exploiter le mouvement relatif pour entrainer des génératrices.

Le **BILBOQUET** illustré sur la Figure 1: système houlogénérateur le BILBOQUET est issue de cette dernière technologie et il est destiné à produire une puissance qui s'élève à 9.6 MW, grâce à quatre génératrices synchrones à aimants permanents de 2.4 MW chacune. Il est constitué d'un flotteur guidé qui se déplace le long d'une colonne flottante et ancrée. Le flotteur est équipé de crémaillères qui actionnent des génératrices situées dans une capsule à la partie supérieure de la colonne par un jeu de pignons et de multiplicateur. La colonne est équipée à sa partie inferieure d'un plateau pesant permettant à la fois un amortissement hydrodynamique et un ancrage par 3 points qui contribue à une grande stabilité verticale. De plus, ce houlogénérateur est relié au réseau via des convertisseurs de type Back-to-Back qui consiste en un ensemble redresseur et onduleur montés dos à dos reliés par un bus continu.

Par ailleurs, le projet BILBOQUET s'inscrit dans le cadre de développement d'un système de récupération de l'énergie des vagues de la mer nécessitant différentes expertises dans des domaines variés notamment l'hydrodynamique, mécanique et génie électrique.

Aussi différents partenaires industriels sont impliqués dans le Projet BILBOQUET:

- Mécanique et conception :
  - ✓D2M
  - ✓CMD
  - ✓CERVAL
- Normes et réglementation d'essais en mer :
  - ✓Bureau VERITAS

- ✓IFREMER
- Electronique de puissance :
  - ✓ADENEO
- Laboratoire de Recherche :
  - ✓AMPERE
  - ✓LBMS

Le laboratoire Ampère est en charge de développement de lois de commande des convertisseurs statiques. Aussi le sujet de thèse a pour objectif de concevoir une loi de pilotage du générateur permettant d'extraire à tout instant le maximum de puissance et du système de surveillance des composantes du système

Du point de vue méthodologie, les travaux de recherches se dérouleront sur différents niveaux .En effet, sur le plan théorique, le travail comprend une modélisation de la partie électrique en vue du développement d'une commande satisfaisant un cahier de charge en termes de transfert de puissance avec le réseau.

Deux modèles seront développés : un modèle de validation prenant en compte le système dans sa globalité et simulant finement la Modulation de largeur d'impulsions et un modèle moyen permettant de mettre au point la commande avec des temps de simulation plus raisonnable. Deux types d'algorithme seront développés: la commande rapprochée (la commande globale est mis au point par le LBMS) et la supervision des composants du système .Les algorithmes ainsi obtenus seront validés en simulation et par la suite testés sur un procédé pilote implanté au laboratoire (Banc <5 kW).

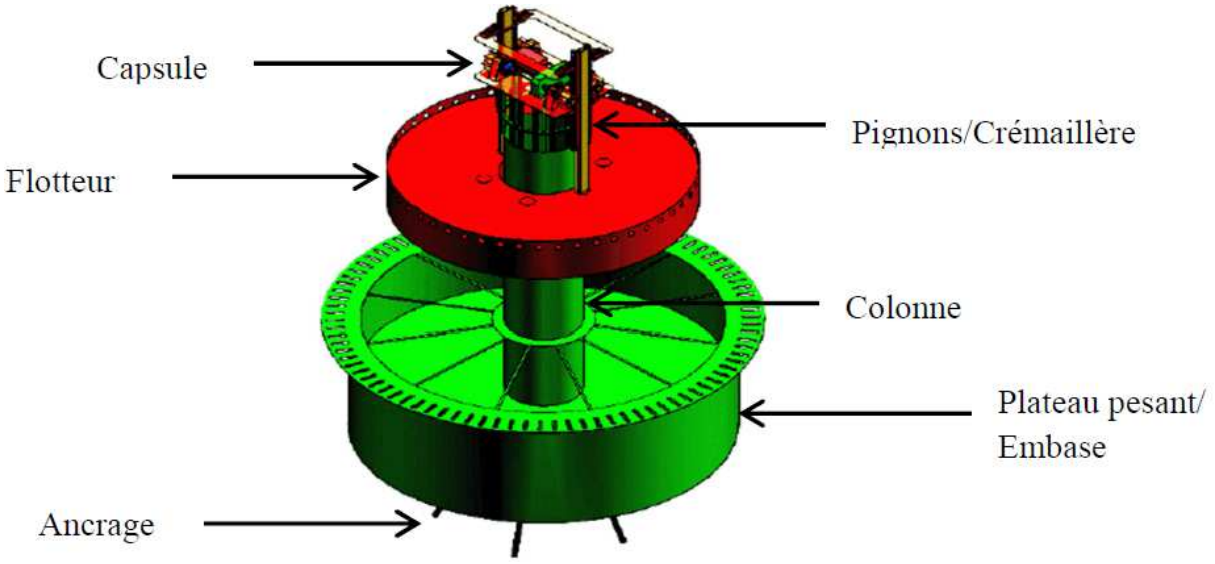


Figure 1: système houlogénérateur le BILBOQUET

2.2. Présentation de la Chaine de conversion de puissance du système houlogénérateur

Le BILBOQUET interagit avec les mouvements d'oscillation, à caractère aléatoire, de la houle en les transformant en un mouvement rotatif qui entraine les génératrices synchrones à aimants permanents. Celles-ci sont chargées de transformer la puissance extraite de la houle en puissance active et/ou réactive qui est transmise au réseau via un convertisseur de puissance de type back-to-back.

Sur la Figure 2, illustrant la chaine de conversion de puissance, nous distinguons deux systèmes essentiels : système mécanique et système électrique.

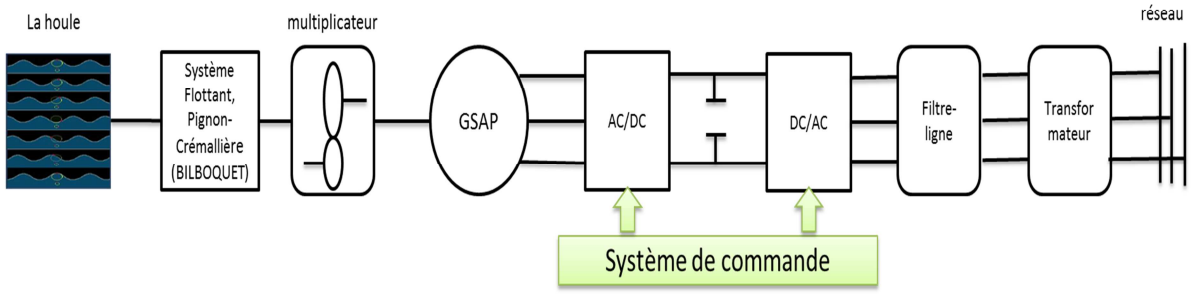


Figure 2 : Représentation de la chaine de conversion de puissance du système houlogénérateur

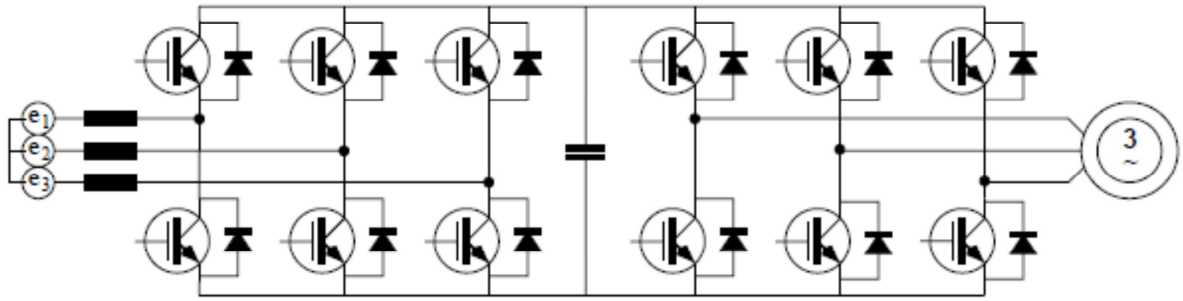


Figure 3: Structure d'un convertisseur de puissance back-to-back

- Le système mécanique : comporte le BILBOQUET excité par la houle et un multiplicateur solidaire à l'arbre de la machine synchrone. Des travaux de modélisation hydrodynamique du système {houle, flotteur, embase} font l'objet de la collaboration avec le laboratoire de recherche (LBMS) de l'école nationale des ingénieurs de Brest. Aussi le modèle développé par LBMS doit fournir la consigne du couple qui permet d'extraire le maximum de puissance. Cette consigne est une

caractéristique couple électromagnétique-vitesse qui dépend des états de mer et de la stratégie de contrôle de la génératrice que nous adopterons.

- Le système électrique : comporte la machine synchrone reliée au réseau par un convertisseur back-to-back, un filtre et transformateur. La Figure 3 illustre le schéma du convertisseur back-to-back, celui-ci est composé d'un onduleur et redresseur pilotés par une modulation de largeur d'impulsions reliés au moyen du bus continu dont la tension doit prendre en compte des deux modes de fonctionnements du redresseur (convertisseur coté machine):

  - ✓ Un fonctionnement en mode redresseur tout diode quand la tension du bus est en dessous d'une valeur limite donnée par $u_{dc_0} = \sqrt{3} \cdot \sqrt{2} V_{1eff}$ avec une MLI vectorielle où V_{1eff} est la valeur efficace du fondamentale de la tension simple de la machine. Ce mode correspond au fonctionnement à vitesse basse de la génératrice.

  - ✓ La possibilité de fonctionner en redresseur à MLI quand la tension du bus continu est en dessus de la valeur seuil $u_{dc_{min}} = \sqrt{2} V_{1eff}$.

Le dimensionnement de la capacité du bus continu d'un convertisseur back-to-back doit permettre de minimiser le taux d'ondulation de la tension du bus continu. En effet, dans [4], l'auteur propose d'assimiler le redresseur à un hacheur quatre quadrants pour simplifier l'analyse du comportement dynamique de la puissance active renvoyée sur le réseau moyennant les hypothèses suivantes :

- ✓ Un système triphasé équilibré.
- ✓ Puissance réactive nulle.
- ✓ Courants de la charge en phase avec la tension composée.

Cette analogie conduit à déterminer la valeur minimale de la capacité du bus continu et elle est donnée par :

$$\boldsymbol{C_{min} = \frac{T_s \cdot P_{charge}}{\Delta u_{dc} \cdot u_{dc}} \left[1 - \frac{e_{dc}}{u_{dc}}\right]} \quad (1)$$

$$\boldsymbol{e_{dc} = \sqrt{2} U_{res}} \quad (2)$$

Où :

U_{res}: La tension composée coté réseau

Δu_{dc}: L'ondulation admise pour la tension du bus continu, T_s et P_{charge} sont la période de découpage et la puissance nominale de machine.

Le système de commande s'appuie sur le redresseur et l'onduleur du convertisseur back-to-back pour réguler le transfert de puissance vers le réseau. L'intérêt du convertisseur back to back est de disposer de plusieurs degrés de liberté vis-à-vis de la commande dans le sens où le redresseur permet de contrôler la génératrice synchrone et l'onduleur permet de contrôler, à priori, la tension du bus continu et la puissance active et réactive transférées au réseau. Aussi nos travaux de thèse, lors de la première année, ont essentiellement porté sur la modélisation des organes du système électrique en vue de conception de loi de commande. Notons que le choix du modèle doit prendre en compte les différentes échelles de temps exhibées par le système houlogénérateur d'où notre choix du modèle moyen pour représenter le comportement dynamique de l'onduleur et redresseur du convertisseur back-to-back loin de tout phénomène de commutation.

2.3. Présentation de la Structure de contrôle

La chaine de conversion de l'énergie à partir de la houle est un système destiné à transférer la puissance du houlogénérateur au réseau. Les convertisseurs statiques utilisés ont pour rôle de réguler les transferts de puissance vers le réseau. Du point de vue commande, on distingue trois niveaux de représentation hiérarchisés (Figure 4)

- Commande rapprochée : elle est conçue pour le pilotage des convertisseurs de puissance, redresseur et onduleur, via la modulation à largeur d'impulsion (SVPWM) pour définir les instants de commutations des IGBT à travers leurs signaux de référence. Elle doit aussi intégrer le fonctionnement en cas de défaut de la génératrice, des convertisseurs ou du réseau.

- Commande haut niveau : elle est conçue pour le contrôle de l'état magnétique et du couple de la machine, de la tension aux bornes du condensateur et de l'énergie réactive et active injectée sur le réseau
- Supervision de puissance : A partir des spécifications d'un cahier de charge, les modes de fonctionnement du système sont définis et les stratégies de replis en cas de défaut sont définies.

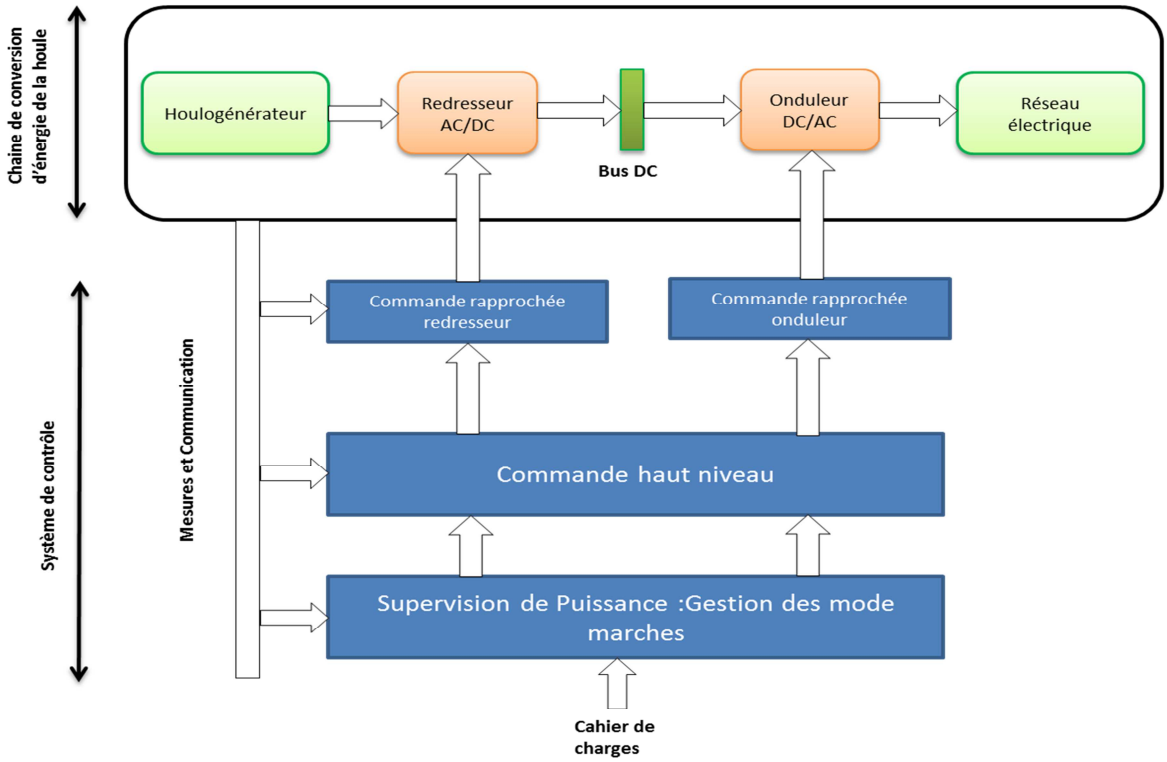


Figure 4: Structure de contrôle hiérarchisé

Dans ce qui a précédé, nous avons mis en contexte notre sujet de thèse dans le cadre du projet Bilboquet puis nous avons présenté la chaine de conversion de puissance du système houlogénérateur en y présentant un bref aperçu du dimensionnement de la capacité du bus continu.

Compte tenu de la structure de contrôle hiérarchisée, les travaux à mener durant cette thèse s'articulent sur le développement des trois étages de la commande (Figure 4) en développant les algorithmes correspondants pour atteindre l'objectif de piloter le houlogénérateur en vue d'extraire le maximum de puissance.

La commande rapprochée nécessite :

- la programmation des modulations de largeur d'impulsions,
- la compensation des temps morts nécessaires à la protection des interrupteurs d'une même branche sur les convertisseurs,
- le découplage des transferts tensions/courants sur la Machine synchrone et le réseau,
- la régulation des courants id et iq dans des repères quadratiques synchrones lié à la position du rotor de la machine synchrone pour le redresseur et lié aux tensions du réseau pour l'onduleur,
- et l'estimation de l'amplitude et des phases des tensions réseau.

Dans un premier temps, nous nous sommes intéressés aux aspects de compensation des temps morts et des non linéarités des composants à semi-conducteur des convertisseurs de puissance, cela nous permet de perfectionner autant que possible les performances des lois de commandes qui seront développée dans le cadre de la commande haut niveau.

3. Etat de l'art des méthodes de compensation

Les onduleurs triphasés pilotés par une Modulation de largeur d'impulsion (MLI) sont à la base de très nombreuses applications industrielles. Les temps morts, introduits pour ne pas court-circuiter la source continue, les temps de fermeture/ouverture des éléments de puissance (IGBT/Diodes)et leurs chutes de tensions sont des non linéarités connues en littérature sous l'appellation **les effets du temps morts**[5].

Les effets des temps morts sont à l'origine de nombreuses anomalies allant de la distorsion du signal du sortie des onduleurs triphasés jusqu'à la dégradation des performances dynamiques des systèmes qui leurs sont associés. En effet, dans [6] et [7], le contenu harmoniques du courant de la machine asynchrone s'enrichit de plus en plus suite à la distorsion de la tension fournie à la charge par l'onduleur. Dans [8] on montre que les performances de l'estimation en ligne des grandeurs physiques, couple, flux et vitesse de la machine asynchrone sont dégradés .Notons également que l'effet des chutes de tensions des composants à semi-conducteur est d'autant plus sévère en basse fréquence (fréquence de fonctionnement de la charge) car les tensions moyennes de commande délivrées par l'onduleur triphasé à la machine deviennent comparables à ces chutes.

Il parait donc indispensable de remédier à ces non linéarités en essayant de compenser les effets des temps morts. L'analyse des effets du temps mort a été largement étudiée en littérature et des méthodes de compensation ont été proposées. Aussi on distingue deux catégories de méthodes de compensation : la première se base sur l'évaluation de la distorsion moyenne, suivant le signe du courant, sur une période de commutation et la deuxième se base sur la compensation impulsion par impulsion au niveau de la commande rapprochée.

La première est la plus répandue [9],[5],[7], [10]… Malgré les résultats satisfaisants obtenus, ses méthodes manquent de précision à cause notamment des difficultés à déterminer le signe du courant au voisinage de zéro. De plus, la détermination de la polarité du courant s'avère d'autant plus délicate avec la présence du phénomène de ''clamping'' qui se définit, lors de son changement de signe, par l'annulation du courant et son maintien à zéro pendant une partie du temps mort [10],[9].

Une autre cause du manque de précision de la première catégorie est la variation des caractéristiques statiques des composants à semi-conducteur en fonction des conditions opératoires notamment la température. C'est dans ce sens que des approches récentes [6]

proposent l'utilisation d'un observateur pour estimer les distorsions du signal de sortie de l'onduleur. Cette dernière solution a l'avantage de se passer de la détermination du signe du courant mais les algorithmes développés deviennent de plus en plus complexes [11], et de fait gourmands en ressources (capacités de calcul, mémoire…) que nous pouvons pas garantir compte tenu des spécifications imposées par les partenaires du projet BILBOQUET fournisseurs des convertisseurs et cartes de contrôles.

Les méthodes de la deuxième catégorie, notamment dans [12], proposent l'élaboration d'un circuit qui détermine le signe du courant et qui modifie les impulsions de la commande rapprochée, mais cette solution reste limitée à cause l'encombrement du à l'ajout de matériel (**hardware**). Une autre approche [13], [14] proposent une solution qui consiste à modifier les impulsions de la MLI (PWM en anglais) en modifiant les commandes des éléments de puissances (IGBT). Cette dernière méthode semble la plus attractive de toutes les méthodes citées ci-dessus compte tenu de la simplicité de sa mise en œuvre et de sa précision.

Dans ce qui suit nous allons présenter l'analyse des effets du temps morts et de méthodes de compensation faisant partie des deux catégories ainsi que les résultats de simulation obtenus par la simulation utilisant Matlab/Simulink.

4. Analyse des effets des temps morts

L'onduleur triphasé de tension étudié (Figure 5) se compose de trois bras indépendants chacun portant deux interrupteurs. Chaque interrupteur est composé d'un transistor IGBT et d'une diode de roue libre montée en antiparallèle. Les transistors qui constituent le même bras sont commandés à la fermeture et à l'ouverture de façon complémentaire pendant une période de commutation T_s .

C'est à travers la modulation de largeur d'impulsion que les temps de conduction et de blocage des IGBT sont définis. Le fonctionnement normal d'un bras d'onduleur requiert un délai appelé temps mort t_d entre les durées de conduction des deux IGBT afin de ne pas court-circuiter le bus continu car les ouvertures et fermetures des composants ne sont pas instantanées.

Les temps de mise en fermeture t_{on} et ouverture t_{off} des transistors caractérisent le comportement dynamique de L'IGBT pendant la commutation et contribuent aux distorsions observées au niveau du signal de sortie de l'onduleur de tension. Les effets du temps mort t_d, des temps de fermeture et d'ouverture augmentent avec la fréquence de commutation et provoquent de différentes anomalies suivant l'application étudiée comme nous l'avons cité plus haut dans l'état de l'art.

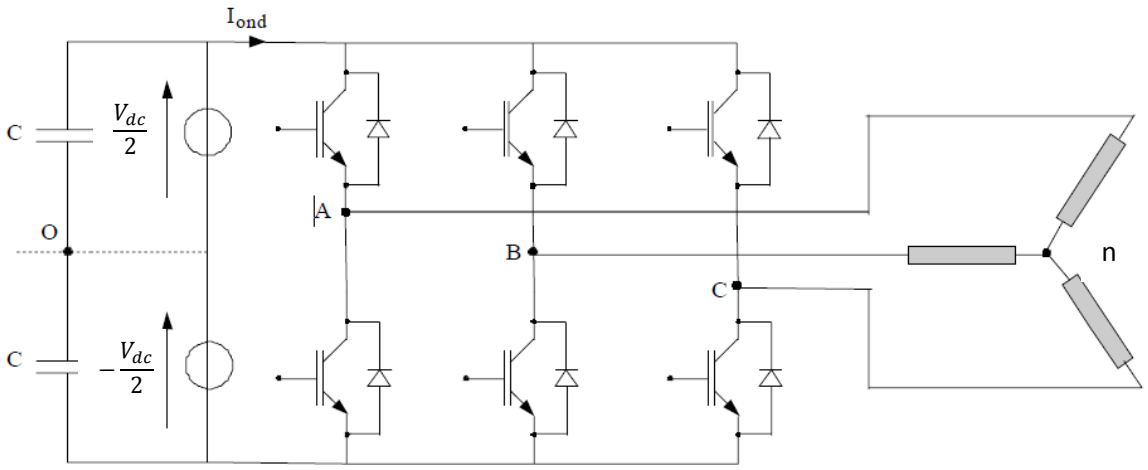


Figure 5: Onduleur triphasé de tension alimentant une charge en étoile

En vue d'analyser les effets des temps morts, les signaux de commandes des éléments de puissance (IGBT) du même bras dans le cas idéal (sans temps mort) et en présence des temps t_d , t_{on} et t_{off} sont illustrés sur la Figure 7

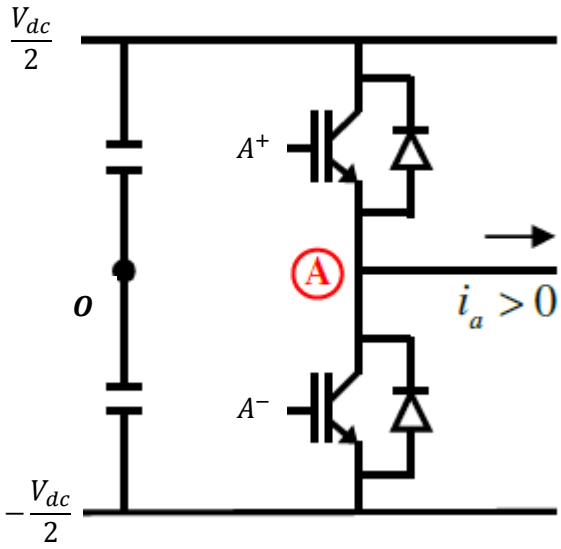


Figure 6 : Un bras d'onduleur triphasé

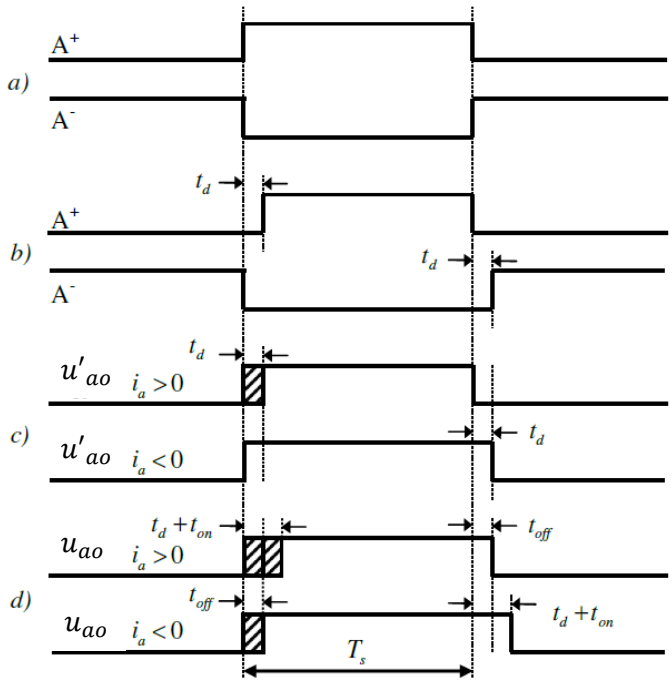


Figure 7 : Impact du temps mort $\boldsymbol{t_d}$, temps de fermeture$\boldsymbol{t_{on}}$ et temps d'ouverture$\boldsymbol{t_{off}}$ sur les signaux de commandes des IGBT [8]

La présence des temps morts induit des modifications sur les durées d'application de la tension de sortie de l'onduleur durant une période de commutation. En effet, ces durées d'application sont déterminées en fonction du signe du courant de la charge et quand le courant est positif (vers la charge) ces durées sont réduites alors qu'elles sont augmentées pour le courant négatif (de la charge).

Pour mettre en relief l'impact des temps mort, la

Figure 8 montre la forme de la tension entre la phase du bras a et neutreu_{an}. Ce tracé est obtenu en simulation aux bornes d'une charge inductive commandée par un onduleur triphasé de tension commandé par une MLI vectorielle pour un temps mort de 5.5 µs et une fréquence de découpage de 3 kHz. L'impact sur la forme de l'onde de la tension devient de plus en plus sévère lorsque la fréquence de découpage de la MLI augmente (

Figure 9) d'où la nécessité de compenser les effets des temps morts. Le développement de certaines méthodes de compensation est abordé dans la partie suivante.

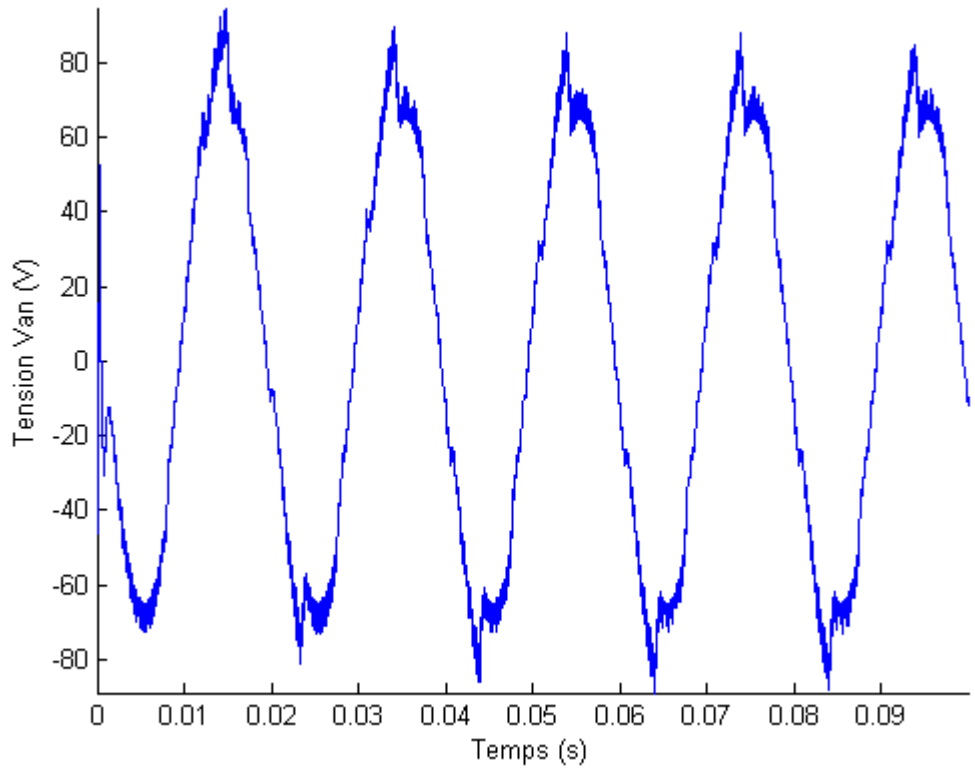


Figure 8: Tension u_{an} pour un temps mort de 5.5µs et une fréquence de découpage 3.5kHz

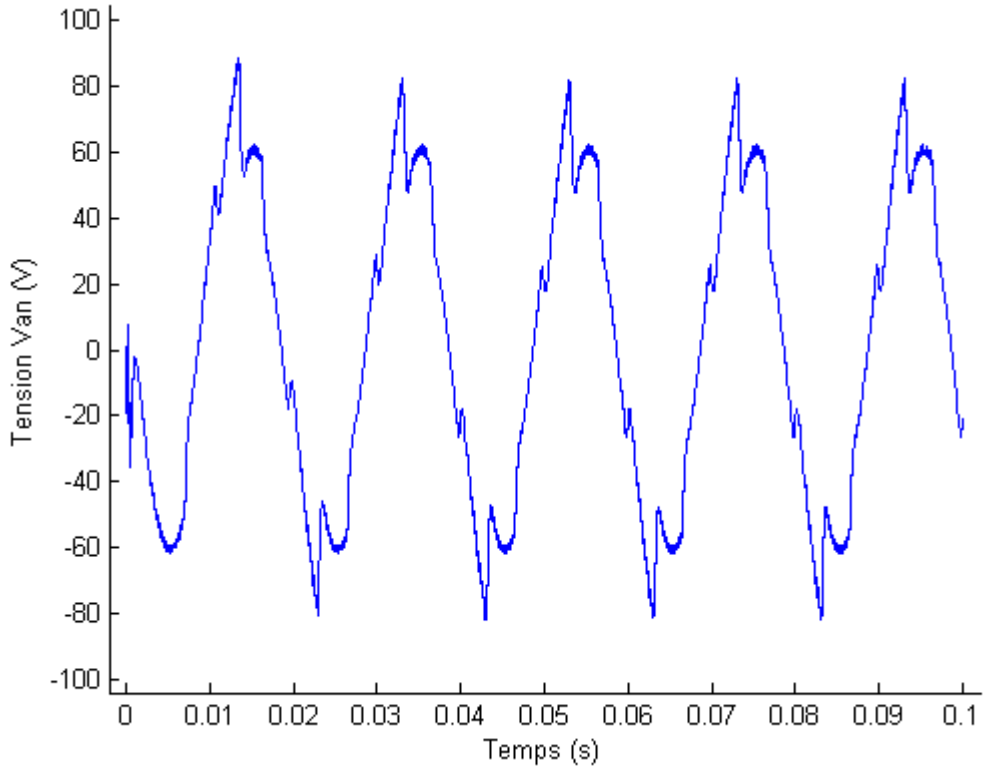


Figure 9 : Tension u_{an} pour un temps mort de 5.5µs et une fréquence de découpage de 5.5 kHz

5. Méthodes de compensation

Comme nous avons cité précédemment, toutes les méthodes de compensation peuvent être classées en deux catégories. Nous présentons d'emblée deux méthodes faisant partie de la catégorie 1.

5.1. Méthodes de compensation de la catégorie 1

5.1.1. Introduction

Les méthodes de compensation appartenant à la catégorie 1 consistent à évaluer en fonction du *courant seulement* la distorsion moyenne du signal de sortie de l'onduleur due à la présence des temps morts et aux chutes de tensions aux bornes des composants à semi-conducteurs (IGBT/Diode) pendant une période de commutation T_s.A cet effet , La fréquence de variation du signe du courant imposé par la charge doit être négligeable devant la fréquence de découpage f_s.

Du point de vue de *l'automatique*, les distorsions de la tension sortie de l'onduleur sont assimilées à des perturbations qu'il est possible de rejeter par anticipation .En effet, la différence entre la tension idéale (sans temps morts) et la tension réellement obtenue est exprimé dans le référentiel diphasé fixe (α, β) puis elle est ajoutée à la tension de référence (Figure 10). Dans des approches plus récentes, ces perturbations sont considérées comme des signaux exogènes estimées en ligne via un observateur mode glissants [6].

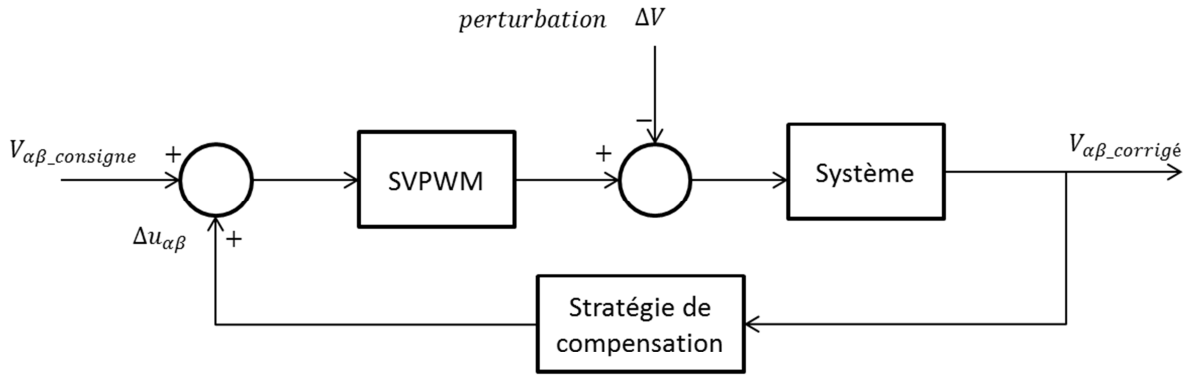


Figure 10: Stratégie de compensation pour les méthodes de la catégorie 1

5.1.2. Méthode de compensation 1

Cette première méthode consiste à quantifier la distorsion due à la présence du temps mort t_d, les temps de fermeture et d'ouverture des éléments de puissance t_{on} et t_{off} ainsi que les chutes de tension aux bornes de l'IGBT et de la diode notées respectivement V_{ce}, V_d.

Dans le référentiel triphasé, la valeur moyenne de la distorsion de tension constatée sur la phase A (tension entre phase et neutre), pendant une période de découpage T_s, est donnée par [5], [7]:

$$\Delta \boldsymbol{u_{an}} = \frac{(t_d+t_{on}-t_{off})}{T_s}.\frac{(V_{dc}-V_{ce}+V_d)}{3}\boldsymbol{Sgn(A)} + \frac{(V_{ce}+V_d)}{6}\boldsymbol{Sgn(A)} \quad (3)$$

$$\boldsymbol{Sgn(A) = 2sgn(i_a) - sgn(i_b) - sgn(i_c)} \quad (4)$$

Où

$Sgn(A)$ est le signe du courant de la phase A qui s'exprime en fonction des signes des courants des trois bras :

$sgn(i_a)$ est la fonction signe du courant parcourant le bras a et elle est défini par :

$$\boldsymbol{sgn(i_a)} = \begin{cases} \boldsymbol{-1,\ i_a < 0} \\ \boldsymbol{1,\ i_a \geq 0} \end{cases} \quad (5)$$

Remarquant dans l'équation (3) que la tension à compenser Δu_{an} est la somme de deux contributions :

$\frac{(t_d+t_{on}-t_{off})}{T_s}.\frac{(V_{dc}-V_{ce}+V_d)}{3}$: Contribution de l'effet des temps mort $t_d + t_{on} - t_{off}$.

$\frac{(V_{ce}+V_d)}{6}$: Contribution de l'effet de chute de tension des composants à semi-conducteur.

Si on définit les signes des courants des deux phases B et C comme suit :

$$\boldsymbol{Sgn(B) = 2sgn(i_b) - sgn(i_a) - sgn(i_c)} \quad (6)$$

$$\boldsymbol{Sgn(C) = 2gn(i_c) - sgn(i_b) - sgn(i_a)} \quad (7)$$

Alors, les expressions des distorsions apparaissant sur les trois phases A, B et C sont données par le système d'équations :

$$\begin{cases} \boldsymbol{\Delta u_{an} = Sgn(A).\frac{(t_d+t_{on}-t_{off})}{T_s}\left(\frac{V_{dc}-V_{ce}+V_d}{3}\right) + Sgn(A).\frac{(V_{ce}+V_d)}{6}} \\ \boldsymbol{\Delta u_{bn} = Sgn(B).\frac{(t_d+t_{on}-t_{off})}{T_s}\left(\frac{V_{dc}-V_{ce}+V_d}{3}\right) + Sgn(B).\frac{(V_{ce}+V_d)}{6}} \\ \boldsymbol{\Delta u_{cn} = Sgn(C).\frac{(t_d+t_{on}-t_{off})}{T_s}\left(\frac{V_{dc}-V_{ce}+V_d}{3}\right) + Sgn(C).\frac{(V_{ce}+V_d)}{6}} \end{cases} \quad (8)$$

L'expression de la distorsion de tension dans le référentiel biphasé fixe $\Delta u_{\alpha\beta}$ est obtenue par la transformation T_{32} suivante de sorte que nous ayons :

$$\boldsymbol{\Delta u_{\alpha,\beta}} = \sqrt{\frac{2}{3}}\begin{bmatrix} \boldsymbol{1} & \boldsymbol{-\frac{1}{2}} & \boldsymbol{-\frac{1}{2}} \\ \boldsymbol{0} & \boldsymbol{\frac{\sqrt{3}}{2}} & \boldsymbol{-\frac{\sqrt{3}}{2}} \end{bmatrix}\begin{pmatrix} \boldsymbol{\Delta u_{an}} \\ \boldsymbol{\Delta u_{bn}} \\ \boldsymbol{\Delta u_{cn}} \end{pmatrix} \quad (9)$$

Où

$$\boldsymbol{T_{32}} = \sqrt{\frac{2}{3}}\begin{bmatrix} \mathbf{1} & -\frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} \\ \mathbf{0} & \frac{\sqrt{\mathbf{3}}}{\mathbf{2}} & -\frac{\sqrt{\mathbf{3}}}{\mathbf{2}} \end{bmatrix} \quad (10)$$

Pour procéder à la compensation des temps morts et des chutes de tension, il suffit de rajouter $\Delta u_{\alpha\beta}$à la tension de référence $V_{\alpha\beta ref}$ qui sert de modulante à la MLI vectorielle (SVPWM) comme c'est indiqué sur le schéma bloc (Figure 10).

5.1.3. Méthode de compensation 2

Dans cette méthode, l'analyse des effets du au temps mort, aux temps de fermeture et d'ouverture (t_d, t_{on}, t_{off}) est la même développée pour la méthode 1.Mais pour analyser les effets des chutes de tensions des composant à semi-conducteur (IGBT/Diode), on introduit un temps équivalent t_{comp} qui traduit qualitativement leurs effets [10] de sorte que : la différence entre les durées de conduction des IGBT dans le cas idéal (sans temps morts) et dans le cas réel (en présence du temps mort) vaut T_{err} dont l'expression est donnée par :

$$\boldsymbol{T_{err} = t_d + t_{on} - t_{off} + t_{comp}} \quad (11)$$

$$\boldsymbol{T_{dead\_a} = sgn(i_a)\left(t_d + t_{on} - t_{off} + t_{comp}\right)} \quad (12)$$

Ou encore :

$$\boldsymbol{T_{dead\_a} = sgn(i_a)T_{err}} \quad (13)$$

$sgn(i_a)$ est la fonction signe du courant de la phase A, t_{comp} est définit comme suit [10]:

$$\boldsymbol{t_{comp}} = \begin{cases} \frac{t_{on}.V_{ce}+t_{off}.V_d}{V_{dc}}, & \boldsymbol{i_a \geq 0} \\ \frac{t_{off}.V_{ce}+t_{on}.V_d}{V_{dc}}, & \boldsymbol{i_a < 0} \end{cases} \quad (14)$$

La distorsion résultante, sur la phase A de l'onduleur, sur une période de découpage T_s est obtenue par l'équation suivante :

$$\boldsymbol{\Delta u_{an} = \frac{T_{dead\_a}}{T_s} \cdot V_{dc}} \quad (15)$$

Par une analyse similaire, nous obtenons les distorsions qui apparaissent sur les phases B et C,

$$\boldsymbol{\Delta u_{bn} = \frac{T_{dead\_b}}{T_s} \cdot V_{dc}} \quad (16)$$

$$\boldsymbol{\Delta u_{cn} = \frac{T_{dead\_c}}{T_s} \cdot V_{dc}} \quad (17)$$

Où

$$\boldsymbol{T_{dead\_b} = sgn(i_b)T_{err}} \quad (18)$$

$$\boldsymbol{T_{dead\_c} = sgn(i_c)T_{err}} \quad (19)$$

Les distorsions sont ensuite exprimées dans le référentiel diphasé fixe (α, β) grâce à la transformation de matriceT_{32}. On peut alors calculer les tensions de compensation Δu_α et Δu_β pour toutes les configurations des signes des trois courants (i_a,i_b,i_c). Les différentes tensions de compensations sont résumées dans Tableau 1: Tensions de compensation dans le repère $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

Secteur	$i_a i_b i_c$	Δu_α	Δu_β
1	**+--**	$2\sqrt{\frac{2}{3}}u_{err}$	0
2	**++-**	$\sqrt{\frac{2}{3}}u_{err}$	$\sqrt{2}u_{err}$
3	**-+-**	$-\sqrt{\frac{2}{3}}u_{err}$	$\sqrt{2}u_{err}$
4	**-++**	$-2\sqrt{\frac{2}{3}}u_{err}$	0
5	**--+**	$-\sqrt{\frac{2}{3}}u_{err}$	$-\sqrt{2}u_{err}$
6	**+-+**	$\sqrt{\frac{2}{3}}u_{err}$	$-\sqrt{2}u_{err}$

Tableau 1: Tensions de compensation dans le repère $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

Les deux méthodes de compensation des temps morts présentées procèdent par quantification, en fonction du courant de la charge, de la distorsion moyenne du signal de sortie de l'onduleur pendant une période de commutation T_s. Ainsi la tension de compensation est rajoutée à la tension de référence pour redéfinir les nouvelles durées de conduction des IGBT qui donnent à la sortie la tension désirée.

Dans ce qui suit nous allons présenter une méthode da la deuxième catégorie accompagnée de résultats de simulation permettant de comparer les méthodes des deux catégories.

6. Une méthode de compensation de la catégorie 2 .

Comme nous avons déjà évoqué dans l'état de l'art sur les méthodes de la compensation, la présente méthode se base sur la compensation impulsion par impulsion au niveau de la commande rapprochée des interrupteurs [13], [14].

Aussi notons que cette méthode est facile à implémenter, il suffit d'apporter des modifications au niveau soft (programme) dans l'algorithme de commande rapprochée qui définit les temps de conduction des IGBT de façon à obtenir la tension de sortie adéquate en prenant en compte les temps morts.

La synthèse de cette méthode repose sur le principe suivant : pendant le temps mort t_{mort} la tension de sortie de l'onduleur ne dépend que du courant de sorte que, pour la phase A :

$$\boldsymbol{u_{an}} = \begin{cases} \boldsymbol{0},\ \boldsymbol{si}\ \ \boldsymbol{i_a} \geq \boldsymbol{0} \\ \boldsymbol{V_{dc}},\ \boldsymbol{si}\ \boldsymbol{i_a} < \boldsymbol{0} \end{cases} \qquad (20)$$

Pour mettre au point cette méthode, considérons deux signaux de commande complémentaires correspondant à deux interrupteurs du même bras (exemple bras A) .

La méthode consiste à créer un temps mort qui ne modifie pas les tensions de sortie qui auraient été obtenue en l'absence des temps morts. On crée donc la commande suivante :

Si le courant est positif (Figure 11):

$$T_{1hon*} = T_{1hon} + t_{mort}$$

$$T_{1loff*} = T_{1hoff} - t_{mort}$$

Si le courant est négatif (Figure12Figure12 : Instants de commutation après compensation $\boldsymbol{i_a} < \boldsymbol{0}$

$$T_{1loff*} = T_{1loff} - t_{mort}$$

$$T_{1lon*} = T_{1lon} + t_{mort}$$

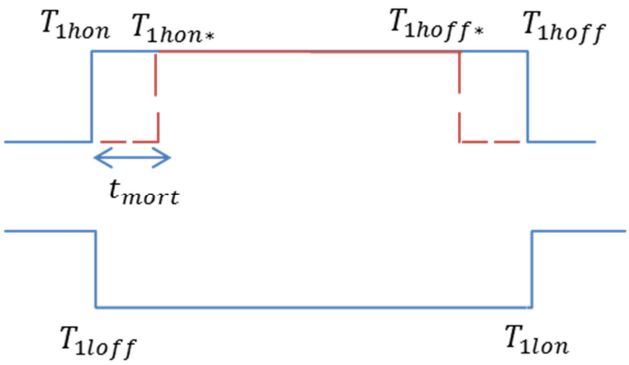


Figure 11 : Instants de commutation après compensation $\boldsymbol{i_a > 0}$

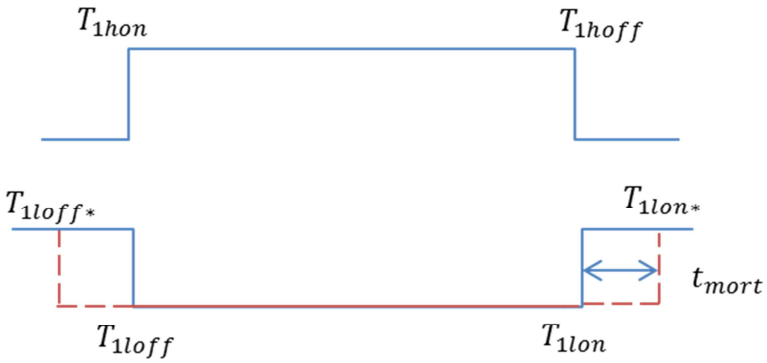


Figure12 : Instants de commutation après compensation $\boldsymbol{i_a < 0}$

Malgré la simplicité de cette méthode, sa mise au point nécessite la prise en compte des éventuelles anomalies qu'elles pourraient introduire .En effet, les nouveaux signaux de commandes générées élargissent ou rétrécissent, parfois d'une manière excessive, les durées de conduction des transistors. Aussi deux butées de saturation pour les rapports cycliques de conduction sont introduites en prenant en compte la présence inévitable du temps morts t_d et une seule commutation par période de découpage T_s [15].

7. Résultats de simulation des méthodes de compensation

Afin de mettre au point au niveau de la simulation les deux méthodes discutées ci-dessus, nous avons considéré un circuit électrique où une charge inductive (R-L) est commandée par un onduleur triphasé de tension à deux niveaux alimenté par une source de tension continue.

Les tensions sont exprimées dans le référentiel biphasé tournant(d, q), tournant par rapport au référentiel(α, β). La stratégie de compensation a été implémentée en langage **C** et intégré dans une S-Function sur Matlab/Simulink suivant le schéma présenté dans Figure 13.

L'onduleur est commandé par une MLI et le modèle de cet onduleur est issu de la bibliothèque *Simpower-Systems*. C'est un modèle à commutation qui tient compte des chutes de tension et des non-linéarités des composants à semi-conducteurs (IGBT/Diode).

Les tensions (V_d,V_q) aux bornes de la charges sont filtrées puis comparées avec celles obtenues sans algorithmes de compensation.

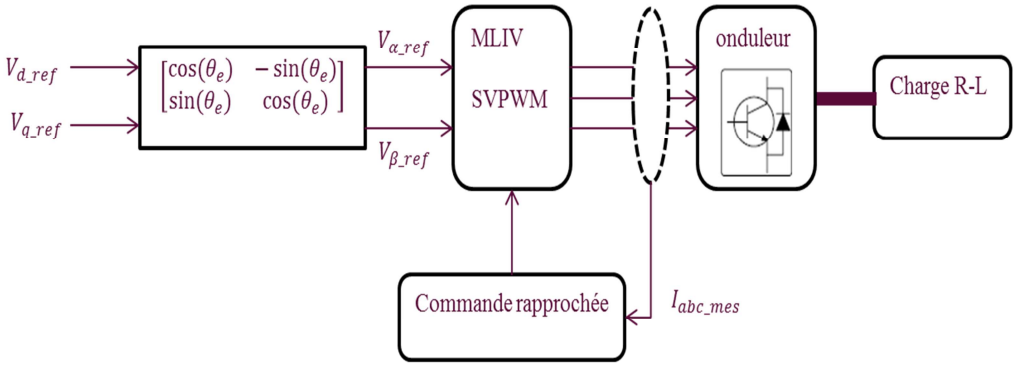


Figure 13: Schéma de simulation des méthodes de compensation

Paramètres de simulation :

- ✓ t_d=10 µs temps mort
- ✓ $t_{on} = t_{off} = 1\ \mu s$
- ✓ $f_s = 3\ kHz$ fréquence de découpage de la MLI vectorielle
- ✓ $T_s = {}^{1}/_{f_s}$ période de la MLI vectorielle.
- ✓ $T_e = 1\ \mu s$ pas de simulation
- ✓ $V_{d\_consigne} = 0 \quad (V)$ tension consigne sur l'axe d
- ✓ $V_{q\_consigne} = 100 \quad (V)$ tension consigne sur l'axe q

La compensation des temps morts et des chutes de tension aux bornes des (IGBT/Diode) doit fournir à la charge les valeurs de la tension de référence sans perte. Aussi les tensions (V_d, V_q) visualisées aux bornes de la charge doivent être en moyenne proches de leurs valeurs consignes V_{d_ref} et V_{q_ref}. Notons que l'intérêt derrière le développement des méthodes de compensation des temps morts c'est de réduire autant que possible les sources d'erreurs qui pourraient apparaitre au niveau de la commande haut niveau et par conséquent permettre d'envisager un haut niveau de performances par des lois de commandes perfectionnées. Ainsi la pertinence des méthodes de compensation est jugée qualitativement par rapport à l'amélioration apportée à l'amplitude et à la forme de la tension corrigée. Les résultats de simulation obtenues, d'une part par les méthodes de la catégorie 1[4], et d'autre parts par la

[4] On notera que les deux méthodes de la catégorie 1 ont été implémentées et testées mais comme les résultats ne présentent pas de différences visibles une seule est conservée pour les tracés.

méthode de la catégorie 2, sont satisfaisants compte tenue des améliorations remarquées sur les tensions corrigées montrées sur les Figure 14 : Tensions ($\boldsymbol{Vd}$,$\boldsymbol{Vq}$) de la charge obtenues par les deux méthodes de la catégorie 1et Figure 15 : Tensions ($\boldsymbol{V_d}$,$\boldsymbol{V_q}$) de la charge obtenues par la méthode de la catégorie 2.

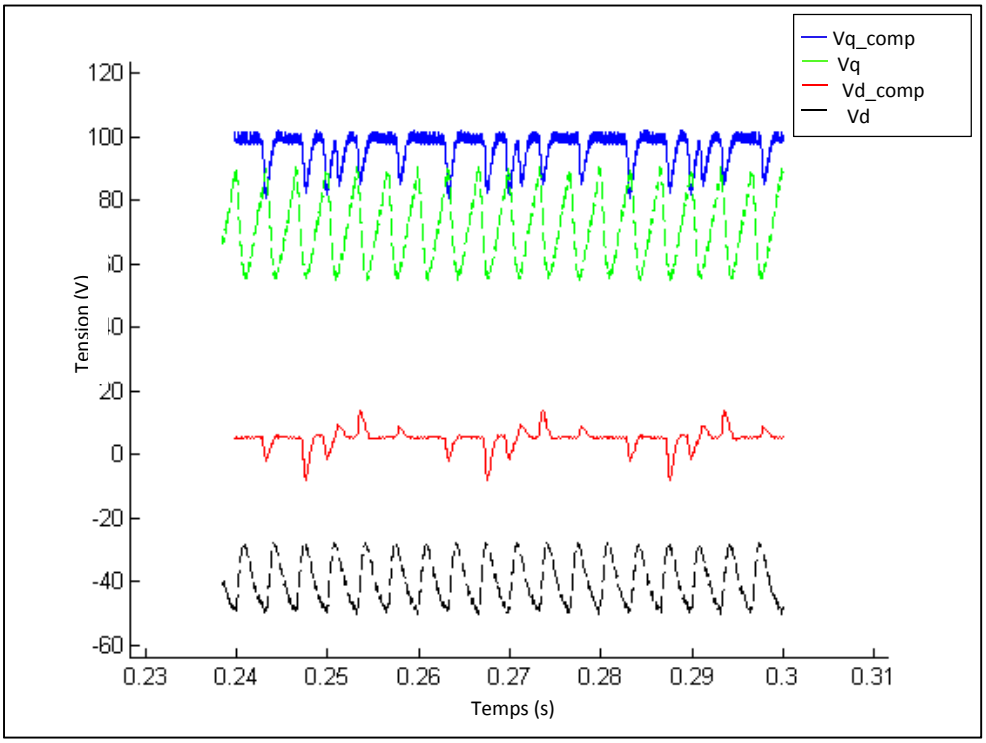


Figure 14 : Tensions ($\boldsymbol{V_d}$,$\boldsymbol{V_q}$) de la charge obtenues par les deux méthodes de la catégorie 1

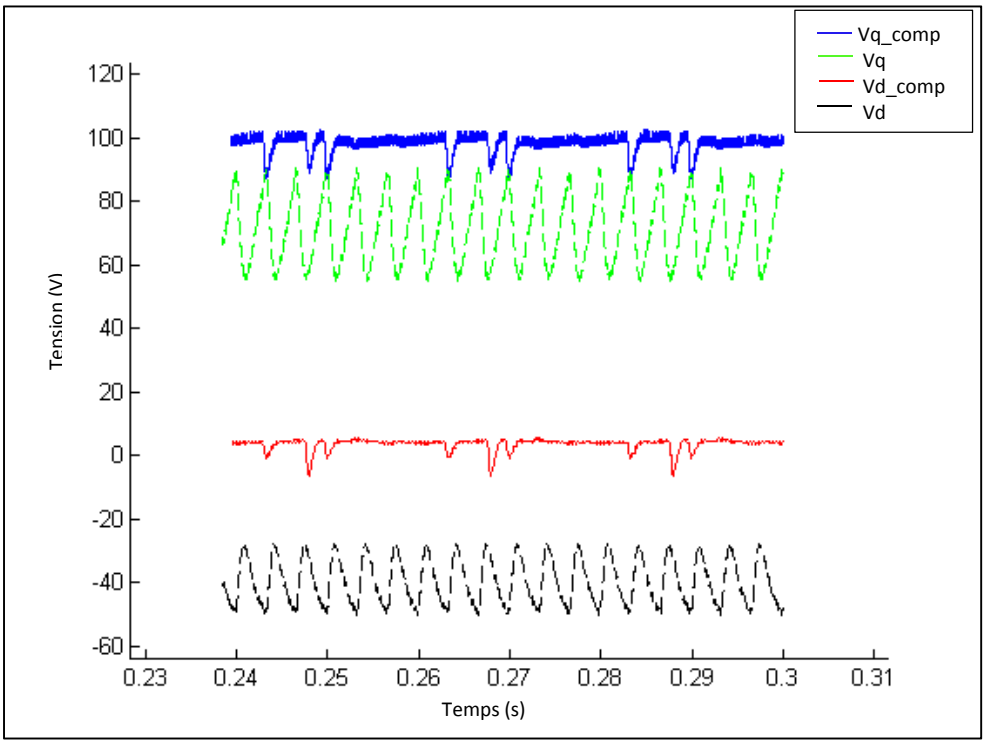


Figure 15 : Tensions (V_d, V_q) de la charge obtenues par la méthode de la catégorie 2

Après avoir présenté un état de l'art des études faites sur l'analyse des effets des temps morts et les méthodes de compensations, nous avons mis en relief l'impact des non-linéarités des convertisseurs en discutant sur le principe des méthodes de compensation et l'amélioration qu'elles apportent sur les tensions corrigées de l'onduleur triphasé de tension. Cette étape était d'une importance primordiale dans le sens où nous ne serons pas amenés à chercher à contrer les effets de temps morts ou les compenser via les lois de commande haut niveau. Par ailleurs, de parts les différentes échelles de temps qui caractérisent la chaine de conversion du système houlogénérateur, les algorithmes de compensation sont couteux en termes de temps de simulation car ils se basent sur des modèles plus fins d'où notre recours au modèle moyen qui assure un compromis entre finesse de représentation et coût de simulation. Dans ce qui suit nous présentons le modèle moyen d'un onduleur triphasé de tension qui tient compte des pertes inhérentes aux composants à semi-conducteur (Diode/IGBT).

8. Modèle moyen d'un onduleur triphasé

8.1. Introduction

Le modèle global de la chaine de conversion du système houlogénérateur nécessite un modèle pour chacune de ses parties. Ces modèles sont souvent de natures différentes suivant l'objectif de l'analyse visée. Aussi l'analyse de la fonction de transfert d'énergie d'un convertisseur statique doit être différente de l'analyse du comportement thermique de ses composants à semi-conducteur pendant la commutation. Certes, tous ces modèles doivent présenter un excellent compromis entre la finesse de la représentation et le coût de simulation. C'est dans ce sens que le modèle moyen a été introduit pour des objectifs de conception de lois de commande du système houlogénérateur [16] [17][18].

Ainsi le modèle moyen d'un onduleur triphasé consiste *à représenter le transfert de l'énergie indépendamment de toute notion de commutation*. De plus cette représentation peut être étendue pour prendre en compte les non linéarités au sein du convertisseur, c'est déjà notre cas de figure dans ce qui va suivre.

8.2. Modèle moyen d'un bras d'onduleur

L'onduleur triphasé de tension considéré consiste en une source de tension continue de forte puissance alimentant trois bras d'onduleur, chacun possédant deux cellules de commutation. Chaque bras, décrit à la Figure **16**, est considéré indépendant à condition de supposer que la tension d'alimentation est constante durant les phases de commutation et en négligeant l'inductance parasite du bus d'alimentation continue [19].

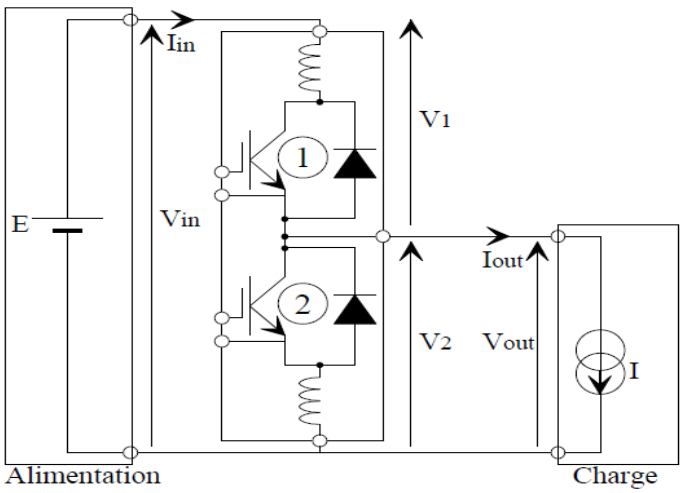


Figure 16:Représentation du bras à l'intérieur de l'onduleur triphasé [19]

La construction du modèle moyen requiert une hypothèse essentielle : la fréquence de commutation des interrupteurs doit être plus grande que les autres fréquences naturelles du système. En pratique cela revient à considérer que le courant dans la charge varie très lentement vis-à-vis de la période de commutation. Aussi le comportement de chaque bras est décrit par un modèle moyen et le modèle global du convertisseur sera composé de trois modèles moyens indépendants. En effet, considérant une séquence d'opérations périodique définie par $S = \{\overline{T_1T_2},\ T_1\overline{T_2},\ \overline{T_1T_2},\ \overline{T_1}T_2\}$ où T_1 et T_2 représentent les valeurs logiques assignées au signal de commande des interrupteurs 1 (interrupteur haut) et 2 (interrupteur bas) du bras d'onduleur montré à la Figure 16: T_i signifie que l'interrupteur est commandé à la fermeture et $\overline{T_i}$ commandé à l'ouverture.

L'élaboration systématique du modèle moyenné d'un bras d'onduleur repose sur les étapes suivantes décrite dans [16] et [19]:

Etape A : ***construction du bloc de commutation***

On définit le bloc de commutation incluant les éléments qui changent de causalité pendant la séquence S qui est représenté par le rectangle au milieu du bras de la Figure 16 contenant les deux cellules de commutation. On cherche donc à modéliser le comportement moyen de ce bloc en tenant compte des non linéarités inhérentes à ses composants à semi-conducteur.

Etape B : ***identification des variables de port externe du bloc de commutation***

Par analyse de causalité, on identifie des variables à port externe du bloc de commutation : V_{in}, I_{in}, V_{out}, I_{out} voir Figure 16.

Etape C : ***identification des variables de port externe du bloc de commutation***

Le bloc de commutation interagit avec la source de tension E qui impose la tension V_{in} et une source de courant qui impose le courant de sortie I_{out}. Aussi les variables d'entrée du bloc de commutation sont $U = \{V_{in},\ I_{out}\}$ et les variables de sortie sont $Y = \{I_{in}, V_{out}\}$

Etape D : ***simplification du circuit à analyser***

La source de tension continue V_{in} peut être remplacée par une source idéale E et la charge est remplacée par une source de courant idéal I sous couvert de l'hypothèse qui considère que la variation de I_{out} et V_{in}, pendant une période de commutation est négligeable devant la fréquence de commutation, sinon ils seraient inclus dans le bloc de commutation.

Etape E : expression des variables de port de sortie du bloc de commutation.

En considérant le circuit simplifié, il est possible d'obtenir une expression pour chaque variable de port de sortie du bloc de commutation y_k qui ne dépend que du vecteur de variable d'entrée U, du vecteur de variables d'état X et de l'état i de la séquence S : $y_k = f_k^i(U, X)$.

On distingue deux modes de fonctionnement de la cellule de commutation de l'onduleur de tension suivant le sens du courant I_{out} durant une séquence S. En effet, si le courant est positif, vers la charge, seuls l'IGBT 1 et la Diode 2 peuvent conduire et quand le courant est négatif l'IGBT 2 et la diode 1 conduisent. Les Tableau 2 : Expression de $\boldsymbol{y_k}$ pour $\boldsymbol{I_{out} > 0}$ et Tableau 3:Expression de $\boldsymbol{y_k}$ pour $\boldsymbol{I_{out} < 0}$ illustrent les expressions des variables de sortie Y en fonction de la séquence S. Notons que $T_s \cdot \rho_1$ représente la durée de commande de la fermeture de l'IGBT 1, $T_s \cdot \rho_2$la durée de la commande de la fermeture de l'IGBT 2 et $T_s \cdot \rho_d$ correspond au temps mort de sorte que : $T_s \cdot \rho_1 + T_s \cdot \rho_2 + 2T_s \cdot \rho_d = T_s$

L'obtention du modèle moyen revient à obtenir les expressions des sorties y_k pondérées de leurs durées pendant chaque période T_s. V_1 et V_2 représentent les caractéristiques statiques des interrupteurs du bras de l'onduleur.

Etape F : ***obtention de la valeur moyenne des variables de port sortie du bloc de commutation***

La valeur moyenne d'une variable de port de sortie est donnée par l'expression :

$$\boldsymbol{Y_k = \frac{1}{T_S}\sum_{i=1}^{N}\int_{t_{i-1}}^{t_i} f_k^i\big(U, X_i(t)\big)dt} \qquad (21)$$

Où

N est le nombre d'états dans la séquence d'opérations de S et T_s la période de commutation.

Séquence$S$	$\overline{T_1T_2}$	$T_1\overline{T_2}$	$\overline{T_1T_2}$	$\overline{T_1}T_2$
Dispositif conducteur	D_2	T_1	D_2	D_2
Durée	$T_s \cdot \rho_d$	$T_s \cdot \rho_1$	$T_s \cdot \rho_d$	$T_s \cdot \rho_2$
V_{out}	V_2	$E - V_1$	V_2	V_2
I_{in}	$0\ A$	I_{out}	$0\ A$	$0\ A$

Tableau 2 : Expression de $\boldsymbol{y_k}$ pour $\boldsymbol{I_{out} > 0}$

Séquence$S$	$\overline{T_1T_2}$	$T_1\overline{T_2}$	$\overline{T_1T_2}$	$\overline{T_1}T_2$
Dispositif conducteur	D_1	D_1	D_1	T_2
Durée	$T_s \cdot \rho_d$	$T_s \cdot \rho_1$	$T_s \cdot \rho_d$	$T_s \cdot \rho_2$
V_{out}	$E - 1$	$E - V_1$	$E - V_1$	V_2
I_{in}	I_{out}	I_{out}	I_{out}	$0\ A$

Tableau 3:Expression de $\boldsymbol{y_k}$ pour $\boldsymbol{I_{out} < 0}$

Compte tenu des étapes développées précédemment, le modèle moyen du bras de l'onduleur triphasé de tension se décline comme suit :

Pour $I_{out} > 0$:

$$\frac{1}{T_s}\int_0^{T_s} \boldsymbol{V_{out}(t)dt} = \boldsymbol{\rho_1[V_{in} - V_{ce}(I_{out})] + (1 - \rho_1)[-V_d(I_{out})]} \quad (22)$$

$$\frac{1}{T_s}\int_0^{T_s} \boldsymbol{I_{in}(t)dt} = \boldsymbol{\rho_1 I_{out}} \quad (23)$$

Pour $I_{out} < 0$:

$$\frac{1}{T_s}\int_0^{T_s} \boldsymbol{V_{out}(t)dt} = \boldsymbol{(1 - \rho_2)[V_{in} + V_d(I_{out})] + \rho_2[V_{dce}(I_{out})]} \quad (24)$$

$$\frac{1}{T_s}\int_0^{T_s} \boldsymbol{I_{in}(t)dt} = \boldsymbol{(1 - \rho_2) I_{out}} \quad (25)$$

On constate que le modèle moyen dépend de la fréquence de découpage $\frac{1}{T_s}$ et des rapports cycliques des IGBT 1 et 2 (ρ_1, ρ_2). Le modèle moyen dépend également des caractéristiques statiques de l'IGBT et de la diode définies par $V_{ce}(I_{out})$, $V_d(I_{out})$ les tensions de chutes aux bornes des composants à l'état passant.

8.3. Validation du modèle moyen

Pour valider le modèle moyen décrivant le comportement de l'onduleur, nous avons comparé les tensions et courants filtrés obtenus d'une part, par le modèle réel de l'onduleur déjà existant sur la bibliothèque ''*Simpower-Systems* ''de Matlab/Simulink, avec ceux obtenus par le modèle moyen développé d'autre part. Le filtrage des deux signaux (tension, courant) est réalisé afin d'éliminer les bruits engendrés par la commutation des interrupteurs à 3kHz.

Force est de constater que les tensions moyennes appliquées à la charge sont bien concordantes avec les tensions fournies par le modèle numérique réel. De plus, la comparaison entre les courants moyens et les courants réels montrent une très légère différence pendant le régime transitoire mais cette différence est due à la présence du filtrage du courant réel. (Figure **17**,Figure**18**).

En guise de perspectives, nous allons utiliser le modèle moyen pour représenter le comportement du convertisseur back-to-back. En effet, le redresseur coté machine et l'onduleur coté réseau seront décrits par le modèle moyen pour obtenir un modèle global qui sera destiné à l'élaboration de lois de commande.

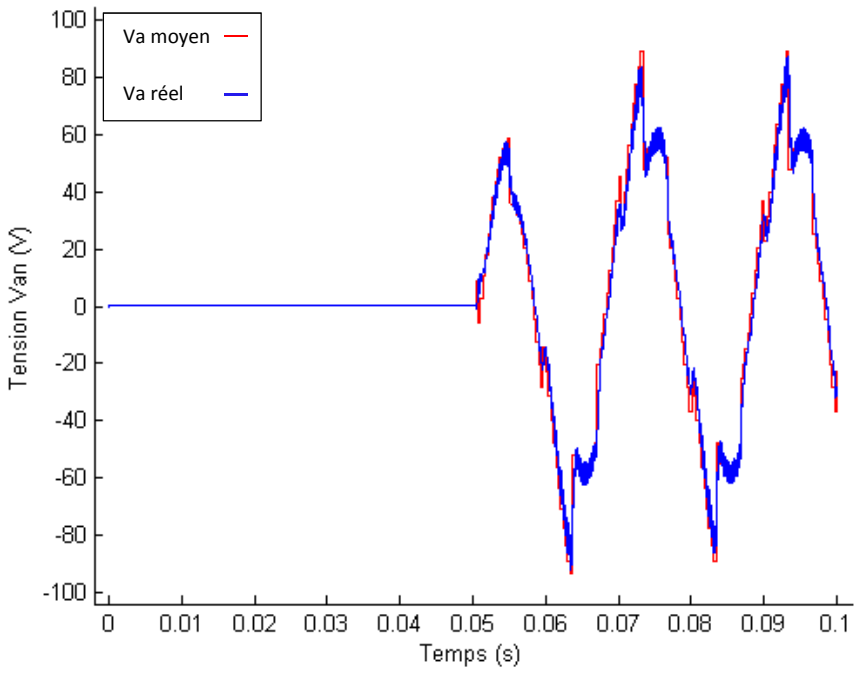


Figure 17:Comparaison tension moyenne et réelle $\boldsymbol{u_{an}}$

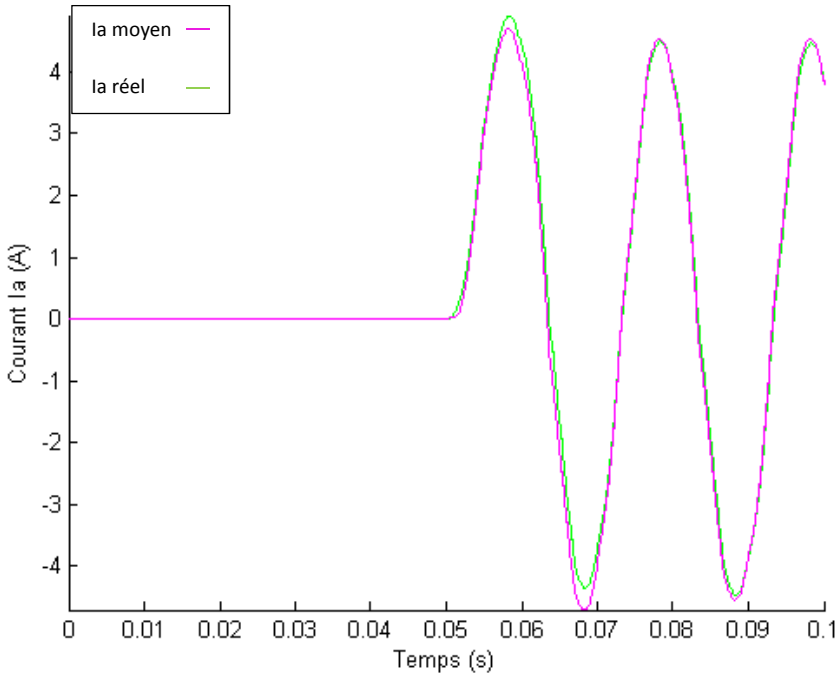


Figure18: Comparaison entre Courant moyen et Courant réel

9. Conclusion et perspectives

9.1.Conclusion

Le présent rapport a abordé non seulement le contexte de notre thèse mais aussi le développement des premières études menées dans le cadre du projet BILBOQUET. En effet, après avoir précisé que notre sujet de thèse s'inscrit dans le cadre de développement d'un nouveau système de récupération d'énergie des vagues de la mer nécessitant différentes compétences dans des domaines variés, nous avons illustré la chaine de conversion de ce système houlogénérateur dont la partie électrique fera l'objet de nos prochaines recherches notamment la modélisation et la commande.

Nous avons adopté une structure de commande hiérarchisée qui nous a permis d'avoir une vision claire sur les actions à mener pendant notre thèse. En effet, on a distingué trois niveaux de représentation du point de vue commande dont le niveau de la commande rapprochée qui regroupe la programmation de la MLI vectorielle et la mise au point des algorithmes de compensation des effets des temps morts et des chutes de tensions des composants à semi-conducteur des convertisseurs de puissance (redresseur et onduleur) dans l'objectif de réduire les sources d'erreur susceptibles d'altérer au performances de lois de la commande haut niveau. Aussi des méthodes de compensation des temps morts et des chutes de tension ont été présentées et validées en simulation.

Le houlogénérateur étant un système à multi-échelles temporel, nous avons opté pour le modèle moyen tenant compte des pertes, afin d'avoir le meilleur compromis entre finesse de représentation et coût de simulation et ce afin de pouvoir aborder convenablement la conception de loi de commande haut niveau ayant pour objectif de piloter le système complet et extraire le maximum de puissance. Aussi nous avons présenté une approche systématique de construction de modèle moyen d'un onduleur triphasé de tension que nous avons validé par comparaison avec le modèle numérique réel existant sur Matlab/Simulink.

9.2.Perspectives

Le transfert d'énergie de la houle au réseau via la chaine de conversion de puissance du système houlogénérateur nécessite une étude multidimensionnelle qui touche plusieurs aspects allant de la modélisation jusqu'à la supervision et le suivi en ligne du système. Aussi nous sommes amenés à définir une vision prospective sur les différentes problématiques que nous devrons aborder dans le cadre de l'étude de la partie électrique du houlogénérateur ainsi que les solutions que l'on pourrait envisager en fonction des contraintes technologiques convenues par les spécifications de projet BILBOQUET.

De point de vue modélisation, le modèle moyen du convertisseur back-to-back sera adopté pour décrire le comportement dynamique de puissance active et réactive injectées au réseau.

Etant contraint par les choix technologiques en terme d'instrumentation, nous ne pourrons pas disposer des mesures directes du courant de la capacité du bus continu du convertisseur back-to-back. Notons que la tension aux bornes de la capacité est bruitée notamment par l'effet du phénomène de la commutation inhérent au fonctionnement de l'onduleur et du redresseur. Par conséquent l'estimation du courant par une dérivation directe (accroissement fini par exemple) entrainerait une amplification du bruit pouvant rendre le résultat inexploitable pour la régulation de la tension du bus continue. Aussi nous proposons d'**estimer en ligne** la tension et le courant du bus continu en y incluant l'étude sur les méthodes de **dérivation** en s'appuyant sur les travaux récents menés à AMPERE par M. DRIDI [20] et L. SIDHOM [21]**.**

Après avoir définit les stratégies de commande en fonction des deux scénarios correspondant à l'état de charge de la capacité (chargée/déchargée) le redresseur et l'onduleur du back-to-back doivent permettre de :

- Contrôler l'état magnétique de la machine synchrone en assurant le découplage des tensions et courants dans le repère tournant (d,q). Le contrôle de ces deux grandeurs se fait par le biais des tensions de modulation V_{dref_red} et V_{qref_red} de la MLI pilotant le redresseur.

- Contrôler la puissance active et réactive renvoyée au réseau en assurant également les découplages des courants et tensions du côté du réseau par le biais des degrés de libertés de la commande de l'onduleur qui sont les tensions de modulation V_{dond_ref} et V_{qond_ref}. Pour atteindre de bonnes performances dans le contrôle de la connexion au réseau, nous devrons synchroniser les tensions et courants renvoyés avec ceux propres au réseau. Aussi une solution par synthèse **d'observateur robuste** permettant d'estimer l'amplitude, la fréquence et la phase des tensions du réseau sera développée en s'inspirant notamment des travaux récents menés au sein d'AMPERE par B. BAYON [22].

En cas de scénarios défavorables (coupure de réseau, très fortes houles, houles « rapides » ...), le système de commande doit être doté de stratégies de commande en mode dégradé qui seront établies en respectant un ensemble de contraintes relatives aux normes fixées par le gestionnaire de réseau et les protections physiques existantes du système houlogénérateur. Cela entre dans le cadre de nos perspectives prévues pour le développement du niveau de la supervision des puissances dans la structure du contrôle hiérarchisée présentée (Figure 4).

En termes d'actions à mener nos travaux se dérouleront en respectant le planning taches qui s'étalent sur les trois années de la thèse.

Tâche 1 : modèle moyen du convertisseur back-to-back

Objectif : décrire le comportement moyen de l'intégralité du convertisseur

Actions :

- Modèle moyen de l'onduleur avec prise en compte des non-linéarités des (IGBT/Diode)
- Modèle moyen du redresseur
- Modèle moyen global de l'ensemble {redresseur, onduleur, bus continu}
- Modèle complet de la partie électrique en vue de la commande

Début : Juin 2013

Deadline : Octobre 2013

Tâche 2 : Commande rapprochée

Objectif : conception de lois de commande des courants coté machine et coté réseau

Actions :

- Développement de lois de commande pour id,iq de la machine via le redresseur
- Développement de lois de commande pour id,iq du réseau via l'onduleur

Début : Octobre 2013

Deadline : fin Janvier 2014

Tâche 3 : Estimation en ligne des tension-courant de la capacité

Objectif : régulation performante de la tension de la capacité du bus continu

Actions :

- Bibliographie sur la problématique de la dérivation numérique et ses solutions
- Synthèse d'observateur pour l'estimation des grandeurs physiques de la capacité (courant, tension)
- Synthèse de loi de commande pour la régulation de la tension de la capacité du bus continu incluant les capteurs logiciels

Début : Février 2014

Deadline : Mai 2014

Tâche 4 : Commande haut niveau

Objectif : conception de lois de commande des grandeurs d'intérêts de la chaîne de conversion électromécanique

Actions :

- Analyse des plages de variations et contraintes sur les grandeurs à contrôler :
  - Couple C_{em} et flux ϕ de machine synchrone,
  - Tension du bus DC V_{dc}
  - Puissance active P et réactive Q côté réseau (appelé et/ou disponible)

- Réflexion sur la stratégie de commande, en collaboration avec le LBMS, en s'appuyant sur l'analyse des mouvements du flotteur imposé par la houle,
- Interprétation des consignes de C_{em}, ϕ, V_{dc}, P et Q en consignes de courant i_d et i_q pour les boucles des régulations des courants du redresseur et de l'onduleur

Début : Juin 2013

Deadline : Février 2014

Tâche 5 : Synchronisation avec les grandeurs du réseau

Objectif : injecter d'une manière saine la puissance produite

Actions :

- Mise en relief de la problématique du synchronisme au réseau dans le cadre l'intégration des énergies renouvelable (intégration au réseau) et les différentes solutions proposées.
- Synthèse d'observateur permettant d'estimer en ligne l'amplitude, la phase et la fréquence du réseau.
- Test des différents algorithmes sur un procédé pilote installé au laboratoire

Début : Mai 2014

Deadline : Octobre 2014

Tâche 6 : supervision du système

Objectif : Réflexion sur la surveillance du système

Actions :

- Définition des modes de marche du système et les stratégies de replis en mode dégradé
- Etude de système de supervision et de surveillance du système

Début : janvier 2015

Deadline : Avril 2015

Tâche 7 : Rédaction de mémoire

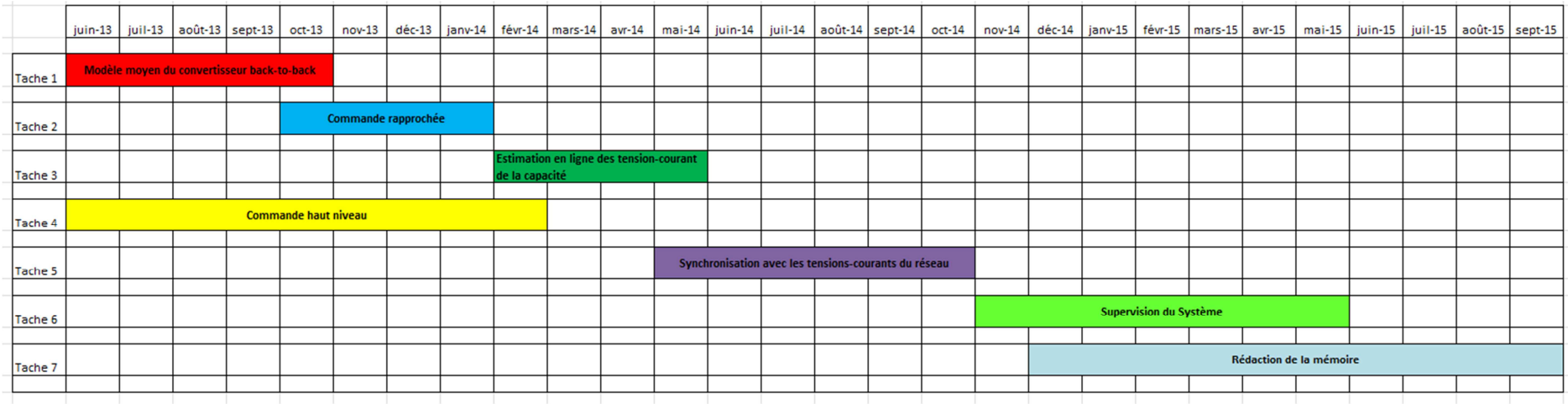
Objectif : réalisation du mémoire de thèse

Actions :

- Définition d'un plan détaillé du manuscrit
- Réalisation d'une première version du mémoire
- Correction du mémoire et envoie au rapporteur

Début : Janvier2015

Deadline : Juin 2015


juin-13
juil-13
août-13
sept-13
oct-13
nov-13
déc-13
janv-14
févr-14
mars-14
avr-14
mai-14
juin-14
juil-14
août-14
sept-14
oct-14
nov-14
déc-14
janv-15
févr-15
mars-15
avr-15
mai-15
juin-15
juil-15
août-15
sept-15
Tache 1
Modèle moyen du convertisseur back-to-back
Tache 2
Commande rapprochée
Tache 3
Estimation en ligne des tension-courant de la capacité
Tache 4
Commande haut niveau
Tache 5
Synchronisation avec les tensions-courants du réseau
Tache 6
Supervision du Système
Tache 7
Rédaction de la mémoire

10. Annexe

Nous proposons, dans ce qui suit de détailler la démonstration de la relation (3) modélisant la distorsion de la tension de la phase A. Pour ce faire, Considérant un onduleur de tension triphasé alimentant une charge montée en étoile (moteur triphasé Figure 5: Onduleur triphasé de tension alimentant une charge en étoile) et raisonnant sur la phase A.

Sur la Figure 7 : Impact du temps mort $\boldsymbol{t_d}$, temps de fermeture$\boldsymbol{t_{on}}$ et temps d'ouverture$\boldsymbol{t_{off}}$, en examinant le signal de commande de l'interrupteurA^+, T_a^* étant sa durée de conduction idéale, en présence des effets de temps morts, la durée de conduction réelle T_a s'exprime :

$$\boldsymbol{T_a = T_a^* - sgn(i_a).T_{dead}} \quad \text{(A. 1)}$$

où

$$\boldsymbol{T_{dead} = \left(t_d + t_{on} - t_{off}\right)} \quad \text{(A. 2)}$$

De plus, une analyse des chutes de tension aux bornes des composants en fonction du courant permet d'exprimer la tension u_{ao} de sorte que Va:

Quand le courant est positif (vers la charge) :

$$\boldsymbol{u_{ao} = \begin{cases} \frac{V_{dc}}{2} - V_{ce}, & S_A = 1 \\ -\frac{V_{dc}}{2} - V_d, & S_A = 0 \end{cases}} \quad \text{(A. 3)}$$

Quand le courant est négatif (de la charge) :

$$\boldsymbol{u_{ao} = \begin{cases} \frac{V_{dc}}{2} + V_d, & S_A = 1 \\ -\frac{V_{dc}}{2} + V_{ce}, & S_A = 0 \end{cases}} \quad \text{(A. 4)}$$

Où S_A est la variable logique associée à l'ordre de commande de l'élément de puissance du haut du bras.

Compte tenu des effets es temps morts et des chutes de tensions des composants à semi-conducteur, la tension entre phase et centre du bras A s'écrit [5]:

$$\boldsymbol{u_{ao} = (V_{dc} + V_d - V_{ce})\left(\frac{T_a}{T_s} - \frac{1}{2}\right) - \frac{1}{2}sgn(i_a)(V_{ce} + V_d)} \quad \text{(A. 5)}$$

Dans *les conditions opératoires normales*, les chutes de tensions des composants à semi-conducteur doivent augmenter linéairement en fonction du courant :

$$\boldsymbol{V_{ce} = V_{ce0} + r_{ce}|i_a|} \quad \text{(A. 6)}$$

$$\boldsymbol{V_d = V_{d0} + r_{d0}|i_a|} \quad \text{(A. 7)}$$

V_{ce0} : Tension seuil du transistor à IGBT,

r_{ce} : Résistance à l'état passant de l'IGBT,

V_{d0} : Tension seuil de la diode,

r_{d0} : Résistance à l'état passant de la diode.

D'une manière similaire, on obtient les expressions de u_{b0} et u_{c0} :

$$\boldsymbol{u_{bo} = (V_{dc} + V_d - V_{ce})\left(\frac{T_a}{T_s} - \frac{1}{2}\right) - \frac{1}{2}sgn(i_b)(V_{ce} + V_d)} \quad \text{(A. 8)}$$

$$\boldsymbol{u_{co} = (V_{dc} + V_d - V_{ce})\left(\frac{T_a}{T_s} - \frac{1}{2}\right) - \frac{1}{2}sgn(i_c)(V_{ce} + V_d)} \quad \text{(A. 9)}$$

L'absence de la connexion du neutre impose :

$$\boldsymbol{i_{an} + i_{bn} + i_{cn} = 0} \quad \text{(A. 10)}$$

Pour toute charge équilibrée, les trois tensions phase –neutre vérifient :

$$\boldsymbol{u_{an} + u_{bn} + u_{cn} = 0} \quad \text{(A. 11)}$$

D'où

$$\boldsymbol{u_{no} = \frac{1}{3}(u_{ao} + u_{bo} + u_{co})} \quad \text{(A. 12)}$$

Où encore,

$$\boldsymbol{u_{no} = \frac{1}{3}(V_{dc} + V_d - V_{ce})\left(\frac{T_a+T_b+T_c}{T_s} - \frac{3}{2}\right) - \frac{1}{6}(sgn(i_a) + sgn(i_b) + sgn(i_c))(V_{ce} + V_d)} \quad \text{(A. 13)}$$

On en déduit la tension

$$\boldsymbol{u_{an} = u_{ao} - u_{no}} \quad \text{(A. 14)}$$

(A.14) s'écrit autrement :

$$\boldsymbol{u_{an} = \frac{1}{3}(V_{dc} + V_d - V_{ce})\left(\frac{2T_a-T_b-T_c}{3T_s}\right) - \frac{1}{6}(V_{ce} + V_d)(2sgn(i_a) - sgn(i_b) - sgn(i_c))} \quad \text{(A. 15)}$$

Compte tenu de la relation (A.1), la distorsion résultante sur la phase A est la différence entre la tension idéale et réelle :

$$\boldsymbol{u^*_{an} - u_{an} = \frac{V_{ce}-V_d}{3}\left(\frac{2T^*_a-T^*_b-T^*_c}{T_s}\right) + \frac{1}{6}(V_{ce} + V_d)Sgn(A) + \frac{V_{dc}+V_d-V_{ce}}{3} \times \frac{T_{dead}}{T_s}Sgn(A)} \quad \text{(A. 16)}$$

Où

$$\boldsymbol{u_{an}^{*} = \frac{2T_a^{*} - T_b^{*} - T_c^{*}}{3T_s} \times V_{dc}} \qquad \text{(A. 17)}$$

En négligeant la contribution du terme $\frac{V_{ce}-V_d}{3}\left(\frac{2T_a^{*}-T_b^{*}-T_c^{*}}{T_s}\right)$, car $\left|\frac{2T_a^{*}-T_b^{*}-T_c^{*}}{T_s} \cdot \frac{V_{ce}-V_d}{3}\right| \leq \frac{2}{3}(V)$

$$\boldsymbol{\Delta u_{an} = \frac{1}{6}(V_{ce} + V_d)Sgn(A) + Sgn(A)\frac{V_{dc}+V_d-V_{ce}}{3} \times \frac{T_{dead}}{T_s}} \qquad \text{(A. 18)}$$

Ce qui achève la démonstration de l'expression de la distorsion du signal de sortie de l'onduleur présentée dans l'équation (3)

.

Références bibliographiques

[1] B. Multon, G. Robin, M. Ruellan, and H. Ben Ahmed, "Situation énergétique mondiale à l'aube du 3ème millénaire. Perspectives offertes par les ressources renouvelables," *revue 3EI*, pp. pp.20–33, Mar. 2004.

[2] B. Multon, H. Clément, Alain, M. Ruellan, J. Seigneurbieux, and H. Ben Ahmed, "Systèmes de conversion des ressources énergétiques marines," in *Les Nouvelles Technologies de l'Energie*, Hermès Publishing, 2006, pp. pp.221–266.

[3] M. Ruellan, "Méthodologie de dimensionnement d'un système de récupération de l'énergie des vagues," École normale supérieure de Cachan - ENS Cachan, 2007.

[4] A. Carlsson, "The back to back converter control and design," Department of Industrial Electrical Engineering and Automation Lund Institute of Technology, Sweden, 1998.

[5] J.-W. Choi and S.-K. Sul, "Inverter output voltage synthesis using novel dead time compensation," *Power Electronics, IEEE Transactions on*, vol. 11, no. 2, pp. 221–227, 1996.

[6] Xin-Yuan Li, "A Novel Method for Dead Time Compensation Based on the Sliding Mode Observer," in *16th International Conference On Mecatronics Technology*, TIANJIN, CHINA, 2012.

[7] Xianqing Cao and Liping Fan, "Dynamic Dead-Time Effect Compensation Scheme for Pmsm Drive," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, pp. 2259–2264, 2012.

[8] M. F. R. Jun Zhang, "Non-Linear Behaviour Compensation of the Converter for Direct Torque Controlled Induction Machines."

[9] J.-W. Choi and S.-K. Sul, "A new compensation strategy reducing voltage/current distortion in PWM VSI systems operating with low output voltages," *Industry Applications, IEEE Transactions on*, vol. 31, no. 5, pp. 1001–1008, 1995.

[10] G. L. Wang, D. G. Xu, and Y. Yu, "A novel strategy of dead-time compensation for PWM voltage-source inverter," in *Applied Power Electronics Conference and Exposition, 2008. APEC 2008. Twenty-Third Annual IEEE*, 2008, pp. 1779–1783.

[11] H.-S. Kim, H.-W. Kim, and M.-J. Youn, "A new on-line dead-time compensation method based on time delay control," in *Industrial Electronics Society, 2001. IECON '01. The 27th Annual Conference of the IEEE*, 2001, vol. 2, pp. 1184–1189 vol.2.

[12] Krishna, M and Raghava Narayanan, G, "A Dead -Time Compensation Circuit for Voltage Source Inverters," presented at the Conference Papers-Electrical, Indian, 2010.

[13] B. Siddharth, "Flux and Torque Estimation in Direct Torque Controlled (DTC) Induction Motor Drive," Faculty of North Carolina State University, North Carolin, 2010.

[14] D. Leggate and R. J. Kerkman, "Pulse-based dead-time compensator for PWM voltage inverters," *IEEE Transactions on Industrial Electronics*, vol. 44, no. 2, pp. 191–197, 1997.

[15] F. Bonnet, "Contribution à l'optimisation de la commande d'une machine asynchrone à double alimentation utilisée en mode moteur," 30-Sep-2008. [Online]. Available: http://ethesis.inp-toulouse.fr/archive/00000679/. [Accessed: 29-Jun-2013].

[16] B. Allard, H. Morel, X. Lin-Shi, and J.-M. Rétif, “Modèle moyen de l’onduleur triphasé de tension pour la conception de lois de commande,” *Revue internationale de génie électrique*, vol. 5, no. 1, pp. 183–201, Mar. 2002.
[17] A. Merdassi, “Outil d’aide à la modélisation moyenne de convertisseurs statiques pour la simulation de systèmes mécatroniques,” Institut National Polytechnique de Grenoble - INPG, 2009.
[18] N. Laverdure, “Sur l’intégration des générateurs éoliens dans les réseaux faibles ou insulaires,” Institut National Polytechnique de Grenoble - INPG, 2005.
[19] P. Lautier, “Modélisation des convertisseurs à découpage pour la conception et la commande : application à l’onduleur,” Institut National des sciences Appliquées de Lyon (INSA), Lyon, 1998.
[20] M. Dridi, " Dérivation numérique : synthèse, application et intégration", Thèse Ecole Centrale de Lyon (ECL), lyon, 2011, http://tel.archives-ouvertes.fr/tel-00655848
[21] L. Sidhom, "Sur les différentiateurs en temps réel : algorithmes et applications", thèse Institut National des Sciences Appliquées (INSA), Lyon, 2011, [tel-00701576 - version 1],
[22] B. Bayon, Estimation robuste pour les systèmes incertains", thèse Ecole Centrale de Lyon (ECL), Lyon, 2012, [tel-00780094 - version 1]

Ecole Centrale de Lyon - INSA de Lyon - Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique,
Microbiologie environnementale et Applications

Mémoire doctorant 1ère année 2012-2013

Doctorant	**Alan CHAUVIN**
Titre de la thèse	**Dimensionnement et gestion optimale du bus de puissance d'une mini-pelle hybride électrique**
Directeur de thèse	Eric Bideaux - Professeur - INSA de Lyon
Co-encadrant	Alaa Hijazi - MCF - INSA de Lyon
Co-encadrant	Ali Sari - MCF - UCBL
Dpt. de rattachement	Méthodes pour l'Ingénierie des Systèmes (MIS)
Date de début	Septembre 2012
Financement	Projet FUI

Laboratoire Ampère - Ecole Centrale de Lyon - 36, avenue Guy de Collongue - 69134 Ecully cedex - France
Tél :+33 (0)4 72 18 60 99 Fax :+33 (0)4 78 43 37 17 http://www.ampere-lab.fr

Table des matières

Résumé

La prise de conscience environnementale et les règlementations incitent les constructeurs à concevoir des engins plus respectueux de l'environnement et capables de disposer de modes de fonctionnement à faible impact environnemental. Dans cette optique, des industriels et le laboratoire Ampère mettent en commun leur expertise autour du projet FUI ELEXC dans le but de mettre au point un prototype de mini-pelle hybride électrique innovant. Ce mémoire présente les travaux de recherche et développement effectués durant la première année de thèse de doctorat. Dans le cadre de la collaboration industrielle, une partie du travail consiste à développer des lois de gestion d'énergie permettant un fonctionnement opérationnel du prototype. Une seconde partie traite de l'état de l'art concernant la démarche de dimensionnement et de commande optimale des véhicules hybrides.

Abstract

The environmental awareness and new restrictions incite manufacturers to design vehicles more eco-aware and able to have operating modes with a low environmental effect. In this perspective, some industrial companies and laboratoire Ampere pool their expert assessment through the project FUI ELEXC in order to develop an innovative hybrid electric mini-excavator prototype. This report presents research and development activities realized during the first year PhD. In the setting of industrial partnership, a part of the work consists to develop energy control laws for a proper functioning of the future prototype. A second part of the thesis deals with the state of the art about the sizing combined to optimal control approach for hybrid vehicles.

1 Introduction

1.1 Contexte

Depuis la fin des années 1990, le secteur du transport de personnes et des marchandises connaît une métamorphose technologique importante. Ces changements proviennent, d'une part, de la modification du comportement des utilisateurs touchés par les problématiques environnementales ainsi que par le coût d'exploitation croissant de leurs véhicules, et d'autre part, de la pression des organisations environnementales et des gouvernements qui mettent en place des législations toujours plus exigeantes concernant les émissions polluantes et l'intégration des véhicules dans l'environnement.

Bien que ces changements s'opèrent principalement dans le domaine du transport routier, les engins non-routiers ne sont pas épargnés. Les législations mises en place par l'US EPA (United States Environmental Protection Agency) et l'Union Européenne, spécifiques à cette catégorie de véhicule, impose aux constructeurs de développer des motorisations toujours moins polluantes.

Ainsi, il a été estimé qu'en 2011, les 2 millions d'engins de construction et de chantiers aux Etats Unis ont consommé près de 26 milliards de litres de diesel et rejeté dans l'atmosphère plus de 75 millions de tonnes de dioxyde de carbone (CO_2) [17]. En outre, d'autres types de particules sont rejetés par les motorisations diesel :
- Oxydes d'azote (NOx)
- Particules diesel
- Monoxyde de carbone (CO)
- Hydrocarbures non brûlés(HC)
- Oxydes de soufre (SOx)

Bien que de nouvelles législations aient été récemment publiées [45], la réduction des émissions polluantes sera réalisée par la modification de l'architecture de puissance de ces engins. La technologie hybride est une solution adaptée car elle permet de mieux gérer les flux énergétiques. Les progrès constants en électronique de puissance et dans les systèmes mécatroniques permettent désormais d'envisager des structures hybrides et électriques sur des engins mobiles non-routiers.

C'est ainsi qu'est né le projet ELEXC (ELEctric EXCavator), labellisé par les pôles de compétitivité ViaMéca et Tenerdis, et sur lequel collaborent plusieurs industriels et le laboratoire Ampère afin de développer les technologies mécatroniques exploitables au sein d'un engin de construction par la réalisation d'un prototype de mini-pelle hybride électrique.

1.2 Problématique

Au delà de l'aspect environnemental, l'hybridation d'un engin de construction permet d'envisager de nouvelles possibilités de mission telles que le travail dans des endroits confinés, ou bien l'emploi de nouvelles lois de pilotage (assistance à la conduite). Ce dernier aspect ne sera pas traité dans ce projet.

Durant ce projet, 3 thèses sont réalisées en parallèle et portent sur les thèmes suivants :
- Modélisation de l'architecture mécanique de la mini-pelle et caractérisation des actionneurs linéaires
- Commande locale des actionneurs électromécaniques

– Dimensionnement et gestion du réseau de puissance

Le travail présenté ici traite du dernier thème. L'objectif de ce projet est de dimensionner et de contrôler de façon optimale le réseau de puissance de l'engin, tout en atteignant des performances similaires à une architecture conventionnelle. La minimisation des pertes énergétiques sera un critère déterminant dans le choix de l'architecture et la réalisation des lois de commande associées.

Durant ce projet, le travail sera partagé en plusieurs étapes :
– Modélisation du réseau de puissance
– Réalisation de benchmarks de performances en boucle ouverte
– Développement des lois de gestion d'énergie pour le prototype
– Optimisation de l'architecture et du dimensionnement du réseau de puissance avec prise en compte des lois de commande optimale

Le travail de la thèse sera décomposé en 2 grandes phases :
– La première partie repose sur le développement du prototype durant laquelle des lois de gestion d'énergie seront testées sur l'architecture existante.
– La seconde partie, orientée méthodologie, vise à optimiser l'architecture et le dimensionnement des composants, afin d'améliorer les performances globales de l'engin.

1.3 Objectifs de la thèse

L'objectif scientifique de cette thèse est de formaliser le lien entre une structure optimale, une optimisation paramétrique et la commande du système. Une méthode à partir de l'optimisation convexe sera étudiée afin de combiner les problèmes de dimensionnement et de commande optimale en un seul et unique problème.

Ce rapport est présenté de la manière suivante :
– Le second chapitre est consacré aux architectures hybrides et à l'hybridation des engins mobiles non-routiers
– Le troisième chapitre présente l'engin étudié et la validation du dimensionnement des composants du prototype
– Le quatrième chapitre est un bref état de l'art des méthodes d'optimisation dynamique appliquées aux véhicules hybrides. Quelques résultats seront présentés
– Dans le cinquième chapitre, la problématique du dimensionnement et du choix de l'architecture est abordée. L'optimisation convexe y est présentée
– Les conclusions sur le travail réalisé cette année et sur les perspectives à court et moyen terme seront exposées à la fin de ce mémoire

2 De l'électrification à l'hybridation des engins mobiles non routiers

2.1 Introduction

Les engins mobiles à usage non routiers ou NRMM (Non-Road Mobile Machinery) sont définis comme des machines mobiles transportant des équipements industriels ou des véhicules non conçus pour le transport de biens ou de personnes par la route [7]. Ils peuvent être décomposés en 4 classes :

- Agriculture
- Construction
- Rail
- Maritime

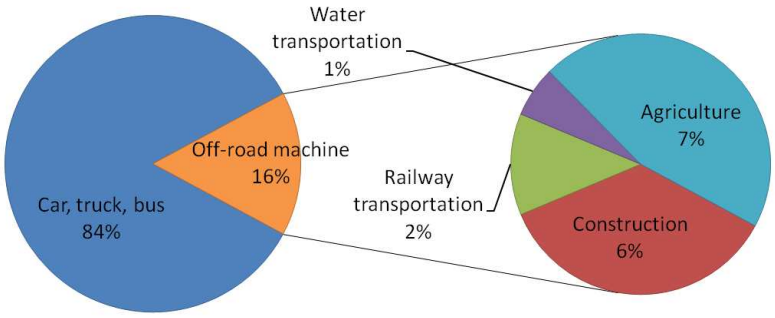


FIGURE 2.1 – Répartition mondiale du nombre de véhicules terrestres et maritimes [24]

Comme indiqué sur la figure 2.1, les engins de construction représentent près de 35% des engins mobiles non-routiers. Cette catégorie regroupe des engins de toute taille et effectuant des tâches très variées comme montré figure 2.2. On peut citer entre autres :

- Les mini-pelleteuses et pelleteuses. Ces engins sont utilisés essentiellement pour des travaux de terrassement et d'assainissement. Un moteur diesel entraine une ou plusieurs pompes hydrauliques qui fournissent la puissance aux différents actionneurs (vérins hydrauliques et moteurs de translation). Les puissances brutes varient d'une dizaine de kW à plusieurs centaines de kW voire plusieurs MW pour les pelles minières. Bien que la plupart de ces machines sont sur chenilles, il existe des versions sur pneumatiques pour les moyennes puissances permettant un déplacement autonome sur routes.
- Chargeuses et minichargeuses. La tâche principal des chargeuses est le chargement de camions ainsi que le déplacement de tas de terre.
- Tractopelle : il s'agit de l'association d'une chargeuse sur pneus et de la structure articulée d'une pelleteuse.
- Bulldozer : ce tracteur à chenilles, chaînes ou pneus dispose d'une lame orientable servant à niveler des terrains, décaper de la terre végétale ou encore effectuer des opérations de charruage (socs installés à l'arrière de l'engin).
- Niveleuse : cet engin, muni d'une lame orientable, sert à façonner les routes. Il peut aussi être employé pour le déneigement des chaussées.
- Tombereaux articulés, camions miniers. Ces camions de transport sont conçus pour évoluer sur des terrains difficiles et transporter de lourdes charges. Les camions miniers peuvent transporter jusqu'à 360 tonnes de charges utile (Liebherr T282C).

FIGURE 2.2 – Présentation de quelques engins mobiles de construction [12]

Notre étude concerne les mini-pelles ou pelles compactes. Les puissances varient entre 10kW et 30kW, le poids en ordre de marche varie quant-à lui de 500kg à moins de 10 tonnes. L'architecture et le fonctionnement sera abordé dans le prochain chapitre.

2.2 Architectures hybrides

Dans une structure conventionnelle, la circulation de l'énergie est unidirectionnelle, c'est-à-dire que les actionneurs se comportent uniquement comme des consommateurs d'énergie. Une architecture hybride permet d'envisager une recirculation d'énergie de telle sorte que les actionneurs renvoient de l'énergie sur le réseau lors de certaines phases de fonctionnement.

D'après Scordia [52], la condition nécessaire pour qu'un véhicule soit qualifié d'hybride est d'avoir deux sources de nature différentes. Pourtant, un véhicule conventionnel peut être assimilé à un système hybride (batterie + moteur thermique). L'International Energy Agency propose la définition suivante [55] :

> Un véhicule hybride a un groupe motopropulseur dans lequel l'énergie peut être transmise par au moins deux dispositifs de conversion d'énergie différents (exemple du moteur à combustion interne, de la turbine à gaz, du moteur Stirling, de la machine électrique, du moteur hydraulique, de la pile à combustible) tirant l'énergie d'au moins deux dispositifs de stockage d'énergie différents (exemple du réservoir à carburant, de la batterie, du volant d'inertie, du supercondensateur, du réservoir de pression). Au moins un des flux, le long duquel l'énergie peut circuler d'un dispositif de stockage d'énergie aux roues, est réversible, tandis qu'au moins un flux est irréversible. Dans un véhicule électrique hybride le dispositif de stockage d'énergie réversible fournit l'énergie électrique.

Bien qu'il existe un grand nombre de configurations possibles, les architectures hybrides peuvent être classées en trois grandes catégories :

- Architecture hybride série
- Architecture hybride parallèle
- Architecture hybride mixte qui est une combinaison des 2 architectures précédentes

Suivant le type d'utilisation envisagé, certaines architectures sont plus appropriées que d'autres. La mini-pelleteuse hybride électrique est considérée comme un système multi-actionneurs. L'architecture

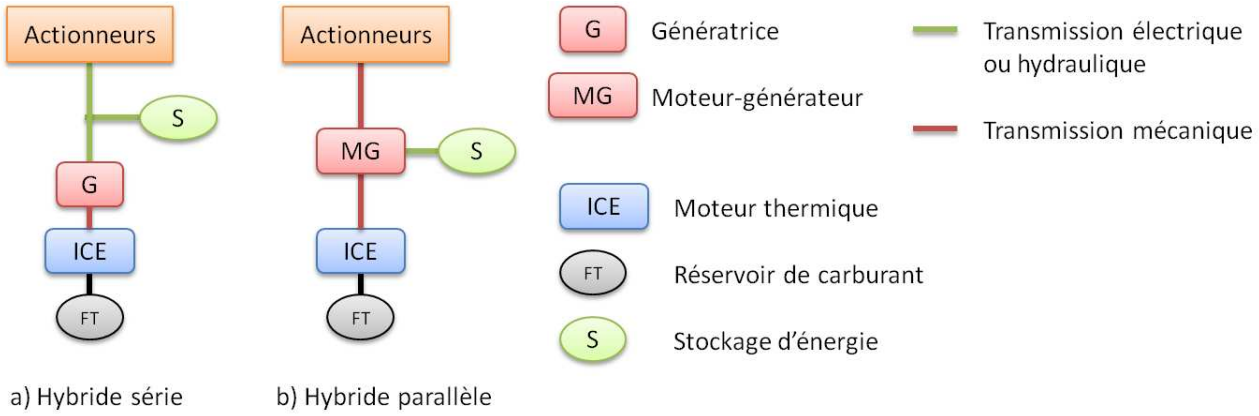


FIGURE 2.3 – Représentation des deux principales architectures hybrides

hybride série semble être la solution la plus adaptée car il n'y aucun intérêt à réaliser un couplage entre la source primaire et les actionneurs (cf figure 2.3). En effet, l'architecture hybride série offre plusieurs atouts majeurs :

1. La source primaire peut fonctionner à son point de rendement optimal, indépendamment de la demande de puissance des actionneurs. La commande de cette source s'en retrouve simplifiée.
2. Dans le cas d'une structure multi-actionneurs et en considérant la source primaire et son convertisseur comme un seul sous-système, il n'existe plus qu'un seul vecteur d'énergie entre les sources de puissances et les actionneurs. Le nombre de composants est limité et les rendements sont améliorés.
3. Le groupe électrogène peut être remplacé aisément par une pile à combustible ou une turbine à gaz. L'architecture est versatile.

2.3 Électrification et hybridation des engins mobiles non routiers

L'architecture conventionnelle d'un véhicule non routier repose traditionnellement sur l'utilisation d'un moteur thermique entraînant soit une transmission mécanique ou hydrostatique (tracteurs agricoles, camions), soit une transmission hydraulique pour les véhicules multi-actionneurs (pelleteuses, tractopelles, chargeuses).

La première étape visant à réduire les émissions de gaz polluants d'un moteur thermique consiste à travailler à régime constant et éviter les zones de fonctionnement à faible rendement. Ainsi, à une vitesse donnée et une puissance requise, le moteur thermique pourra se placer au point de rendement qui minimisera la consommation et/ou le rejet d'émissions de gaz polluants. La seconde étape concerne la gestion des flux énergétiques. En implémentant des actionneurs de puissance réversibles et un système de stockage d'énergie, il est possible de récupérer de l'énergie lors de phases de "freinage" et ainsi améliorer le bilan énergétique global du véhicule.

De nombreuses études ont été réalisées concernant l'électrification et l'hybridation des véhicules à usage spécifique. Nous présenterons ici quelques exemples de travaux.

2.3.1 Electrification d'un camion minier [14]

Les camions miniers sont des véhicules de transport de très grandes tailles et grosse puissance utilisés principalement dans les mines à ciel ouvert pour convoyer du minerai. Les plus gros camions pèsent jusqu'à

200 tonnes à vide et peuvent transporter près de 360 tonnes de charge utile comme montré sur la figure 2.4 (à gauche).

Cet engin ne dispose pas d'une structure hybride mais en possède les bases. Sur l'architecture conventionnelle, un moteur thermique transmet la puissance aux roues par l'intermédiaire d'une transmission mécanique (boite de vitesse + différentiel).

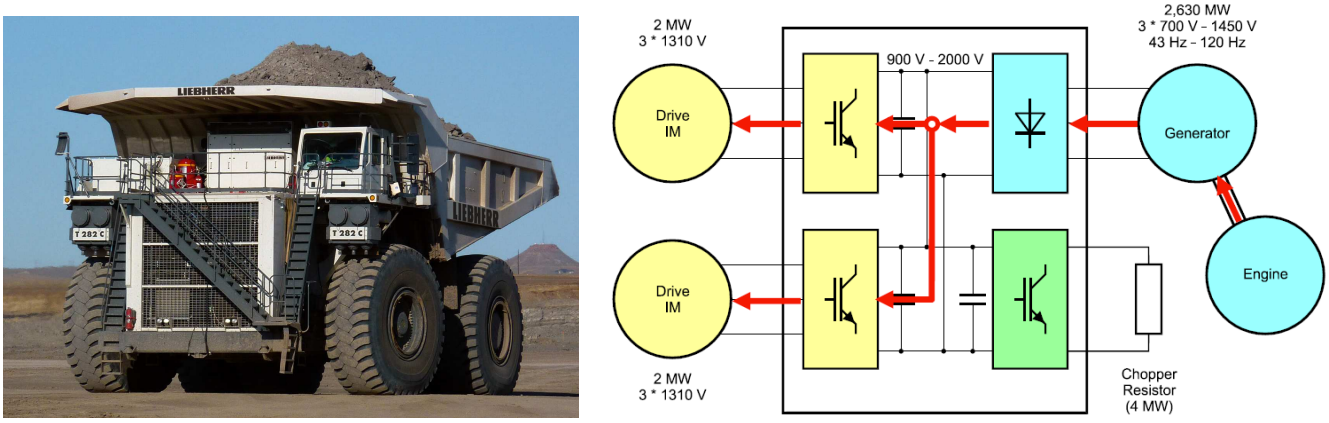


FIGURE 2.4 – Vue globale (à gauche) et réseau de puissance (à droite) du camion minier T282 de la société Liebherr [24]

Sur la nouvelle architecture électrifiée présentée figure 2.4 (à droite), le moteur thermique est couplé à un générateur électrique. L'ensemble fonctionne comme un groupe électrogène. La tension alternative en sortie du générateur est convertie en tension continue, laquelle alimente 2 moteurs asynchrones d'une puissance unitaire de 2MW. Le freinage est assuré en partie par les moteurs. L'énergie récupérée est dissipée au travers de résistances de freinage (chopper resistor). Cette configuration existe notamment dans le domaine maritime pour la propulsion de navires à plusieurs hélices (exemple du paquebot Queen Mary 2).

Au delà d'une diminution de la consommation de carburant, les lois de commande des moteurs de traction permettent un contrôle plus précis du véhicule dans des conditions difficiles tels que le démarrage en pente ou les phénomènes de réduction d'adhérence (antipatinage et antidérapage dans les virages).

2.3.2 Hybridation d'un tracteur agricole [58]

Les machines agricoles représentent la moitié des véhicules mobiles non-routiers dans le monde (voir figure 2.1). Le marché des machines agricoles repose essentiellement sur les tracteurs agricoles, une machine considérée comme polyvalente. Au fil des années, les constructeurs proposent des gammes de plus en plus puissantes (>100kW) tandis que les gammes de faible puissance tendent à disparaître.

Un prototype de tracteur électrique a été mis au point par Roland Schmetz en 1996 [51]. A partir d'une architecture série, un moteur diesel 6 cylindres d'une puissance nominale de 101kW entraîne une génératrice connectée via un bus DC à un moteur électrique.

Puis la société Case New Holland a présenté au SIMA 2009 (Salon International des Machines Agricoles), le NH2, un tracteur à pile à combustible, dérivé du modèle T6000. L'intérêt d'utiliser une pile à combustible est le lien d'indépendance énergétique d'une exploitation agricole. En effet, le dihydrogène peut être récupéré à partir du biogaz produit au sein de l'exploitation.

Dans les travaux de Tritschler [58], une architecture hybride série est étudiée. Une pile à combustible de 85kW fournit la puissance nominale sur le bus de puissance. Une batterie de puissance a été ajoutée afin de rendre le système plus dynamique car la pile à combustible ne peut répondre rapidement aux pics de puissance exigés lors des conditions d'utilisation de la machine.

2.3.3 Hybridation d'une excavatrice hydraulique [60]

La technologie hydraulique est adaptée pour transmettre de très fortes puissances avec des actionneurs compacts. Sur les pelleteuses de fortes puissances (>100kW), des pressions de 250 à 350 bars au sein du circuit hydraulique sont couramment utilisés. Les efforts et couples transmis par les actionneurs et moteurs hydrauliques atteignent respectivement plusieurs centaines de kN et kNm.

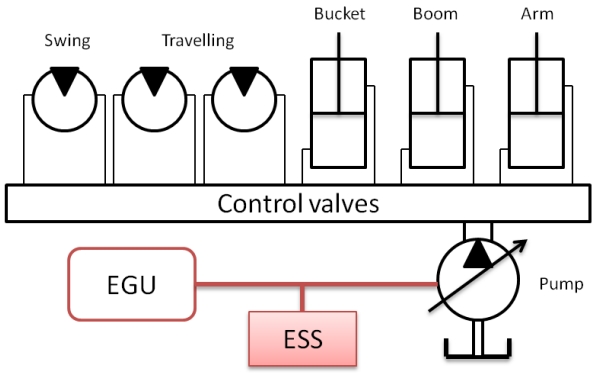


FIGURE 2.5 – Architecture d'une excavatrice hydraulique hybride [60]

L'hybridation proposée par Wang [60] concerne l'alimentation en énergie de la pompe d'entrée du circuit hydraulique d'une pelleteuse de la catégorie 5 tonnes (figure 2.5).

Dans cet article, les architectures hybrides parallèle et série sont analysées et comparées avec une structure conventionnelle. Il apparaît qu'en terme d'économie de carburant, les 2 structures hybrides sont équivalentes. Mais en terme de coût d'investissement, l'architecture parallèle est plus avantageuse [60].

2.3.4 Hybridation d'une chargeuse sur pneus [44]

Dans l'étude réalisée par Ochiai [44], le problème traite de l'hybridation d'une chargeuse sur pneus. Le circuit hydraulique permettant la manutention du godet (vérins hydrauliques) est conservée. La partie traction est électrifiée (voir figure 2.6)

Un moteur thermique entraîne en parallèle une pompe hydraulique et une génératrice électrique. La pompe hydraulique assure la transmission de puissance aux vérins hydrauliques. La génératrice alimente le moteur de traction permettant le déplacement de l'engin. Lorsque le conducteur freine, le moteur électrique se comporte comme un générateur électrique et recharge la batterie placée sur le bus de tension. Lors de fortes accélérations, la batterie assiste la génératrice pour limiter les accélérations au niveau du moteur thermique. Les tests expérimentaux montrent une réduction de la consommation de carburant entre 25 et 30% par rapport à un système conventionnel.

2.4 Hybridation hydraulique et hybridation électrique

Au travers de ces différents exemples et de l'analyse présentée par Rydberg [49], il apparaît que deux technologies se concurrencent : l'hybridation hydraulique et l'hybridation électrique. Ces 2 technologies

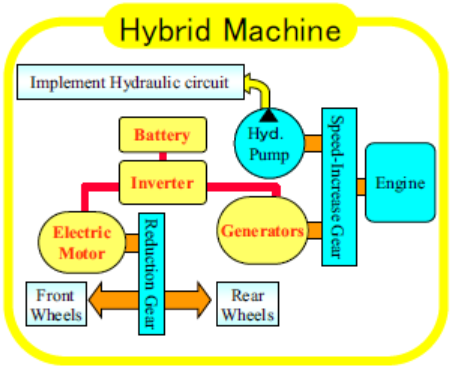


FIGURE 2.6 – Architecture hybride de la chargeuse sur pneus et représentation du véhicule

se différencient essentiellement par le vecteur énergétique employé pour le stockage d'énergie. Ainsi, l'excavatrice hydraulique de Wang repose sur une technologie hybride électrique bien que la transmission de puissance soit réalisée par le biais d'un circuit hydraulique. Dans l'étude présentée par Hippalgaonkar [26], il s'agit d'une hybridation hydraulique où l'énergie hydraulique est stockée sous pression dans des accumulateurs. Enfin, on peut citer les récents travaux de Immonen [27] dans lesquels est étudiée l'hybridation électrique du générateur de puissance d'un engin mobile de construction suivant plusieurs architectures possibles. L'architecture hybride série offre de meilleurs résultats en terme de consommation de carburant mais le retour sur investissement de l'architecture parallèle est plus rapide.

Chaque technologie hybride possède des points forts et des limites. Le tableau 2.1 propose une comparaison succincte des performances de trois types de transmission.

	Electrique	Mecanique	Hydraulique
Puissance massique actionneur	+	+	++
Puissance volumique actionneur	-	+	++
Transmission d'énergie	++	+	+
Stockage d'énergie	++	-	+
Intégration de conception	++	+	++
Coût	-	+	-

TABLE 2.1 – Comparaison des transmissions de puissance [24]

Bien que les actionneurs hydrauliques développent une très forte puissance massique par rapport à un actionneur électrique, la densité de stockage énergétique d'une batterie de type lithium-ion est bien meilleure [13] .

Les récents travaux sur l'hybridation des engins de travail mobiles tendent vers une hybridation électrique couplée à une transmission hydraulique. Une alternative consiste à électrifier certains actionneurs comme c'est le cas sur le modèle PC200-8 de Komatsu [32], [33], [63].

Dans le cadre du projet ELEXC, il a été choisi d'utiliser des actionneurs électromécaniques, afin de proposer des solutions en rupture technologique dans ce domaine d'application. Ce type d'actionneurs existe déjà dans l'aéronautique [29], [34]. La thèse traitant de la commande locale des actionneurs se focalise sur les problèmes de contrôle liés à leur utilisation sur ce type de véhicules.

3 Cahier des charges de la mini-pelle et dimensionnement

3.1 Description de la machine de référence

L'engin étudié est une mini-pelle de catégorie 2.5 tonnes comme montré figure 3.1. Dans un premier temps, la structure et le principe de fonctionnement d'une mini-pelle conventionnelle sont présentés. L'architecture du réseau de puissance est illustrée par la figure 3.2.

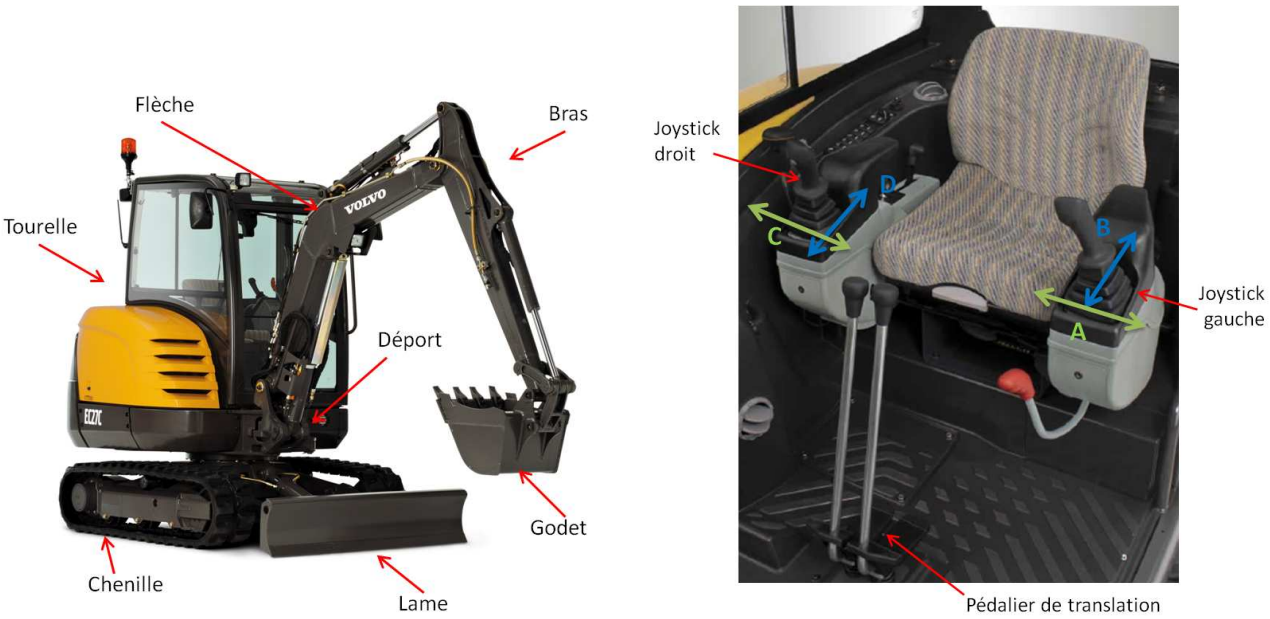


FIGURE 3.1 – Vue extérieure d'une mini-pelle hydraulique (à gauche) et vue de la cabine de pilotage (à droite) - modèle EC27C de Volvo Construction Equipment

La machine est composée d'un châssis inférieur où sont fixées 2 chenilles contrôlées indépendamment. Chaque chenille dispose d'un moteur hydraulique entraînant un pignon (barbotin). Ces moteurs sont contrôlés par l'opérateur à l'aide d'un pédalier de translation (voir figure 3.1). Une lame réglable en hauteur par l'intermédiaire d'un vérin hydraulique permet de stabiliser l'engin pendant certaines opérations ou pousser la terre lors des phases de translation.

La tourelle tourne sur le châssis inférieur grâce à un moteur hydraulique et un engrenage à denture intérieure (couronne) fixée sur la partie inférieure de la tourelle. L'opérateur contrôle la vitesse de rotation grâce à la commande A du joystick gauche. La structure d'actionnement placée devant la mini-pelle est appelée équipement. Elle est composée d'un godet, un bras (appelé aussi balancier) et une flèche, reliés entre eux par des liaisons pivot. La flèche est reliée au châssis de la tourelle. L'équipement peut tourner autour d'un axe vertical, indépendamment de la rotation de la tourelle grâce à l'actionneur de déport.

L'actionnement de l'équipement est réalisé comme suit. La commande B du joystick gauche contrôle l'actionneur du bras, la commande C du joystick droit contrôle l'actionneur de godet et la commande D du joystick droit contrôle l'actionneur de flèche. Toutes ces commandes permettent un contrôle en vitesse

des actionneurs dans la limite de puissance de la pompe hydraulique.

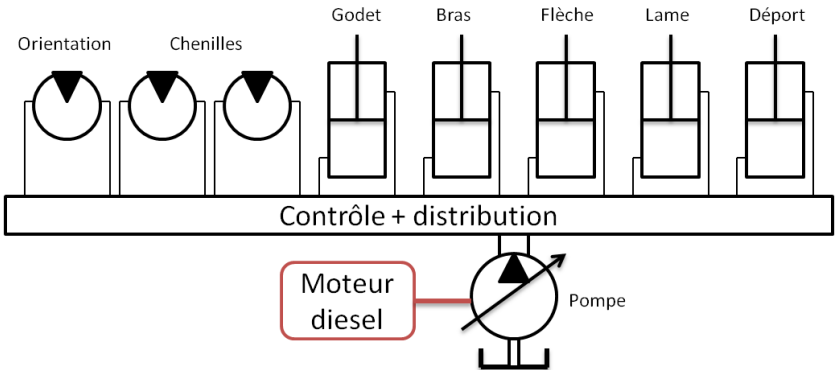


FIGURE 3.2 – Structure du réseau de puissance d'une mini pelle hydraulique

Les performances d'une pelleteuse sont caractérisées par 2 paramètres : la force d'excavation et la force d'arrachement, liés aux efforts transmis par l'actionneur de godet et l'actionneur de bras. Pour le modèle présenté, ces valeurs sont respectivement de 2549daN et 1664daN/1805daN (bras court/bras long).

Afin de fournir la puissance nécessaire à l'actionnement des vérins et moteurs hydrauliques, le circuit hydraulique de puissance est mis sous pression à 250 bars par une pompe à cylindrée variable (voir figure 3.2). Cette pompe est entraînée par un moteur thermique de 20kW.

Dans le projet ELEXC, la structure mécanique de la mini-pelle est conservée. Tous les actionneurs ainsi que les circuits hydrauliques et le moteur thermique sont retirés.

3.2 Architecture électrique du réseau de puissance

Le choix a été fait d'utiliser une architecture hybride série entièrement électrique. De ce fait, toute l'énergie électrique utilisée pour alimenter les actionneurs transite sur un même bus de puissance.

Une des caractéristiques de ce réseau de puissance est la versatilité des sources de puissance, c'est-à-dire que l'utilisateur peut choisir quelle configuration choisir (voir figure 3.3). La structure de base comporte un bus continu haute tension (600V DC) sur lequel sont connectés les actionneurs par le biais d'un boitier appelé Motor Drive System. Un convertisseur basse tension DC-DC alimente le réseau électrique de la cabine et tous les calculateurs. De plus, un mode plug-in permet de connecter la machine sur le réseau électrique. La partie source de puissance se différencie comme suit :

- configuration range extender thermique : un groupe électrogène diesel (range extender thermique) associé à une batterie
- configuration range extender hydrogène : une pile à combustible associé à une batterie
- configuration tout électrique : deux batteries couplées en parallèle

3.3 Modes de fonctionnement

La multiplicité des sources permet d'envisager plusieurs modes de fonctionnement de l'engin. La version tout électrique est particulière car elle se rapporte à un système mono source. Les batteries sont rechargées pendant ou après les opérations de travail.

Dans les 2 versions avec range extender, plusieurs modes sont possibles :

- Mode hybride : le range extender fournit une puissance moyenne et la batterie compense les pics de puissance

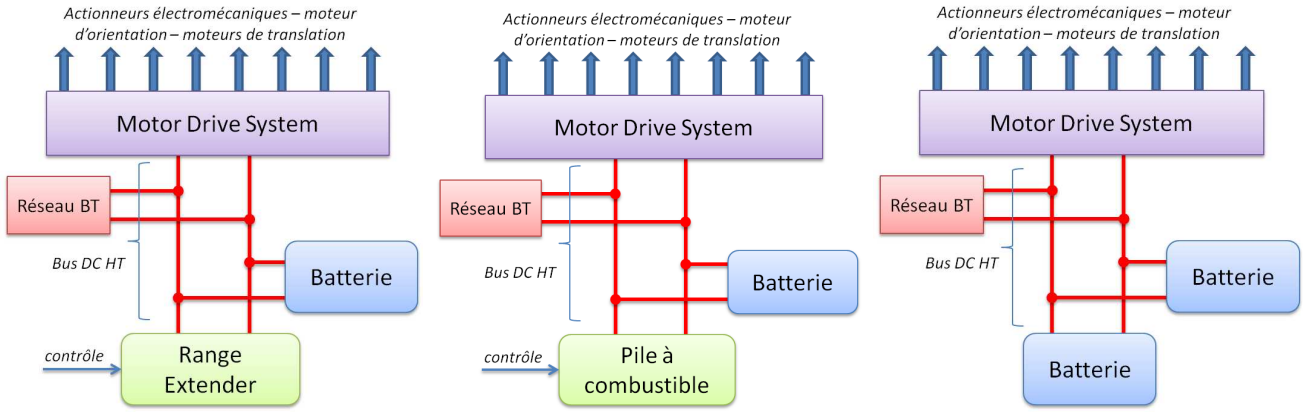


FIGURE 3.3 – Configurations des sources de puissance de la mini-pelle hybride

- Mode Zéro Émission : le range extender est éteint, utilisation de la batterie uniquement
- Mode Range Extender forcé : l'opérateur impose le fonctionnement du range extender afin de recharger la batterie
- Mode Eco : Disponible quelque soit la (ou les) source(s) utilisée(s), la puissance fournie aux actionneurs est limitée à un seuil prédéfini
- Mode Plug-in : le range extender est éteint, le convertisseur AC-DC provenant du réseau électrique assure la puissance moyenne sur le réseau de bord

3.4 Cycles de mission

Bien qu'une mini-pelle soit considérée comme une machine polyvalente apte à réaliser toutes sortes de tâches dans la limite de ses performances, elle est avant tout dédiée à des tâches spécifiques. Ces tâches peuvent être classées en 4 grandes catégories :

- Creuser une tranchée (<0.8m de profondeur)
- Creuser un fond de fouille (>0.8m de profondeur)
- Autres tâches : niveler le sol, remblayer une tranchée, ...
- Se déplacer de façon autonome

Chaque catégorie citée auparavant comprend des cycles spécifiques. Par exemple, 5 cycles différents ont été définis pour la mission "creuser une tranchée" :

- **Cycle A1** : Creuser une tranchée et mettre la terre à côté de la tranchée (25% du temps d'utilisation totale de la machine)
- **Cycle A2** : Creuser une tranchée et charger la terre dans un camion (15% du temps)
- **Cycle A4** : Creuser une tranchée le long d'un mur et charger la terre dans le camion avec le vérin de déport uniquement (3% du temps)
- **Cycle A5** : Idem cycle A4 avec moteur de tourelle d'orientation uniquement (3% du temps)
- **Cycle A6** : Idem cycle A5 avec vérin de déport rentré (3% du temps)

Ces cycles peuvent être considérés comme identique entre 2 engins de même gabarit car les performances globales sont similaires. Mais avec des machines de taille/puissance différentes, les efforts aux niveau des actionneurs sont différents.

Afin de dimensionner correctement les actionneurs et les sources de puissance de la mini-pelle hybride électrique, une campagne de mesure a été réalisée sur le modèle de référence hydraulique. Des capteurs de pression au niveau des chambres de vérins hydrauliques et des capteurs de déplacement de la tige de

piston ont permis de récupérer les efforts et déplacements des actionneurs. Pour les moteurs d'orientation et de translation, les capteurs ont permis de récupérer les valeurs des couples et vitesses suivant différents cycles.

On souhaite obtenir des performances statiques et dynamiques similaires avec la version hybride électrique. La configuration de l'équipement polyarticulé est légèrement différente par rapport à la structure conventionnelle, un changement de repère géométrique et dynamique a permis de convertir les mesures pour correspondre aux efforts/déplacements appliqués réellement sur les actionneurs électromécaniques.

Un cycle de travail tel que le creusement d'une tranchée est une succession de cycles de creusement et de translation temporaire. A partir de tous ces cycles, on définit un cycle de travail journalier de référence (voir annexe A) qui permettra d'évaluer les besoins énergétiques de la mini-pelle au niveau du bus de puissance.

3.5 Validation du dimensionnement de l'architecture et cahier des charges

Les composants ayant déjà été choisis par le constructeur, il s'agit ici de valider le dimensionnement des sources de puissance, des stockages d'énergie et éventuellement d'apporter des modifications pour obtenir un meilleur comportement du réseau électrique embarqué.

Dans la version tout électrique, l'engin doit pouvoir réaliser une journée de travail complète sans avoir besoin d'utiliser une borne de recharge. Il s'agit donc de connaître les besoins énergétiques du système en fonction d'un cycle de mission de référence. Pour cela, les modèles utilisés et leur inversion ont permis, grâce aux mesures effort/déplacement de chaque actionneur, de remonter à la puissance électrique instantanée. En supposant que le pack batterie est à un niveau de 95% en début de cycle, le pack batterie atteint un niveau de charge de 15% en fin de cycle.

Dans le cas de la version avec range extender (thermique ou pile à combustible), le générateur de puissance doit atteindre au minimum la puissance moyenne délivrée sur un cycle de travail afin de pouvoir recharger la batterie. Plus la puissance délivrée sera importante, plus vite la batterie sera rechargée. En contrepartie, le fournisseur des batteries limite le courant de recharge et décharge pour des raisons de durabilité.

Il existe plusieurs façons de connecter deux sources sur un même bus de puissance [50]. Dans la version range extender (thermique et pile à combustible), 4 topologies sont possibles et présentées figure 3.4. Pour la version Range Extender thermique, l'EGU (Groupe électrogène) est présenté comme un assemblage de 3 sous-sytèmes : un moteur à combustion interne, un générateur synchrone et un convertisseur de tension AC-DC (redresseur). Dans la version Pile à combustible, l'EGU représente la pile à combustible complète.

Pour des raisons d'encombrement, il a été choisi de n'implémenter aucun convertisseur DC-DC sur la version du range extender thermique (topologie A). Dans ce cas, la tension du bus DC HT varie en fonction de l'état de charge de la batterie et du courant débité au travers de celle-ci. Chaque variateur de puissance dispose d'un système de filtrage de la tension continue avant l'étage de commande de l'actionneur. L'architecture du range extender thermique (EGU) est décrite dans la section B.6. Toutefois cette architecture pose des problèmes d'ondulations de courant sur le bus de puissance. Pour atténuer ces ondulations, on propose d'intégrer un filtre passif à la sortie du range extender thermique.

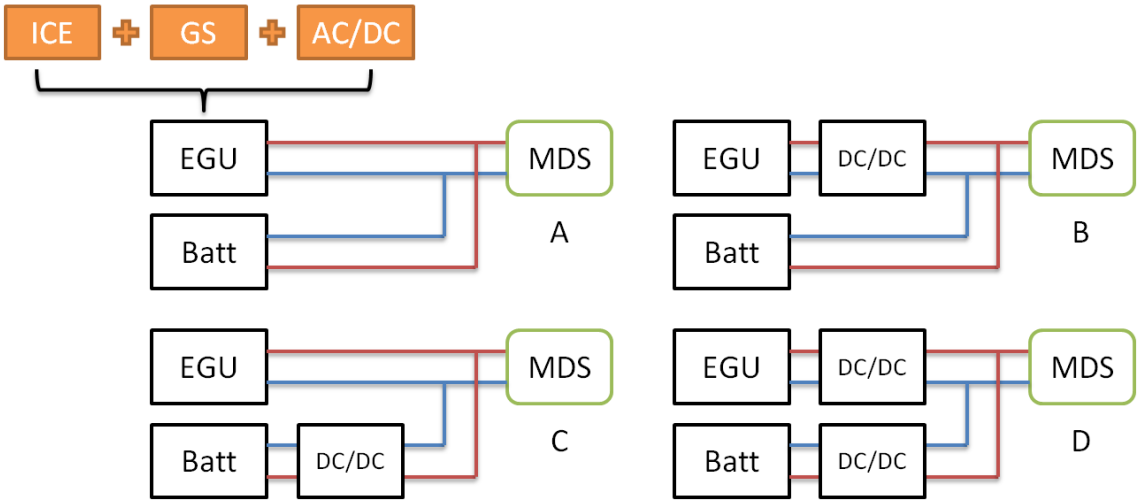


FIGURE 3.4 – Topologie du bus DC HT pour la version range extender thermique

Dans le cas du range extender à pile à combustible, la topologie B est envisagée car la pile à combustible n'est pas capable de fournir une tension de 600V DC en sortie de pile. De plus, le convertisseur DC-DC protège la pile contre d'éventuels retours de courant ou des changements de courant/tension trop brutaux sur la pile.

Si d'un point de vue énergétique, la topologie A apparaît comme une architecture optimale pour limiter les pertes énergétiques (plus de pertes à travers les convertisseurs DC-DC), la commande de ces systèmes est complètement différente. Si l'on souhaite faire travailler le range extender thermique à puissance constante, sa commande sera liée aux conditions d'utilisation de la batterie et donc à la demande de puissance des actionneurs. La connaissance des consignes de vitesse de l'opérateur (ordres joystick) ainsi que la mise en place du modèle des actionneurs permettra d'estimer le besoin en temps réel des actionneurs pour pouvoir anticiper la demande de puissance sur le réseau haute tension (voir synopsis annexe F).

4 Gestion d'énergie

4.1 Introduction

L'intérêt d'une architecture hybride sur un véhicule est de pouvoir répartir la demande de puissance des actionneurs entre la source de puissance et le stockage d'énergie dans l'objectif de minimiser la consommation de carburant. D'autres critères peuvent être utilisés comme la minimisation des émissions de CO_2 ou autres polluants.

L'objectif final est de pouvoir implémenter des stratégies de gestion d'énergie en temps réel sur le prototype. Pour cela, le travail est partagé en 2 phases. Dans un premier temps, à partir d'un cycle de mission connu, on réalise une optimisation globale de l'ensemble. Cette étape est valable uniquement hors-ligne car elle nécessite la connaissance du cycle complet.

Les résultats issus de cette optimisation globale serviront de référence pour l'étude des performances de la solution en-ligne. Par ailleurs, la lecture des résultats peut permettre de créer des bases de règles qui sont la base des systèmes experts et de la logique floue.

De nombreuses méthodes d'optimisation ont été développées dans le cadre de la commande de véhicules hybrides. Scordia [52] classe ces méthodes en 2 catégories :
- Les méthodes exactes issues de la théorie de la commande optimale
- Les méthodes heuristiques permettant de trouver rapidement une solution proche de la solution optimale

Les méthodes exactes telles que la programmation dynamique et le principe du maximum de Pontryaguine ont été largement employées ces dernières années, [48], [56], [11] ou [23]. L'intérêt majeur de ces méthodes est de fournir un optimum global [8]. Toutefois, les temps de calcul et l'espace mémoire requis limitent le nombre de variables et la taille du problème à résoudre.

Les méthodes heuristiques (ou méta heuristiques) regroupent des méthodes permettant de résoudre des problèmes avec des temps d'exécution variable et avec des solutions proches de l'optimum global. Ces méthodes sont inspirées du monde réel et de l'intelligence collective. Les algorithmes du recuit simulé (ou méthode de Monte-Carlo) [9], les méthodes évolutives du type algorithme génétique ou encore l'optimisation par essaims particulaires [18] ont été testées sur des problèmes de commande de véhicules hybrides.

D'autres méthodes issues de la recherche opérationnelle ont été appliquées à des véhicules hybrides [21]. Ces méthodes font appel aux algorithmes de Branch and Bound, et leurs dérivées, Branch and Cut et Branch and Price. L'algorithme de Branch and Bound est présenté dans la section 4.2.3.

La plupart des méthodes citées précédemment ne sont pas adaptées pour la gestion d'énergie en ligne. Les stratégies avec des bases de règles exigent une parfaite connaissance du système, propre à chaque véhicule étudié bien que les stratégies avec logique floue offrent des performances convenables [6], [61]. Les stratégies à base de réseaux de neurones nécessitent aussi une période d'apprentissage du système pour obtenir de bons résultats [52]. La théorie de la commande optimale par le principe du maximum de Pontryaguine peut être implémenté en ligne. La variation en temps-réel du facteur de Lagrange sur

un horizon de temps glissant est une solution envisageable sur tout type d'architecture hybride [10], [30]. Au contraire, la programmation dynamique stochastique [3] implémentée par Johannesson [28], dérivée de la méthode hors-ligne, permet d'obtenir des performances proches de l'optimum global si la mission est connue à l'avance (ligne de bus, cartographie GPS). D'autres stratégies de gestion d'énergie telles que l'ECMS (Equivalent Consumption Minimization Strategy) [42], [48] ou encore le MPC (Model Predictive Control) [53], [31] sont d'autres méthodes adaptées en temps réel.

4.2 Méthodes d'optimisation globale

4.2.1 Programmation dynamique

La programmation dynamique est une méthode d'optimisation basée sur un algorithme d'exploration combinatoire. Cette méthode est adaptée pour minimiser un coût cumulatif sur une trajectoire. Dans la plupart des cas, on travaille sous forme discrète.

Pour un système dynamique qui évolue dans le temps, on associe 3 variables :
- la variable de temps, noté t, prenant des valeurs discrètes et comprise dans un intervalle $[0, T]$.
- la variable d'état du système, notée x, représentant un point à l'instant t de la trajectoire du système dynamique
- la variable de commande ou de décision, notée u

Le système est régi par une équation d'état, exprimée sous forme discrète :

$$x(t+1) = f(x(t), u(t), t) \tag{4.1}$$

Suivant la trajectoire $x(t)$ proposée par la commande $u(t)$, la performance du système déterminée par un critère de coût est impactée. Ce coût est noté J :

$$J(u) = \sum_{t=0}^{T-1} g(x(t), u(t), t)) \tag{4.2}$$

où g représente une fonction de coût instantané qui ne dépend que de l'état de la commande u à l'instant t. Pour minimiser le critère J, on se ramène à résoudre :

$$minJ(x_0, u(0), u(1), ..., u(T)) \tag{4.3}$$

De ce modèle J, il est donc possible d'appliquer les équations d'Hamilton-Jacobi-Bellman par le biais du principe d'optimalité de Bellman [2] énoncé comme suit :

> Dans un processus d'optimisation dynamique, une suite de décisions est optimale si, quels que soient l'état et l'instant considérés sur la trajectoire qui lui est associée, les décisions ultérieures constituent une suite optimale de décisions pour le sous-problème dynamique ayant cet état et cet instant comme conditions initiales.

En considérant u^* comme la séquence des commandes optimales permettant de minimiser le critère J, alors le critère devient :

$$J^* = J(u^*) = min_{u(0),u(1),...,u(T-1)} \sum_{t=0}^{T} g(x(t), u(t), t) \tag{4.4}$$

équivalente à

$$J(u^*) = min_{u(1),u(2),\ldots,u(T-1)} \left(g(x(0), u(0), 0) + \sum_{t=1}^{T} g(x(t), u(t), t) \right) \tag{4.5}$$

De façon récursive, la résolution du problème est basée sur la résolution du dernier terme en T.

La réalisation pratique est illustrée sur la figure 4.1 où l'on cherche à minimiser le plus court chemin entre A et G. En partant de G, on remonte progressivement l'arbre des solutions antérieures.

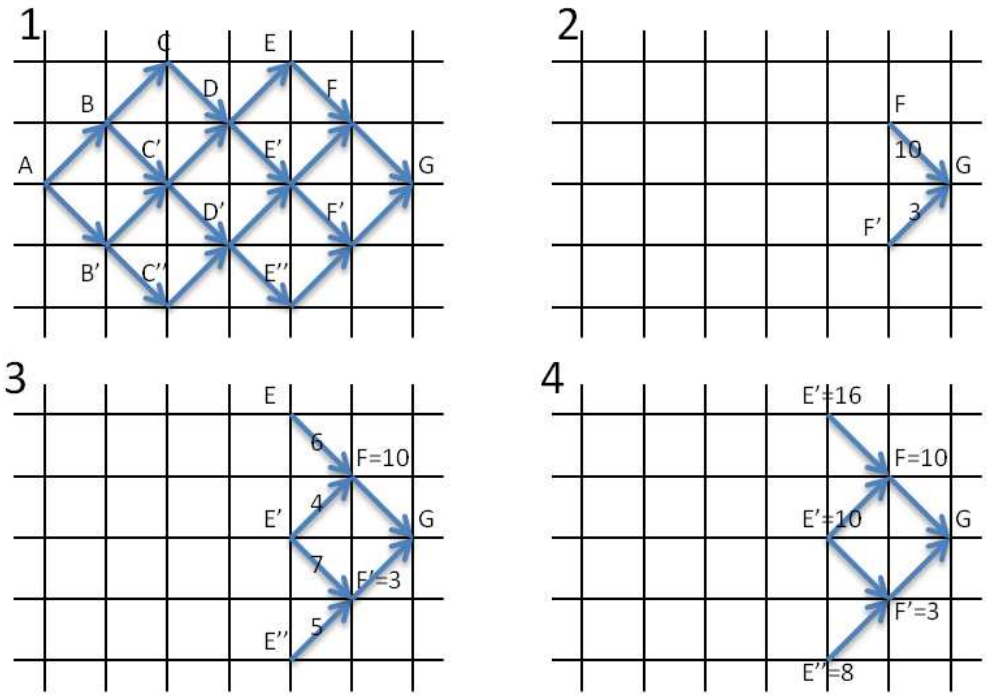


FIGURE 4.1 – Recherche du plus chemin le plus court entre A et G par le principe de la programmation dynamique

Bien que cette méthode soit simple à mettre en oeuvre et permette d'obtenir un optimum global, [8] et [52] indiquent qu'il existe un certain nombre de points limitants à cette méthode :

- Tous les chemins étant testés, il y a explosion combinatoire dans le cas de problèmes de grande taille et système multivariables (cas d'une architecture parallèle avec boite de vitesses). L'espace mémoire et la puissance de calcul sont les facteurs limitants pour la résolution du problème
- Le choix de la discrétisation de l'espace d'état. La commande étant elle aussi discrétisée par l'intermédaire de l'espace d'état, certains chemins optimaux ne sont pas pris en compte
- L'étude de cas aux frontières qui imposent de négliger certains points car non atteignables

4.2.2 Principe du maximum de Pontryaguine

Considérons un système évoluant suivant l'équation (4.7) sur un horizon de temps fini et défini par un coût noté J que l'on souhaite minimiser :

$$J(u) = \int_{t_0}^{t_1} g(x(t), u(t), t) \tag{4.6}$$

$$\dot{x} = f(x(t), u(t), t), x(t_0) = x_0 \tag{4.7}$$

Le point de vue de Lagrange permet de résoudre ce problème de minimisation en associant un multiplicateur à chaque contrainte du système. On note λ le multiplicateur associé à l'équation (4.7). La fonction coût devient alors :

$$min_{x(t),u(t)} J(u) = \int_{t_0}^{t_1} \{g(x(t),u(t),t) + \lambda[\dot{x} - f(x(t),u(t),t)]\}dt \tag{4.8}$$

On décide de choisir λ d'après l'équation (4.9) :

$$\dot{\lambda} = -f_x(x,u,t) \cdot \lambda + g_x(x,u,t), \lambda(t_1) = 0 \tag{4.9}$$

On introduit la fonction de Pontryaguine décrite équation (4.10) appelée par extension Hamiltonien et notée $\mathcal{H}$ et pour laquelle la condition d'optimalité devient $\mathcal{H}_u(x,u,t) = 0$.

$$\mathcal{H}(x,u,\lambda,t) = \lambda \cdot f(x,u,t) - g(x,u,t) \tag{4.10}$$

Dans le cas où g est convexe et f est linéaire, $\mathcal{H}$ est concave. Cette condition d'optimalité correspond à un maximum de $\mathcal{H}$ d'où le nom du maximum de Pontryaguine. Par ailleurs, l'équation dynamique de l'état s'écrit $\dot{x} = f(x,u,t) = \mathcal{H}_\lambda$ et l'équation de λ est $\dot{\lambda} = -\mathcal{H}_x$.

Dans notre cas, le signe de la contrainte de l'Hamiltonien est inversé et donc l'objectif sera d'obtenir le minimum de Pontryaguine. Pour obtenir une minimisation du critère J il faut trouver la commande u telle qu'elle puisse satisfaire les conditions nécessaires suivantes :

$$\mathcal{H}_u(x,u,t) = 0; x = \mathcal{H}_\lambda, \lambda = -\mathcal{H}_x \tag{4.11}$$

4.2.3 Algorithme de Branch and Bound

Issue elle aussi de l'optimisation combinatoire, la méthode de Branch and Bound, appelée aussi méthode de séparation et évaluation, consiste à discrétiser l'espace de recherche sous forme arborescente et définir des propriétés mathématiques permettant de décider si une branche peut oui ou non contenir la solution optimale. Pour les problèmes de grande taille, l'intérêt de cette méthode est de limiter l'énumération de toutes les solutions possibles.

Soit F l'ensemble des solutions admissibles d'un problème, appelé aussi racine.

$$\begin{aligned} \text{minimiser} \quad & c^T x \\ \text{contraint par} \quad & x \in F \end{aligned}$$

Des procédures de bornes inférieures et supérieures sont appliquées à la racine. La technique de la relaxation linéaire est une méthode généralement employée pour ce type de problème. Si ces deux bornes sont égales, alors une solution optimale est trouvée et la procédure est stoppée. Sinon, l'ensemble des solutions est divisée en deux ou plusieurs sous-problèmes, devenant des enfants de la racine (voir figure 4.2). La méthode est ensuite appliquée récursivement à ces sous-problèmes, engendrant ainsi une arborescence. Si une solution optimale est trouvée pour un sous-problème, elle est réalisable, mais pas nécessairement optimale, pour le problème de départ.

Comme elle est réalisable, elle peut être utilisée pour éliminer toute sa descendance : si la borne inférieure d'un nœud dépasse la valeur d'une solution déjà connue, alors on peut affirmer que la solution optimale globale peut être contenue dans le sous-ensemble de solution représenté par ce nœud. La recherche continue jusqu'à ce que tous les nœuds sont soit explorés, soit éliminés.

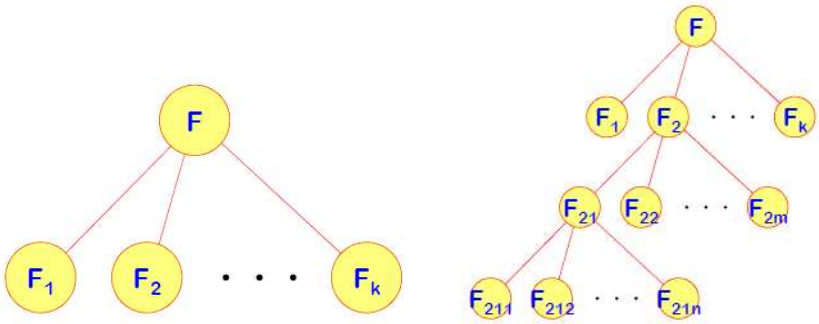


FIGURE 4.2 – Représentation schématique de la séparation

On utilise des bornes sur le coût optimal pour éviter d'explorer certaines parties de l'ensemble des solutions admissibles. De ce fait, la performance de cette méthode dépend essentiellement de la capacité de cette fonction à exclure rapidement des solutions partielles.

La mise en place de cet algorithme sera réalisée sur la base des travaux de Merakeb [35] et Fontchastagner [16]. Les résultats présentés par Gaoua [19] et Guemri [21] montrent de meilleurs résultats en terme de minimisation de la fonction coût et une réduction du temps de calcul par rapport à la programmation dynamique.

4.3 Formulation du problème

Notre étude se focalise sur la mini-pelle électrique en version range extender thermique. La répartition de puissance au niveau du réseau de puissance est schématisée figure 4.3

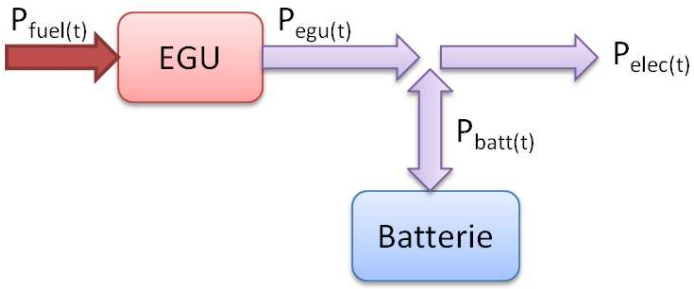


FIGURE 4.3 – Schématisation du réseau de puissance

Le problème d'optimisation peut se décomposer comme suit [8] :
- Critère de coût
- Equation d'état
- Conditions limites
- Contraintes instantanées
- Contraintes d'état
- Commande

On discrétise l'espace de temps en n pas fixes ΔT.

Critère de coût : Il s'agit de minimiser la consommation de carburant. Ce coût dépend du cycle et du dimensionnement des sources.

$$J(cycle, dim) = \sum_{i=0}^{n-1} P_{fuel}(i)\Delta T \tag{4.12}$$

$P_{fuel}(i)$ représente la puissance équivalente de carburant consommé par le range extender thermique.

Equation d'état : On fixe l'état de charge de la batterie comme variable d'état. L'évolution de l'état de charge s'écrit :

$$SOC(i+1) = SOC(i) - \frac{I_{batt}(i)}{Q_{nom}} \tag{4.13}$$

où $I_{batt}(i)$ et Q_{nom} représentent respectivement le courant débité par la batterie et sa capacité nominale.

Conditions limites : L'objectif proposé est d'utiliser au maximum le mode sur batterie au cours du cycle journalier et de recharger sur le secteur la batterie durant les phases de repos nocturnes. On suppose donc que l'état de charge de la batterie est maximal en début de cycle et minimal en fin de cycle.

$$SOC(init) = SOC_{max}; SOC(fin) = SOC_{min} \tag{4.14}$$

Contraintes instantanées de type égalité : Le cycle de mission journalier est défini par $P_{elec}(t)$ grâce au modèle inverse de tous les actionneurs et la puissance du réseau basse tension. On suppose qu'à tout instant t, le réseau fournit la demande de puissance exacte :

$$P_{elec}(t) = P_{egu}(t) + P_{batt}(t) \tag{4.15}$$

où $P_{batt}(t) = (OCV_{batt}(SOC) - R_{batt}I_{batt}(t))I_{batt}(t)$

Certains composants sont limités de par leur dimensionnement tel que la batterie et le range extender thermique.

$$0 \leq P_{egu}(t) \leq P_{max,egu} \tag{4.16}$$

$$I_{min,batt} \leq I_{batt}(t) \leq I_{max,batt} \tag{4.17}$$

Contraintes d'état : Bien qu'en théorie l'état de charge de la batterie varie entre 0 et 100%, les valeurs extrêmes d'utilisation sont restreintes pour des raisons de durabilité et de sécurité.

$$SOC_{min} \leq SOC(t) \leq SOC_{max} \tag{4.18}$$

Commande : La répartition de puissance est contrôlée par la consigne de puissance sur le range extender thermique. Il est possible de faire varier cette commande entre 0 et $P_{max,egu}$. On suppose qu'une demande de puissance nulle est équivalente à un arrêt du range extender. De plus, un redémarrage impose un surcoût. Dans la version prototype, la commande du range extender thermique est une consigne d'activation permettant d'éteindre ou d'activer le moteur thermique à une puissance constante notée $P_{opt,egu}$ et considérée comme une consigne de puissance permettant de travailler au point de fonctionnement optimal ($P_{opt,egu} \neq P_{max,egu}$).

4.4 Résultats de l'optimisation

La résolution du problème d'optimisation a été réalisée par le principe de la programmation dynamique. Les résultats présentés en annexe D concernent le cas où le Range Extender fonctionne uniquement à puissance fixe (solution du prototype). La variable de contrôle est la commande d'activation du Range Extender.

5 De la commande optimale à l'optimisation structurelle

5.1 Problématique

Guzzella et Sciaretta [22] ont défini trois niveaux d'optimisation dans la conception d'un véhicule hybride (voir figure 5.1).

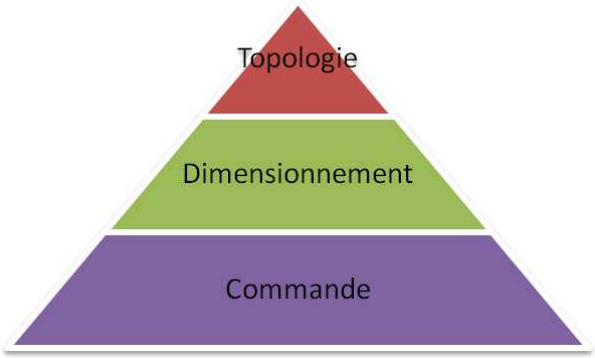


FIGURE 5.1 – Niveaux de conception d'un véhicule

L'optimisation structurelle a pour objectif de trouver la meilleure configuration de transmission de puissance (architecture). L'optimisation paramétrique vise à trouver le meilleur dimensionnement pour une architecture fixée. Enfin, la commande optimale concerne la supervision du système. Tous ces niveaux sont optimisés suivant un ou plusieurs critères à atteindre.

La commande d'une architecture hybride sera véritablement considérée comme optimale lorsque le dimensionnement sera lui aussi optimisé. Un surdimensionnement des composants engendre des lois de commande sous-optimales pour l'architecture considérée. Cependant, il ne s'agit plus de considérer uniquement la minimisation d'un critère énergétique mais de prendre en compte plusieurs critère liés au système global (optimisation multi-objectifs). La figure 5.2 provenant de l'analyse de Silvas [54] résume la complexité du problème de conception/commande d'un véhicule hybride si l'on prend en compte, en plus, le choix technologique des composants. En effet, contrairement aux véhicules légers où la masse influe sur la consommation d'énergie, la minimisation de carburant de la mini-pelle tendrait vers un range extender thermique de puissance nulle et une batterie capable d'effectuer entièrement le cycle journalier. Des contraintes de coût sur le système doivent être ajoutées puis rapportées à un coût d'utilisation (retour sur investissement).

Plusieurs approches pour le dimensionnement optimal ont été proposées. La démarche proposée par Scordia [52] consiste à évaluer la consommation sur un panel de sources de puissance de taille différente après validation des performances du véhicule suivant le cahier des charges. Cette méthode est aussi employée par Filipi [15]. Dupriez-Robin [11] a pour sa part réalisé une optimisation locale des sources de puissance à partir de résultats d'optimisation globale obtenus sur un véhicule de référence.

Wu [62] a quant-à lui englobé la fonction coût comme le coût total de la chaine de traction du véhicule sous contraintes de performances. L'étude a été réalisée à partir de la méthode PCOA (Parallel Chaos

Optimization Algorithm) sans s'occuper dans l'aspect commande.

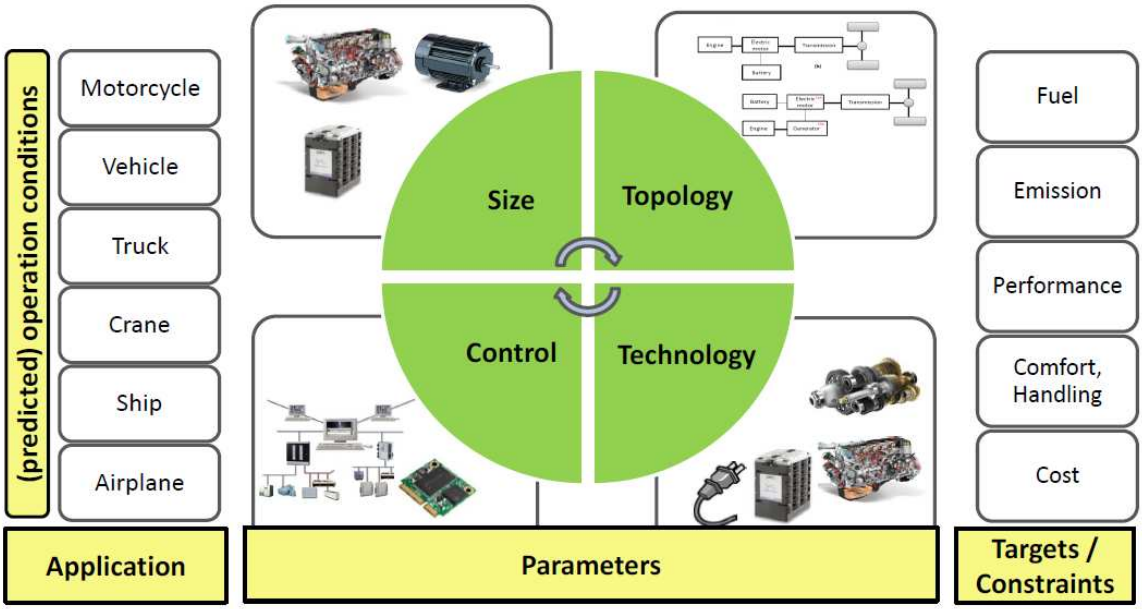


FIGURE 5.2 – Problématique de conception du groupe de puissance d'un véhicule [54]

Une autre approche proposée par Akli [1] tient compte de la dynamique du système et réalise pour cela un dimensionnement par une gestion énergétique "fréquentielle". Akli considère le groupe électrogène comme une source à dynamique très lente. Les éléments de stockage rapide assurent les composantes de haute fréquence de la mission et les générateurs d'énergie opèrent à puissance nominale dès que possible, le reste est assuré par les batteries.

L'optimisation multiobjectifs vise à combiner deux termes de nature différente et à les minimiser (ou maximiser). Par exemple, Patil [46] cherche à minimiser les émissions de CO_2 et le coût d'exploitation du véhicule. Pour cela, il utilise un algorithme développé sur le principe de la programmation dynamique multiobjectifs. Ensuite un nouvel algorithme combinant le dimensionnement et la commande optimale ont été développés. Cette méthode reprend la structure développée pour une application similaire [47]. Pour un dimensionnement fixé, le problème de la commande et les contraintes de dimensionnement sont calculés séparément. Ensuite, un algorithme d'optimisation évalue l'optimalité du système. Si toutes les conditions d'optimalité ne sont pas réunies, un nouveau dimensionnement est proposé et le problème est recalculé.

5.2 Optimisation convexe

5.2.1 Introduction

L'optimisation convexe est un outil applicable à ce genre de problématique [4]. Murgovski [39] propose de réaliser un dimensionnement optimal combiné à une stratégie de gestion d'énergie hors-ligne en reformulant le problème sous forme convexe. Pour cela, la fonction objectif à minimiser comprend le coût des composants. Les modèles de comportement des composants sont approximés par des fonctions de types convexes.

5.2.2 Généralités et propriétés de l'optimisation convexe

Dans cette partie sont abordés les concepts d'optimisation convexe, les propriétés des propriétés associées. On notera dom f le domaine de f et $\mathbb{R}$ l'ensemble des nombres réels. Les propriétés énoncées proviennent de l'ouvrage de Boyd et Vandenberghe [4].

Définition 1 : L'ensemble $\mathcal{C} \subseteq \mathbb{R}^n$ est convexe si le segment entre deux points $x, y \in \mathcal{C}$ est inclus dans $\mathcal{C}$, c'est-à-dire $\theta x + (1-\theta) \cdot y \in \mathcal{C}$ pour θ tel que $0 \leq \theta \leq 1$.

Définition 2 : Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est convexe si le domaine de f est un ensemble convexe et $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ pour tout $x, y \in$ dom f et quelque soit θ tel que $0 \leq \theta \leq 1$. De plus, la fonction f est dite concave si $-f$ est convexe.

Définition 3 : Soit le problème

$$\begin{array}{rll} \text{minimiser} & f_0(x) & \\ \text{contraint par} & f_i(x) \leq 0, & i = 1, ..., m \\ & h_j(x) = 0, & j = 1, ..., p \\ & x \in \mathcal{X} & \end{array}$$

est convexe si $\mathcal{X} \subseteq \mathbb{R}^n$ est un ensemble convexe, $f_i(x), i = 0, ..., m$ sont des fonctions convexes et $h_j(x), j = 1, ..., p$ sont affines dans l'espace des variables de décision x.

Propriété 1 : Une fonction affine $f(x) = ax + b$ est convexe et concave.

Propriété 2 : Une fonction quadratique $f(x) = ax^2 + bx + c$ avec dom $f \subseteq \mathbb{R}$ est convexe si $a \geq 0$

Propriété 3 : Une fonction $f(x, y) = x^2/y$ avec dom $f = \{(x, y) \in \mathbb{R}^2 | y > 0\}$ est convexe

Propriété 4 : Une fonction $f(x, y) = \sqrt{xy}$ avec dom $f = \{(x, y) \in \mathbb{R}^2 | x \geq 0, y \geq 0\}$ est concave.

Propriété 5 : Un produit $f(x, y) = xy$ n'est généralement pas une fonction convexe.

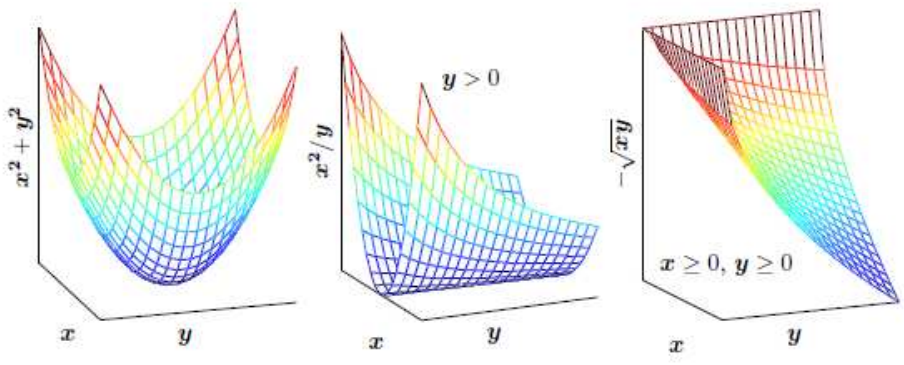


FIGURE 5.3 – Exemples de fonctions convexes

Théorème 1 : Une intersection $\mathcal{S} = \cap \mathcal{S}_i$, d'ensembles convexes $\mathcal{S}_i$, est un ensemble convexe.

Théorème 2 : Une fonction cumulative non négative $f = \sum \omega_i f_i$ avec $\omega_i \geq 0$, de fonctions convexes f_i, est une fonction convexe. Cette propriété peut s'étendre aux sommes et intégrales infinies.

Théorème 3 : Un maximum local $f(x) = max\{f_1(x), ..., f_m(x)\}$ de fonctions convexes $f_i(x), i = 1, ..., m$ est une fonction convexe. De façon similaire, un minimum local $f(x) = min\{f_1(x), ..., f_m(x)\}$ de fonctions concave $f_i(x), i = 1, ..., m$ est une fonction concave.

5.3 Démarche d'optimisation

La résolution d'un problème convexe peut être réalisée de façon relativement simple grâce à des solveurs tels que SeDuMi ou SDPT3 [39]. La ToolBox CVX développée à Stanford University [20] et implémentable sous MatLab permet de traiter ce type de problème.

Cependant, une limitation de l'optimisation convexe est qu'il n'est pas possible d'intégrer des variables discrètes. Dans les problèmes d'optimisation liés à la gestion d'énergie de véhicules hybrides, des variables binaires d'activation de sources de puissance, ou le choix du rapport de boite de vitesses, d'une transmission sont des variables décisionnelles inclues dans ce type de problème. Une approche proposée par Murgovski [41] consiste à choisir les variables discrètes par des routines heuristiques et approximer le problème comme un sous-problème convexe. Dans le cas étudié par Murgovski, l'optimisation concernait un bus hybride parallèle électrique disposant de points de recharge rapide aux stations d'arrêt. Les variables discrètes concernaient l'activation du moteur thermique et le choix du rapport de la boite de vitesse. De plus, les modèles de comportement des composants doivent être approximés par des fonctions convexes. Enfin, Murgovski a pris en compte l'aspect thermique des composants pour le dimensionnement des composants [40].

La démarche employée est décomposée en 5 étapes :

1. Choix du cycle de référence
2. Définir une architecture pour le problème
3. Boucle 1 : Décider l'état d'activation du moteur thermique par une routine heuristique
4. Boucle 2 : Décider le rapport de transmission à engager par une routine heuristique
5. Résoudre le sous-problème convexe à chaque itération de la boucle imbriquée

Conclusions et perspectives

Durant cette première année de thèse, une phase de recherche bibliographique consacrée à l'hybridation des véhicules et en particulier aux engins mobiles non routiers a permis d'appréhender la problématique de gestion d'énergie et la modélisation des composants. Un état de l'art des principales méthodes d'optimisation a permis de mettre en avant les points forts et faiblesses de quelques méthodes avec comme élément central le fait que le dimensionnement et la commande sont la plupart du temps étudiés sous forme de 2 boucles imbriquées. Dans la plupart des cas, l'optimum global est garantit au prix de temps de calcul prohibitifs. L'optimisation convexe est une méthode appropriée pour traiter ce type de problème à condition de rendre le problème convexe.

En parallèle de cette partie recherche, un travail de développement a consisté à traiter les données fournies par le constructeur et les différents partenaires industriels afin de préparer l'intégration du prototype de mini-pelle hybride électrique. Après une phase de traitement du signal, les résultats issus des simulations en programmation dynamique ont servi à établir une base de règles qui sera implémentée dans l'ECU (Electronic Control Unit) du prototype. Cette partie de développement sera poursuivie au cours du second semestre 2013 lors de la phase d'intégration des algorithmes de gestion d'énergie et de limitation de puissance. Cette version n'est que temporaire car non optimale. La version définitive sera issue des différents travaux de recherche effectués durant la thèse.

Dans la suite de la thèse, plusieurs axes seront abordés. Tout d'abord, le lien entre le choix d'une architecture et sa commande seront traités. Pour cela, le langage Bond Graph sera une piste étudiée afin de proposer une méthode générale pour l'étude et l'analyse des réseaux de puissance hybrides embarqués. On se focalisera essentiellement sur les interactions entre les sources de puissance afin d'évaluer l'impact d'une architecture sur sa commande en tenant compte des pertes énergétiques globales du système.

Un autre travail consistera à lier le dimensionnement et la commande optimale dans un même problème comme l'a traité Murgovski [39] avec l'optimisation convexe. Cette méthode sera approfondie ou servira de référence pour tester une autre démarche.

Une dernière partie de la thèse concernera la mise en place d'une stratégie de gestion d'énergie implémentable en ligne pour la mini-pelle hybride. Un contrôle de type MPC (Model Predictive Control) est un choix envisageable et cohérent par rapport à la dynamique en temps réel du système global (transmission de données par réseau CAN) comme montré en annexe F.

Bibliographie

[1] C. R. Akli. *Conception systémique d'une locomotivve hybride autonome. Application à la locomotive hybride de démonstration et d'investigations en énergétique LHyDIE développée par la SNCF.* PhD thesis, Institut National Polytechnique de Toulouse, 2008.

[2] R. Bellman. *Dynamic Programming.* Princeton University Press, first edition, 1957.

[3] D. P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, third edition, 2001.

[4] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[5] D. Candusso. *Hybridation du groupe électrogène à pile à combustible pour l'alimentation d'un véhicule électrique.* PhD thesis, Institut National Polytechnique de Grenoble, 2002.

[6] S. Caux, W. Hankache, M. Fadel, and D. Hissel. On-line fuzzy energy management for hybrid fuel cell systems. *International Journal of Hydrogen Energy*, 35(5) :2134–2143, 2010.

[7] J. Chan. Explanatory memorandum to the non-road mobile machinery regulations 2006, Consulté en Juin 2013. `http://www.legislation.gov.uk/uksi/2006/29/pdfs/uksiem_20060029_en.pdf`.

[8] J.-C. Culioli. *Introduction à l'Optimisation.* Ellipses Marketing, 2012.

[9] S. Delprat. *Evaluation de stratégies de commande pour véhicules hybrides parallèles.* PhD thesis, Université de Valenciennes et du Hainaut-Cambresis, 2002.

[10] A. Dubray. *Adaptation des lois de gestion d'énergie des véhicules hybrides suivant le profil de mission suivi.* PhD thesis, Institut National Polytechnique de Grenoble, 2002.

[11] F. Dupriez-Robin. *Dimensionnement d'une propulsion hybride de voilier, basé sur la modélisation par les flux de puissance.* PhD thesis, Université de Nantes, 2010.

[12] Volvo Construction Equipment. Site internet, Consulté en Mai 2013. `http://www.volvoce.com`.

[13] M. Erkkilä, F. Bauer, and D. Feld. Universal energy storage and recovery system - a novel approach for hydraulic hybrid. In *The 13th Scandinavian International Conference on Fluid Power*, 2013.

[14] O. Fenker. Diesel electric trucks with drive power in the MW range. In *ECPE Seminar - More Electric vehicles*, 2009.

[15] Z. Filipi, L. Louca, B. Daran, C.-C. Lin, U. Yildir, B. Wu, M. Kokkolaras, H. Assanis, D. Peng, P. Papalambros, J. Stein, D. Szkubiel, and R. Chapp. Combined optimisation of design and power management of the hydraulic hybrid propulsion system for the 6x6 medium truck. *International Journal of Heavy Vehicle Systems*, 11(3/4) :372–402, 2004.

[16] J. Fontchastagner. *Résolution du problème inverse de conception d'actionneurs électromagnétiques par association de méthodes déterministes d'optimisation globale avec des modèles analytiques et numériques.* PhD thesis, Institut National Polytechnique de Toulouse, 2007.

[17] C. Frey, W. Rasdorf, and P. Lewis. Comprehensive field study of fuel use and emissions of non-road diesel construction equipment. *Transportation Research Record : Journal of the Transportation Research Board*, 2158(1) :69–76, 2010.

[18] W. Gao and C. Mi. Hybrid vehicle design using global optimisation algorithms. *Int. Journal of Electric and Hybrid Vehicles*, 1(1) :57–70, 2007.

[19] Y. Gaoua, S. Caux, and P. Lopez. A Combinatorial Optimization Approach for the Electrical Energy Management in a Multi-Source System. In *Proceedings of the 2nd International Conference on Operations Research and Enterprise Systems*, pages 55–59, Barcelona, Spain, 2013.

[20] M. Grant, S. Boyd, and Y. Ye. Cvx : Matlab software for disciplined convex programming, 2008. Online accès : `http://cvxr.com/cvx/`.

[21] M. Guemri, S. Caux, S. U. Ngueveu, and F. Messine. Heuristics and lower bound energy management in hybrid-electric vehicles. In *9th International Conference of Modeling, Optimization and Simulation*, 2012.

[22] L. Guzzella and A. Sciaretta. *Vehicle propulsion systems - Introduction to modeling and optimization*. Springer Verlag, 2005.

[23] W. Hankache. *Gestion optimisée de l'énergie électrique d'un groupe électrogène hybride à pile à combustible*. PhD thesis, Institut National Polytechnique de Toulouse, 2008.

[24] T. Herlitzius. System integration and benefits of electrical solutions in mobile machines. In *ECPE Seminar - More Electric vehicles*, 2009.

[25] A. Hijazi. *Modélisation électrothermique, commande et dimensionnement d'un système de stockage d'énergie par supercondensateurs avec prise en compte de son vieillissement : application à la récupération de l'énergie de freinage d'un trolleybus*. PhD thesis, Université Claude Bernard Lyon I, 2010.

[26] R. Hippalgaonkar and M. Ivantysynova. A series-parallel hydraulic hybrid mini-excavator with displacement controlled actuators. In *The 13th Scandinavian International Conference on Fluid Power*, 2013.

[27] P. Immonen. *Energy efficiency of a diesel-electric mobile working machine*. PhD thesis, Lappeenranta University of Technology, 2013.

[28] L. Johannesson. *Predictive control of hybrid electric vehicles on prescribed routes*. PhD thesis, Chalmers University of Technology, 2009.

[29] W. Karam. *Générateurs de forces statiques et dynamiques à haute puissance en technologie électromécanique*. PhD thesis, Institut National des Sciences Appliquées de Toulouse, 2007.

[30] J. T. B. A. Kessels. *Energy management for automotive power nets*. PhD thesis, Technische Universiteit Eindhoven, 2007.

[31] M. Koot, J. T. B. A. Kessels, B. de Jager, W. P. M. H. Heemels, P. P. J. Van den Bossche, and M. Steinbuch. Energy management strategies for vehicular electric power systems. *IEEE transactions on vehicular technology*, 54(3) :771–782, 2005.

[32] T.-S. Kwon, S.-W. Lee, S.-K. Sul, B.-I. Kang, M.-S. Hong, C.-G. Park, and N.-I. Kim. Power control algorithm for hybrid excavator with super capacitor. *Indsutrial applications society annual meeting, IEEE*, pages 1–8, 2008.

[33] T. Lin, Q. Wang, B. Hu, and W. Gong. Development of hybrid powered hydraulic construction machinery. *Automation in construction*, 19 :11–19, 2010.

[34] J. Liscouet. *Conception préliminaire des actionneurs électromécaniques - Approche hybride directe/inverse*. PhD thesis, INSA de Toulouse, 2010.

[35] A. Merakeb. *Optimisation multicritères en contrôle optimal : application au véhicule électrique*. PhD thesis, Université Mouloud Mammeri, Tizi-Ouzou, 2011.

[36] V. Mester. *Conception optimale systémique des composants des chaînes de traction électrique*. PhD thesis, Ecole Centrale de Lille, 2007.

[37] M. Montaru. *Contribution à l'évaluation du vieillissement des batteries de puissance utilisées dans les véhicules hybrides selon leurs usages*. PhD thesis, Institut National Polytechnique de Grenoble, 2009.

[38] R. Mosdale. Transport électrique routier - batteries pour véhicules électriques. *Les Techniques de le l'Ingénieur*, D5565 :1–20, 2003.

[39] N. Murgovski. *Optimal Powertrain Dimensioning and Potential Assessment of Hybrid Electric Vehicles*. PhD thesis, Chalmers University of Technology, 2012.

[40] N. Murgovski, L. Johannesson, A. Grauers, and J. Sjoberg. Dimensioning and control of a thermally constrained double buffer plug-in HEV powertrain. In *51st Annual Conference on Decision and Control*, pages 6346–6351. IEEE, 2012.

[41] N. Murgovski, L. Johannesson, J. Söberg, and B. Egardt. Component sizing of a plug-in hybrid electric powertrain via convex optimization. *Mechatronics*, 22 :106–120, 2012.

[42] C. Musardo, G. Rizzoni, Y. Guezennec, and B. Staccia. A-ECMS : An adaptive algorithm for hybrid electric vehicle energy management. *European Journal of Control*, 11(4-5) :509, 2005.

[43] M. Neuman, H. Sandberg, B. Wahlberg, and A. Folkesson. Modelling and control of series HEVs including resistive losses and varying engine efficiency. *SAE International*, 2008.

[44] M. Ochiai and S. Ryu. Hybrid in construction machinery. In *Proceeding of the 7th JFPS International Symposium on Fluid Power*, pages 41–44, September 2008.

[45] Official Journal of the European Union. Commission directive 2012/46/eu. Technical report, The European Commission, 2012.

[46] R. M. Patil. *Combined design and control optimization : application to optimal PHEV design and control for multiple objectives*. PhD thesis, University of Michigan, 2012.

[47] J. A. Reyer and P. Y. Papalambros. Combined optimal design and control with application to an electric DC motor. *Journal of Mechanical Design*, 124 :183–191, 2002.

[48] G. Rousseau. *Véhicule hybride et commande optimale*. PhD thesis, Ecole des Mines ParisTech, 2008.

[49] K.-E. Rydberg. Energy efficient hydraulic hybrid drives. In *The 11th Scandinavian International Conference on Fluid Power*, 2009.

[50] R. Saisset. *Contribution à l'étude systémique de dispositifs énergétiques à composants électrochimiques. Formalisme Bond Graph appliqué aux piles à combustible, accumulateurs Lithium-Ion, Véhicule Solaire*. PhD thesis, Institut National Polytechnique de Toulouse, 2004.

[51] R. Schmetz. Stepless changing with diesel-electric power. *Profi test*, 1999.

[52] J. Scordia. *Approche systématique de l'optimisation du dimensionnement et de l'élaboration de lois de gestion d'énergie de véhicules hybrides*. PhD thesis, Université Henry Poincaré - Nancy I, 2004.

[53] L. Serrao. *A comparative analysis of energy management strategies for hybrid electric vehicles*. PhD thesis, The Ohio State University, 2009.

[54] E. Silvas, T. Hofman, and M. Steinbuch. Review of optimal design strategies for hybrid electric vehicles. In *IFAC Workshop on Engine and Powertrain Control, Simulation and Modeling*, volume 3, pages 57–64, 2012.

[55] R. T. M. Smokers, A. J. J. Dijkhuizen, and R. G. Winkel. Worldwide developments and activities in the field of hybrid road-vehicle technology. *IEA Implementing Agreement for Hybrid and Electric Vehicle Technologies and Programmes*, 2000.

[56] O. Sundström. *Optimal control and design of hybrid-electric vehicles*. PhD thesis, ETH Zürich, 2009.

[57] B. Tounsi. *Etude comparative de groupes électrogènes embarqués à large gamme de vitesse variable associant machines à aimants permanents et conversion statique*. PhD thesis, Institut National Polytechnique de Toulouse, 2006.

[58] P. Tritschler. *Optimisation de l'architecture électrique et gestion d'énergie pour un système à pile à combustible embarquée dédié à l'application agricole.* PhD thesis, Institut National Polytechnique de Grenoble, 2010.

[59] M. Urbain. *Modélisation électrique et énergétique des accumulateurs au lithium. Estimation en ligne du SOC et du SOH.* PhD thesis, Institut National Polytechnique de Lorraine, 2009.

[60] D. Wang, C. Guan, S. Pan, M. Zhang, and X. Lin. Performance analysis of hydraulic excavator powertrain hybridization. *Automation in construction*, 18(1) :249–257, 2009.

[61] D. Wang, X. Lin, and Y. Zhang. Fuzzy logic control for a parallel hybrid hydraulic excavator using genetic algorithm. *Automation in construction*, 20 :581–587, 2011.

[62] X. Wu, B. Cao, X. Li, and X. Ren. Component sizing optimization of a plug-in hybrid electric vehicles. *Applied energy*, 88 :799,805, 2011.

[63] H. Yao and Q. Wang. Development of power train of hybrid power excavator. In *The 13th Scandinavian International Conference on Fluid Power*, 2013.

Table des figures

Liste des tableaux

A
Définition d'un cycle de travail journalier

Le constructeur de la mini-pelle a établi un cycle de travail journalier basé sur une succession de cycles de référence (cycles présentés dans la section 3.4). Un exemple typique représenté sur la figure A.1 concerne le creusement d'une longue tranchée. La mission est réalisée comme suit :

- L'opérateur va positionner l'engin sur le point de travail → translation 300m
- Pour creuser une tranchée d'une longueur de 3m, l'opérateur répète le cycle A1 24 fois
- L'opérateur déplace la machine de 3m le long de la tranchée
- Le cycle A1 est répété à nouveau pour creuser 3m de tranchée supplémentaire. Cette phase est répétée pendant 1h30
- L'opérateur fait une pause pendant 15 minutes. La machine est toujours allumée mais les actionneurs sont au repos
- Environ 3 heures après le début du cycle, l'opérateur fait une pause longue de 45 minutes, la machine est arrêtée. Le mode de recharge autonome est disponible
- Les cycles A1 et G sont à nouveau répétés pendant 3 heures
- Lorsque le travail est terminé, l'opérateur repositionne la machine dans l'espace de parking où la batterie pourra être rechargée durant la phase de repos

Référence	**Travail**	**Utilisation**	**Commentaire**
A1	Creuser une tranchée	25%	Mettre la terre sur le côté
A2	Creuser une tranchée	15%	Charger la terre dans un camion benne
A4	Creuser une tranchée le long d'un mur	3%	Chargement avec vérin de déport uniquement
A5	Creuser une tranchée le long d'un mur	3%	Chargement avec le moteur d'orientation de la tourelle
A6	Creuser une tranchée le long d'un mur	3%	Chargement avec vérin de déport complètement rentré
B1	Creuser un fond de fouille	4%	Mettre la terre sur le côté
B2	Creuser un fond de fouille	8%	Charger dans un camion benne
C1	Décapage	8%	
D1	Nivelage	5%	
E1	Remblayage	2%	
F	Translation	18%	Petite vitesse - lame au sol
G	Translation	5%	Grande vitesse

Table A.1 – Cycles de référence

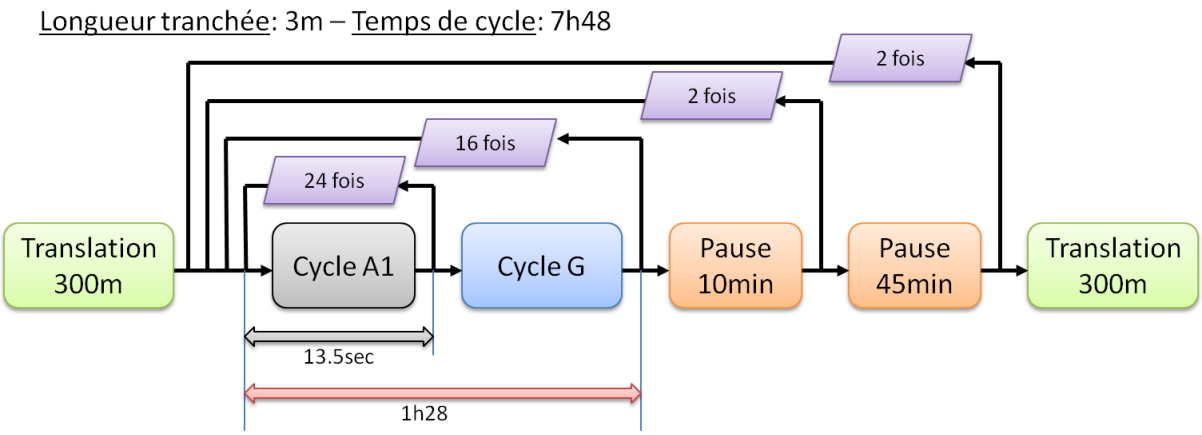


FIGURE A.1 – Synopsis d'un cycle de travail journalier - Tâche : Creuser une tranchée

B Modélisation

La modélisation permet de représenter et simuler le comportement d'un système. Ces modèles décrivent des comportements dynamiques ou énergétiques. Ils sont plus ou moins complexes suivant les hypothèses réalisées et le niveau de détails recherché.

Dans ce projet, deux types de modèles sont développés :
- Un modèle moyen énergétique simplifié permettant de caractériser les pertes énergétiques de chaque composant.
- Un modèle dynamique plus détaillé permettant d'implémenter des lois de commande

B.1 Modèle direct et modèle inverse

Le premier modèle cité plus haut sert à valider le dimensionnement des composants et fournir un profil de mission du point de vue du réseau de puissance. Une fois le dimensionnement fixé, une commande est implémentée et testée en prenant en compte le comportement réel du système comme montré par Hijazi [25] et Mester [36]. Ces modèles sont définis par des entrées et sorties comme montré sur la figure B.1.

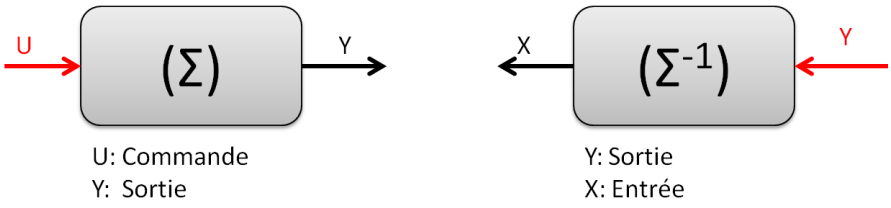


FIGURE B.1 – Représentation de modèle direct et modèle inverse

Modèle direct : Soit un système noté Σ ayant pour entrée la variable U et pour sortie la variable Y. La variable U représente la consigne appliquée au système. Le système répond par le biais du modèle implémenté. Le système peut être bouclé afin de contrôler la sortie Y face aux perturbations ressenties en temps réel.

Modèle inverse : Soit un système noté Σ^{-1} ayant pour entrée la variable X et pour sortie la variable Y. Connaissant la variable de sortie Y (mesures, performances d'un cahier des charges à atteindre), on cherche les valeurs de la variable d'entrée U. Ce type de modèle est surtout utilisé pour dimensionner des systèmes et évaluer des besoins en entrée.

B.2 Actionneurs électromécaniques

Les actionneurs électromécaniques convertissent une puissance électrique en une puissance mécanique. Ces actionneurs sont caractérisés par une dynamique et un rendement qui dépendent des conditions d'utilisation du composant.

La figure B.2 représente la chaîne de puissance d'un actionneur électromécanique. Celui-ci se compose d'un moteur électrique et de son convertisseur de puissance, un réducteur mécanique et une transmission par vis permettant de convertir un mouvement de rotation en mouvement de translation.

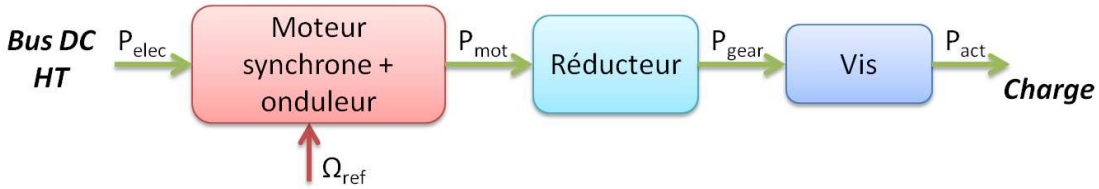


FIGURE B.2 – Structure d'un actionneur électromécanique

Le convertisseur de puissance convertit la tension du bus continu en une tension triphasée aux bornes du moteur. Les pertes énergétiques d'un moteur synchrone peuvent être comparées à celles d'un modèle équivalent DC monophasé. Le réducteur mécanique est caractérisé par un rendement global et une inertie équivalente. La vis est un système complexe faisant intervenir de nombreuses variables qui dépendent des conditions d'utilisation du système. Dans un premier temps, nous ferons l'hypothèse d'un rendement global constant et d'une inertie essentiellement liée à la rotation de la vis.

Soit le modèle de la vis représenté figure B.3. L'équation dynamique de la vis peut être écrite de la façon suivante :

$$\begin{bmatrix} V_{act}(t) \\ F_{act}(t) \end{bmatrix} = \frac{2\pi}{pas} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\eta_{screw} J_{screw} & \eta_{screw} \end{bmatrix} \begin{bmatrix} \Omega_{screw}(t) \\ \dot{\Omega}_{screw}(t) \\ C_{screw}(t) \end{bmatrix} \tag{B.1}$$

où $J_{screw}[kg.m^2]$ est l'inertie de la vis, le $pas[m]$ est un paramètre géométrique et $\eta_{screw}[-]$ est le rendement global de la vis qui dépend de l'effort et de la vitesse appliqué (supposé constant dans un premier temps). Nous ferons l'hypothèse que ce rendement est constant dans un premier temps. $V_{act}(t)[m/s]$ et $F_{act}(t)[N]$ représentent la vitesse et l'effort appliqués à l'extrémité de l'actionneur, au niveau de la charge.

FIGURE B.3 – Variable du modèle de la vis

FIGURE B.4 – Variables du réducteur mécanique

Le réducteur est connecté directement à la vis donc $C_{screw}(t) = C_{rot}(t)$ et $\Omega_{screw}(t) = \Omega_{rot}(t)$. La dynamique du réducteur est décrite par l'équation :

$$\begin{bmatrix} \Omega_{screw}(t) \\ C_{screw}(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{ratio} & 0 & 0 \\ 0 & -J_{red}\eta_{red} & \eta_{red} ratio \end{bmatrix} \begin{bmatrix} \Omega_{mot}(t) \\ \dot{\Omega}_{mot}(t) \\ C_{mot}(t) \end{bmatrix} \tag{B.2}$$

où $C_{mot}(t)[Nm]$ est le couple fourni par le moteur, $J_{red}[kg.m^2]$ est l'inertie équivalente du réducteur sur l'axe rapide et $\eta_{red}[-]$ représente le rendement global du réducteur. $ratio$ est le rapport de réduction entre l'arbre d'entrée et l'arbre de sortie. Les variables d'entrée et de sortie du système sont présentés figure B.4.

Le moteur synchrone peut être modélisé sous forme d'un moteur équivalent moyen DC monophasé. Seules les pertes joules dans les bobinages du stator et les pertes par frottement sont prises en compte.

$$\begin{bmatrix} \Omega_{mot}(t) \\ C_{mot}(t) \end{bmatrix} = \frac{1}{k_e} \begin{bmatrix} 1 & 0 & -R_s & 0 & 0 \\ -b_m & -J_{mot} & k_t k_e + b_m R_s & b_m L_s - R_s & -L_s \end{bmatrix} \begin{bmatrix} U_{mot}(t) \\ \dot{U}_{mot}(t) \\ I_{mot}(t) \\ \dot{I}_{mot}(t) \\ \ddot{I}_{mot}(t) \end{bmatrix} \tag{B.3}$$

où $J_{rotor}[kg.m^2]$ est l'inertie du rotor, $b_{mot}[Nm.s/rad]$ est le coefficient de frottement visqueux. $K_t[Nm/A]$ et $K_e[V.s.rad^{-1}]$ sont respectivement la constante de couple et la constante contre électromotrice du moteur électrique. $R_s[\Omega]$ et $L_s[H]$ représentent la résistance et l'inductance de bobinage au stator. Les variables d'entrée et de sortie sont présentés figure B.5

FIGURE B.5 – Variables du moteur synchrone

FIGURE B.6 – Variables du convertisseur

Le convertisseur du moteur comprend plusieurs sous-systèmes. La tension d'entrée provenant du bus HT est filtrée. Un étage de puissance commandé fait varier la fréquence et l'amplitude des tensions aux bornes des phases du stator. On suppose dans un premier temps que le rendement du convertisseur est constant.

$$P_{elec,mot}(t) = \frac{P_{elec,mot}}{\eta_{converter}} \tag{B.4}$$

où $\eta_{converter}[-]$ est le rendement global du convertisseur (supposé constant).

B.3 Moteur d'orientation de la tourelle

Pour le moteur d'orientation de la tourelle, la vis est remplacée par un réducteur mécanique supplémentaire comme montré sur la figure B.7. Le pignon de l'arbre de sortie du réducteur 2 engrène l'engrenage à denture intérieure (couronne) qui fait partie intégrante de la tourelle. Ce système assure un rapport de réduction supplémentaire pour assurer des vitesses de rotation très faible de la tourelle. Le rapport de réduction globale atteint un ratio de 300 pour cette configuration.

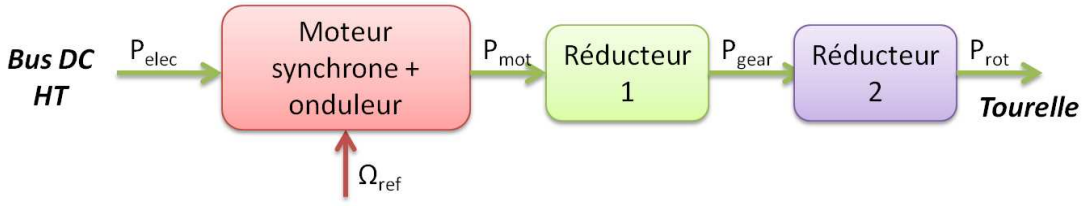


FIGURE B.7 – Structure du moteur d'orientation de la tourelle

Les 2 réducteurs peuvent être modélisés comme un seul réducteur possédant 2 étages de réduction parallèles :

$$\begin{bmatrix} \Omega_{rot}(t) \\ C_{rot}(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{ratio} & 0 & 0 \\ 0 & -J_{red}\eta_{red} & \eta_{red} ratio \end{bmatrix} \begin{bmatrix} \Omega_{mot}(t) \\ \dot{\Omega}_{mot}(t) \\ C_{mot}(t) \end{bmatrix} \tag{B.5}$$

où $J_{red}[kg.m^2]$ est l'inertie équivalente du réducteur et $\eta_{red}[-]$ représente le rendement global. $ratio$ est le rapport de réducteur entre l'arbre d'entrée et l'arbre de sortie du réducteur. Les variables d'entrée et de sortie du système sont décrites sur la figure B.4.

B.4 Moteurs de translation

Les moteurs de translation ont un fonctionnement identiques au moteur d'orientation de la tourelle. Le réducteur planétaire permet d'obtenir un grand rapport de réduction dans une structure compacte.

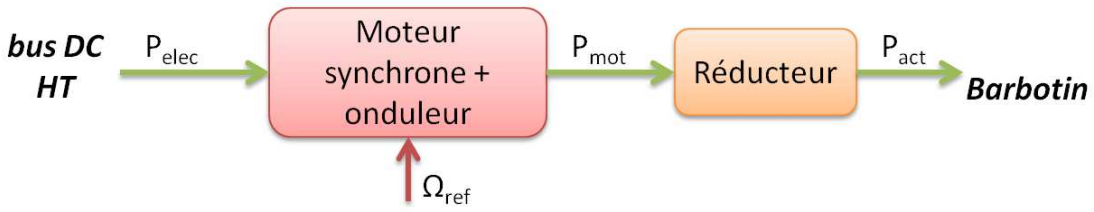


FIGURE B.8 – Structure d'un motoréducteur

Les équations dynamiques du système ne sont pas détaillées ici car identiques aux équations précédentes.

B.5 Système de stockage d'énergie

Le système de stockage d'énergie (SSE) est au cœur du fonctionnement de la mini-pelle hybride électrique. Quelque soit la configuration utilisée, le SSE doit être capable d'assurer l'alimentation en énergie des actionneurs. Il existe plusieurs technologies de SSE de type accumulateur électrochimique. Quelques données techniques sont décrites dans le tableau B.1.

Type	**Energie massique [Wh/kg]**	**Energie volumique [Wh/L]**	**Puissance massique [W/kg]**	**Nombre de cycles**	**Coût [€/kWh]**
Plomb-Acide	35	70	110	600-1000	150
Nickel-Cadmium	50	85	175	1500-2000	600
Nickel-Hydrure de métal	70	175	200	1500	250
Zinc-air	180	-	125	400	125
Lithium-Ion (cathode $LiCoO_2$)	120	200	250	1000	400
Lithium-Fer phosphate	150	200	300	2000	500-1000

TABLE B.1 – Performances générales de quelques technologies de batteries [38]

Le fonctionnement d'un accumulateur électrochimique repose sur des réactions chimiques entre deux couples oxydo-réducteurs, Ox_1/Red_1 et Ox_2/Red_2 se déroulant à deux électrodes différentes [37]. Dans le cas d'accumulateurs électrochimiques réversibles, appelés aussi batteries, la réaction se déroule dans les 2 sens.

$$Ox_2 + Red_1 \underset{decharge}{\overset{charge}{\rightleftharpoons}} Red_2 + Ox_1 \tag{B.6}$$

Ces dernières années, la technologie lithium fer phosphate (LiFePO4) est particulièrement bien adaptée au stockage d'énergie dans les systèmes embarqués car elle offre une forte densité d'énergie, un coût raisonnable et une meilleure sécurité comparée à d'autres technologies au lithium.

Une batterie est constituée d'un assemblage de cellules. Un modèle statique de type RC [59] est suffisant pour le dimensionnement de la batterie. Ces cellules ont un potentiel à l'équilibre compris entre 2 et 4 volts. La tension de la batterie doit atteindre la tension de bus DC car c'est elle qui fixe la tension

sur le réseau. On fait l'hypothèse que toutes les cellules sont identiques. La tension U_{batt} délivrée par la batterie sur le bus s'écrit :

$$U_{batt}(t) = OCV_{batt}(SOC) - R_{batt}I_{batt}(t) \tag{B.7}$$

où $OCV_{batt}(SOC)$ est la tension d'équilibre en circuit ouvert (Open Circuit Voltage) dépendant de l'état de charge de la batterie, R_{batt} est la résistance équivalente et quantifie les pertes joules en charge et décharge. Enfin I_{batt} est le courant traversant la batterie.

L'évolution de la tension d'équilibre $OCV_{batt}(SOC)$ est une donnée fournie par le fabricant de la batterie. L'état de charge de la batterie dépend de sa capacité nominale $Q_{nom}[Ah]$ et du courant I_{batt} :

$$SOC = 100\left(1 - \frac{q}{Q_{nom}}\right) \tag{B.8}$$

$$q = \int I_{batt}dt \tag{B.9}$$

Au delà de l'aspect énergétique, la batterie doit être capable de fournir la puissance nécessaire aux actionneurs. Hors une des caractéristiques principales de la technologie LiFePO4 est de pouvoir accepter des pics de courant d'une valeur de 10C en décharge et 3C en charge. Mais ces valeurs de courant sont largement supérieures aux limitations de puissance rencontrées sur la machine hydraulique. Sur la version hydraulique de la mini-pelle, la puissance totale délivrée par tous les actionneurs en même temps est supérieure à la puissance maximale fournie par la pompe. Ainsi, lorsque plusieurs actionneurs travaillent en même temps, la pompe est le point limitant du système. Sur une architecture hybride électrique, si la batterie est surdimensionnée en terme de puissance, il n'y a plus aucune limitation et cela peut entrainer des problèmes de sécurité et de surconsommation d'énergie.

Un algorithme sera développé pour limiter la puissance totale fournie aux actionneurs en modifiant les consignes de vitesse de l'opérateur. Le synopsis est présenté annexe E.

B.6 Range Extender thermique

Dans les architectures hybrides série électrique munies d'un moteur thermique, on cherche le plus souvent à faire travailler ce dernier aux points de fonctionnement où le rendement est optimal [10]. Le range extender thermique peut être considéré comme un groupe électrogène (Engine Generator Unit).

Des modèles analytiques de groupes électrogènes ont été développés [43]. Il a aussi été montré que la minimisation de la consommation de carburant est meilleure si l'on travaille à régime variable [57].

Dans le projet ELEXC, le range extender thermique a été décomposé en 3 sous-systèmes : le moteur thermique, le générateur synchrone et le redresseur (voir figure B.9). Un pilotage du moteur thermique à régime fixe a été choisi par le constructeur. Le moteur thermique est réglé à vitesse constante par un régulateur proportionnel mécanique. La sortie du vilebrequin entraîne le rotor d'un générateur synchrone à rotor bobiné. La tension triphasée en sortie du générateur est redressée par un redresseur à pont de diodes. Le range extender thermique est piloté par l'intermédiaire du régulateur AVR (Automatic Voltage Regulator) du générateur synchrone.

La courbe de rendement théorique notée $\eta_{P_{EGU}(t)}$ et illustrée sur la figure B.10 est une fonction telle que décrite par Neuman [43]. En négligeant la consommation liée aux régimes transitoires (approche quasi-statique), on peut supposer que la relation entre la puissance en sortie du générateur et la puissance de carburant équivalente est :

$$P_{EGU}(t) = \eta_{P_{EGU}(t)}P_{fuel}(t) \tag{B.10}$$

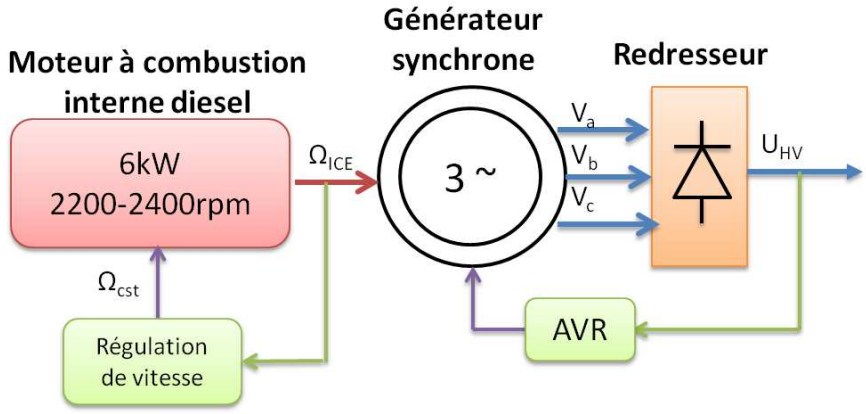


FIGURE B.9 – Architecture du range extender thermique

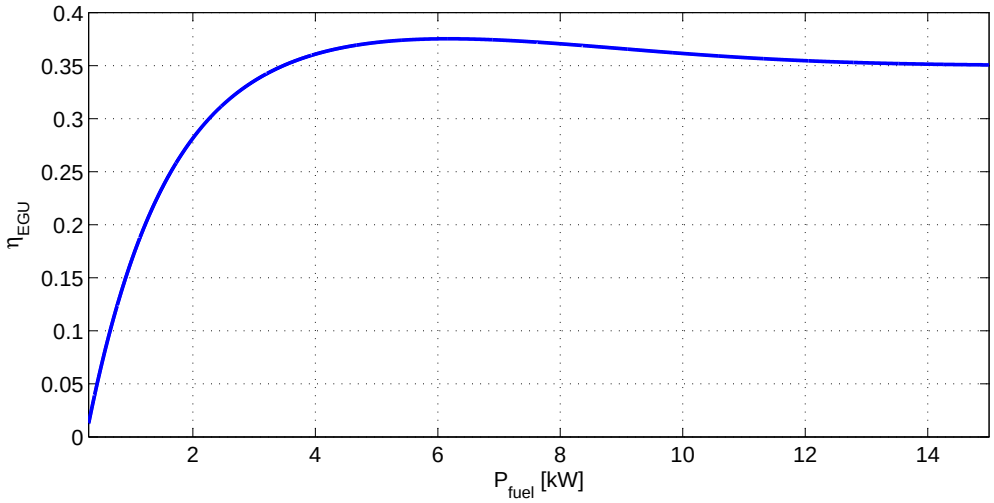


FIGURE B.10 – Rendement théorique d'un groupe électrogène de 5kW[43]

B.7 Pile à combustible

Le principe de fonctionnement de la pile à combustible a été découvert en 1839 par William Grove. Les premières applications firent leur apparition dans l'aérospatiale lors des programmes Apollo de la NASA où des piles à combustible alimentaient le réseau de bord des véhicules habités.

La technologie de la pile à combustible a trouvé de l'intérêt dans le transport terrestre en tant qu'alternative aux carburants fossiles. Il existe plusieurs technologies liées aux composants et aux températures de fonctionnement de ces systèmes. Dans le domaine de l'embarqué, la technologie PEMFC (Proton Exchange Membrane Fuel Cell) est la plus employée [5].

Une pile à combustible est équivalente à un convertisseur électrochimique où le dihydrogène réagit avec le dioxygène suivant l'équation suivante :

$$2H_2 + O_2 \longrightarrow 2H_2O + \text{electricite} + \text{chaleur}$$

Dans notre application, le produit de la réaction qui nous intéresse est l'électricité. L'efficacité de conversion est le rapport entre l'énergie de dihydrogène consommée par le système et l'énergie électrique produite en sortie de la pile à combustible. Le rendement moyen théorique d'une pile à combustible de type PEMFC est représenté figure B.11.

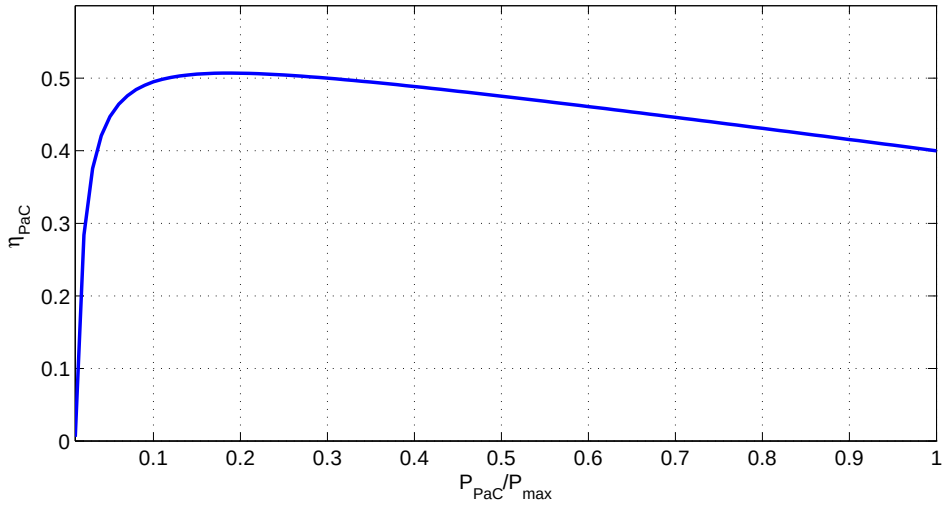


FIGURE B.11 – Rendement théorique moyen d'une pile à combustible de type PEMFC [6]

$$P_{PaC}(t) = \eta_{P_{PaC}(t)} P_{H2}(t) \tag{B.11}$$

B.8 Réseau auxiliaire basse tension

Le réseau BT configuré en 24V permet d'alimenter tous les auxiliaires de vie du véhicule :
- Radio
- HVAC
- Joysticks et autres contrôles
- Projecteurs de travail et signalisation

En outre, les convertisseurs et les unités de calcul (Electronic Control Unit) sont aussi alimentés en 24V, de même que les systèmes de management batterie (BMS et MBMU).

En phase de travail, la puissance demandée par tous ces systèmes est supposée constante. Les fluctuations de puissance sont faibles. On fait l'hypothèse que le convertisseur DC-DC 600-24V fonctionne à régime constant avec un rendement global noté $\eta_{bt} = 96\%$.

La puissance fournie sur le puissance de puissance pour alimenter le réseau basse tension peut s'écrire :

$$P_{elec,bt}(t) = \frac{P_{bt}(t)}{\eta_{bt}} \tag{B.12}$$

C
Validation du dimensionnement des actionneurs

Dans cette annexe sont présentés les résultats de dimensionnement des actionneurs issus des modèles développés et présentés en annexe B.

Pour chaque cycle de référence, on dispose du profil effort-déplacement de l'actionneur. A partir du cycle de travail journalier défini annexe A, on construit le profil complet effort-déplacement de chaque actionneur pour le cycle global. L'utilisation des équations dynamiques issues des modèles développés en annexe B permettent de remonter à la puissance requise sur le réseau HT du véhicule.

C.1 Profil de puissance réseau HT

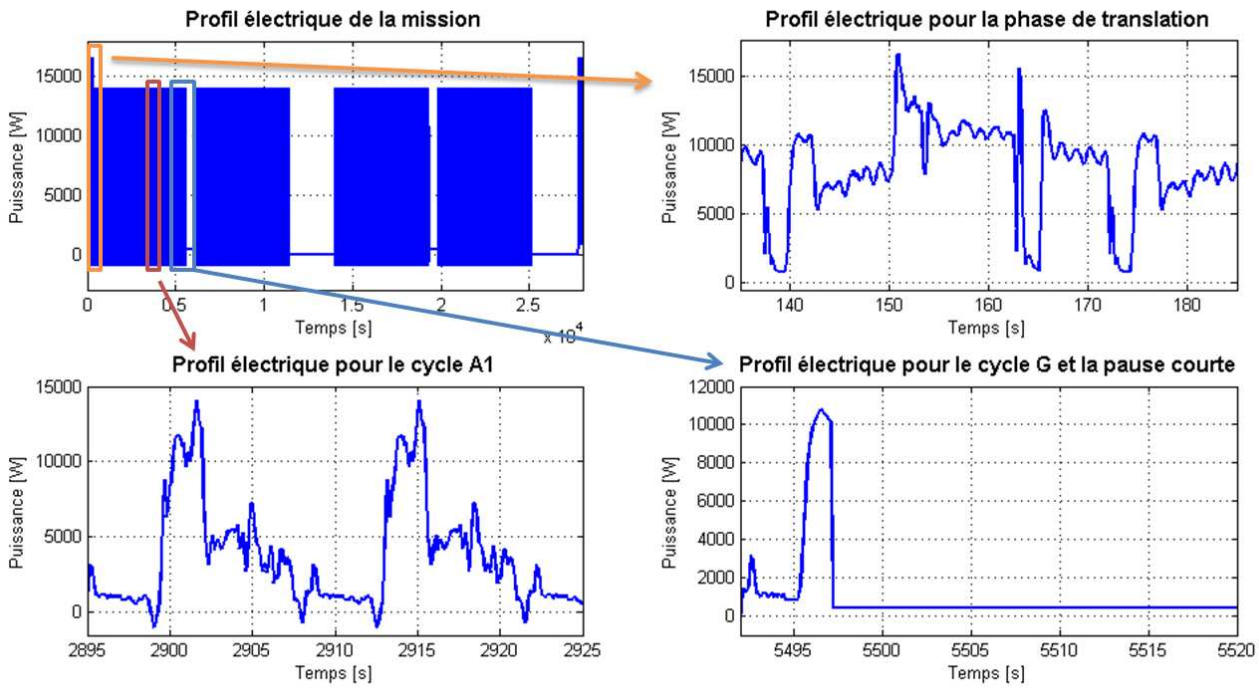


FIGURE C.1 – Profil de puissance électrique sur le bus de puissance HT pour un cycle de mission journalier

C.2 Zones de fonctionnement des actionneurs

Pour le cycle de mission journalier, les valeurs couple/vitesse des moteurs de chaque actionneur sont tracés pour vérifier leur dimensionnement vis-à-vis des performances demandées. La frontière en ligne forte représente la zone de fonctionnement continue du moteur électrique, les frontières en pointillées représentent les zones de fonctionnement intermittentes.

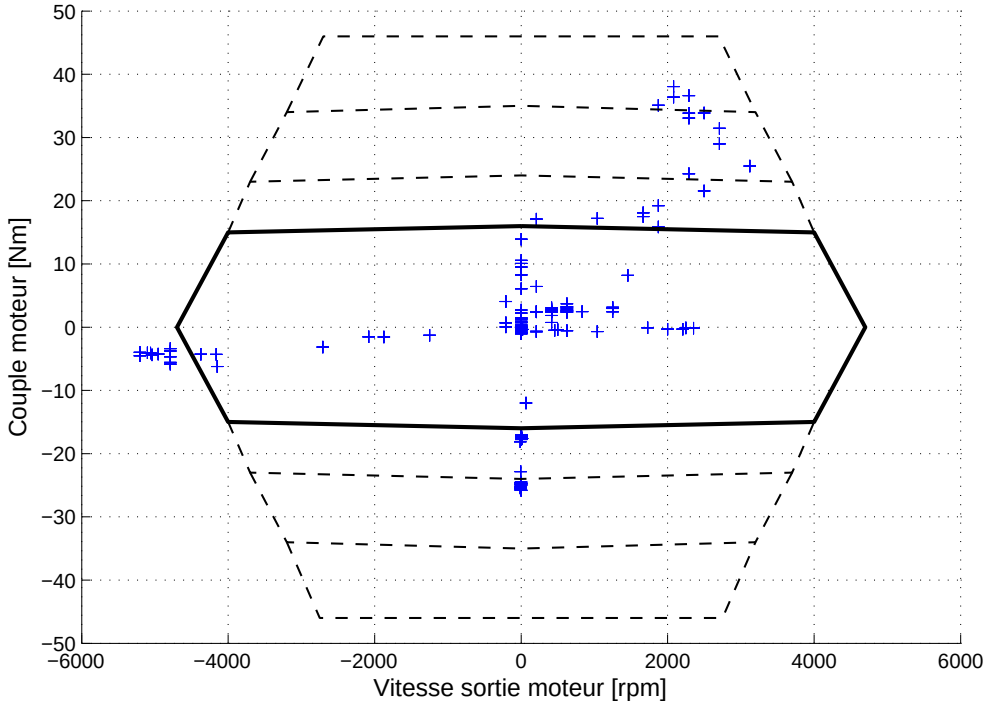


FIGURE C.2 – Zone de fonctionnement du moteur de l'actionneur de godet

L'actionneur de godet ne pourra pas atteindre la vitesse maximale demandée dans le troisième quadrant.

Les moteurs électriques des actionneurs de bras et de flèche ainsi que le moteur d'orientation de la tourelle sont correctement dimensionnés.

Pour le moteur de translation, il apparaît un manque de couple dans le second quadrant, lorsque le moteur se comporte en mode générateur. La trajectoire laisse supposer une phase de transition dynamique liée à la machine hydraulique.

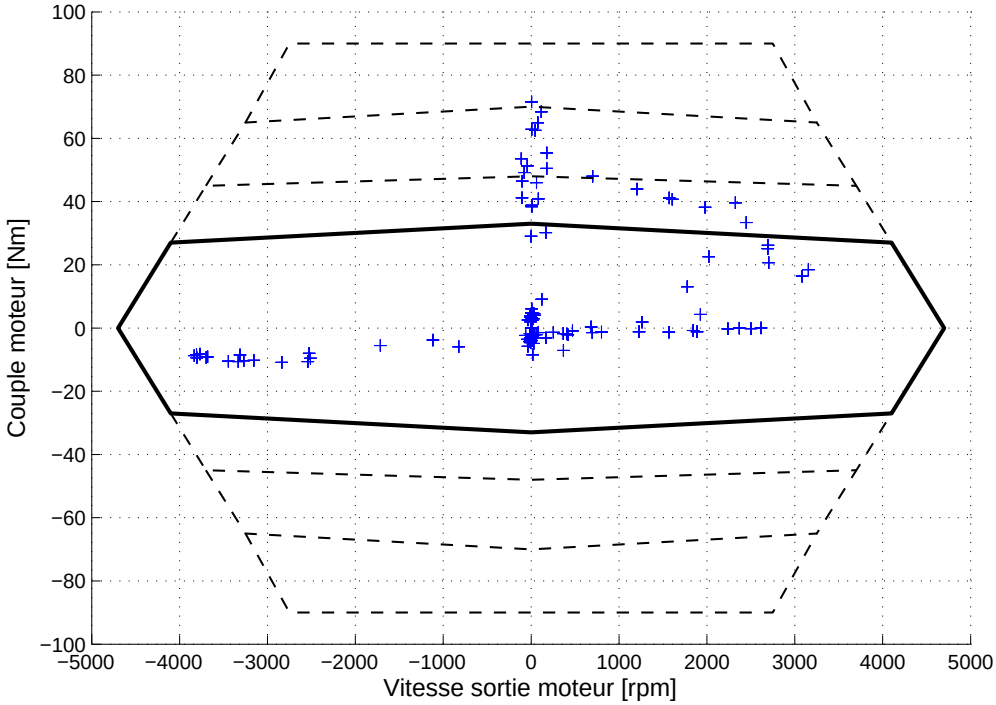


FIGURE C.3 – Zone de fonctionnement du moteur de l'actionneur de bras

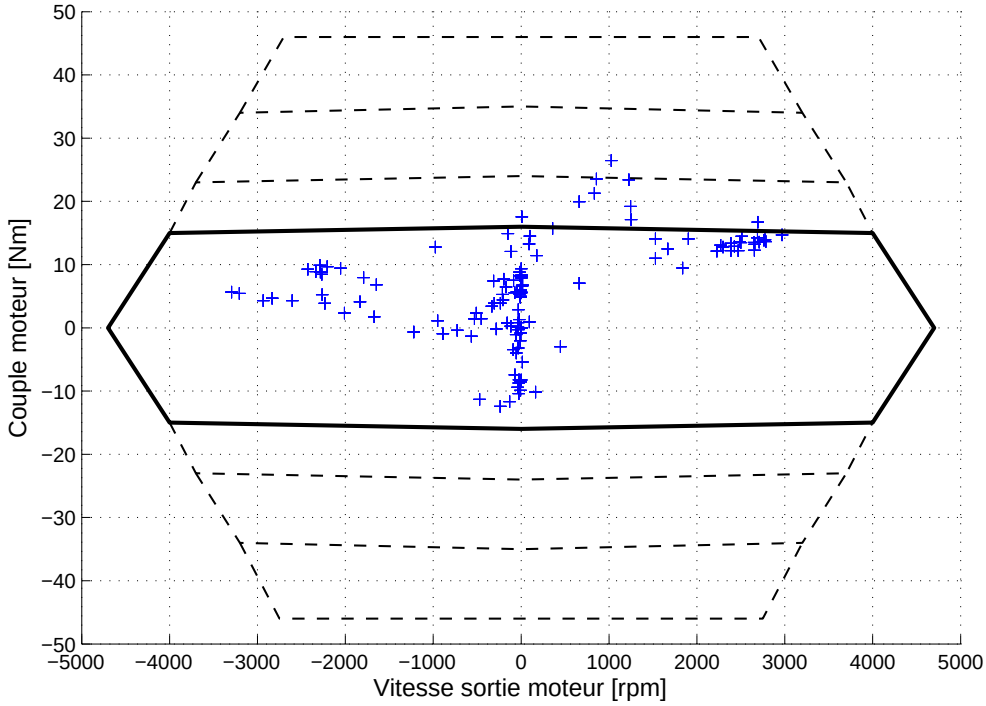


FIGURE C.4 – Zone de fonctionnement du moteur de l'actionneur de flèche

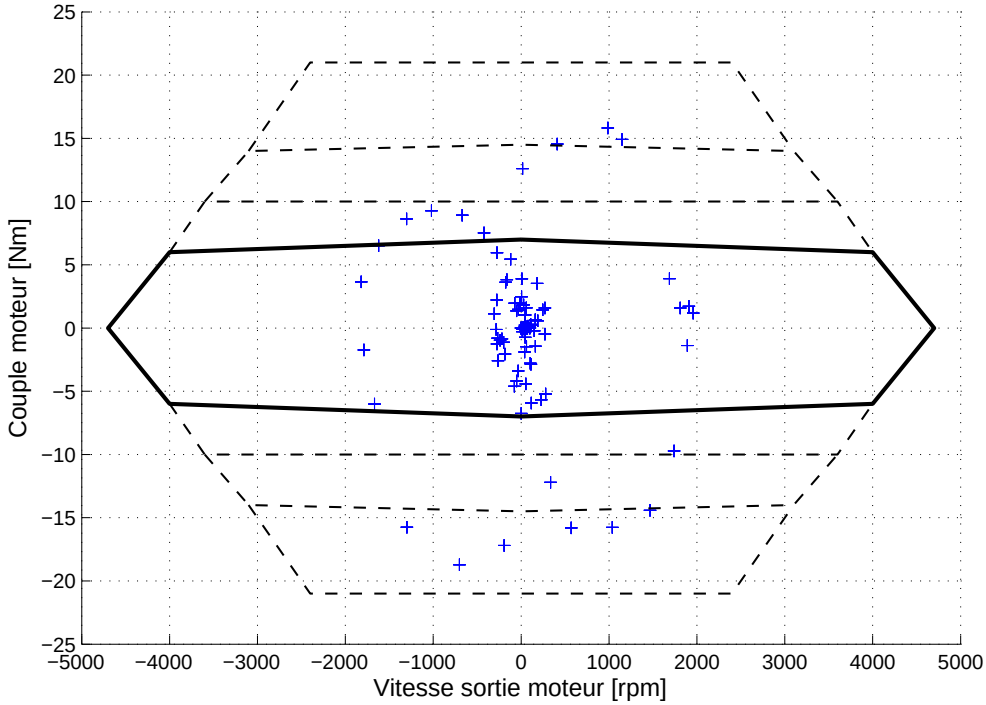


FIGURE C.5 – Zone de fonctionnement du moteur d'orientation de la tourelle

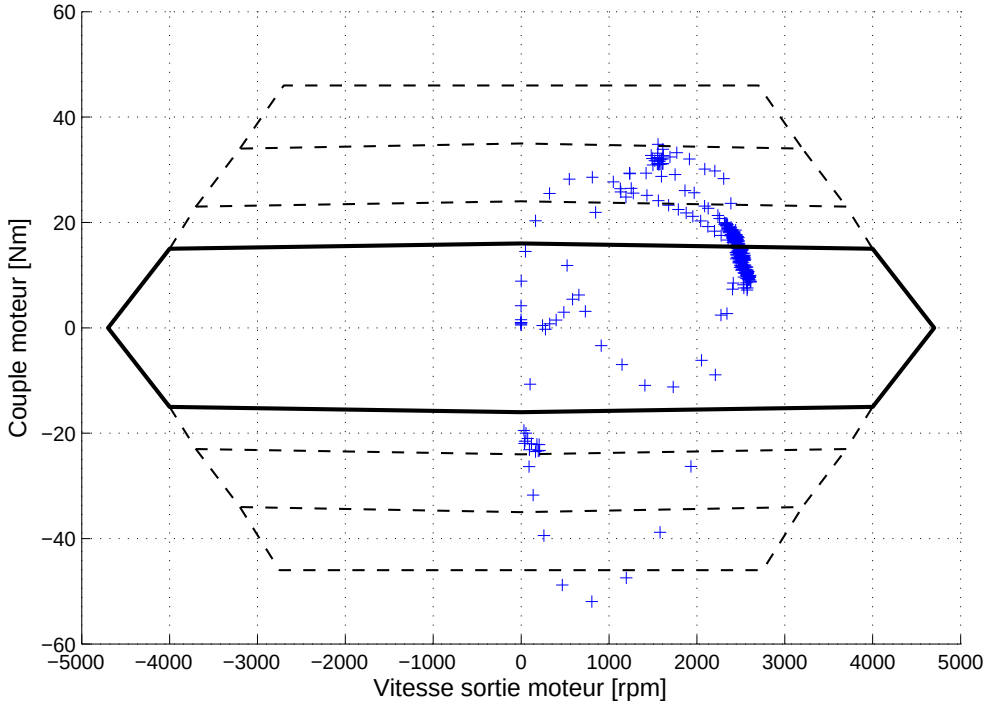


FIGURE C.6 – Zone de fonctionnement du moteur de translation

D
Résultats sur la gestion d'énergie hors-ligne

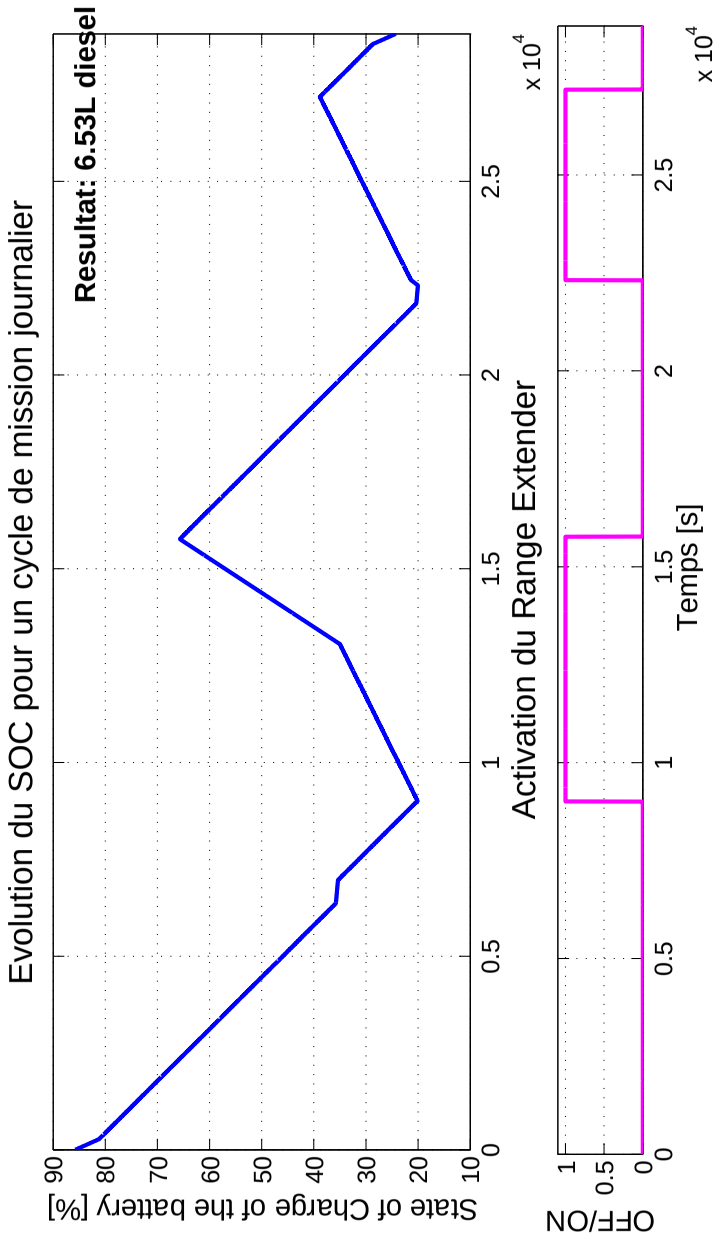


FIGURE D.1 – Résultats issus de la Programmation Dynamique

E Limitation de puissance de la mini-pelle hybride électrique

Comme indiqué dans le paragraphe B.5, il est nécessaire de limiter la puissance totale disponible aux actionneurs en phase de fonctionnement multi-actionneurs. La batterie étant dimensionnée au-delà de la puissance nécessaire, il faut réaliser une limitation par la consigne.

Pour garantir la sécurité du système pendant les opérations en charge, les actionneurs doivent disposer d'un couple moteur supérieur ou équivalent au couple de charge. Pour limiter la puissance à fournir aux actionneurs, on limitera la vitesse de déplacement de l'actionneur.

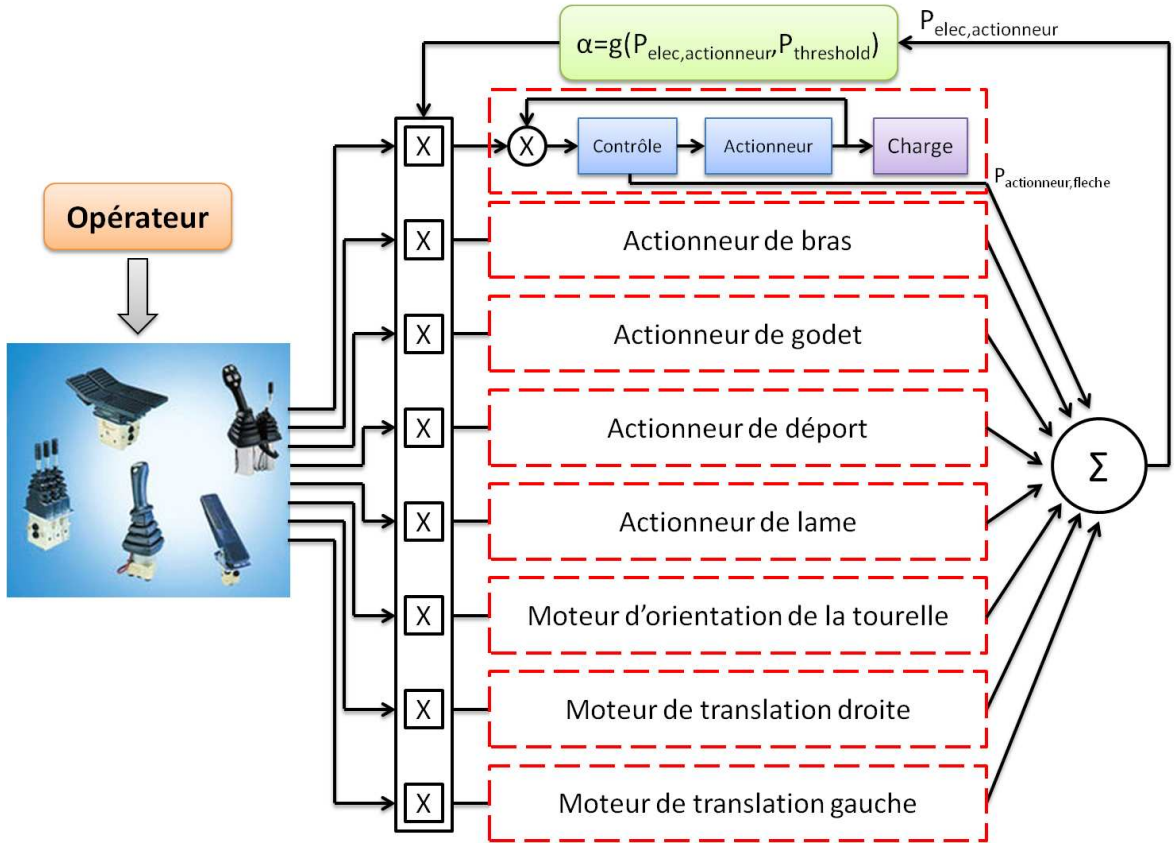


FIGURE E.1 – Diagramme de limitation de puissance des actionneurs

Dans une structure classique, la consigne de vitesse provenant des commandes (joysticks, pédalier) est directement envoyée au contrôleur de l'actionneur. Dans la structure proposée ici, si la puissance électrique globale des actionneurs atteint le seuil fixé, un paramètre noté α est calculé et envoyé sur la consigne de vitesse afin de réduire la consigne de référence. Cette opération multiplicative est réalisée sur tous les actionneurs. De ce fait, les actionneurs à l'arrêt et sous charge ne sont pas impactés par cette réduction de consigne de vitesse. Lorsque les actionneurs sont sous charge à vitesse nulle, la puissance électrique requise correspond aux pertes joules dans les bobinages et liée au courant nécessaire pour fournir un couple moteur opposé au couple de charge.

F Structure de commande de la mini-pelle hybride électrique

La structure de commande pour la gestion d'énergie de la mini-pelle hybride électrique avec la configuration Range Extender est présentée ci-dessous.

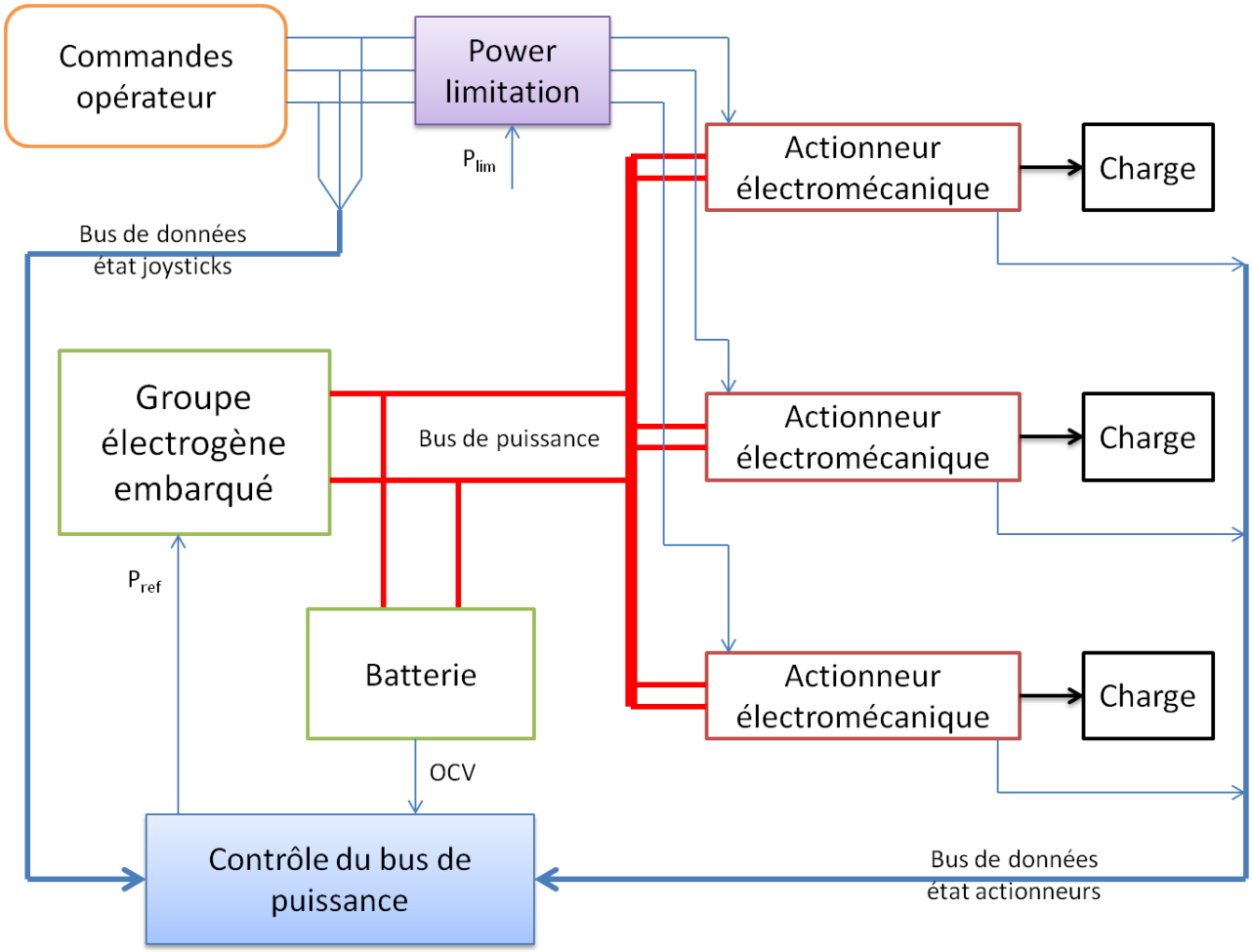


FIGURE F.1 – Diagramme de commande de la mini-pelle hybride électrique en configuration Range Extender thermique

Suivant les consignes envoyées par l'opérateur (commandes joysticks et modes de fonctionnement), l'état du réseau de puissance (batterie) et l'état des actionneurs, il s'agit de contrôler la puissance du groupe électrogène afin de répondre aux contraintes du cahier des charges. En temps réel, il existe des temps de latence liés au protocole de communication du bus de données. Dans les applications embarquées possédant une approche en multiplexage, le réseau CAN (Controller Area Network) est couramment utilisé.

Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

Mémoire doctorant 1ère année

2012 -2013

Nom - Prénom	**Herzig Nicolas**
Titre de la thèse	**Recherche et développement autour d'un nouveau Simulateur pour l'Apprentissage des Gestes de l'Accouchement**
Directeur de thèse	**Tanneguy Redarce**
Co- encadrants	**Richard Moreau**
Dpt. de rattachement	**Méthodes pour l'ingénierie des systèmes**
Date début des travaux	**01/02/2013**
Type de financement	**ANR**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

Résumé

Ce document fait le rapport de mes six premiers mois de travaux sur le développement et la recherche autour d'un nouveau simulateur pour l'apprentissage des gestes de l'accouchement. Ces travaux s'intègrent dans le cadre du projet SAGA financé par l'Agence Nationale de la Recherche. Mon travail gravite autour de trois axes, le premier consiste à augmenter la mobilité de la tête fœtale de la partie haptique du simulateur et, ainsi, modéliser la nouvelle dynamique de ce système. Dans un second temps, il est nécessaire d'implémenter de nouvelles lois de commande pour piloter le simulateur. À l'heure actuelle, les lois de commande envisagées sont des lois de contrôle en raideur. Enfin, le dernier axe concerne la mise en relation entre la partie physique du simulateur et les simulations numériques d'accouchement, développées par des laboratoires partenaires.

Abstract

This report summarize the first six months of my work on a childbirth simulator. This work, financed by the Agence Nationale de la Recherche, is a part of the SAGA project. My work can be divided in three parts. At first, a new version of the simulator, with more mobility of the simulator's foetal head, have to be developped. Then, I will deduce the dynamique model of the haptic interface. In the second part, I will implement new control law for the simulator. Currently, a stiffness control should be chosen. Finally, a link between the physical part and the numerical simulation have to be implement in order to converge the work of the different collaborators.

Table des matières

1. Introduction

1.1 Le contexte

De nos jours, les simulateurs jouent un rôle de plus en plus important dans notre quotidien. En effet, on trouve, aujourd'hui, des simulateurs dans différents domaines, notamment l'aéronautique, le nucléaire et le médical. Les simulateurs sont des outils qui permettent de reproduire de manière plus ou moins fidèle des situations, un environnement ou des phénomènes réels. Ils sont ainsi utilisés dans des cas où, ce qu'ils simulent est difficilement reproductible pour des raisons de coût, de risques ou de rareté d'occurrence.

Dans le domaine du médical, les simulateurs sont de plus en plus employés. En effet, ces simulateurs sont aussi bien utilisés dans le cadre de la formation initiale du personnel médical, que pour le maintien et la mise à niveau de leurs compétences au cours de leur carrière. Si les simulateurs, il y a quelques années, étaient pour la plupart anatomiques, avec un degré de réalisme plus ou moins important, on remarque, depuis peu un développement important des simulateurs virtuels ou instrumentés. Cet accroissement est notamment dû aux avancées techniques et technologiques de ces dernières années. C'est dans le but de proposer un nouveau simulateur plus fidèle et permettant de reproduire différentes situations que le projet Simulateur pour l'Apprentissage des Gestes de l'Accouchement (SAGA) a vu le jour.

1.2 Le projet SAGA

Le projet Simulateur pour l'Apprentissage des Gestes de l'Accouchement (SAGA), financé par l'Agence Nationale de la Recherche (ANR), a pour but de concevoir et développer un nouvel outil de formation pour les obstétriciens et les sages-femmes. Ce projet fait suite aux différents travaux réalisés par les partenaires du projet, notamment le projet BirthSIM qui sera présenté plus en détail par la suite (*cf.* 2.1.1). Ainsi, le projet SAGA a l'ambition de couvrir 3 axes d'amélioration dans le but de fournir un outil de qualité adapté à la formation du personnel médical. En effet, les différents partenaires cherchent à adapter au mieux les différents scénarios pris en compte par le simulateur pour proposer une formation avec des évaluations objectives sur les gestes et le comportement des utilisateurs. Les deux autres axes d'amélioration sont la qualité de modélisation par l'utilisation d'un modèle numérique pilote, et un retour physique pour l'utilisateur plus réaliste au travers d'une interface haptique améliorée.

Les différents partenaires intervenant sur le projet sont les suivants :

- L'équipe SAARA du LIRIS
- l'équipe GMCAO du Laboratoire TIMC-IMAG
- Le centre de Robotique (CAOR) d'ARMINES / MINES ParisTech
- L'équipe TFAP du LSE
- L'équipe ACM du Laboratoire Ampère
- La société Didhaptic
- L'association « Naissance et Connaissance »
- L'école de Sages-Femmes de Grenoble

Ces partenaires mettent, ainsi, leurs connaissances en commun, dans le but de proposer au final un simulateur pertinent.

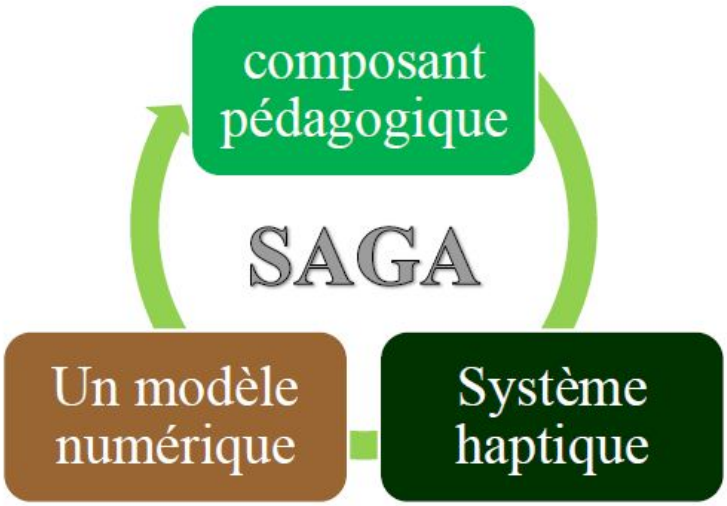


FIGURE 1.1 – Diagramme de présentation du projet SAGA

1.3 Ma participation au projet

Les recherches de mes prédécesseurs Ruimark Silveira, Olivier Dupuis, Osama Olaby, Richard Moreau et Romain Buttin ont abouti à la réalisation de diverses versions d'un simulateur d'accouchement baptisé BirthSIM. La dernière version de ce simulateur sera décrite avec plus de détails dans la partie 2.1.1. Actuellement, le BirthSIM est composé d'une interface haptique comprenant les mannequins anatomiques d'un bassin maternel et d'une tête fœtale. Cette dernière est actionnée par un vérin pneumatique. Le simulateur comprend aussi une interface graphique et une instrumentation permettant d'évaluer certains gestes de l'utilisateur, notamment, les manœuvres liée à l'extraction par forceps.

Mon travail consiste à réaliser une nouvelle version de la partie haptique. En effet, dans un premier temps, il est nécessaire, afin d'augmenter la fidélité du simulateur, d'augmenter le nombre de degrés de liberté pilotés de la tête fœtale. En effet, l'objectif est de passer d'un seul degré liberté piloté sur la version précédente à 4 mobilités commandées. La refonte de l'architecture physique du BirthSim impose une étude approfondie de la cinématique et de la dynamique. Cette étude a pour but à la fois de dimensionner les différents capteurs et actionneurs à intégrer, mais aussi de modéliser le système. Le second axe de mes travaux consiste à implémenter de nouvelles lois de commande permettant de retranscrire différents scénarios, mais aussi d'améliorer le rendu haptique de l'interface. Dans cette optique, il semble intéressant d'implémenter un contrôle en raideur combiné à l'utilisation d'actionneurs pneumatiques qui apportent une compliance naturelle. Enfin, la dernière partie consistera à mettre en place un couplage entre cette interface haptique et le modèle numérique développé par les partenaires du projet. Ainsi, la simulation permettrait de piloter l'interface et en contrepartie l'interface fournirait des informations telles que les efforts appliqués par l'utilisateur, afin que le modèle numérique puisse les prendre en compte.

2. État de l'art et cahier des charges du simulateur

2.1 État de l'art sur les simulateurs d'accouchement

Une étude récente[12] pour la Haute Autorité de la Santé fait l'état de l'art des simulateurs utilisés dans la santé, et conclut sur le fait que les simulateurs pour le médical sont de véritables outils permettant le « développement personnel continu ». Dans le domaine de l'obstétrique, les simulateurs sont utilisés depuis des années. En effet, Angélique du Coudray fit réaliser en 1759 un mannequin anatomique représentant le corps maternel, ainsi que le fœtus. Ce mannequin avait pour but à l'époque de transmettre son expérience sur les différents accouchements qu'elle avait pu rencontrer[22]. Des évolutions en terme d'anthropomorphisme ont été réalisées depuis, et à l'heure actuelle, on trouve, un nombre important de simulateurs d'accouchement différents sur le marché.

D'autre part, on peut classer ces simulateurs en différentes catégories. En effet, on distingue les simulateurs anatomiques, les simulateurs automatisés et les simulateurs virtuels. Les simulateurs anatomiques sont des simulateurs qui se concentrent sur la représentation physique de la parturiente ainsi que celle du fœtus. Ils sont en général purement passif, c'est-à-dire qu'ils ne sont ni actionnés, ni instrumentés. Ce sont pour la plupart des mannequins articulés, qui modélisent plus ou moins précisément l'anatomie humaine. Les simulateurs automatisés sont des simulateurs pilotés et instrumentés. Le fait que le simulateur soit instrumenté permet souvent de mesurer ou étudier les gestes et opérations réalisés par l'utilisateur. Enfin, les simulateurs virtuels sont des simulateurs où les anatomies de la parturiente et du fœtus sont modélisées virtuellement. Cela a, généralement, pour but de favoriser la compréhension des différentes interactions qu'il peut y avoir entre le fœtus et le bassin maternel. Ces simulateurs offrent la possibilité de voir ce qu'il se passe à l'intérieur du bassin, en revanche ils ne permettent pas au médecin de retrouver les repères anatomiques du fait qu'il ne soit pas possible de toucher. Par souci de synthèse, je ne présenterai que les simulateurs automatisés les plus avancés.

2.1.1 Les simulateurs

Noelle commercialisé par la société Gaumard

Le premier simulateur Noelle a été développé en 2000 par la société Gaumard. Ce simulateur est instrumenté et a la particularité de pousser relativement loin le réalisme. En effet, ce simulateur est composé d'un mannequin complet représentant la parturiente et de différents mannequins représentant des fœtus à différentes maturités. D'autre part, Noelle simule la plupart des comportements physiologiques de la parturiente et du fœtus, par exemple les contractions, la respiration, les battements de cœur, *etc.* Ce simulateur a été développé dans le but de servir de base de formation à différents cas possibles d'accouchement, il permet, ainsi de s'exercer aussi bien sur des gestes fondamentaux que sur des cas critiques d'accouchements difficiles. Ce simulateur est muni d'interfaces permettant de le piloter et permettant la visualisation en temps réel de certaines informations liées à l'accouchement [1]. En revanche, il ne dispose pas d'interface qui permette de voir, en temps réel, la position de la tête fœtale et d'éventuels instruments employés. Il est souvent compliqué de connaitre précisément les capteurs et actionneurs utilisés sur les modèles industriels, toutefois il semblerait que Noelle intègre différents actionneurs électriques permettant de reproduire la trajectoire de la tête fœtale et la rotation intra-pelvienne du fœtus. Enfin, aucun capteur ne semble mesurer les efforts appliqués par l'utilisateur sur la tête fœtale.

1. Site de Gaumard [en ligne] http://www.gaumard.com/ (consulté le 11/06/2013)

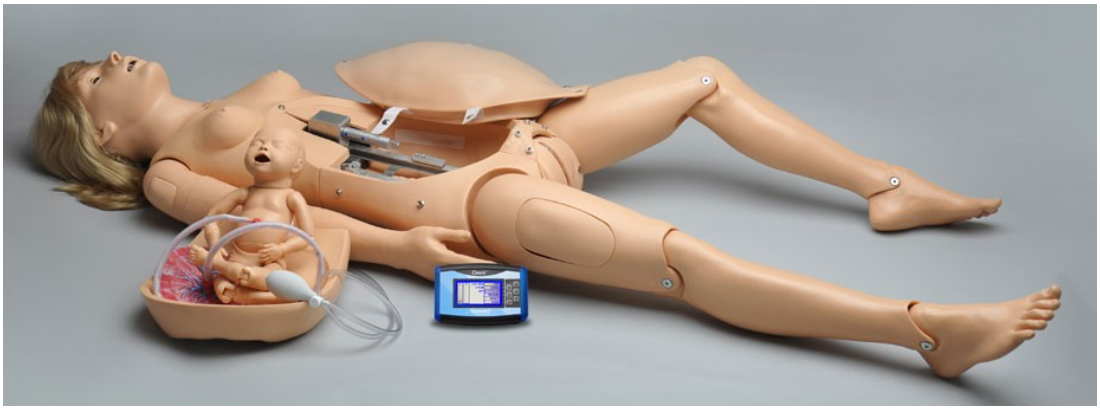


FIGURE 2.1 – Le simulateur Noelle

SimMom développé par Limbs & Things et Laerdal

SimMom est, au même titre que Noelle de Gaumard, un simulateur instrumenté relativement complet. En effet, il est constitué d'un mannequin de femme complet et articulé, ainsi que d'un mannequin de fœtus articulé aussi. Une interface permet de simuler et de représenter en temps réel les caractéristiques vitales de la parturiente, notamment les caractéristiques cardiaques et respiratoires. Ce simulateur permet aux équipes médicales de s'expérimenter à de nombreux cas d'accouchement [2]. Ainsi, il est possible de simuler une extraction instrumentale, une présentation du siège, une dystocie des épaules, *etc.* En revanche, les capteurs présents sur le simulateur ne permettent pas de mesurer les efforts exercés sur la tête fœtale et ce dernier ne simule pas non plus les efforts expulsifs. Ainsi, lors d'une formation ou d'un exercice, le formateur ne peut qu'évaluer les gestes de l'opérateur en fonction de ce qu'il voit, il ne dispose pas d'élément lui permettant une analyse plus approfondie et réellement quantitative.

FIGURE 2.2 – Le simulateur SimMom

Le simulateur SIMone développé par 3B Scientific

Le simulateur SIMone est un simulateur instrumenté relativement complet. En effet, il est constitué d'une partie anatomique, représentant le bassin maternel et d'une tête fœtale dont le déplacement est piloté. Il dispose aussi d'une interface permettant la visualisation de différentes informations simulées sur l'accouchement en cours, ainsi que la position de la tête fœtale dans le bassin maternel. Un

2. Site de Laerdal [en ligne] http://www.laerdal.com/fr/ (consulté le 11/06/2013)

capteur d'effort, présent au niveau de la nuque du fœtus, mesure l'effort appliqué sur la tête. Ce capteur permet aussi de mesurer l'orientation de l'effort appliqué, et cette orientation est affichée dans l'interface graphique du simulateur [3]. Les différences marquantes avec le simulateur BirthSIM sont notamment la prise en compte des différentes opérations médicales réalisées (la médicamentation par exemple) et la simulation sonore de la douleur et de la respiration de la parturiente. En revanche, il n'est pas possible avec ce simulateur de visualiser la position des instruments à l'intérieur du bassin.

FIGURE 2.3 – Le simulateur SIMone

Le simulateur LM-095 commercialisé par Koken MPC

Le simulateur LM-095 ou Virtual Reality Vaginal Exam Model est un simulateur instrumenté, composé d'un mannequin de bassin maternel, muni de différents accessoires dans le but de représenter des dilatations de col et des engagements de la tête fœtale différents. À ce mannequin s'ajoute une interface permettant de visualiser l'intérieur du bassin en 3 dimensions, ainsi que la position des doigts de l'opérateur. En effet, un jeu de capteurs est fourni afin d'instrumenter les doigts de l'opérateur [4]. Ces capteurs sont des capteurs de position électromagnétiques Trackstar commercialisés par Ascension Technology Corporation [5].

Le simulateur développé à l'Université Johns Hopkins, USA

Ce simulateur instrumenté a été développé dans le but de mesurer les différents efforts mis en jeu sur la tête fœtale lors d'une présentation avec dystocie des épaules. En effet, l'objectif était de comparer 3 manœuvres différentes d'extraction de dystocie des épaules, à savoir la manœuvre de Mac Roberts, celle de Rubin antérieure et celle de Rubin postérieure[2]. Ce mannequin est composé d'un mannequin représentant les membres inférieurs de la parturiente (bassin et jambes) et d'un mannequin fœtal instrumenté. Les capteurs utilisés sont une jauge de contrainte permettant de mesurer l'effort

3. Site de 3B Scientific [en ligne] http://www.3bscientific.fr/ (consulté le 11/06/2013)
4. Site de Koken MPC [en line] http://www.kokenmpc.co.jp/english/ (consulté le 11/06/2013)
5. Site d'Ascension Technology [en line] http://www.ascension-tech.com/ (consulté le 11/06/2013)

FIGURE 2.4 – Le simulateur LM-095

appliqué sur la tête fœtale, un potentiomètre rotatif pour la mesure de la rotation de la tête et enfin un potentiomètre linéaire qui mesure l'élongation du plexus brachial. Les efforts de traction exercée sont dans les trois manœuvres, de l'ordre de 100N[13][14]. Les opérations ont aussi été réalisées avec des gants instrumentés, toujours pour valider les efforts exercés sur la tête fœtale. Deux versions de gants ont été conçues, les premiers avec des capteurs piezoresistifs, les seconds avec des capteurs piézoélectriques. Il est à noter que le mannequin maternel simule les contractions utérines qui expulsent le mannequin fœtal. En effet, un compresseur pneumatique est utilisé pour la mise sous pression de ballons, la variation de pression dans ces ballons modélise les contractions utérines et permettent au mannequin fœtal d'avancer.

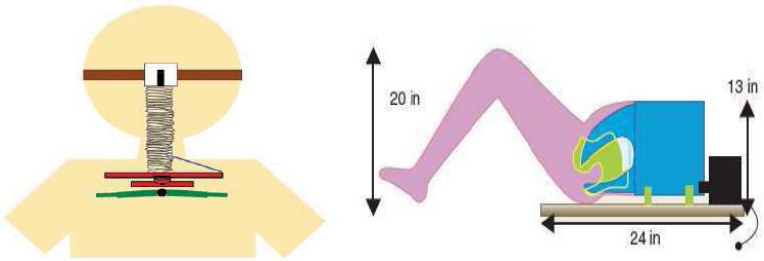


FIGURE 2.5 – Le simulateur développé à l'Université Johns Hopkins

Le simulateur développé à School of Computing Sciences, UK

Ce simulateur est un compromis entre un simulateur instrumenté et un simulateur virtuel. En effet, ce simulateur est composé d'un mannequin d'un bassin osseux uniquement. La représentation du corps fœtal est obtenue par incrustation en réalité augmentée[16]. Il est possible d'interagir avec cette interface de réalité augmentée à l'aide de forceps équipés de traqueurs optiques. Ces forceps permettent ainsi de déplacer le modèle de fœtus dans le bassin osseux. En revanche, ce simulateur ne gère ni la déformation de la tête fœtale due à l'extraction par forceps, ni le retour haptique des efforts mis en jeu lors d'une telle extraction. En parallèle du simulateur, une modélisation de la tête fœtale par éléments finis a été réalisée. Cette modélisation a pour but d'affiner le simulateur pour l'avenir[15].

Le simulateur breveté par Riener et *al.* (US 7241145 B2)

Ce simulateur est constitué d'un mannequin de bassin maternel et d'un mannequin fœtal (ce mannequin peut être réduit à la tête seulement). Ce mannequin fœtal peut être piloté par diverses chaines cinématiques, mais dans la littérature, on ne trouve que des publications d'essais effectués à l'aide d'un robot 6 axes. Une interface permet la visualisation de la position de la tête dans le bassin maternel. Dans le brevet, il est aussi proposé d'instrumenter la tête fœtale à l'aide de capteurs de force et de couple situés au niveau de la nuque, et de capteurs de pression, afin de mesurer les efforts de contact exercés par les forceps lors d'une extraction instrumentale[24][21].

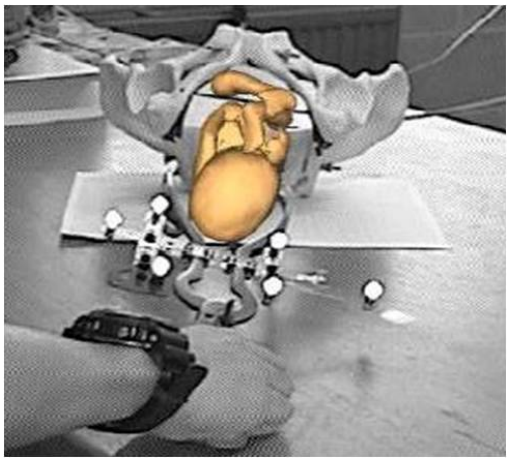


FIGURE 2.6 – Le simulateur développé à School of Computing Sciences

Une seconde version du simulateur a été réalisée (mais elle n'apparait pas dans le brevet), en intégrant une interface de réalité augmentée. Ainsi Sielhorst et *al.* ont conçu cette interface constituée d'un casque avec des lunettes qui permet de déterminer les mouvements de la tête de l'utilisateur et des traqueurs disposés sur les forceps. La technologie utilisée pour le tracking des différents mouvements est exclusivement optique[28][27]. La position de la tête fœtale est quant à elle obtenue grâce aux différents capteurs intégrés au robot 6 axes. Cette évolution a permis d'étudier des méthodes d'analyse des gestes d'utilisation de forceps en terme de trajectoire. Les deux méthodes de comparaisons de trajectoire employées sont les méthodes de Longest Common Subsequence et de Dynamic Time Warping[26]. Sielhorst et *al.* concluent par la remarque suivante, il serait plus pertinent d'étudier le déplacement de 3 points non colinéaires, plutôt que la position et l'orientation d'un repère fixe de l'instrument.

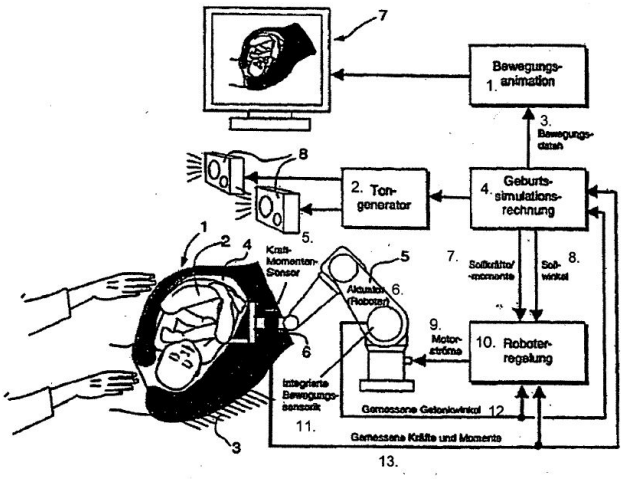


FIGURE 2.7 – Le simulateur breveté par Riener et *al.*

Le simulateur BirthSIM développé par le laboratoire Ampère

Le simulateur BirthSIM est un simulateur instrumenté qui a pour but de former les obstétriciens et sages femmes à l'extraction instrumentale lors d'un accouchement. Il est, notamment, optimisé pour l'utilisation des forceps. En effet, ce simulateur est composé d'un mannequin de bassin maternel, ainsi qu'un mannequin de tête fœtale, cette dernière étant pilotée et instrumentée. À l'aide de forceps instrumentés, il est possible de visualiser, en temps réel et via une interface graphique, la position de la tête fœtale et des deux cuillères composant le forceps. Ce simulateur est paramétrable et propose, ainsi, différentes situations. L'interface permet aussi l'apprentissage des gestes liés à l'utilisation de

forceps et permet de quantifier la qualité d'exécution des opérations de l'utilisateur[19]. Un brevet a été déposé sur ce simulateur par Dupuis et *al.* en 2005 (FR 2858453 A1)[11].

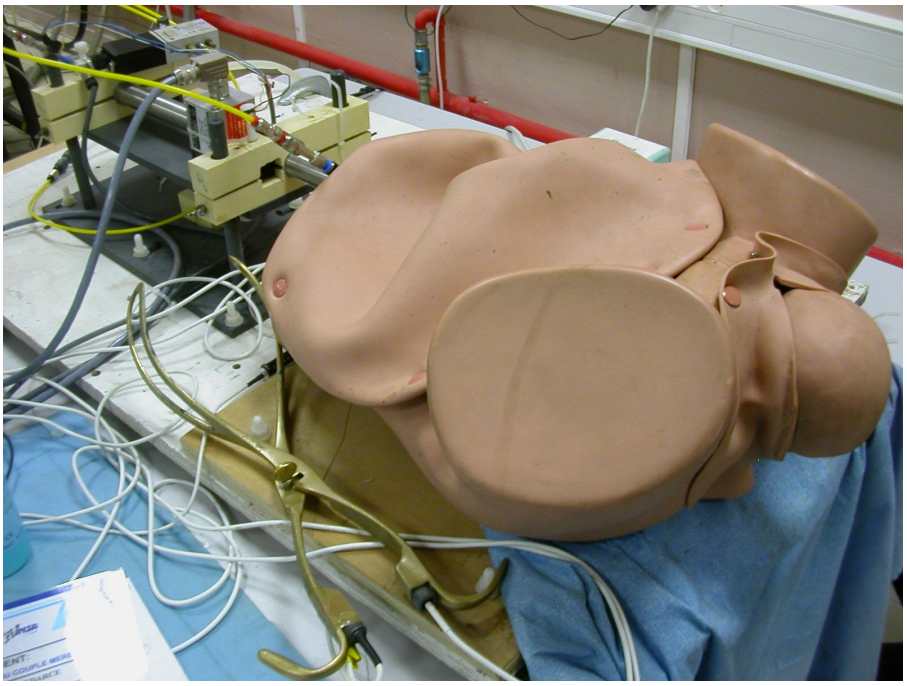


FIGURE 2.8 – Le simulateur BirthSIM

2.1.2 Comparaison des différents simulateurs

On peut comparer ces différents simulateurs sur plusieurs critères, notamment les types de présentations possibles, la mobilité de la tête fœtale, la présence d'interface de visualisation, *etc.* Un tableau comparatif de ces simulateurs, ainsi que la description des différents critères est disponible en Annexe B. Pour résumer, on remarque que s'il y a de nombreux simulateurs, peu d'entre eux offrent la possibilité de combiner à la fois une interface de visualisation des instruments et une interface haptique. D'autre part, aucun ne semble avoir réussi à prendre en compte la déformation des corps mous du bassin. Enfin, si certains simulateurs permettent la mobilité de la tête fœtale en rotation, dans la majorité des cas, elle n'est ni pilotée ni instrumentée.

2.2 Cahier des charges

2.2.1 Analyse fonctionnelle

Fonction principale

Il est important de rappeler que l'objectif est de fournir au corps médical un outil fonctionnel et adapté à la formation sur les gestes de l'accouchement. Ainsi, le simulateur du projet SAGA doit faire le lien entre l'utilisateur et le savoir, les connaissances et les compétences dans le domaine de l'obstétrique. La fonction principale qui en découle est par conséquent la suivante : **Assurer l'acquisition d'un savoir, de connaissances ou de compétences à l'utilisateur**.

Comme il a été précisé précédemment, le projet SAGA poursuit 3 axes. On peut décomposer, dans un premier temps, cette étude selon ces 3 axes.

Les outils pédagogiques

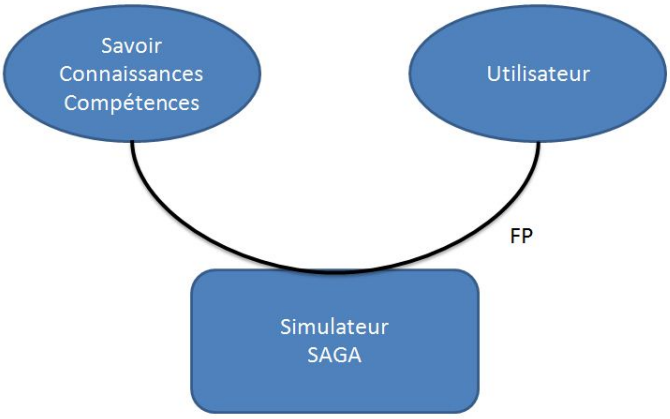


FIGURE 2.9 – Bête à cornes

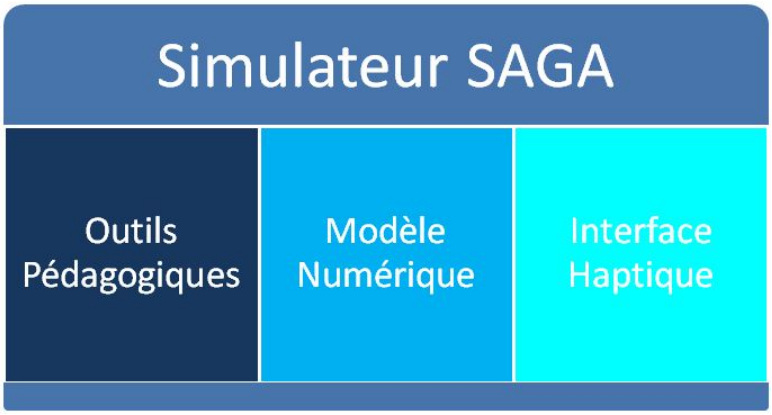


FIGURE 2.10 – 3 axes d'approfondissement

Les outils pédagogiques intégrés au projet SAGA ont pour but de rendre efficiente l'utilisation du simulateur. En effet, ces outils amènent une analyse sur le transfert de connaissances et permettent d'adapter au mieux le simulateur aux besoins d'apprentissage des différents corps médicaux.

Cet axe d'approfondissement intervient par exemple lors du choix et de la rédaction des différents scénarios. En effet, ces scénarios incorporés au simulateur devront être pertinents et valoriser la formation des différents utilisateurs. Le simulateur devra aussi proposer une méthode d'évaluation des gestes et du comportement des utilisateurs permettant d'estimer leur progression de manière qualitative et quantitative.

Le modèle numérique

Le modèle numérique joue un rôle primordial pour le simulateur. En effet, dans le cadre du projet, le modèle numérique est le pilote du système haptique. Ainsi, il devra être réaliste et devra fournir des informations suffisamment rapidement pour qu'elles puissent être exploitables par la partie physique du simulateur.

L'avancée technologique due à l'utilisation de ce modèle numérique réside dans le fait que contrairement aux autres simulateurs présents dans le commerce, la trajectoire n'est pas imposée selon un scénario, mais est calculée en temps réel en fonction des informations fournies par les capteurs. D'autre part, le modèle prend aussi en compte les déformations des tissus afin de déterminer leur influence durant l'accouchement. Cette simulation permettra aussi de détecter une éventuelle déchirure des tissus.

Le système haptique

Le système haptique est la partie physique du simulateur. En effet, il sert d'interface entre l'utilisateur et le modèle numérique. Dans le projet SAGA, la partie haptique devra retranscrire finement les sensations que peut ressentir un obstétricien ou une sage-femme lors de la réalisation de différents gestes.

C'est au travers de cette interface qu'il sera possible de mesurer, à l'aide d'une métrique adaptée, les différentes opérations de l'utilisateur. Ainsi, les efforts appliqués par l'opérateur et les trajectoires des différents instruments semblent être des paramètres pertinents dans l'évaluation des gestes.

Enfin, pour que le simulateur soit fidèle aux phénomènes qu'il modélise, il est nécessaire que le système haptique retranscrive un maximum de cas possibles de présentations. Toutefois, même si le simulateur du projet se veut à la fois fidèle et complet en termes de scénarios, il est, malheureusement, impossible d'être exhaustif et de proposer l'ensemble des situations pouvant se produire lors d'un accouchement. Par conséquent, des concessions doivent être faites de manière à simplifier la conception, mais aussi l'utilisation du simulateur.

2.2.2 Vers un nouveau simulateur

Augmentation de la mobilité de la tête

Comme on peut le constater dans les comparaisons précédentes (*cf.* 2.1.2), il semble primordial d'augmenter les mobilités de la tête fœtale du système haptique. En effet, dans l'objectif de retranscrire au mieux le comportement et les sensations d'un véritable accouchement, les rotations de la tête doivent être intégrées. De plus, pour un ressenti plus juste, il est nécessaire que certaines mobilités soient pilotées afin que le système haptique puisse fournir un retour d'effort. Ainsi, après discussion avec Olivier Dupuis, un obstétricien partenaire du projet, nous sommes arrivés à un classement par ordre de priorité des mobilités à implémenter. Voici ce classement :

1. trajectoire de la tête fœtale
2. orientations de présentations (OP, OS, OIGA, OIDA, OIGP, *etc.*)
3. flexion
4. asynclitisme

Il est à noter que s'il semble que les trois premières doivent être pilotées, la dernière pourrait rester passive à condition qu'un indexage initial soit effectué, afin de garantir une certaine reproductibilité. Ces mobilités seront détaillées dans la partie 3.1.1 et des figures illustrant ces mobilités sont présentées en Annexe C.

Interface de visualisation de la tête et du bassin mou

L'utilisation d'un modèle numérique permet d'obtenir des données difficiles à mesurer, notamment les différentes déformations des corps mous. Ainsi, l'implantation d'une interface permettant de visualiser les déformations de la tête fœtale ou du plancher pelvien permettrait à l'utilisateur de réévaluer son geste et de prendre conscience des efforts mis en jeu durant l'accouchement.

Augmentation du nombre de scénarios proposés

L'augmentation du nombre de degrés de liberté (ddl) permet de prendre en charge de nouveaux scénarios. Il sera nécessaire à la fois de les expliciter, mais aussi de les paramétrer de manière à ce qu'ils soient jouables par le simulateur. En effet, il faudra, par exemple, rédiger les scénarios en relation avec les différentes variétés de présentations de la tête fœtale.

Augmentation du degré d'anthropomorphisme

L'augmentation du degré d'anthropomorphisme consiste à rendre plus réaliste anatomiquement le simulateur. Divers axes d'amélioration ont été suggérés, parmi ces derniers certains semblent intégrables facilement. Par exemple, l'ajout des muscles pelviens qui permettrait de rendre plus réalistes

les procédures de toucher vaginal, ou l'ajout des épaules qui permettrait d'effectuer le diagnostic des épaules en début d'accouchement. D'autres améliorations, en revanche, semblent plus complexes à implémenter. On peut donner l'exemple du remplacement des matériaux représentant la peau pour une sensation de toucher plus réaliste. Ou alors, le fait de piloter les plaques osseuses du crâne fœtal pour simuler un éventuel chevauchement.

Couplage numérique/haptique

Le couplage entre partie haptique et modèle numérique est primordial. En effet, ces deux parties sont développées par des partenaires différents, mais la bonne réussite du projet dépend de la mise en relation de ces dernières. Si ce couplage ne pourra être réalisé, physiquement, qu'à la fin, il est toutefois nécessaire de prendre en compte ce couplage au fur et à mesure de l'avancement des travaux. C'est ainsi que les informations échangées entre la partie haptique et la simulation numérique devront, par exemple, être définies dès le début.

2.3 Choix technologiques

2.3.1 Mannequin Anatomique

Afin de retranscrire l'anatomie de la parturiente et du fœtus, le simulateur intègre un mannequin anatomique que l'on trouve dans le commerce. L'annexe D présente des images et un tableau qui permet de comparer, en termes de qualité, coût et délai, différents modèles présents sur le marché.

Durant une réunion avec les différents partenaires, nous avons fait en sorte de pouvoir tester ces différents mannequins. Il semblerait, d'après l'obstétricien présent, que le meilleur choix en termes de qualité serait de combiner le bassin maternel de Prompt et la tête fœtale de l'Obstetrical Mannikin. En effet, les muscles pelviens présents sur le mannequin de Laerdal apportent le rendu le plus proche de la réalité d'un point de vue anatomique. D'autre part, la tête du mannequin fœtal de Simulaids est l'une des rares sur laquelle il est possible d'appliquer une ventouse, mais en plus le choix des matériaux utilisés apporte une certaine rigidité qui, au toucher, est pertinente.

2.3.2 Les capteurs de positions

Dans le simulateur, différentes informations de positions vont être nécessaires. En effet, dans un premier temps, il est nécessaire de connaitre la position de la tête fœtale. Cette information est à la fois utile pour les utilisateurs du simulateur, mais aussi pour le système dans le cadre d'un asservissement en position. Dans un deuxième temps, il est aussi nécessaire de connaitre la position des instruments ou des mains de l'utilisateur. Ces informations sont utiles dans le cadre de l'analyse du geste et l'interface de visualisation de la position des instruments.

Pour caractériser la position d'un objet dans l'espace, 6 informations sont nécessaires. En effet, les 3 premières sont les projections sur une base orthonormée de la distance entre un point de cet objet et un point d'origine, les 3 suivantes correspondent aux 3 angles de rotation du repère lié à l'objet par rapport au repère de base. On parle, ainsi, de 6 degrés de liberté (DDL). Pour obtenir ces informations de positions, il existe différentes technologies. Les capteurs proposés sur le marché sont les suivants :

- Les capteurs mécaniques avec contact
- Les capteurs mécaniques sans contact
- Les capteurs magnétiques
- Les capteurs à ultrason
- Les capteurs optiques

Le tableau 2.1 fait un comparatif de ces différentes technologies.

Pour l'application sur le simulateur, les technologies les plus adaptées semblent être les capteurs magnétiques pour la mesure de position des instruments ou mains de l'utilisateur. En effet, la compacité de ces capteurs rend leur utilisation moins invasive et déstabilisante pour l'opérateur. On pourra remarquer que c'est le choix qui a été fait pour le simulateur LM-095, alors que pour le simulateur

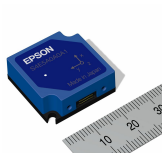
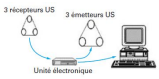
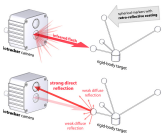
Technologie	Mécanique avec contact	Mécanique sans contact	Magnétique	Ultrason	Optique
Précision	++	++	++	−	+
Coût†	−	+	−	++	=
Rapidité	++	++	−	=	+
Remarques	Possibilité de retour d'efforts (câbles) Peu ergonomique	Dérive Encombrant	Sensible à un environnement métallique	Possibilité d'occultation Sensible à l'environnement (Température)	Possibilité d'occultation Nécessite une puissance de calcul importante
Image					

TABLE 2.1 – Comparatif des technologies de capteurs de position

† : Pour faciliter la lecture du tableau, les « + » correspondent à un coût faible, ce qui est un avantage. Les « − » signifie donc un coût élevé.

breveté par Riener les capteurs choisis sont des capteurs optiques (*cf.* 2.1.1). Pour la mesure de la position de la tête, le plus adapté semblerait être plusieurs capteurs mécaniques avec contact. En effet, l'analyse de la cinématique en partie 3.1 nous permet de déterminer la position et l'orientation de la tête fœtale en fonction de la position de chaque actionneur.

2.3.3 Les capteurs d'efforts et de pression de contact

Nous avons vu précédemment qu'il serait intéressant de pouvoir mesurer les efforts appliqués par l'opérateur sur la tête fœtale. On peut distinguer deux types d'efforts, les efforts résultants appliqués à la tête, c'est-à-dire les efforts qui vont avoir tendance à déplacer la tête, et les efforts de contact entre les instruments ou les mains de l'opérateur et la tête. Ces derniers auront plutôt pour impact de déformer la tête.

Deux types de capteurs semblent être adaptés pour ce type de mesure, les capteurs piezorésistifs et les capteurs piézoélectriques. Les piezorésistifs sont des éléments dont la résistance électrique varie selon l'effort qu'on leur applique. Les piézoélectriques sont des matériaux pour lesquels une tension électrique est générée lorsqu'on les déforme. Les capteurs utilisant des piezorésistifs semblent être plus adaptés au simulateur, de par la dynamique lente du système. En effet, les piézoélectriques sont précis, mais ne fonctionnent pas dans le cas d'une utilisation en statique.

Dans le cas de notre simulateur, il reste un choix à faire pour la mesure de pression de contact. Ce choix consiste à définir s'il faut instrumenter la tête ou les instruments. Dans le cas de l'instrumentation de la tête, l'avantage est dû au fait qu'il soit possible de mesurer ces efforts, quels que soient les instruments utilisés. En effet, il existe différents types d'instruments utilisés pour l'extraction lors de l'accouchement, notamment les mains, les forceps de Tarnier, de Suzor ou de Thierry et les ventouses. En revanche, l'augmentation du nombre de capteurs, présents dans ou autour de la tête, peut poser

un problème d'encombrement. En contrepartie, la pose de capteurs sur les instruments peut gêner l'utilisateur et impose de rééquiper chaque nouvel instrument utilisé. On trouve dans la littérature un capteur qui est souvent utilisé pour la mesure de pression de contact. Ce capteur est le capteur Flexiforce de Tekscan. Ce capteur est aussi bien utilisé sur des instruments, par exemple une pince pour chirurgie micro-invasive[7], que sur des simulateurs de corps mous comme un simulateur pour le diagnostic mammaire[29].

Pour la mesure des efforts résultants, deux solutions sont envisageables. La première consiste à placer un capteur d'effort multi-axes utilisant des éléments piezoresistifs. La seconde consiste à mesurer les efforts transmis aux actionneurs. En effet, il est possible d'estimer les efforts appliqués à la tête à partir d'informations relevées au niveau des actionneurs. Pour cela, il est nécessaire de connaitre le modèle dynamique du système et les efforts reçus sur chaque actionneur. La mesure des efforts transmis aux actionneurs correspond à une mesure de différence de pression dans les chambres pour les actionneurs pneumatiques et une mesure d'intensité pour les actionneurs électriques.

2.3.4 Les actionneurs

Les travaux de mes prédécesseurs ont montré l'avantage de l'utilisation d'actionneur pneumatique pour la réalisation du simulateur. En effet, les actionneurs pneumatiques possèdent une compliance naturelle, ce qui permet de retranscrire plus facilement un rendu haptique mou. Ainsi je souhaite conserver un actionnement pneumatique linéaire pour le déplacement de la tête fœtale selon l'axe longitudinal. La cinématique du simulateur sera décrite plus en détail dans la partie 3.1, toutefois, elle sera constituée de 2 actionneurs linéaires et 2 actionneurs rotatifs.

Les actionneurs linéaires utilisés seront pneumatiques afin d'optimiser le rendu haptique mou. En revanche pour les actionneurs rotatifs deux possibilités s'offrent à nous, utiliser des vérins pneumatiques ou utiliser des moteurs à courant continu. Ces deux technologies sont souvent utilisées en robotique[6] et ont l'avantage d'être relativement compactes. En effet, les actionneurs rotatifs seront placés en bout de bras ce qui impose l'utilisation d'actionneurs peu encombrants.

Les deux autres contraintes qui permettent de dimensionner ces actionneurs sont les efforts fournis et les courses. Pour ce qui est des efforts fournis, les actionneurs devront permettre de bloquer les mouvements effectués par l'opérateur. L'objectif étant de simuler un maximum de cas de présentations ; dans certains de ces cas, l'extraction instrumentale est impossible et le praticien doit savoir renoncer et choisir d'effectuer une césarienne. Les efforts maximaux admissibles, pour une extraction instrumentale, donnés dans la littérature sont de l'ordre de 200 N[20][17] en traction et de 7 N.m[2] pour la torsion. Ces efforts peuvent être considérés comme importants pour de petits actionneurs. Pour ce qui est de la course, pour pouvoir simuler l'ensemble des orientations de présentation l'actionneur tournant autour de l'axe longitudinal doit avoir une course de 360°, et celui modélisant la flexion doit avoir une course de l'ordre de 270°.

Les actionneurs pneumatiques ont pour avantages de pouvoir fournir des efforts importants et possèdent aussi une compliance naturelle ce qui simplifie la démarche pour offrir un retour haptique fidèle. En contrepartie, ils sont relativement encombrants et ont des courses relativement limitées. En effet, pour les vérins rotatifs à palette (les plus compacts), on dépasse rarement les 270° de course, ce qui est insuffisant pour couvrir l'ensemble des orientations. De plus, pour les dimensions souhaitées, les vérins rotatifs ne peuvent pas fournir l'effort souhaité. Or un étage de réduction ne permet pas d'augmenter à la fois la course et les efforts. Par conséquent, si les vérins rotatifs sont choisis, il faudra faire un compromis entre diversité des orientations proposées et efforts reproduits.

Les moteurs électriques en revanche n'ont pas de limite de course, mais développent un couple relativement faible. Ainsi, un étage de réduction sera nécessaire pour pouvoir atteindre les couples souhaités. Ceci peut entrainer l'irréversibilité de la cinématique, ce qui peut être un inconvénient lors de la mise en place d'un contrôle en raideur[23]. D'autre part, les moteurs à courant continu sont source de champs magnétiques qui peuvent perturber les capteurs de position dans le cas du choix de capteurs magnétiques (*cf.* 2.3.2)

2.3.5 Protocole de communication

Comme il a été dit précédemment, le simulateur SAGA repose sur la combinaison d'un système haptique et d'un modèle numérique. Toutefois si l'on souhaite que cette combinaison fonctionne bien, il est primordial de prendre en compte, et ce dès la conception, les informations échangées entre la partie physique et la partie numérique.

Pour le simulateur, le modèle numérique est pilote, ainsi les informations transmises au système haptique peuvent être considérées comme une commande. Voici une liste des informations qu'il serait intéressant de transmettre au système haptique :

- Positions et orientations de la tête
- Vitesses et accélérations de la tête
- Raideur des muscles
- Couples et efforts appliqués sur la tête
- Direction de l'axe de poussée

D'autre part, l'instrumentation du système haptique permet de mesurer et de prendre en compte les opérations de l'utilisateur. Ainsi, afin que le modèle numérique prenne en compte ces informations il serait intéressant de mesurer :

- Positions et orientations réelles de la tête
- Les efforts appliqués sur la tête
- Positions et orientations des instruments et des mains
- Pressions de contact des instruments sur la tête

3. Recherches

3.1 Cinématique

3.1.1 Étude de la cinématique d'un accouchement

L'augmentation du nombre de degrés de liberté du système haptique du simulateur ne peut être justifiée qu'après une étude du mouvement de la tête fœtale. On trouve dans la littérature de nombreuses informations sur les mouvements qui interviennent lors d'un accouchement. Malheureusement, ces informations sont rarement jointes de valeurs numériques. Nous verrons dans un premier temps les différents mouvements de la tête fœtale au cours d'un accouchement. Puis nous étudierons les valeurs obtenues dans le cadre d'une simulation numérique d'accouchement sans trajectoire imposée[5], réalisée par l'équipe SAARA du laboratoire partenaire.

Trajectoire de la tête fœtale

La forme du bassin joue un rôle primordial dans la trajectoire de la tête fœtale. Lors de l'accouchement, la descente de la tête se fait en suivant la forme du sacrum concave[8]. Cette concavité impose donc de suivre une trajectoire courbe. Cette trajectoire semble être très importante dans le processus de l'accouchement, et c'est pourquoi nous avons décidé de la rendre prioritaire dans l'implémentation des mobilités (*cf.* 2.2.2).

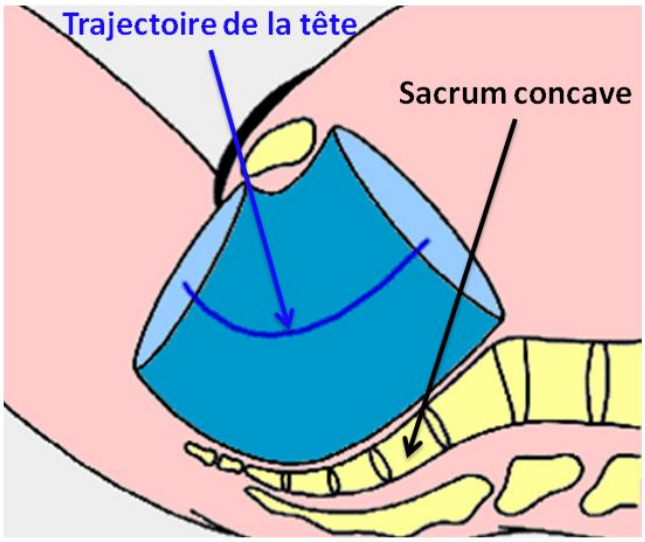


FIGURE 3.1 – Trajectoire de la tête et sacrum concave

Orientation de présentation et rotation intra-pelvienne

On recense 8 orientations de présentation céphaliques. En effet, il existe d'autres types de présentations (présentation du siège, de l'épaule, *etc.*), mais le simulateur se focalise pour l'instant sur les cas d'accouchement ou la tête se présente en premier. Ces présentations représentent tout de même 95% des accouchements. L'orientation de présentation est déterminée à l'aide de l'orientation des sutures et fontanelles de la tête fœtale, lors d'un toucher vaginal. Un degré de mobilité en rotation est donc nécessaire pour simuler ces différentes possibilités. Les 8 orientations céphaliques sont les suivantes :

- Occipito-Pubienne (OP)
- Occipito-Saccrée (OS)

- Occipito-Iliaque Droite Antérieure (OIDA)
- Occipito-Iliaque Gauche Antérieure (OIGA)
- Occipito-Iliaque Droite Postérieure (OIDP)
- Occipito-Iliaque Gauche Postérieure (OIGP)
- Occipito-Iliaque Droite Transverse (OIDT)
- Occipito-Iliaque Gauche Transverse (OIGT)

D'autre part, la tête est amenée à tourner au cours de sa descente. En effet, si l'on prend les cas les plus fréquents, le fœtus commence son engagement avec une orientation OIDA ou OIGA pour finir en OP. La tête a donc effectué 45° de rotation durant la descente[9][10]. Des figures illustrant ces mobilités sont présentées en Annexe C.

Flexion et asynclitisme

La flexion est la rotation de la tête qui permet de basculer la tête d'avant en arrière. Dans la plupart des accouchements, la tête à tendance à se fléchir de plus en plus au cours de la descente[8].

L'asynclitisme est le degré d'inclinaison latérale de la tête. Dans la plupart des cas, l'engagement se fait de manière synclite (pas de rotation). On parle d'asynclitisme postérieur si la rotation rapproche la tête vers le pubis, et d'asynclitisme antérieur lorsque la tête se rapproche du sacrum. Des figures illustrant ces mobilités sont présentées en Annexe C.

Étude d'une simulation numérique

L'équipe SAARA, partenaire du projet, travaille sur la modélisation numérique d'un accouchement. Dans le cadre de sa thèse, Romain Buttin, doctorant dans cette équipe, a simulé la descente de la tête fœtale sans trajectoire imposée[5][4]. Cette simulation a permis, donc, d'estimer la trajectoire de la tête fœtale dans le bassin maternel. La figure 3.2 définit le repère utilisé, alors que la figure 3.3 illustre l'évolution d'un point de la tête fœtale dans le plan sagittal du bassin. Le point suivi sur cette trajectoire correspond au sommet de la tête, c'est-à-dire le point le plus avancé dans le canal pelvien au début de la simulation. Il est à noter que la simulation prend en compte les déformations de la tête et du bassin mou, ce qui explique les variations selon l'axe antéro-postérieur. Ces résultats permettent, d'une part, de mettre en avant la trajectoire selon 2 axes de la tête, et ainsi de justifier l'augmentation du nombre de degrés de liberté du simulateur. D'autre part, ils permettent d'estimer le déplacement selon ces axes. Ainsi, d'après cette simulation, on mesure un déplacement de la tête de l'ordre de 200mm selon l'axe longitudinal et de 60mm selon l'axe antéro-postérieur, lors de la phase d'expulsion.

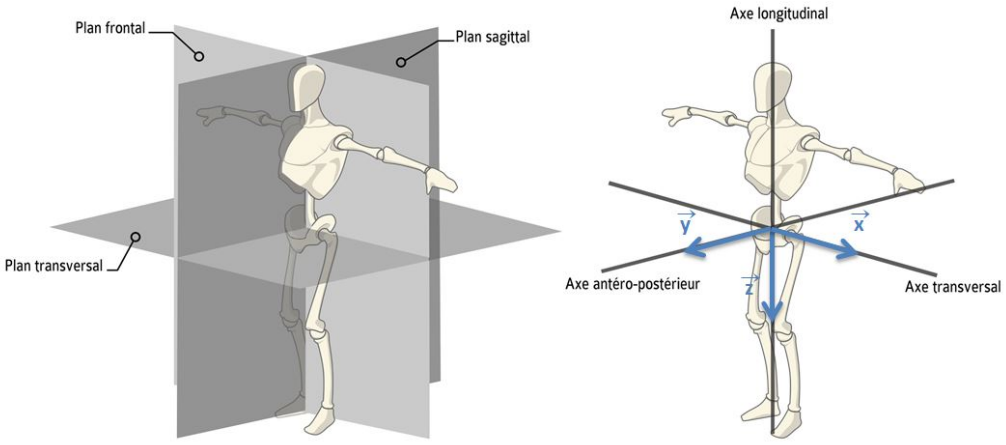


FIGURE 3.2 – Plans et axes anatomiques

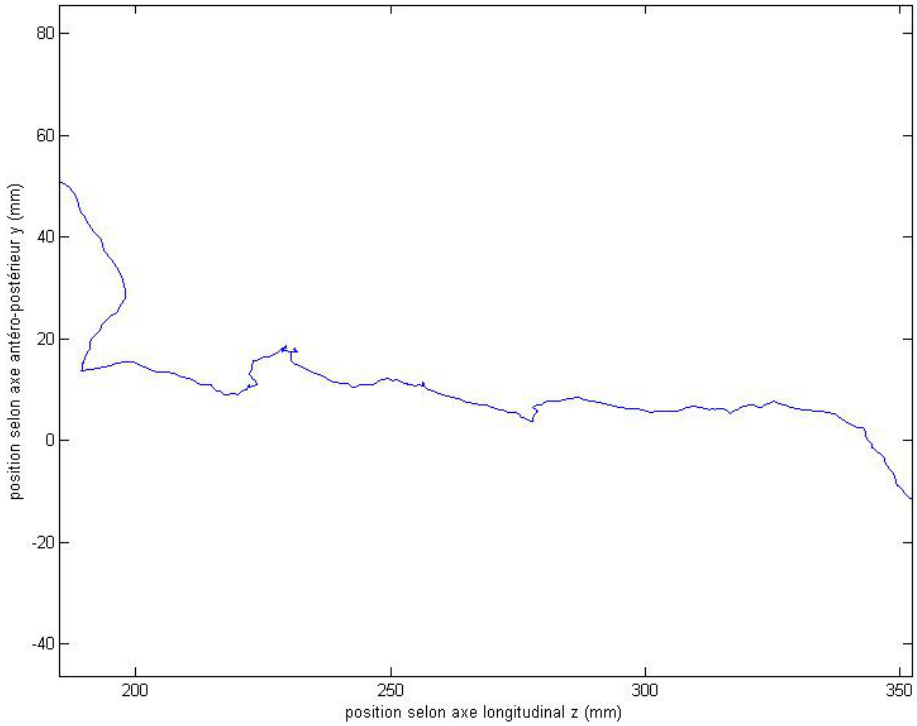


FIGURE 3.3 – Déplacement d'un point de la tête fœtale dans le plan sagittal maternel

3.1.2 Modélisation

Afin d'augmenter le réalisme du simulateur, on souhaite parvenir à une interface haptique avec 4 degrés de liberté pilotés. Afin d'étudier le système et de dimensionner ses composants, j'ai modélisé la mécanique de la partie haptique et j'en ai déduit le modèle cinématique. Ainsi, les figures 3.4 présentent le modèle cinématique envisagé du simulateur. Différentes remarques peuvent être faites sur cette modélisation.

De prime abord, on remarque que les degrés de liberté pilotés sont deux translations et deux rotations. Cela correspond aux deux vérins linéaires et deux moteurs ou vérins rotatifs étudiés dans la partie 2.3.4. Voici une nomenclature simplifiée des sous ensemble :

{0} : Bâti et bassin

{1} : Corps du premier vérin linéaire

{2} : Tige du premier vérin linéaire

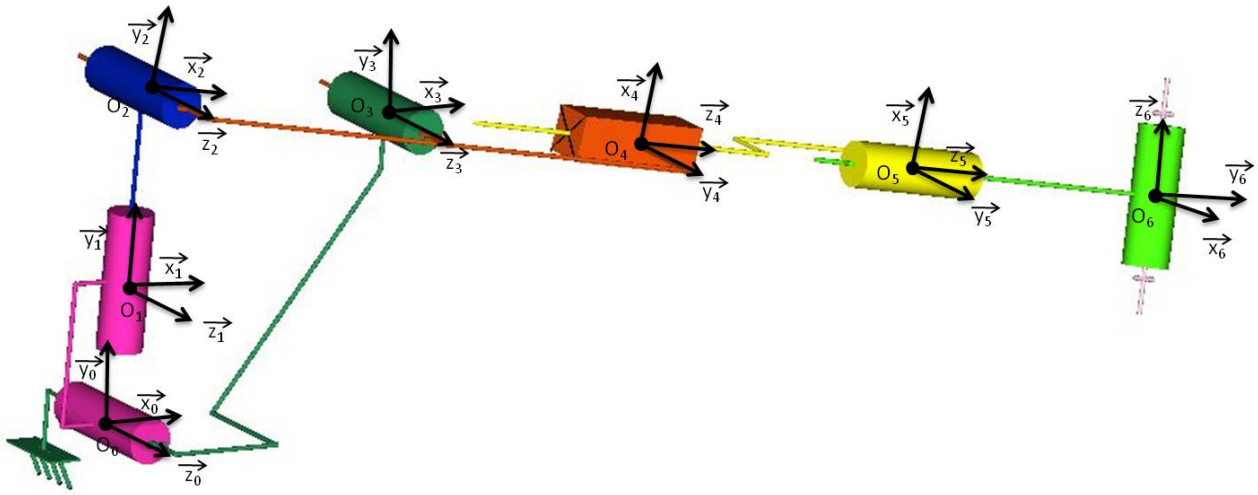
{3} : Corps du second vérin linéaire

{4} : Tige du second vérin linéaire et stator du premier moteur

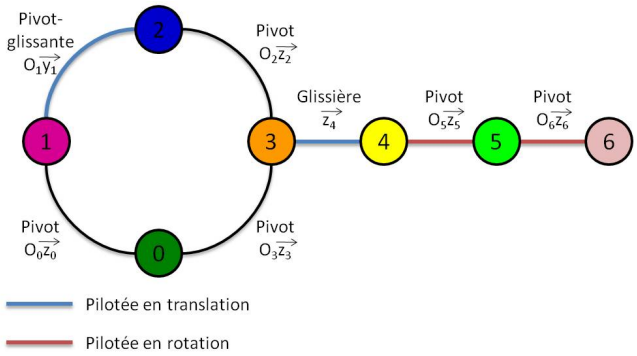
{5} : Rotor du premier moteur et stator du second moteur

{6} : Rotor du second moteur et tête fœtale

En outre, on peut remarquer que la liaison entre les sous-ensembles {3} et {4} est modélisée par une glissière, ce qui implique un blocage de la rotation de la tige du vérin linéaire. Ceci peut être obtenu de différentes manières, une des possibilités est d'utiliser un vérin à tige carrée. Enfin, on peut en déduire que les deux translations pilotées permettront de reproduire la trajectoire de la tête fœtale, alors que, les rotations entre les sous-ensembles {4} et {5} puis {5} et {6} permettrons de reproduire respectivement la rotation intra-pelvienne et la flexion de la tête (*cf.* 3.1.1).



(a) Schéma cinématique



(b) graphe des liaisons

FIGURE 3.4 – Modélisation du système haptique

D'autre part, le système peut être scindé en deux parties. La première est une partie bouclée constituée des sous-ensembles {0}, {1}, {2} et {3} et une partie série composée des sous-ensembles {0}, {3}, {4}, {5} et {6}. La boucle permet, après fermeture géométrique, d'obtenir une relation entre la translation selon l'axe $\overrightarrow{y_1}$ de la liaison entre {1} et {2}, qui est piloté, et la rotation autour de $\overrightarrow{z_3}$ de la liaison entre {0} et {3}. Une fois cette relation obtenue, la partie série permet d'obtenir le modèle géométrique du système haptique du simulateur. Deux paramétrisations différentes ont été utilisées pour définir les repères. En effet, pour la partie bouclée, une paramétrisation permettant de simplifier au maximum la fermeture géométrique a été utilisée, alors que pour la partie série, la paramétrisation de Denavit-Hartenberg conventionnelle a été utilisée.

Enfin, la dernière remarque concerne les repères des sous-ensembles {0} et {6}. Ils sont respectivement liés au bassin maternel et à la tête fœtale, mais il faudra tout de même effectuer un changement de base pour se rapporter au repère anatomique défini sur la figure 3.2 dans la partie précédente.

Étude de la partie bouclée

Si l'on isole la partie bouclée, le modèle équivalent est représenté sur la figure 3.5. L'objectif est d'obtenir une relation entre la sortie du premier vérin linéaire d_1 et l'angle de rotation du corps du second vérin linéaire par rapport au bâti θ_2

On pose les notations suivantes :

$O_0O_1 = a_0$

$O_1O_2 = d_1$

$O_2O_3 = a_2$
$\overrightarrow{O_0O_3}.\overrightarrow{x_3} = c_h$
$\overrightarrow{O_0O_3}.\overrightarrow{y_3} = c_v$
$(\widehat{\overrightarrow{x_2}, \overrightarrow{x_3}}) = \theta_2$
$(\widehat{\overrightarrow{O_3O_2}, \overrightarrow{O_3O_0}}) = \beta$
$(\widehat{\overrightarrow{O_0O_3}, \overrightarrow{y_0}}) = \delta$
Ainsi les paramètres géométriques sont a_0, a_2, c_h, c_v et δ et les paramètres de mouvement sont d_1, θ_2 et β.
On a d'après le théorème de Pythagore

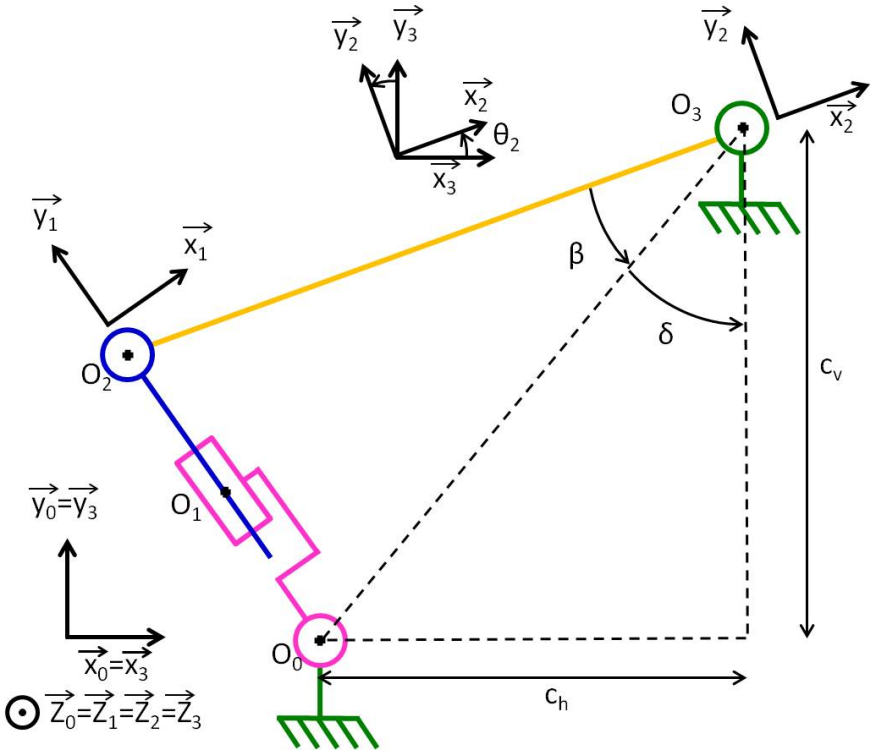


FIGURE 3.5 – Modélisation de la partie bouclée

$$\overline{O_0O_3} = \sqrt{c_h^2 + c_v^2} \tag{3.1}$$

D'où si l'on applique le théorème d'Al-Kashi au triangle $O_0O_2O_3$

$$(d_1 + a_0)^2 = a_2^2 + c_h^2 + c_v^2 - 2a_2\sqrt{c_h^2 + c_v^2}\cos\beta \tag{3.2}$$

on en déduit

$$\beta = \arccos\left(\frac{a_2^2 + c_h^2 + c_v^2 - (d_1 + a_0)^2}{2a_2\sqrt{c_h^2 + c_v^2}}\right) \tag{3.3}$$

D'autre part, $\theta_2 + \beta + \delta = \frac{\pi}{2}$ et $\delta = \arctan\left(\frac{c_h}{c_v}\right)$, d'où

$$\theta_2 = \frac{\pi}{2} - \arccos\left(\frac{a_2^2 + c_h^2 + c_v^2 - (d_1 + a_0)^2}{2a_2\sqrt{c_h^2 + c_v^2}}\right) - \arctan\left(\frac{c_h}{c_v}\right) \tag{3.4}$$

Or $\frac{\pi}{2} + \arccos x = \arcsin x$, on obtient après simplification

$$\boxed{\theta_2 = \arcsin\left(\frac{a_2^2 + c_h^2 + c_v^2 - (d_1 + a_0)^2}{2a_2\sqrt{c_h^2 + c_v^2}}\right) - \arctan\left(\frac{c_h}{c_v}\right)} \tag{3.5}$$

Étude de la partie série

Si l'on isole la partie série, le modèle équivalent est représenté sur la figure 3.6. Après la paramétrisation Denavit-Hartenberg classique, on obtient le tableau 3.1. Il est à noter que le lien 4' correspond à un changement de repère, car la convention de Denavit-Hartenberg ne permet pas de définir directement un angle entre $\overrightarrow{y_{n\text{-}1}}$ et $\overrightarrow{y_n}$. D'autre part, le θ_2 est défini exactement de la même manière que pour la partie bouclée. On retrouve, ainsi, les 4 variables pilotes, θ_2 qui dépend directement de d_1 et les variables d_4, θ_5 et θ_6. Ces variables correspondent bien aux deux translations et deux rotations pilotées.

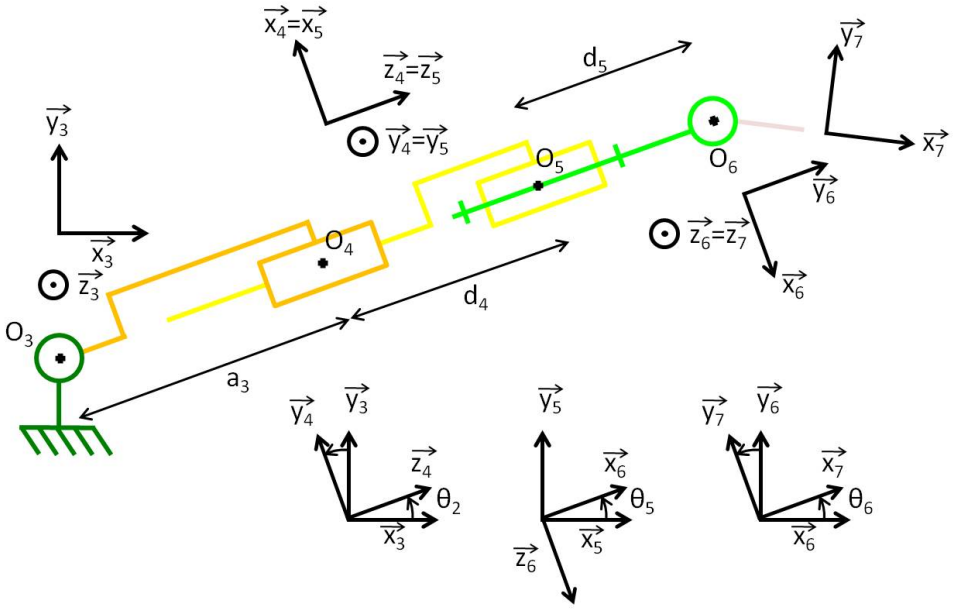


FIGURE 3.6 – Modélisation de la partie série

Link	θ	d	a	α	R ou P
3	θ_2	0	a_3	0	R
4'	$\frac{\pi}{2}$	0	0	$\frac{\pi}{2}$	N/A
4	$\frac{\pi}{2}$	d_4	0	0	P
5	θ_5	d_5	0	$\frac{\pi}{2}$	R
6	θ_6	0	0	0	R

TABLE 3.1 – Paramétrisation de Denavit-Hartenberg de la partie série

Les matrices de transformation sont les suivantes :

$${}^0T_3 = \begin{pmatrix} \cos\theta_2 & -\sin\theta_2 & 0 & a_3\cos\theta_2 \\ \sin\theta_2 & \cos\theta_2 & 0 & a_3\sin\theta_2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$ {}^{3}T_{4'} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

$$ {}^{4'}T_{4} = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & d_4 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

$$ {}^{4}T_{5} = \begin{pmatrix} \cos\theta_5 & 0 & \sin\theta_5 & 0 \\ \sin\theta_5 & 0 & -\cos\theta_5 & 0 \\ 0 & 1 & 0 & d_5 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

$$ {}^{5}T_{6} = \begin{pmatrix} \cos\theta_6 & -\sin\theta_6 & 0 & 0 \\ \sin\theta_6 & \cos\theta_6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

On obtient donc le modèle géométrique direct suivant, qui représente la position et l'orientation de la base $(O_6, \overrightarrow{x_7}, \overrightarrow{y_7}, \overrightarrow{z_7})$ dans la base $(O_3, \overrightarrow{x_3}, \overrightarrow{y_3}, \overrightarrow{z_3})$:

$$ {}^{0}T_{6} = \begin{pmatrix} c\theta_2 s\theta_6 + c\theta_6 s\theta_2 s\theta_5 & c\theta_2 c\theta_6 - s\theta_6 s\theta_2 s\theta_5 & -c\theta_5 s\theta_2 & (a_3 + d_4 + d_5) c\theta_2 \\ s\theta_2 s\theta_6 - c\theta_2 c\theta_6 s\theta_5 & c\theta_6 s\theta_2 + c\theta_2 s\theta_6 s\theta_5 & c\theta_2 c\theta_5 & (a_3 + d_4 + d_5) s\theta_2 \\ c\theta_5 c\theta_6 & -c\theta_5 s\theta_6 & s\theta_5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

avec $\theta_2 = \arcsin\left(\dfrac{a_2^2 + c_h^2 + c_v^2 - (d_1 + a_0)^2}{2a_2\sqrt{c_h^2 + c_v^2}}\right) - \arctan\left(\dfrac{c_h}{c_v}\right)$

Remarque : $\cos\theta_i$ et $\sin\theta_i$ sont respectivement remplacés par $c\theta_i$ et $s\theta_i$

3.2 Les lois de commande

3.2.1 Les commandes de la version précédente du BirthSIM

Le second axe de recherche, sur lequel je me suis concentré, est l'implémentation de lois de commande. En effet, l'objectif est de mettre en place des lois de commande qui fournissent un rendu haptique fidèle. Mes prédécesseurs ont mis en place 4 lois de commande différentes dont l'utilisation dépend du scénario que l'on joue. En effet, une dizaine de scénarios ont été implémentés sur la version précédente du BirthSIM. Ces scénarios permettent de simuler des situations d'accouchement différent, par exemple un accouchement eutocique ou un accouchement nécessitant une extraction instrumentale. Les lois de commande implémentées sont :

- un suivi de trajectoire en effort reconstruit
- un asservissement en position par retour d'état
- une commande hybride effort reconstruit/position
- une commande hybride effort/vitesse.
- un asservissement en position et vitesse avec un gain variable

La commande avec asservissement de position est utilisée pour des scénarios où l'on souhaite positionner avec précision la tête fœtale. Pour un entrainement aux touchers vaginaux par exemple. La commande de suivi de trajectoire en effort reconstruit est utilisée pour les scénarios d'extractions instrumentales où les efforts pour déplacer la tête sont faibles. Enfin, l'asservissement en position et vitesse avec un gain variable est utilisé sur des scénarios d'extractions instrumentales où les efforts à fournir pour déplacer la tête sont importants.

3.2.2 Les commandes en raideur dans la littérature

Retranscrire la sensation de toucher des corps mous est primordial dans le cas de notre simulateur. En effet, si, pour les robots industriels, on cherche, en général, à avoir une rigidité maximum, pour le simulateur d'accouchement, on doit essayer de reproduire la souplesse des tissus humains. Les précédentes lois de commande se reposaient pour cela sur la compliance naturelle du vérin pneumatique. Une autre approche qu'y est celle que je vais poursuivre est celle d'implémenter des lois de commande qui permettent un contrôle en raideur. En effet, on trouve dans la littérature différentes études en ce sens que je vais présenter dans cette partie.

La raideur est le rapport entre l'effort et le déplacement. Ainsi, implémenter une loi de commande en raideur permet de donner une impression de toucher de corps mous dans le cas où la raideur est faible. En faisant varier cette raideur, on rigidifie plus ou moins le système.

Un même degré de liberté ne peut pas être piloté à la fois en effort et en position. Il y a, ainsi, deux approches qui permettent d'augmenter le nombre de degrés de liberté. La première est d'utiliser deux actionneurs pour piloter un seul mouvement. La seconde concerne les actionneurs pneumatiques et consiste plutôt à utiliser deux pré-actionneurs.

Dans le premier cas, on trouve par exemple les travaux de Christine Prelle[23]. Durant sa thèse, elle a implémenté une loi de commande en raideur par retour d'état et retour d'effort explicite en utilisant deux soufflets métalliques. La commande par retour d'effort explicite permet notamment de modifier de manière linéaire la raideur du système en modifiant simplement un gain.

D'autre part, on trouve aujourd'hui des études sur des actionneurs à raideur variable[3]. Ces actionneurs sont constitués de deux sous-actionneurs, le premier contrôle la position initiale (sans contrainte), alors que le second modifie directement la raideur du système. En définitive, dans cette solution la raideur est apportée de manière mécanique en contraignant plus ou moins un ressort.

La seconde méthode profite du fait que les vérins pneumatiques sont mus par un différentiel de pression entre deux chambres séparées par un piston. Ces chambres sont alimentées en fluide par un ou plusieurs pré-actionneurs. Dans la plupart des cas, ces pré-actionneurs sont commandés de manière à avoir une action symétrique sur les deux chambres, c'est-à-dire que le débit d'entrée du fluide dans la première chambre est égal au débit de fluide sortant de l'autre chambre. L'idée consiste donc à piloter les deux chambres séparément de manière à avoir un degré de liberté supplémentaire. C'est par exemple le sujet des travaux effectués par Xiangrong Shen[25]. Il a développé une commande qui permet de piloter un vérin pneumatique en effort et en raideur, en utilisant 2 servo-distributeurs. Cette commande repose sur un retour par mode glissant.

Au laboratoire Ampère, les récents travaux de l'équipe ACM ont permis de proposer une nouvelle loi de commande en raideur. En effet, les travaux de Frédéric Abry[1] prennent en compte les fortes non-linéarités récurrentes dans l'utilisation de vérin pneumatique. Ainsi, une loi de commande se basant sur le backstepping a été implémentée sur un banc d'essai et les résultats sont positifs.

Enfin, une autre approche a été abordée par Mahvash et *al.*[18]. En effet, une loi de commande en raideur a été implémentée sur un robot continuum. Cette fois, les actionneurs sont électriques et la commande repose sur le calcul de la déflation du robot dû à une action extérieure. Ainsi, l'effort fourni par l'actionneur est recalculé en fonction de l'écart par rapport à la position de référence. La connaissance du modèle cinématique est par contre nécessaire pour implémenter cette loi. Cette démarche fonctionne bien en statique et elle permet de définir des raideurs variables selon la direction des efforts appliqués.

4. Conclusion et perspectives

Le développement d'un simulateur d'accouchement est un projet qui couvre un large domaine de compétences. En effet, de nombreux choix technologiques s'imposent et ces choix impacteront directement les performances de ce simulateur. Ainsi, la poursuite de mes travaux dépend essentiellement des choix qui vont être faits lors de la validation de l'architecture. Une fois l'architecture de la nouvelle version du simulateur figée, il sera possible d'implémenter les lois de commande et d'effectuer les premières expérimentations. C'est pourquoi mes travaux à l'heure actuelle se sont essentiellement focalisés sur la partie développement. En effet, ces six premiers mois de thèse m'ont permis de réaliser un cahier des charges, étudier la mécanique de l'accouchement et ainsi, obtenir le modèle géométrique de la partie haptique du simulateur. Ce modèle va me permettre de déterminer les relations cinématiques. Enfin, l'obtention du modèle dynamique me permettra de finir le dimensionnement des actionneurs et pré-actionneurs. La connaissance de ces modèles sera aussi un départ pour le développement des nouvelles lois de commande.

D'autre part, la réalisation du simulateur se fera certainement en deux étapes. La première consistera à implémenter et commander les deux vérins linéaires permettant de simuler la trajectoire de la tête fœtale. Cela permettra de simplifier le système et me permettra d'obtenir de premiers résultats sur l'intégration des lois de commande. C'est seulement dans un second temps que les deux actionneurs rotatifs simulant la flexion et la rotation intra-pelvienne seront intégrés. Ainsi, cela me laisse plus de temps pour pouvoir figer la technologie employée et poursuivre mes recherches sur cette partie, notamment sur les lois de commande en raideur d'actionneurs électriques.

En outre, je devrais bientôt recevoir d'autres jeux de données provenant des simulations numériques d'accouchements des différents partenaires. Ces données me permettront, notamment, d'obtenir plus d'information sur l'évolution de la raideur des muscles pelviens au cours de l'accouchement. Ces informations sont nécessaires pour le dimensionnement des pré-actionneurs alimentant les vérins. Ces données m'informeront aussi sur la rotation de la tête au cours de l'accouchement.

Enfin, la dernière partie de mes travaux qui concerne la mise en place de l'interaction entre la simulation numérique et la partie haptique sera prise en compte dans la suite de mes travaux. Toutefois, cette mise en relation ne pourra être implémentée physiquement que lorsque les travaux respectifs des différents partenaires auront suffisamment avancé. Un dialogue continu entre les participants au projet assurera la convergence de nos travaux.

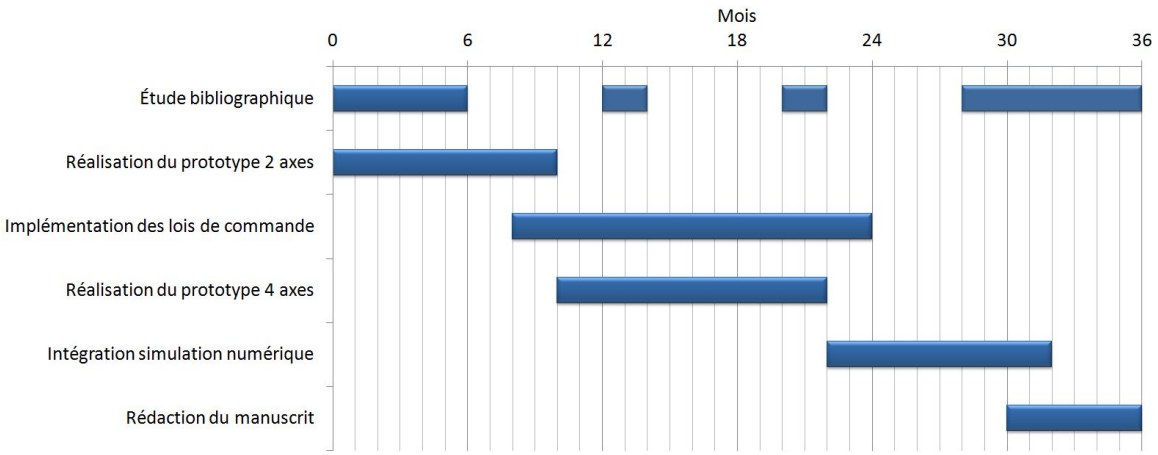


FIGURE 4.1 – Planning prévisionnel

Bibliographie

[1] F. Abry, X. Brun, S. Sesmat, and E. Bideaux. Non-linear position control of a pneumatic actuator with closed-loop stiffness and damping tuning. In *European Control Coference*, 2013. [**under press**].

[2] R.H. Allen, B.R. Bankoski, C.A. Butzin, and D.A. Nagey. Comparing mechanical fetal response during descent, crowning, and restitution among deliveries with and without shoulder dystocia. *American Journal of Obstetrics and Gynecology*, 171(6) :1621–1627, 2007.

[3] D. Braun, M. Howard, and S. Vijayakumar. Optimal variable stiffness control : formulation and application to explosive movement tasks. *Autonomous Robots*, 33(3) :237–253, 2012.

[4] R. Buttin. *Modélisation biomécanique du système reproductif féminin et du foetus humain pour la réalisation d'un simulateur virtuel d'accouchement.* PhD thesis, Université Claude Bernard Lyon 1, 2010. Thèse de Doctorat en Informatique.

[5] R. Buttin, F. Zara, B. Shariat, T. Redarce, and G. Grangé. Biomechanical simulation of the fetal descent without imposed theoretical trajectory. *Computer Methods and Programs in Biomedicine*, 2013.

[6] R. Caen, M. Lajoie-Mazenc, and B. Trannoy. Actionneurs en robotique. *Techniques de l'ingénieur*, 1990.

[7] M.O. Culjat, C.H. King, M.L. Franco, C.E. Lewis, J.W. Bisley, E.P. Dutson, and W.S. Grundfest. A tactile feedback system for robotic surgery. In *Engineering in Medicine and Biology Society*, pages 1930–1934. IEEE, 2008.

[8] Collège National des Gynécologues et Obstétriciens Français (CNGOF). *Engagement.* 2006.

[9] Collège National des Gynécologues et Obstétriciens Français (CNGOF). *Rotation intra-pelvienne de la tête foetale.* 2006.

[10] Comité éditorial pédagogique des Écoles de sages-femmes de France. *Étude de l'accouchement en présentation du sommet : l'accouchement (troisième temps de la deuxième étape du travail).*

[11] O. Dupuis, M. Betemps, G. Delhomme, A. Dittmar, H.T. Redarce, and R.C. Silveira. Simulateur fonctionnel et anatomique d'accouchement, 2005. Brevet WO 2005/013229 A3.

[12] J-C Granry and M-C Moll. Rapport de mission, état de l'art (national et international) en matière de pratiques de simulation dans le domaine de la santé dans le cadre du développement professionnel continu (dpc) et de la prévention des risques associés aux soins. Technical report, Haute Autorité de Santé, 10 janvier 2012.

[13] E.J. Kim, R.H. Allen, J.H. Yang, M.K. McDonald, W. Tam, and E.D. Gurewitsch. Simulating complicated human birth for research and training. In *Engineering in Medicine and Biology Society*, volume 26, pages 2762–2766. IEEE, 2004.

[14] E.J. Kim, P. Theprungsirikul, M.K. McDonald, E.D. Gurewitsch, and R.H. Allen. A biofidelic birthing simulator. *IEEE Engineering in Medicine and Biology Magazine*, 24(6) :34–39, 2005.

[15] R.J. Lapeer, M.S. Chen, and J.G. Villagrana. An augmented reality based simulation of obstetric forceps delivery. In *Acm International Symposium on Mixed and Augmented Reality*, pages 274–275. IEEE, 2004.

[16] R.J. Lapeer, M.S. Chen, and J.G. Villagrana. Simulating obstetric forceps delivery in an augmented environment. In *Augmented environments for Medical Imaging including Augmented Reality in Computer-aided Surgery*, 2004.

[17] K.K. Leslie, P. Dipasquale-Lehnerz, and M. Smith. Obstetric forceps training using visual feedback and the isometric strength testing unit. *American College of Obstetricians and Gynecologists*, 105(2) :377–382, 2005.

[18] M. Mahvash and P.E. Dupont. Stiffness control of surgical continuum manipulators. *IEEE Transactions on Robotics*, 27(2) :334–345, 2011.

[19] R. Moreau. *Le simulateur d'accouchement BirthSIM : un outil complet pour la formation sans risque en obstétrique*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 2007. Thèse de Doctorat en Automatique industrielle.

[20] R. Moreau, M.T. Pham, X. Brun, T. Redarce, and O. Dupuis. Simulation of an instrumental childbirth for the training of the forceps extraction : control algorithm and evaluation. *IEEE Transactions on Information Technology in Biomedicine*, 15(3) :364–372, 2011.

[21] T. Obst, R. Burgkart, E. Ruckhaberle, and R. Riener. The delivery simulator : a new application of medical VR. In *Medicine Meets Virtual Reality*, volume 98, pages 281–287, 2004.

[22] J. Petitcolas. Le mannequin de mme du coudray : ou comment former les accoucheuses au XVIIIe siècle. *La revue du praticien*, 56(2) :226–229, 2006.

[23] C. Prelle. *Contribution au contrôle de la compliance d'un bras de robot à actionnement électro-pneumatique*. PhD thesis, Institut National des Sciences Appliquées de Lyon, 1997. Thèse de Doctorat en Automatique industrielle.

[24] R. Riener and B. Rainer. Birth simulator, 2007. Brevet US 7241145 B2.

[25] X.R. Shen and M. Goldfarb. Simultaneous force and stiffness control of a pneumatic actuator. *Journal of Dynamic Systems Measurement and Control-Transactions*, 129(4) :425–434, 2007.

[26] T. Sielhorst, T. Blum, and N. Navab. Synchronizing 3d movements for quantitative comparison and simultaneous visualization of actions. In *International Symposium on Mixed and Augmented Reality*, pages 38–47. IEEE, 2005.

[27] T. Sielhorst, T. Obst, R. Burgkart, R. Riener, and N. Navab. An augmented reality delivery simulator for medical training. In *Augmented environments for Medical Imaging including Augmented Reality in Computer-aided Surgery*, 2004.

[28] T. Sielhorst, J. Traub, and N. Navab. The AR apprenticeship : Replication and onmidirectional viewing of subtle movements. In *Acm International Symposium on Mixed and Augmented Reality*, pages 290–291. IEEE, 2004.

[29] Tekscan. Force sensors for design. *Machine Design Custom Media*, 2013. **[en ligne]** http://insidepenton.com/machinedesign/nl/TekScanEbook.pdf (visité le 24/06/2013).

Annexe A : Glossaire

Dans cette annexe, quelques définitions de mots utilisés dans mon rapport sont données. La plupart de ces mots proviennent du vocabulaire médical.

Anthropomorphisme : Caractéristique d'un mécanisme, d'une structure rappelant l'humain.

Asynclitisme : Déviation de l'axe de la tête du fœtus par rapport à l'axe du bassin.

Bassin mou : Ensemble des tissus mou présent autour du bassin osseux.

Bassin osseux : Le bassin osseux est une partie du squelette, en forme d'entonnoir, constitué des deux os coxaux latéraux, du coccyx et du sacrum en arrière. C'est la ceinture pelvienne, constituant la jonction entre la colonne vertébrale mobile (axe du tronc) et les membres inférieurs.

Compliance : Capacité d'un manipulateur à avoir un comportement souple, à s'adapter à son environnement.

Dystocie des épaules : Se produit lors d'un accouchement où les épaules du fœtus restent bloquées par le bassin de la parturiente.

Eutocique (accouchement) : Un accouchement qui se déroule dans des conditions normales.

Extraction instrumentale : Dans le cadre d'un accouchement, l'extraction instrumentale correspond à un accouchement pour lequel des outils ont été nécessaires pour extraire le fœtus. Les outils fréquemment utilisés sont les forceps ou les ventouses.

Flexion : Mouvement où une partie du corps (partie d'un membre, *etc.*) forme un angle avec la partie voisine, position résultant de ce mouvement. Dans le cas de la tête, cela correspond au mouvement de rotation permettant de basculer la tête d'avant en arrière.

Fontanelle : Chacun des espaces cartilagineux compris entre les os du crâne des jeunes enfants, avant son ossification complète, aux points de jonction des sutures osseuses.

Haptique : Qui concerne le sens du toucher.

Médicamentation : Action de prescrire ou de prendre des médicaments.

Muscles pelviens : Ensemble des muscles constituant le bassin mou.

Parturiente : Femme qui accouche.

Plancher pelvien : Voir muscles pelviens.

Plexus brachial : Plexus formé des quatre derniers nerfs cervicaux et du premier nerf dorsal, qui gouvernent la motricité et la sensibilité du membre supérieur.

Présentation céphalique : Correspond à un accouchement dans lequel le fœtus présente sa tête en premier.

Présentation du siège : Correspond à un accouchement dans lequel le fœtus présente son bassin avant sa tête.

Sacrum : Os formé par la soudure des cinq vertèbres sacrées, à la partie inférieure de la colonne vertébrale, s'articulant avec les os iliaques pour former le bassin.

Suture : Articulation immobile dentée entre deux os réunie par du tissu fibreux.

Annexe B : Comparaison des différents simulateurs

B.1 Critères

Niveau d'anthropomorphisme

Il est question, ici, de comparer les degrés de réalisme des différents simulateurs. Pour définir cela, une note de 1 à 3 a été attribuée selon différents critères. Par exemple, le mannequin est-il complet, ou ne représente-t-il qu'une section ? Les textures au toucher sont-elles réalistes ? Les fluides sont-ils présents ? (liste non exhaustive) La note 3 étant un degré de réalisme maximal et la note 1 étant la moins bonne.

Possibilité de réalisation des gestes fondamentaux

Ce critère a pour but d'évaluer la possibilité de réalisation des gestes fondamentaux (toucher vaginal, diagnostic...), dans le but d'éventuelles formations ou entrainements à la réalisation de ces gestes.

Types d'accouchements simulés

Ce critère détermine les différents types d'accouchement que le simulateur prend en charge. En effet, les différents types d'accouchements que peut prendre en compte le simulateur sont les suivants :

- eutocique
- instrumental
- siège
- dystocique
- césarienne

Interface de visualisation de la position de la tête fœtale

Il s'agit de savoir s'il y a une interface qui permet de connaitre en temps réel la position de la tête fœtale.

Interface de visualisation de la position des instruments et/ou des mains

L'interface permet-elle de visualiser la position des instruments ou des mains de l'utilisateur ?

Simulation des efforts mis en jeu lors de l'accouchement avec la possibilité de régler certains paramètres

Lors de l'accouchement, différents efforts viennent s'appliquer sur le fœtus, et plus particulièrement sur sa tête. Les différents efforts qui s'appliquent permettent ou s'opposent à l'avancée du fœtus dans le canal pelvien. Ces efforts sont notamment la force d'expulsion volontaire due aux contractions abdominales, la force d'expulsion involontaire due aux contractions utérines, la force d'expulsion instrumentale due à l'utilisation d'outils obstétriques et enfin la force résistive des muscles pelviens.

La répétabilité des opérations

L'objectif est de savoir s'il est possible de répéter plusieurs fois les mêmes opérations. En effet, à partir de mêmes conditions initialement déterminées, le simulateur reproduit-il le même comportement ?

Possibilité de formation et d'évaluation des gestes médicaux

Le simulateur donne-t-il la possibilité d'évaluer de manière qualitative et quantitative les opérations effectuées par ces derniers ?

Possibilité d'expérimenter de nouvelles techniques

Le simulateur permet-il d'expérimenter de nouvelles méthodes d'extraction ou de diagnostic et ainsi de les valider ?

Transportabilité

Le simulateur peut être susceptible d'être déplacé entre différents hôpitaux ou cliniques, ainsi ce critère détermine si l'on a la possibilité de le transporter.

Possibilité de simuler les différents niveaux et orientations de présentation

On s'intéresse, ici, à la possibilité de simuler les différents cas pouvant se manifester durant un accouchement. Le niveau de présentation est la distance relative entre l'extrémité de la tête fœtale et les épines sciatiques du bassin de la parturiente. Le niveau évolue généralement entre -5 et +5 cm. L'orientation de présentation est l'orientation qu'a la tête fœtale par rapport à l'axe du canal pelvien. On recense 8 cas de figure possibles, OP, OS, OIGA, OIDA, OIGP, OIDP, OIDT et OIGT.

Mobilité de la tête fœtale dans le bassin maternel

Naturellement, la tête fœtale a la possibilité de bouger selon plusieurs axes. Le simulateur prend-il en compte cette possibilité ? Cette information est d'autant plus importante lors d'entrainement à l'utilisation de forceps, car les manipulations pour de certaines orientations de présentation consistent à réorienter le fœtus dans le bassin maternel.

B.2 Tableau comparatif des simulateur en fonction des critères

Récapitulatif des critères :

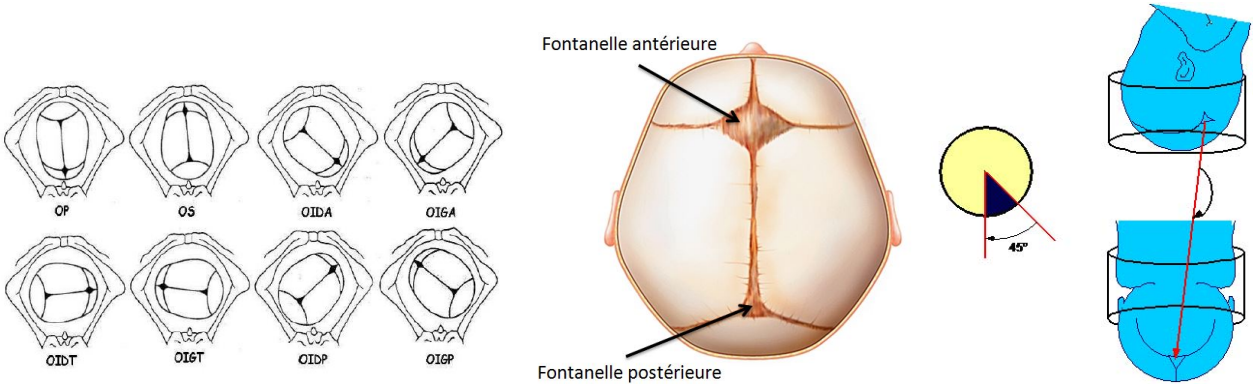
1. Niveau d'anthropomorphisme
2. Possibilité de réalisation des gestes fondamentaux
3. Types d'accouchements simulés
4. Interface de visualisation de la position de la tête fœtale
5. Interface de visualisation de la position des instruments et/ou des mains
6. Simulation des efforts mis en jeu lors de l'accouchement avec la possibilité de régler certains paramètres
7. La répétabilité des opérations
8. Possibilité de formation et d'évaluation des gestes médicaux
9. Possibilité d'expérimenter de nouvelles techniques
10. Transportabilité
11. Possibilité de simuler les différents niveaux et orientations de présentation
12. Mobilité de la tête fœtale dans le bassin maternel

critères / simulateurs	1	2	3†	4	5	6	7	8‡	9	10	11	12
Noelle	3	x	e,i,s,d			x	x	q	x	x	x	x
SimMom	3	x	e,i,s,d					q	x	x	x	x
SIMone	2	x	e,i	x		x	x	x	x		x	x
LM-095	2	x	e	x	x		x	q	x	x	x	x
Simulateur Université Hopkins	2	x	e,d			x	x	x	x		x	x
Simulateur School of Computing Sciences	1		i	x		x	x			x	x	x
Simulateur breveté par Riener et *al.*	2	x	e,i	x	x	x	x	x	x		x	x
BirthSIM	2	x	e,i	x	x	x	x	x	x		x	

TABLE B.1 – Comparatif des simulateurs

† : e : eutocique, i : instrumental, s : siège, d : dystocie des épaules
‡ : q : qualitatif uniquement, x : qualitatif et quantitatif

Annexe C : Mobilités de la tête fœtale



(a) Orientations de présentations (b) Crâne fœtal et fontanelles (c) Rotation intra-pelvienne

FIGURE C.1 – Orientations et rotation intra-pelvienne
[source : CNGOF]

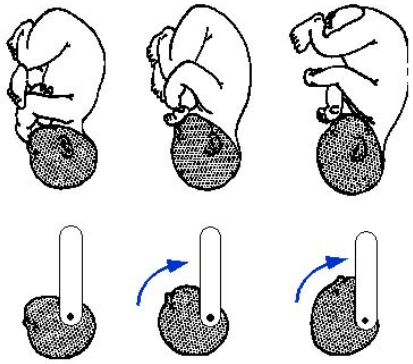
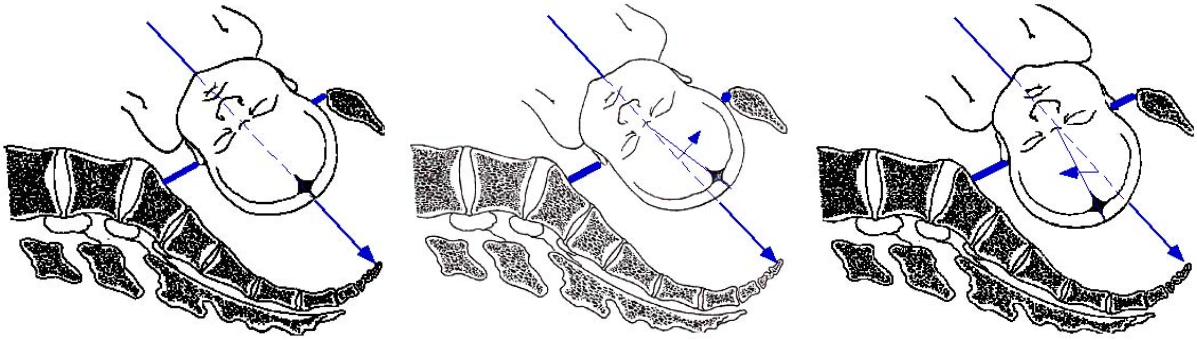


FIGURE C.2 – Flexion céphalique complémentaire lors de l'engagement
[source : CNGOF]



(a) Présentation Synclite (b) Assynclitisme postérieur (c) Assynclitisme antérieur

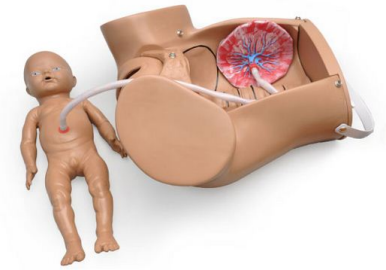
FIGURE C.3 – Assynclitisme
[source : CNGOF]

Annexe D : Comparatif de mannequins anatomiques présents sur le marché

Mannequin	Prompt	S500	AK063B	Obstetrical Mannikin	VG395
Fabricant	Laerdal	Gaumard	Adam Rouilly	Simulaids	3B Scientific
Bassin	X	X	X	X	X
Fœtus	X	X	X	X	X
Muscles pelviens	X				
Prix approximatif	4350€	500€	4600€	750€	1400€
Délai d'approvisionnement	5 semaines	3 semaines	5 semaines	3 jours	n/c
Utilisé pour	SimMom	Noelle	LM095 (Koken MPC)	BirthSIM	

TABLE D.1 – Comparatif des mannequins anatomiques

(a) Prompt



(b) S500

(c) AK063B

(d) Obstetrical Manikin

(e) VG395

FIGURE D.1 – Mannequins anatomiques

Table des figures

Liste des tableaux

Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

Mémoire doctorant 1ère année

2012 -2013

Nom - Prénom	**Loudière – Kevin**
Titre de la thèse	**Modélisation prédictive d'une chaîne d'entrainement complète en milieu aéronautique. Vers une optimisation d'ensemble du système**
Directeur de thèse	**Christian Vollaire**
Co- encadrants	**Costa François (SATIE)**
Dpt. de rattachement	**Méthode pour l'ingénierie des systèmes**
Date début des travaux	**01 Avril 2013**
Type de financement	**CIFRE**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

Table des matières

Chapitre 1

Introduction

1.1 Contexte de la thèse

Cette thèse s'inscrit dans le contexte du développement de l'avion plus électrique. La recherche de l'amélioration de la qualité et de la sécurité des vols a entrainé une augmentation de la puissance électrique embarquée due à la multiplication des sources et actionneurs électriques (des écrans vidéos aux freins du train d'atterrissage en passant par le système de ventilation et les gouvernes de vol...). L'ajout d'un système électrique dans un réseau de n systèmes entraine potentiellement n fois plus d'interactions électromagnétiques indésirables. Il est donc important de connaître ces interactions pour pouvoir les éliminer ou au moins les réduire. Cet axe de recherche constitue ce que l'on appelle la compatibilité électromagnétique (CEM). Il ne faut également pas oublier le contexte économique actuel qui tend à réduire les coûts au maximum, ce qui se traduit par un cahier des charges très complexe et peu modulable.

La thèse a débuté en Avril 2013. Elle s'effectue dans le cadre d'un contrat CIFRE entre la société Labinal Power Systems (anciennement Hispano-Suiza), le laboratoire Ampère et le laboratoire SATIE. Labinal Power Systems est une entreprise appartenant au groupe SAFRAN et est spécialisée dans l'électronique de puissance embarquée dans le milieu aéronautique. La thèse se fait dans le département de Méthodes pour l'Ingénierie des Systèmes sous la direction de Christian Vollaire pour le laboratoire Ampère et dans le département Composants et Systèmes pour l'Energie Electrique sous la direction de François Costa pour le laboratoire SATIE.

1.2 Rappel du sujet de thèse

L'objectif de la thèse est de créer des modèles d'une chaîne d'entraînement électrique typique en prenant en compte le fonctionnement électrotechnique et la Compatibilité ElectroMagnétique (CEM) de l'ensemble connexion réseau continu - filtre d'entrée - onduleur - filtre de sortie - harnais - moteur électrique - plan de masse (composite). Les modèles seront orientés vers une approche prédictive de la chaîne sur un domaine de validité de 10 kHz à 50 MHz. Plusieurs phénomènes vont être ajoutés au modèle existant : la température et les phénomènes de saturation des matériaux magnétiques. Cette approche sera corrélée avec les mesures qui seront effectuées sur une chaîne d'entrainement disponible à Labinal Power Systems. La finalité de ce travail est d'aider un ingénieur électronicien non spécialiste à concevoir un produit avec un bon comportement CEM dès les premières phases de la conception. Cela implique de connaître les paramètres influents sur la

chaîne, connaître leur impact et savoir comment les optimiser pour pouvoir réduire les perturbations électromagnétiques tout en restant dans le cadre des objectifs électriques, thermiques, mécaniques et économiques propres au milieu aéronautique. Cette optimisation se veut globale et concernera donc l'ensemble de masse et du volume de la chaîne. L'aspect réduction des perturbations et l'aspect filtrage seront étudiés en parallèle pour pouvoir analyser le paramètre masse du filtre en le comparant au paramètre câble ou au paramètre pertes thermiques. Les résultats obtenus seront ensuite compilés dans un outil ingénieur qui aidera à la conception des produits d'un point de vue CEM. Cet outil devra être rapide, fiable et modulable.

1.3 Points clés retenus

Partie	**Modélisation prédictive du système**	**Réduction des perturbations et filtrage**	**Création d'outils ingénieurs**
Objectifs	– Augmentation du domaine de validité de la simulation de 15 MHz à 50 MHz – Prise en compte de la température – Prise en compte de la saturation des matériaux magnétiques	– Réduction des perturbations au niveau de la source de bruit – Optimisation du filtrage – Optimisation globale de la chaîne	– Optimisation des méthodes de prédiction et de dimensionnement – Protocole de renseignement des différents modèles – Outil d'aide à la prise de décision
Mots clés	– Modélisation – Prédiction – Température – Etudes paramétriques	– Perturbations électromagnétiques – Filtrage – Optimisation	– Rapidité – Robustesse – Simplification – Modularité

Table 1.1 – Points clés de la thèse

Chapitre 2

Etat de l'art

2.1 Compatibilité électromagnétique

La compatibilité électromagnétique est autant un problème de mécanicien que d'électricien : les problèmes rencontrés sont créés par des sources électriques et se propagent par des chemins créés par les mécaniques ou les interconnexions électriques. En effet, un système a beau créer un maximum d'interférences, si celles-ci ne peuvent pas se propager, il n'y a aucun souci de compatibilité électromagnétique. Il y a donc deux solutions pour réduire les perturbations électromagnétiques : les réduire à la source, c'est le travail de l'électronicien, ou les empêcher d'atteindre leurs victimes en modifiant leur chemin de propagation, c'est le travail du mécanicien. On introduit ici les notions fondamentales de source et de victime qui sont reliées aux notions d'émissivité (source) et de susceptibilité (victime). Ces notions de source et victime ne s'excluent pas : une source peut également être une victime. Les perturbations peuvent se propager de deux manières différentes : conduite (câbles, composants passifs, blindages et plus généralement tout matériau permettant le passage d'un courant) et rayonnée (généralement dans l'air).

Les études électriques d'un système concernent généralement le courant qui circule dans les conducteurs, la tension aux bornes d'un dipôle... En CEM, elles sont menées avec une autre représentation. Cette base est utilisée pour pouvoir distinguer les différents types de perturbations : les perturbations de mode différentiel (MD) et les perturbations de mode commun (MC). Le mode différentiel concerne les phénomènes existants entre deux phases du système, par exemple entre le bus positif et le bus négatif d'un onduleur. Le mode commun concerne les perturbations circulant entre le système et sa masse électrique.

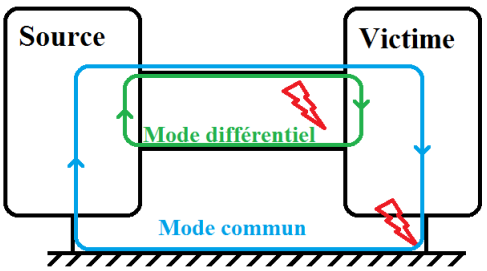


FIGURE 2.1 – Chemin de propagations Mode commun / Mode différentiel

Pour comprendre pourquoi cette représentation est possible, il est nécessaire d'expliquer com-

ment les perturbations peuvent circuler d'une phase à une autre ou entre une phase et la masse. Les chemins sont créés par les éléments parasites existants au niveau de chaque conducteur ou isolant. Ces imperfections permettent l'apparition de phénomènes de couplages inductifs ou capacitifs qui sont autant de chemins de propagation différents pour les perturbations. Il est donc très important, lors d'une modélisation, de savoir où sont ces éléments parasites et comment les représenter.

Dans le milieu aéronautique, la conception d'un point de vue CEM d'un produit est basée sur l'utilisation d'une norme choisie par les organismes de certification (. Dans le contexte de l'aviation civile, il s'agit de la norme DO-160F. Elle impose des niveaux de courants sur la gamme [150kHz ;30MHz] pour le conduit et de champ sur la bande [30MHz ;6GHz] pour le rayonné. Le respect de la norme devrait a priori permettre de ne rencontrer aucun problème de CEM mais il ne faut pas oublier que, d'un prototype initial dans un laboratoire à l'intégration finale du produit dans l'avion, l'environnement va changer et donc les perturbations aussi. Les mesures sont effectuées dans un cadre strict et un dispositif nommé "Réseau Stabilisateur d'Impédance de Ligne (RSIL)" est couplé à l'alimentation. Le convertisseur testé ne voit plus l'impédance de la source de puissance mais voit celle du RSIL qui correspond à l'impédance typique du réseau de bord d'un avion. Cela permet une reproductibilité des mesures indépendamment de l'alimentation utilisée, les mesures pouvant s'effectuer sur une sortie 50 Ω du RSIL. Il est important de noter qu'un RSIL sera calibré pour les mesures correspondantes à une norme mais qu'il est possible que ce même RSIL ne convienne pas pour une autre norme (utilisation d'un RSIL aéronautique pour une application automobile par exemple).

2.2 Modélisation prédictive

La modélisation prédictive est complexe en CEM car il s'agit de prévoir toutes les interactions qui vont exister au coeur de la chaîne électrique. Dans les simulations classiques, un modèle idéal d'interrupteur peut suffire mais si l'on souhaite connaître le comportement CEM de la chaîne, il va être nécessaire de modéliser de manière fine le semiconducteur. Il en va de même pour le reste de la chaîne : l'impédance d'un câble n'est pas la même à 10 kHz ou à 10 MHz, une capacité n'est plus capacitive au-delà d'une certaine fréquence...

De manière générale, pour étudier le comportement d'une chaîne complète, la solution consiste à modéliser indépendamment chacun des éléments de la chaîne dans un premier temps : le RSIL, les semiconducteurs, le module de puissance, l'actionneur et la connectique. L'étape suivante consiste à modéliser les interactions entre les différents éléments. Le RSIL est en théorie l'élément du système le plus simple à modéliser : il s'agit d'un dispositif calibré par le constructeur donc il est possible d'en obtenir un bon modèle jusqu'à sa fréquence maximale d'utilisation à partir des données techniques fournies. Les autres éléments (câbles, convertisseur, moteur) sont plus complexes à modéliser.

Le câble de transmission de puissance peut être modélisé de différentes manières. La méthode la plus ancienne consiste à utiliser les équations dites "des télégraphistes" établies à partir des équations de Maxwell. Les premières modélisations de câbles étaient idéales et indépendantes de la fréquence, ce qui est problématique car, par exemple, l'effet de peau aura un impact certain sur le comportement CEM du système. Des corrections ont ensuite été mises en place pour corriger ces phénomènes liés à la fréquence. Une méthode proposée par Moreau [1] consiste à mesurer les

impédances de mode commun et de mode différentiel des trois conducteurs d'une section du câble triphasé et à les modéliser ensuite en utilisant un réseau de composants R, L et C en cascade. Cette méthode est relativement simple à mettre en place mais ne permet pas de faire une modélisation prédictive au sens strict car, pour établir le modèle, un prototype du câble est au moins nécessaire. Elle permet néanmoins de pouvoir moduler la longueur du câble en variant le nombre de sections de câble dans les simulations. Doorgah [2] a utilisé une variante de cette modélisation. Il a déterminé que les perturbations de mode commun circulent en majorité dans le blindage du câble. Il en conclut que le blindage du câble ne doit pas être considéré comme idéal et qu'il faut le modéliser de la même manière que les conducteurs du triphasé. Weens [3] simule le comportement d'un câble en utilisant une modélisation à élément fini avec le logiciel FEMM. Cette méthode utilise les caractéristiques physiques (permittivité, conductivité...) et mécaniques (longueur, épaisseur...) du câble. Il souligne l'importance du positionnement des conducteurs à l'intérieur du câble qui impacte fortement sur la cohérence entre les modèles simulés et mesurés. De manière similaire, Genoulaz [4] propose de simuler le câble avec la méthode des moments surfaciques en utilisant le logiciel Comsol.

La modélisation prédictive d'un moteur électrique est un point relativement complexe. Dans de nombreuses études (Revol [5], Labrousse[6], Vermaelen [7]), le moteur est considéré comme un élément fixé de la chaîne, disponible pour effectuer des mesures d'impédance et ainsi créer un modèle comportemental. Le domaine fréquentiel de validité de ce modèle est alors directement lié à l'étendue de la plage de mesure lorsque Labrousse [6] utilise les résultats sous forme de matrices d'impédance ou à la complexité du modèle R, L, C et G des moteurs utilisés par Revol [5] ou Vermaelen [7]. L'inconvénient de cette technique de modélisation est que la mesure des impédances s'effectue avec un moteur à l'arrêt, déconnecté du reste du système, dans une position donnée. Toutes les mesures effectuées sont donc faites avec l'hypothèse que l'impédance du moteur est la même quel que soit le point de fonctionnement. Doorgah [2] utilise un modèle hybride : une partie BF prédictive, élaborée à partir des équations électromécaniques et une partie HF construite à partir des mesures d'impédance. D'autres travaux sont menés en ce moment par Boucenna [8] sur la modélisation prédictive du moteur en HF. Les premiers résultats portent sur le mode commun mais permettent déjà de mieux appréhender les chemins de propagation dans le moteur et donc d'améliorer la modélisation du moteur dans son environnement.

La modélisation prédictive du convertisseur peut être envisagée de différentes façons. Genoulaz [4] utilise un modèle simple constitué d'une résistance et d'une inductance en série à l'état passant et d'une capacité pour l'état bloqué du transistor. La simulation est ensuite effectuée en utilisant les diverses impédances selon les ordres de commandes. Une solution que propose Doorgah [2] consiste à modéliser séparément la carte du convertisseur et les semiconducteurs. La modélisation de la carte avec le logiciel InCa3D permet d'obtenir les résistances et les inductances propres et mutuelles des pistes de manière prédictive grâce à la méthode PEEC. Le logiciel ne permettant pas encore de modéliser les capacités parasites, deux solutions sont proposées :
- évaluer les capacités théoriquement à partir des grandeurs géométriques et électriques. Cela reste prédictif.
- mesurer les impédances avec un analyseur d'impédance. La modélisation est alors plus comportementale que prédictive car dans ce cas, le module existe déjà physiquement.

La modélisation des semiconducteurs est effectuée avec l'aide d'un outil intégré au logiciel SABER, le SaberModel Architect. Cet outil permet de créer un modèle du composant à partir des datasheets du fabricant. C'est là l'inconvénient de cette méthode car tous les fabricants ne

fournissent pas les mêmes données et elles sont parfois insuffisantes pour modéliser le module avec l'outil. Une autre solution consiste à étudier de manière intensive le composant et pouvoir ainsi prévoir son comportement à partir de ses caractéristiques géométriques et mécaniques. Cette approche a été étudiée par Morel et al.[9] et appliquée à un JFET SiC pour des simulations SABER. Vermaelen [7] souligne le problème de ce type de simulations qui s'effectuent dans le domaine temporel : le temps de calcul. Il propose deux autres solutions qui consistent à modéliser l'interrupteur de manière idéale (ce qui est inadapté à la CEM à cause des temps de commutation) ou à modéliser les interrupteurs à partir de sources de tension. Utiliser des sources de tensions permet d'avoir des temps de calculs courts et d'envisager des optimisations sur la chaîne. Le problème est que ces sources de tensions représentent le comportement de l'onduleur à un point de fonctionnement donné et donc, en dehors de ce point, les résultats ne seront plus aussi bons. La solution que propose Labrousse [6] pour régler ce problème consiste à modéliser non pas le composant mais l'ensemble de la cellule de commutation sous la forme d'une fonction de transfert avec, en entrée l'ordre de commutation et, en sortie la tension et comme paramètre la résistance de grille, la tension de bus ou le courant de ligne par exemple. Labrousse [6] a appliqué cette méthode à un MOSFET mais elle est envisageable sur n'importe quel type de composant. Cette méthode a notamment été approfondie par Hrigua [10].

2.3 Réduction des perturbations et filtrage

Comme cela a été précisé dans la section sur la compatibilité électromagnétique, le but des études CEM lors de la conception est de s'assurer que le produit pourra être intégré à son environnement. La solution la plus simple consiste à respecter les normes demandées par le client. Dans le cas d'un produit ne respectant pas la norme, deux solutions sont possibles. La première consiste à réduire les perturbations sur le système en changeant des paramètres qu'ils soient électriques (ralentir les fronts de commutation), géométriques (changer le positionnement des câbles) ou mécaniques (changer de matériau). Cette solution est la plus intéressante à mettre en oeuvre en aéronautique car elle se fait, en théorie, sans ajout de composant donc elle améliore la fiabilité sans augmenter ni le volume, ni la masse de la chaîne. Cependant, sa difficulté de mise en oeuvre peut décourager les ingénieurs non spécialistes. La solution vers laquelle ils se tournent le plus souvent réside dans le filtrage des perturbations entre le système "polluant" et son environnement "pollué". L'avantage du filtrage est qu'il est réalisé sur une chaîne connue et donc modélisable. Dans ce contexte et hormis pour des applications précises, les études (Foissac [11], Genoulaz [12], Jettanasen [13], Bishnoï [14]) sont faites à partir de modèles dits de "boites noires" établies de manière comportementale et qui correspondent à l'état nominal de la chaîne. Ce type de représentation rend le système beaucoup plus simple et plus rapide à optimiser. Bishnoï [14] utilise cette forme de représentation à partir d'une méthode quadripolaire pour représenter la chaîne. Il mesure les impédances et les courants dans la chaîne avec des conditions choisies (shunt, haute impédance) et il procède ensuite à des optimisations pour avoir un modèle précis jusqu'à environ 40 MHz.

Lorsque la modélisation de la chaîne est réalisée de manière précise, l'étape de conception du filtre peut être envisagée. Plusieurs techniques de filtrage existent sous la forme de 3 grandes familles : les filtres passifs, actifs et hybrides. Les filtres actifs agissent en mesurant les courants dans la chaîne électrique et en réinjectant des courants pour corriger les perturbations (Ogasawara [15]). Les filtres hybrides (Ali [16], Biela[17]) sont constitués d'une partie active et d'une partie passive, les deux parties filtrent généralement des plages de fréquences différentes. L'accent va être mis sur

les techniques de filtrage passif. Le principe d'un filtre passif est d'offrir un nouveau chemin de propagation aux perturbations. La structure la plus commune consiste à ajouter une capacité qui créé le nouveau chemin grâce à sa faible impédance sur la plage de fréquence où l'on souhaite filtrer et une inductance qui va bloquer les perturbations grâce à sa forte impédance sur cette même plage de fréquence. En pratique, le dimensionnement d'un filtre est plus compliqué. Le cahier des charges dans le domaine aéronautique est très contraignant : les matériaux doivent fonctionner sur une large gamme de température (-55°C à 200°C) et sous des conditions de pression étendues (du sol à l'altitude de croisière), plusieurs études ont été menées dans ce sens (Robutel [18], Ho[19]). Les filtres doivent également être aussi légers et petits que possible, ce qui oriente le choix de matériau vers ceux qui possèdent la plus grande densité de puissance et un fonctionnement proche de la saturation pour les matériaux magnétiques et du claquage pour les capacités par exemple. De plus, les composants utilisés pour les filtres ne sont pas parfaits, leur comportement capacitif ou inductif existe seulement sur une certaine gamme de fréquence. Au-delà, les éléments parasites existant deviennent prépondérants et l'atténuation du filtre s'en trouve réduite. Pour éviter cela, on peut alors limiter les inductances parasites par un couplage judicieux entre les conducteurs ou réduire les capacités parasites en réduisant la surface des conducteurs...

An Zhou [20] propose une méthode pour optimiser le placement de la capacité et de l'inductance du filtre. De Oliveira [21] a montré que le positionnement le plus simple dans un filtre n'est pas le plus optimisé. Il souligne également la possibilité d'améliorer le comportement du filtre en optimisant la géométrie de l'ensemble convertisseur+filtre dans une vision globale de la chaîne. Les éléments constitutifs du filtre ne sont pas toujours discrets. Mandray [22] propose par exemple de modifier la géométrie du busbar pour faire apparaître des capacités de mode commun et de mode différentiel. Robutel [18] crée des capacités de mode commun directement dans le module de puissance. Cette méthode est particulièrement intéressante car elle permet de réduire au minimum les chemins de propagation des perturbations HF et donc les perturbations qu'il reste à filtrer.

2.4 Création des outils ingénieurs

Les sections précédentes ont montré que les études CEM sont complexes et longues à mener. Un des buts de la thèse est de construire des outils qui devront permettre à l'ingénieur de prédire le comportement d'une chaîne et si nécessaire d'élaborer les filtres de MC et de MD en cas de perturbations trop importantes. Toure [23] propose une méthode générique à partir d'une interface logicielle qui se veut modulable grâce à des bibliothèques de composants et de sources de perturbations évolutives. Les filtres sont orientés Basse Fréquence mais les aspects Haute Fréquence ne sont pas oubliés. L'auteur souligne l'importance des protocoles d'optimisation pour avoir des calculs rapides. L'inconvénient de cette méthode généraliste est que la structure est fixée. Les améliorations du filtre citées précédemment ne peuvent ainsi pas être appliquées. Certains aspects comme la fréquence de commutation ou les matériaux utilisés sont notamment fixés.

Chapitre 3

Présentation de la chaîne électrique étudiée

3.1 Chaîne électrique existante

Le système étudié (schéma simplifié figure 3.1) est composé d'un onduleur triphasé, d'un harnais et d'un moteur. L'ensemble de la chaîne est alimenté par un réseau HVDC de 540 Vdc. L'onduleur fonctionne en boucle ouverte et est commandé par un ordre de fréquence de rotation du moteur. Le harnais de 2m est blindé et contient les 3 conducteurs du triphasé. Le moteur est une machine synchrone double étoile à aimant permanent.

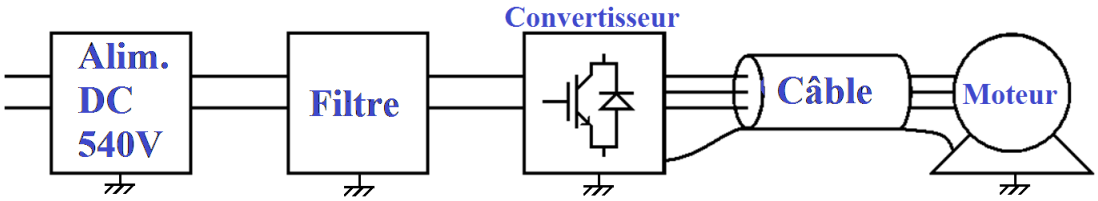


FIGURE 3.1 – Chaîne électrique simplifiée

Le module de puissance utilisé a été développé par Microsemi, il s'agit de l'APTGT25x120t3G. C'est un module triphasé bidirectionnel qui peut fonctionner jusqu'à 1200V avec 25A à 80°C. Il contient 6 IGBT avec leurs diodes antiparallèles. Ce module est avantageux car il est facilement intégrable sur un PCB, a des inductances parasites minimisées et fonctionne jusqu'à une fréquence de commutation de 20 kHz. Les diodes et les IGBT sont caractérisés à partir de la datasheet de ce module. Ce module est piloté par 6 drivers qui étaient utilisés sur d'anciens projets de Label Power System. Les 6 drivers sont eux commandés par une carte de commande spécifique. La commande réalisée est une modulation par largeur d'impulsion vectorielle en boucle ouverte qui permet le pilotage du moteur à la fréquence de rotation souhaitée via un rapport tension/fréquence constant. La fréquence de commutation des IGBT est fixée par la carte à 7,5 kHz ou 15 kHz. Les drivers sont reliés au module de puissance via un circuit imprimé et des fils de connexion. Une capacité de découplage de 10 μF est placée sur le PCB de la carte de puissance au plus proche du module de puissance.

Au niveau du circuit de puissance, le module est relié aux RSIL via des fils non blindés d'une vingtaine de centimètres du côté réseau du convertisseur. La connexion entre la sortie du convertisseur et la connectique du câble de puissance est faite avec trois câbles non blindés également. L'absence de blindage de ces fils est nécessaire pour pouvoir faire les mesures des courants de mode commun et de mode différentiel aux extrémités de l'onduleur à l'aide de sondes de courant. Les sondes utilisées sont une F-51 de Fisher Custom Communication avec une bande passante de 10 kHz à 500 MHz et une 6741-1 de Solar Electronic avec une bande passante de 10 kHz à 100 MHz. Les RSIL sont fabriqués par Fischer Custom Communication, le modèle utilisé est le FCC-LISN-5-10-1-01-DO-160. La valeur des capacités côté réseau a changé, il est donc nécessaire d'ajouter une capacité à l'extérieur du boitier du RSIL pour avoir la valeur de capacité souhaitée. Le point de masse du RSIL est connecté à la masse globale du système : le plan de masse en cuivre.

Le moteur utilisé a été développé lors de précédents projets de Labinal Power Systems. Sa tension nominale est de 540V, le courant maximal admissible en continu est de 10A, il peut tourner jusqu'à 4000 tours par minute et développe un couple nominal de 4,4Nm. L'utilisation d'un moteur double étoile permet une réduction de l'ondulation du couple et l'utilisation d'aimants permanents au niveau du rotor permet de concentrer la majorité des pertes au niveau des enroulements du stator, d'où un refroidissement plus facile.

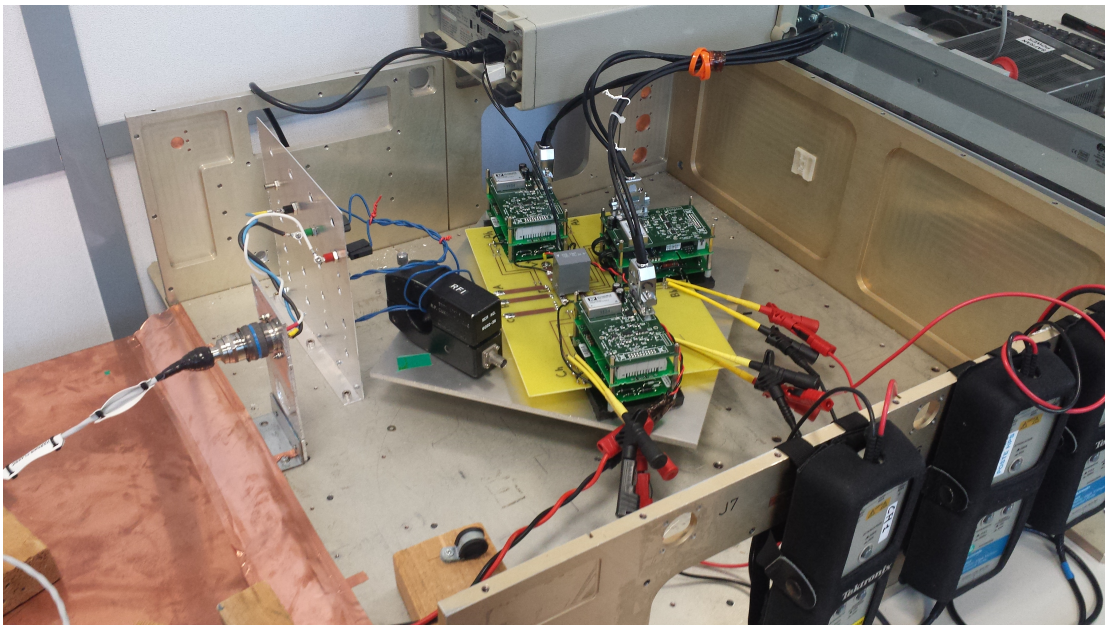


FIGURE 3.2 – Convertisseur utilisé

3.2 Simulations réalisées sur SABER

Ce type de chaîne électrique a déjà été l'objet de précédentes recherches par Doorgah [2]. Le système étudié a été modélisé à l'aide du logiciel SABER. Selon l'auteur, la modélisation est bonne jusqu'à 20 MHz. Au-delà de ces fréquences, la modélisation des sous-systèmes doit être améliorée, notamment celle du transistor et celle de la machine synchrone double étoile (MSDE). Il souligne également l'impact des algorithmes de calculs internes au logiciel de calcul de circuit qui peuvent

avoir une influence sur les résultats obtenus par simulation. Ces travaux constitueront la base des études qui vont être menées par la suite afin de ne pas repartir de zéro. Dans un premier temps, nous analyserons les techniques existantes pour modéliser les sous-systèmes et, dans un second temps, ces techniques seront utilisées pour améliorer la précision de la modélisation et étendre la bande de fréquence de validité du système.

La source de puissance est modélisé à partir d'une source idéale de tension. Les RSIL sont représentés à partir des données constructeurs qui ont été validées par la mesure. La connexion entre les RSIL et le module de puissance est réalisée par une association série d'une résistance de 60 mΩ et d'une inductance de 110nH sur chaque bus (valeurs obtenues par une mesure). La capacité de découplage est placée en parallèle du bloc convertisseur, les éléments parasites de cette connexion sont présents dans ce bloc. La capacité est représentée avec ces éléments parasites : résistance et inductance (figure 3.3).

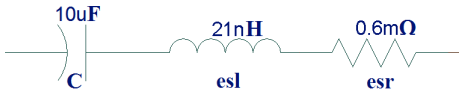


FIGURE 3.3 – Modèle de la capacité de découplage

La modélisation du module de puissance se fait avec des inductances parasites entre les différents éléments constitutifs de l'onduleur triphasé : les 6 IGBT et les 6 diodes (figure 3.4). Les semiconducteurs sont modélisés via le SaberModelArchitect.

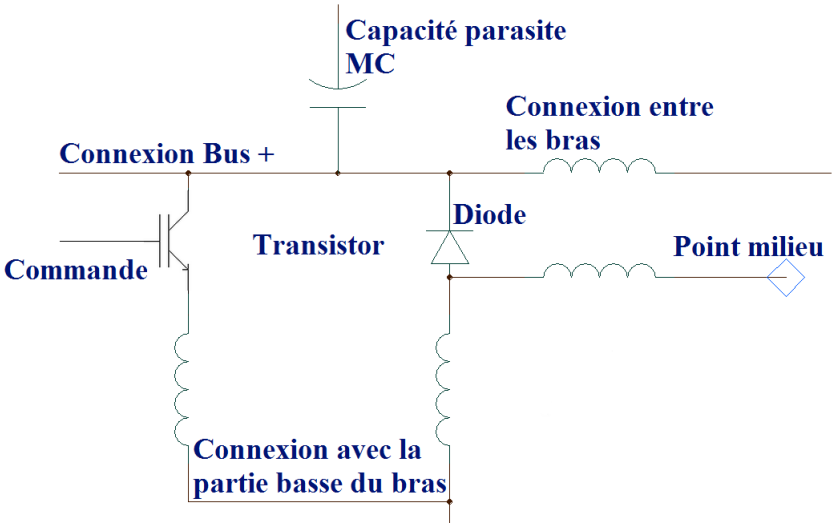


FIGURE 3.4 – Modélisation de la partie haute d'un des bras de l'onduleur

La modélisation du câble de 2 mètres est faite à partir de 10 cellules de 20 centimètres. On utilise une association cellules de longueur très inférieure à la longueur d'onde de la fréquence maximale considérée pour tenir compte des phénomènes de propagation dans le câble. Dans la cellule, chaque conducteur est représenté par une association résistance-inductance (Rp et Lp pour les phases, Rb et Lb pour le blindage). Les intéractions entre chaque phase sont représentées par des capacités Cpp

et des mutuelles Mpp et entre les phases et le blindage par des capacités Cpb et des mutuelles Mpb.

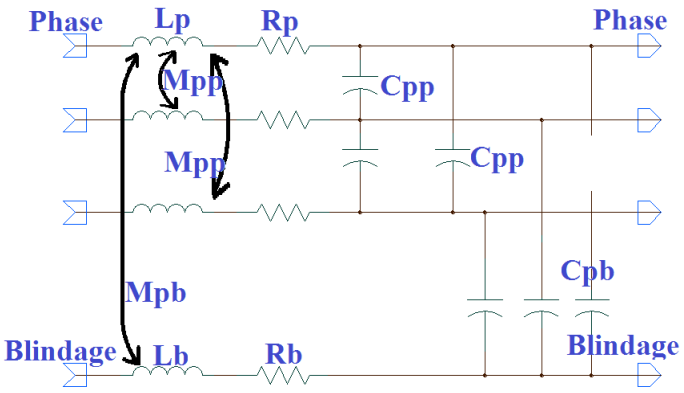


FIGURE 3.5 – Modèle d'une cellule du câble

Le moteur est représenté avec deux partie distinctes (figure 3.6).La partie basse fréquence est modélisée par la résistance r et l'inductance l donnée par le constructeur associés à la force électromotrice calculée à partir des équations mécaniques. La partie HF est modélisée par 2 impédances de mode commun Zmc et une impédance de mode différentiel Zmd. Ces impédances sont déterminées par des mesures effectuées sur le moteur.

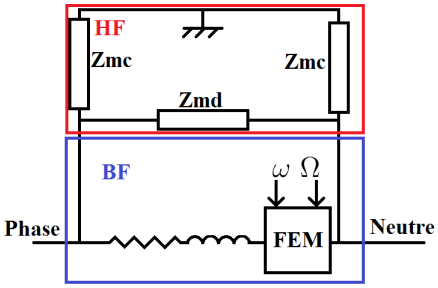


FIGURE 3.6 – Modèle simplifié du moteur utilisé

3.3 Choix des paramètres de simulations

Le calcul des grandeurs sous SABER se fait grâce à un Algorithme dit de Newton-Raphson en fonctionnement normal. Pour la majorité des points, il n'y a pas de problème de convergence lors du calcul mais parfois, pour certaines commutations, les résultats divergent. Ces divergences apparaissent le plus souvent au moment des commutations et bien que ces phénomènes soient reproductibles, il n'est pas possible de savoir à l'avance si la simulation va converger. Pour résoudre ce problème, il faut arrêter la simulation avant ces points problématiques et relancer une nouvelle simulation à partir du dernier point calculé. Cela permet de calculer la zone problématique avec des points différents et donc de converger.

La durée totale des simulations est de 50 ms pour atteindre le régime permanent. En effet, malgré l'utilisation de conditions initiales proches des valeurs du régime final au niveau des in-

ductances et des capacités, il existe toujours des phénomènes transitoires. Ce problème devra être résolu pour pouvoir lancer les protocoles d'optimisation. Le moteur fonctionnant à 100 Hz lors des simulations, la dernière simulation doit durer au minimum 10 ms pour avoir une période complète. Le pas de temps utilisé est variable mais le maximum est fixé à 10 ns pour avoir des spectres corrects jusqu'à 50 MHz (théorème de Shannon-Nyquist). Lorsque les calculs sont terminés, les fichiers de points sont écrits dans des fichiers textes puis interpolés sur Matlab pour avoir un nombre de points équivalent à 2^n sur un nombre entier de périodes électriques (dans notre cas, une simulation de 10 ms avec une période d'échantillonnage de 10 ns représente 100k points donc il faut interpoler le signal à 2^{17} soit 131072 points). Ce nombre de points permet une meilleure précision lors du calcul de la transformée de Fourier du signal temporel. De plus, le pas de temps n'est pas constant lors des simulations, il est donc nécessaire de rééchantillonner le signal temporel à pas constant.

Chapitre 4

Etude paramétrique de la source de bruit

4.1 Contexte de l'étude

L'état de l'art a mis en avant l'importance de la modélisation de la source de bruit. Les premiers travaux de cette thèse ont été menés dans ce sens. L'objectif des simulations et des mesures est de pouvoir identifier quel est l'impact de différents paramètres du convertisseur et de la commande sur les perturbations électromagnétiques. Nous pourrons alors envisager une modélisation sous forme de "boite noire" du composant ou de la cellule de commutation comme cela a été mené dans les travaux de Labrousse [6] et de Hrigua [10] par exemple. Les travaux devront permettre de trouver une façon simple de tirer les paramètres constitutifs de la représentation à partir d'une mesure ou des datasheets des composants. Costa [24] propose de faire l'étude d'une tension simplifiée sous la forme d'un trapèze. Cet article souligne l'impact de la fréquence de commutation et des temps de montée sur les harmoniques de la tension. Les paramètres étudiés sont les suivants :

- Fréquence de commutation (1/T)
- Rapport cyclique (α)
- Temps de montée (Tmontée)
- Temps de descente (Tdescente)
- Amplitude du signal (V)
- Ordre de la fréquence électrique du moteur dans le cas de la mesure (Hz)

Dans le cadre de mes travaux de thèse, des calculs ont été effectués sur un trapèze théorique avec Matlab, sur un hacheur simple sans les éléments parasites avec SABER. Des mesures ont été effectuées en faisant commuter un bras de l'onduleur à vide en mode hacheur et en mode onduleur.

4.2 Signaux étudiés

4.2.1 Etude théorique

Le signal étudié représente la tension aux bornes d'un interrupteur qui commute. Cette tension est représentée par un trapèze : l'état bas vaut 0V et l'état haut vaut 200V, les obliques du trapèze représentent les temps de montée et de descente et la largeur du trapèze représente le rapport cyclique (figure 4.1). On va utiliser une source de bruit témoin qui sera la base des calculs qui vont suivre : les paramètres étudiés varieront autour des valeurs de cette source témoin.

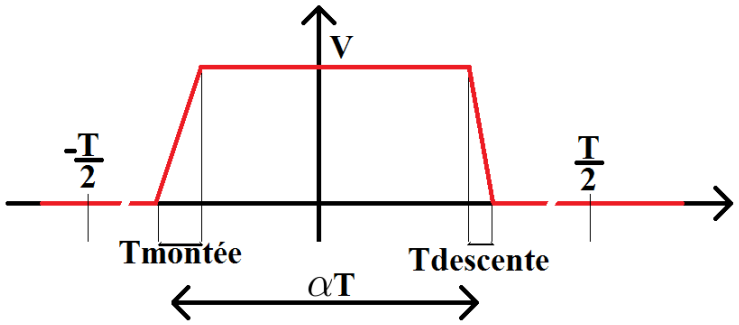


FIGURE 4.1 – Modèle du signal étudié

Les paramètres de la source de bruit témoin sont donc les suivants :
- Fréquence de commutation (f_{com}=1/T) : 10 kHz
- Rapport cyclique (α) : 1/2
- Temps de montée (Tmontée) : 100 ns
- Temps de descente (Tdescente) : 100 ns
- Amplitude du signal (V) : 200 V

Ce signal est créé sur Matlab avec les paramètres sélectionnés. Le passage en fréquentiel est assuré via une FFT avec le nombre de points suffisant pour que la bande de fréquence de validité s'étende au moins jusqu'à 50 MHz (en réalité, le spectre s'étend au-delà car on calcule la FFT à partir d'un signal temporel composé de 2^n points). Le spectre obtenu est ensuite traité de manière à obtenir seulement l'enveloppe du spectre. Le fichier Matlab utilisé est contenu dans l'annexe 1.

4.2.2 Calculs avec SABER

Nous étudions la tension aux bornes d'un transistor dans un environnement plus complexe (figure 4.2). Nous utilisons un hacheur Boost que nous faisons avec un rapport cyclique α à une fréquence fcom. La valeur de l'inductance est fixée à 5mH, la capacité est de 200 μF et la résistance de 100 Ω. La diode utilisée est une diode idéale (elle conduit parfaitement dans un sens, pas du tout dans l'autre).

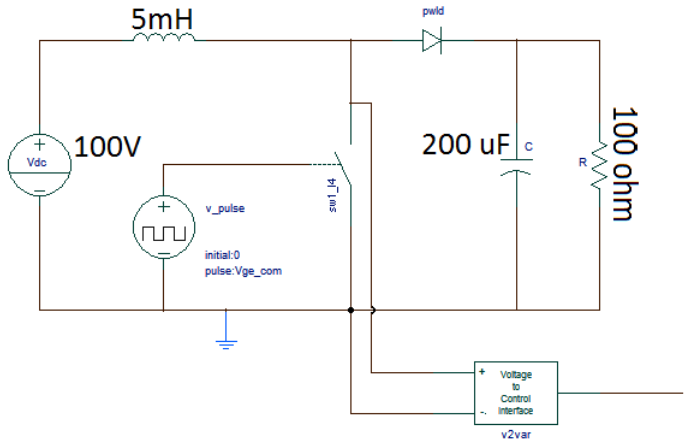


FIGURE 4.2 – Schéma du circuit électrique étudié sous SABER

On dispose des mêmes paramètres d'étude que dans l'étude théorique. Les paramètres du hacheur Boost "témoin" sont les suivants :

- Fréquence de commutation (1/T) : 10 kHz
- Rapport cyclique (α) : 1/2
- Temps de montée (Tmontée) : 100 ns
- Temps de descente (Tdescente) : 100 ns
- Amplitude du signal (V) : 200 V

On obtient via SABER des fichiers de points contenant les allures des tensions en temporel. Ces fichiers de points sont ensuite interpolés sous Matlab et on obtient finalement les spectres des tensions.

4.2.3 Mesures effectuées sur le banc d'essai

La mesure est effectuée en commutant un seul bras. La carte de commande permet de choisir la fréquence de commutation, le rapport cyclique dans le cas d'un fonctionnement en mode hacheur et la fréquence électrique du fondamental dans le cas d'un fonctionnement en mode onduleur. L'installation est faite avec un module driver sur la phase A.

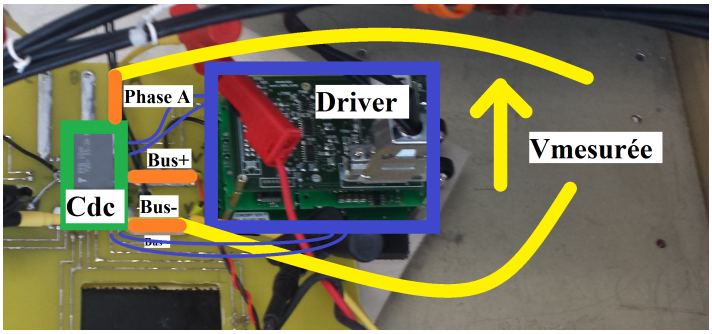


FIGURE 4.3 – Convertisseur utilisé pendant les essais

Les temps de commutation sont réglés par l'ajout d'une résistance supplémentaire de grille à l'extérieur du module de driver (figure 4.4). Deux résistances de grille existent déjà dans la carte driver : une résistance d'enclenchement de 1.3 Ω et une résistance de déclenchement de 10 Ω.

4.2.4 Comparaison entre les différentes méthodes

Avant de commencer l'étude paramétrique, nous avons vérifié que dans les 3 cas les signaux ont la même forme. Nous nous sommes placés dans le cas où la tension au bornes du transistor est de 100 V, la fréquence de découpage est de 7,5 kHz, les temps de commutation d'environ 10 ns et le rapport cyclique égal à 0,5. Les calculs théoriques et les simulations effectuées avec le hacheur donnent des résultats similaires. On note qu'à partir de la dizaine de MHz, le spectre de la tension mesurée n'a plus la forme des deux autres. Il y a donc des phénomènes qui n'ont pas été pris en compte dans les calculs et les simulations.

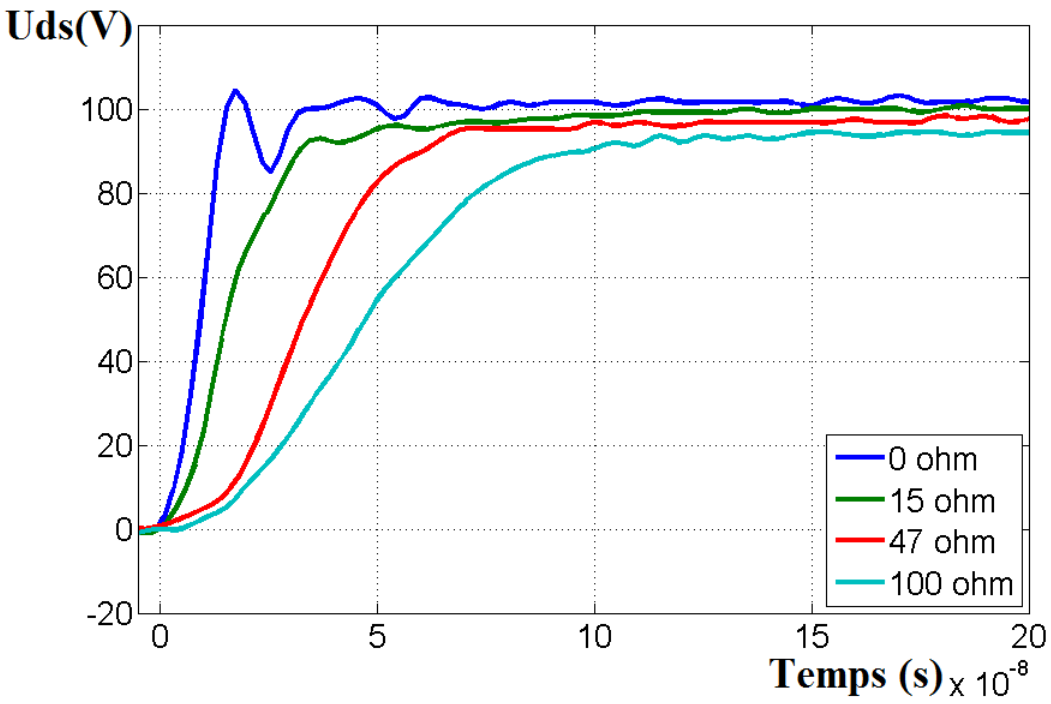


FIGURE 4.4 – Impact de la résistance de grille additionnelle

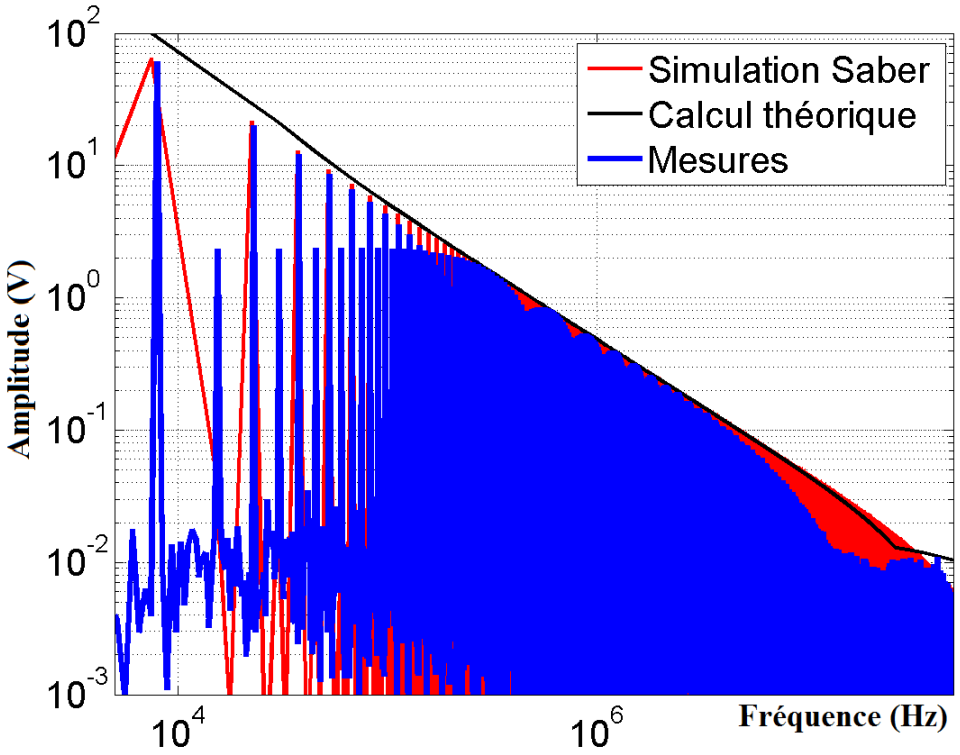


FIGURE 4.5 – Comparaison des spectres avec les 3 méthodes

4.3 Etude de l'influence de la fréquence de commutation

Ce paragraphe porte sur l'étude de l'influence de la fréquence de commutation sur la source de bruit.

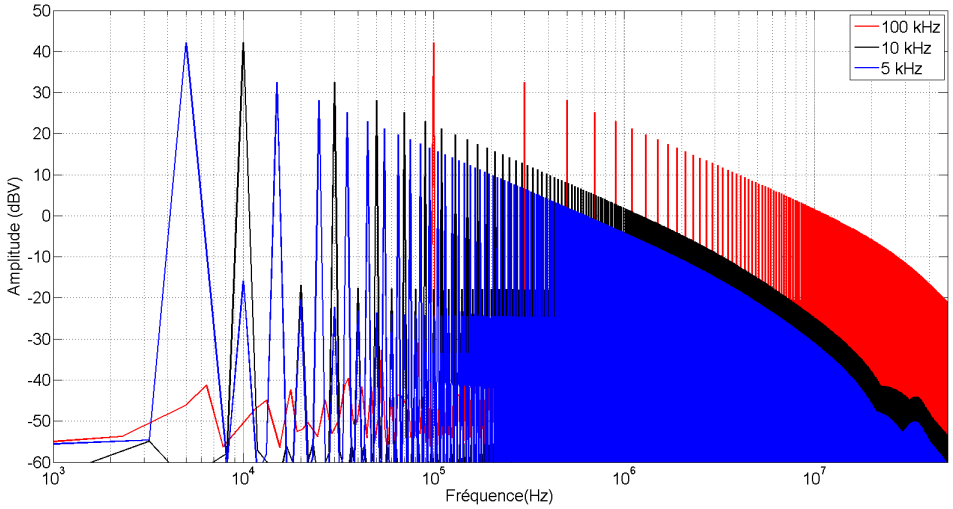


FIGURE 4.6 – Impact de la fréquence de commutation (obtenu sous SABER)

Les premières harmoniques des 3 signaux ont la même valeur. La décroissance observée est de -20dB par décade jusqu'à une fréquence donnée qui est la même pour chacune des tensions. Cette valeur de pente correspond exactement à celle du spectre d'un signal créneau parfait, c'est-à-dire avec des temps de commutation nuls. Les tensions mesurées sur le convertisseur en mode hacheur (Figure 4.7) corroborent ces observations. Pour mieux exposer ce phénomène, le spectre du signal commutant à 7,5 kHz a été décalé sur l'axe des fréquences.

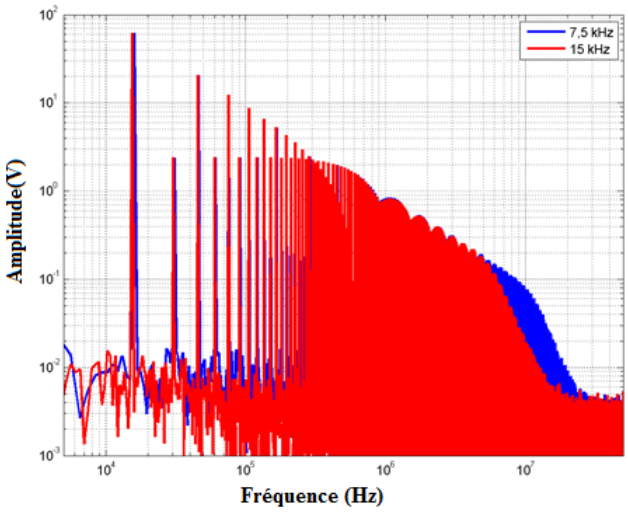


FIGURE 4.7 – Influence de la fréquence de commutation (mesures)

4.4 Etude de l'influence des temps de commutation

Ce paragraphe porte sur l'étude de l'influence des temps de commutation sur la source de bruit. Dans un premier temps, les temps de montée et de descente sont égaux et dans un second temps, nous étudions l'impact d'une dysimétrie entre ces deux temps.

4.4.1 Temps de commutation égaux

L'observation du spectre de la tension calculée avec Matlab montre une rupture de pente à 310 kHz. Avant cette fréquence, la pente est de -20 dB/décade comme observé précédemment mais après cette fréquence la pente est de -40 dB/décade. Les calculs théoriques indiquent que cette fréquence est inversement proportionnelle au produit de π et du temps de commutation.

$$f_{cassure} = \frac{1}{\pi * T_{commutation}} \tag{4.1}$$

Dans notre cas, le temps de montée est de 1μs et donc la fréquence de cassure est de 318 kHz, cela correspond bien à ce que nous observons sur le spectre (figure 4.8).

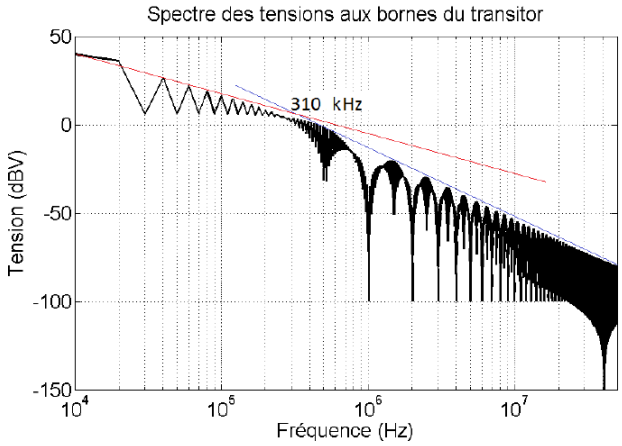


FIGURE 4.8 – Asymptote du spectre à 10 kHz

L'observation des spectres indique que plus les temps de commutation sont longs, plus les fréquences de cassures sont faibles et donc plus l'amplitude en HF est faible par rapport à un signal possédant des temps de commutation plus rapides. Cette observation n'est pas étonnante quand on sait que des temps de commutation plus rapides induisent des variations de tensions plus rapides et donc plus de perturbations.

Le temps de montée qui est utilisé dans SABER est différent de celui utilisé dans les calculs théoriques. Les temps qui sont renseignés dans SABER correspondent à un délai et au temps de commutation effectif. Pour les simulations qui vont être présentées, nous avons utilisé les temps de montée et de descente qui permettent d'avoir l'allure souhaitée en temporel (figure 4.9). Nous allons étudier ce problème pour comprendre quelle en est la cause et quelles sont les solutions envisageables dans le logiciel SABER.

Nous observons les spectres des signaux avec les différents temps de commutation (environ 1 μs, 100 ns, 10 ns sur la figure 4.10). Les remarques faites sur les calculs théoriques sont ici confirmées.

On voit que l'influence du temps de montée s'exerce après une certaine fréquence et que celle-ci dépend du temps de montée. Avant cette fréquence, la décroissance est de -20 dBV/décade et après, la décroissance est de -40 dBV/décade. Nous observons bien que la fréquence de coupure est entre 300 kHz et 500 kHz. Il faut néanmoins noter que le signal n'a pas strictement la forme d'un trapèze. C'est pourquoi il est donc intéressant de voir quel est l'impact de la discontinuité qui est créée au début et à la fin de la commutation.

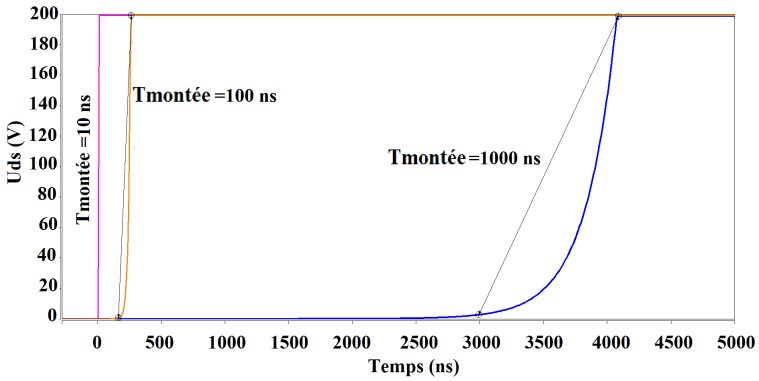


FIGURE 4.9 – Tension drain-source lors des commutations du transistor

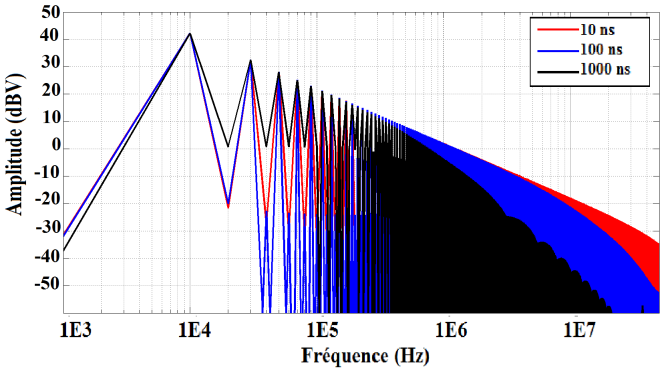


FIGURE 4.10 – Influence des temps de commutation (noir : 1000 ns, bleu : 100 ns, rouge : 10 ns)

4.4.2 Temps de commutation différents

Nous étudions les spectres de la tension aux bornes du transistor obtenus en introduisant une différence entre les temps de montée et de descente. Dans notre cas, le temps de montée est fixé et le temps de descente est paramétrisé de la manière suivante (nous obtenons les mêmes résultats si on paramètrise le temps de montée ou le temps de descente) :

- Condition 1 : 10 ns
- Condition 2 : 100 ns
- Condition 3 : 1000 ns

La différence est relativement importante par rapport au régime de fonctionnement normal d'un transistor mais cela permet de mieux observer l'impact de ce paramètre. Le rapport cyclique est égal à 1/2, l'amplitude du signal temporel est fixée à 200V, la fréquence de commutation est de 10 kHz et le temps de montée est égal à 100 ns.

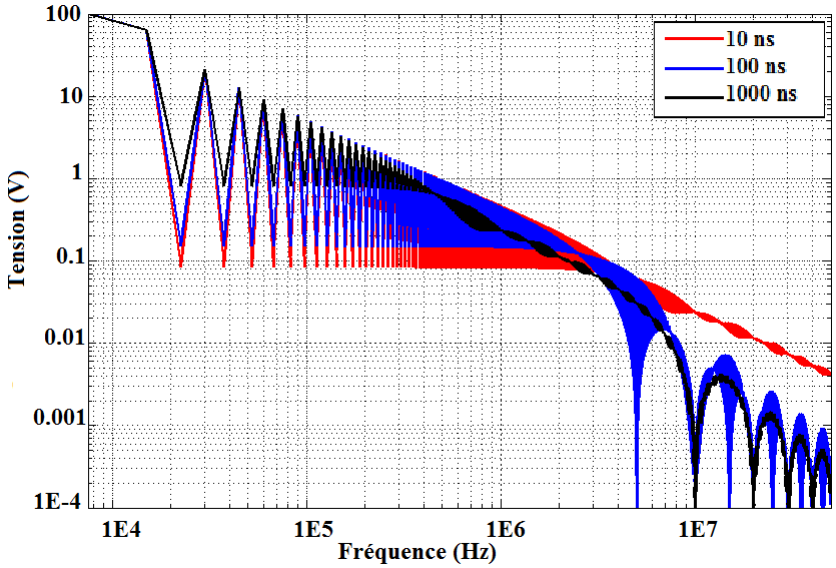


FIGURE 4.11 – Influence de la différence entre les temps de montée et de descente (noir : 1000ns, bleu : 100ns, rouge : 10ns)

On note que le niveau du spectre est déterminé par le temps de commutation le plus rapide entre le temps de montée et le temps de descente (figure 4.11). En effet, le signal possédant le temps de descente le plus rapide (condition 1) est plus perturbé que le signal témoin (condition 2) et le signal possédant le temps de descente le plus lent (condition 3) : entre 8 et 30 dBV de plus. On remarque de plus que, bien que le signal (condition 3) soit moins perturbé que le signal témoin, cette différence avec le signal témoin (condition 2) est beaucoup moins marquée qu'avec la condition 1.

4.4.3 Mesures effectuées

Pour modifier la vitesse de commutation des IGBT, il faut modifier la résistance de grille au niveau de la commande. L'augmentation de la résistance va réduire le courant de charge dans le composant et donc ralentir la commutation. Lors des essais, les valeurs de résistances ajoutées sont de 0 Ω, 15 Ω, 47 Ω et 100 Ω. Cela correspond à des temps de commutation de respectivement 10 ns, 25 ns, 45 ns et 80ns environ.

L'observation des spectres (figure 4.12) indique que la vitesse de commutation et donc la résistance de grille impactent sur l'allure des spectres. Le ralentissement de la commutation diminue bien la fréquence de coupure mais la décroissance du spectre n'est pas exactement de -40 dB/décade contrairement à la théorie. Le signal temporel n'est pas exactement un trapèze : les ruptures de pentes sont moins marquées et la tension lors de la commutation temporelle n'est pas de la forme d'une rampe. Dans ce papier de Costa [24], il est montré que l'adoucissement des ruptures de pentes change la forme du spectre. Cela explique peut-être la forme des spectres obtenus par la mesure.

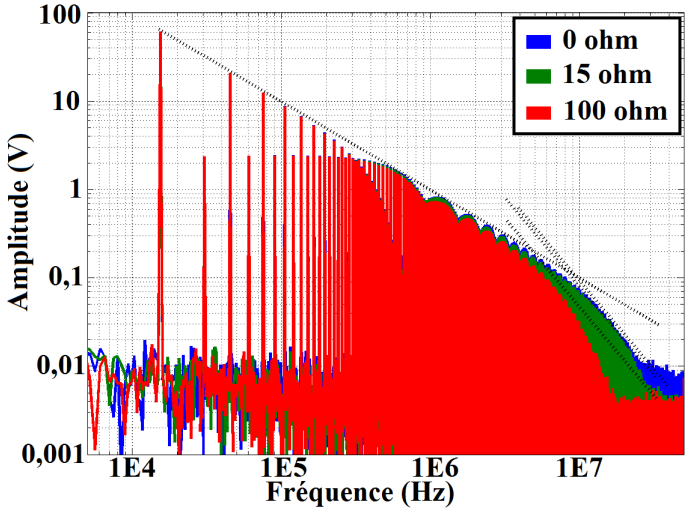


FIGURE 4.12 – Spectres des tensions Uds avec résistances de grilles différentes 0 Ω, 15 Ω et 100 Ω

4.5 Etude de l'influence du rapport cyclique

4.5.1 Impact théorique

On observe les spectres de la tension aux bornes du transistor obtenus en faisant varier le rapport cyclique.

- Condition 1 : 0,25
- Condition 2 : 0,5
- Condition 3 : 0,75

L'amplitude du signal temporel est fixée à 200V, la fréquence de commutation est de 10 kHz et les temps de commutation sont égaux à 100 ns.

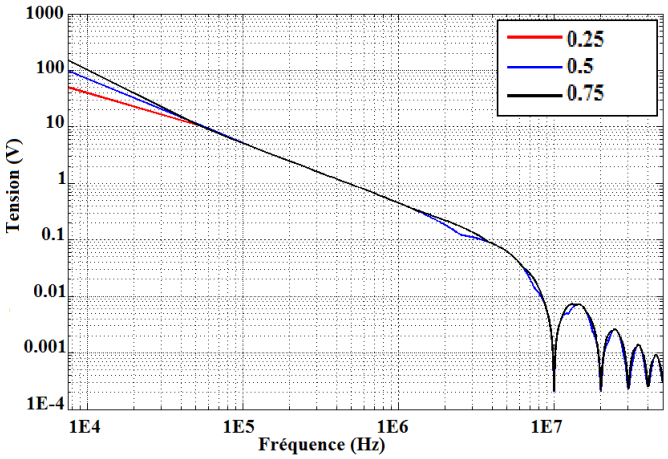


FIGURE 4.13 – Influence du rapport cyclique

Nous n'observons presque aucune différence entre les spectres mesurés sur la gamme de fréquence étudiée (figure 4.13). La seule différence est qu'avec un rapport cyclique de 0,5, il y a un signal qui commute au niveau de 20 kHz ($2^{*}f_{com}$).

4.5.2 Tensions mesurées

L'utilisation du convertisseur en mode hacheur permet de choisir le rapport cyclique à appliquer. Les mesures ont donc été effectuées avec des rapports de 0,10, 0,30, 0,50, 0,70 et 0,90. La figure 4.14 montre qu'en Basse Fréquence (jusqu'à quelques MHz), il n'y a effectivement pas de différence entre les différentes tensions. En Haute Fréquence (au-delà de 10MHz), on note une différence de niveau. Il semble que le fonctionnement avec un rapport cyclique proche de 0,50 rejette plus de perturbations que le fonctionnement à 0,10.

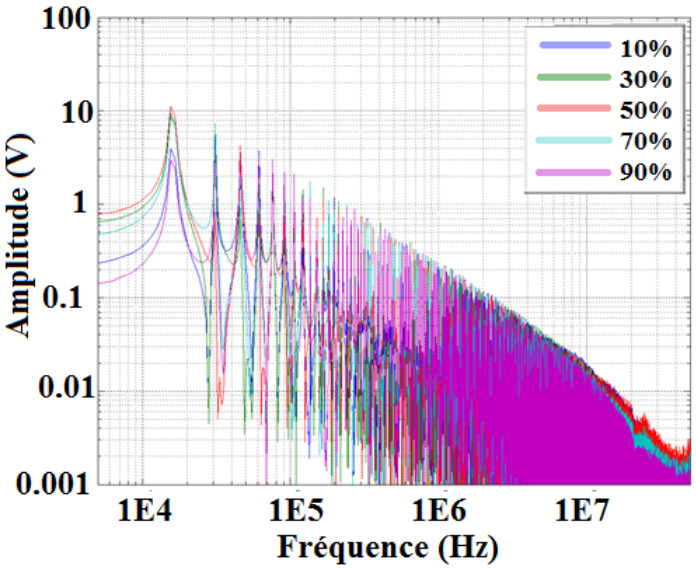


FIGURE 4.14 – Spectres des tensions mesurées avec différents rapports cycliques

4.6 Conclusions

Les deux paramètres étudiés les plus impactants sont la fréquence de commutation et le temps de commutation : le temps de commutation impacte sur le comportement HF du spectre alors que la fréquence de commutation impacte de deux manières différentes. En basse fréquence, la fréquence de commutation va décaler les harmoniques mais elles vont garder la même amplitude. Au-delà de la fréquence de coupure (qui est fonction du temps de commutation), la forme du spectre est la même mais elle est décalée d'une constante égale à $20 * log(f_{com2}/f_{com1})$. Lorsque le temps de montée et le temps de descente sont différents, c'est la durée la plus courte qui va imposer la forme du spectre. Par exemple, si le temps de montée est de 1 μs et que le temps de descente est de 100 ns, la fréquence de coupure du spectre sera de 3,18 MHz et non de 318 kHz dans le cas d'un signal commutant à 1 μs. En théorie, le rapport cyclique ne semble pas impacter sur la forme du spectre de la tension. En revanche les mesures réalisées font apparaître que ce n'est pas exactement le cas dans la réalité.

Chapitre 5

Perspectives

5.1 Simulations

Une première approche théorique a été effectuée et, pour confirmer ces aspects théoriques, nous avons réalisé des mesures. Il est maintenant important de vérifier que le modèle existant sur SABER se comporte dans le sens de ce qui a été observé. Nous avons déterminé plusieurs objectifs à court terme :

- Diminution de la durée de simulation via la réduction des phénomènes transitoires ;
- Meilleure modélisation de la source de perturbation (amélioration de la stabilité, de la précision...) ;
- Analyse du modèle du moteur (Utilité de la partie Basse Fréquence).

A moyen terme, il va falloir comprendre quels sont les éléments importants du modèle. Il s'agit notamment de comprendre quelle gamme de fréquence est impactée lorsque nous simulons la chaîne sans modéliser la carte de puissance et quel est l'impact sur les niveaux de perturbations. Le paramètre température va aussi être intégré dans la chaîne dans le but de comprendre son importance vis-à-vis des perturbations CEM. Nous allons par exemple modéliser la variation de capacité en fonction de la température pour la capacité de découplage ou faire varier les temps de commutation des transistor en fonction de ce paramètre. Ces simulations seront couplées à des mesures effectuées sur la chaîne pour comprendre où sont les divergences entre les résultats et pouvoir les limiter voire les supprimer. Enfin, l'objectif final est d'avoir un modèle qui réagit correctement jusqu'à 50 MHz. Un second modèle plus simple est envisagé afin de réaliser les optimisations nécessaires dans le cadre des outils ingénieurs. Ce modèle devra être précis dans la gamme de fréquence où le besoin d'atténuation est critique mais sera simplifié dans les autres zones grâce aux études paramétriques réalisées.

5.2 Mesures

L'objectif à court terme va être d'analyser le comportement du module IGBT dans le cadre d'un onduleur monophasé. La charge utilisée devra être modélisée et avoir un comportement fréquentiel sans trop de résonances pour simplifier le modèle qui sera réalisé sous SABER. Dans un premier temps, cette étude devrait nous permettre d'analyser l'impact du paramètre courant dans la charge et donc d'améliorer notre connaissance du module. Dans un second temps, l'impact des résistances de grilles sur les perturbations CEM et sur les pertes thermiques au niveau de l'onduleur monophasé sera étudié. Ces mesures vont permettre de procéder à l'optimisation globale de la chaîne d'un point

de vue CEM et notamment de comprendre si l'approche qui consiste à privilégier l'optimisation thermique via la masse du dissipateur par rapport à l'optimisation CEM via la masse du filtre est à nuancer. Les résultats obtenus seront vérifiés sur la chaîne complète avec l'onduleur triphasé à IGBT et sur un autre onduleur, cette fois à base de JFET. Des mesures pour étudier le paramètre température sont également prévues : l'étude va être cantonnée au module de puissance et à la capacité de découplage. Pour fixer le paramètre température, deux options sont envisagées : mettre le convertisseur dans une enceinte de type étuve ou envoyer un flux d'air chaud sur le convertisseur avec une girafe.

5.3 Création de l'outil ingénieur

A moyen terme, les paramètres influents sur le modèle prédictif devraient être connus. Cela permettra de simplifier le modèle donc de procéder à l'optimisation de la chaîne en terme de CEM. Un outil de création de filtre complètera cette optimisation. Il sera réalisé à partir de bases de données existantes à Labinal Power Systems. Ces bases de données devraient permettre de choisir la structure du filtre, ses éléments constitutifs et potentiellement le type de câble à utiliser. L'objectif final est de pouvoir prédire le comportement de la chaîne avec les éléments les plus simples possibles et de guider l'ingénieur qui souhaite améliorer son produit (cas d'un système existant) ou dimensionner un système complet avec des performances CEM acceptables dès les premières réalisations.

Bibliographie

[1] Maxime Moreau, Nadir Idir, and Philippe Le Moigne. Modeling of Conducted EMI in Adjustable Speed Drives. *IEEE Transactions on Electromagnetic Compatibility*, 51(3) :665–672, August 2009.

[2] Naraindranath Doorgah. *Contribution à la modélisation prédictive CEM d'une chaine d'entrainement*. PhD thesis, Université de Lyon, 2012.

[3] Yannick Weens. *Modélisation des câbles d'énergie soumis aux contraintes générées par les convertisseurs électroniques de puissance*. PhD thesis, Université des Sciences et Technologies de Lille, 2006.

[4] Jerome Genoulaz. *Contribution à l'Étude du Rayonnement des Câbles Soumis aux Signaux de l'Électronique de Puissance dans un Environnement Aéronautique*. PhD thesis, Université des Sciences et Technologies de Lille, 2008.

[5] Bertrand Revol. *Modélisation et optimisation des performances CEM d'une association variateur de vitesse -machine asynchrone*. PhD thesis, Université Joseph Fourier, 2004.

[6] Denis Labrousse. *Amélioration des techniques d'estimation des perturbations conduites. Application à une chaîne de traction de véhicule électrique*. PhD thesis, Ecole Normale Supérieure de Cachan, 2011.

[7] C Vermaelen. *Contribution à la modélisation et à la réduction des perturbations conduites dans les systèmes d'entraînement à vitesse variable*. PhD thesis, Ecole Normale Supérieure de Cachan, 2003.

[8] Nidhal Boucenna. Etude des chemins de propagation des courants de mode commun dans les parties métalliques des machines à induction. In *JCGE 2013*. SATIE, ENS Cachan, CNRS, 2012.

[9] H. Morel, Y. Hamieh, D. Tournier, R. Robutel, F. Dubois, D. Risaletto, C. Martin, D. Bergogne, C. Buttay, and R. Meuret. A multi-physics model of the vjfet with a lateral channel. pages 1–10, 2011.

[10] Slim Hrigua. Modélisation par fonctions de transfert du comportement transitoire d'une cellule de commutation équipée de semi- conducteurs SiC. In *JCGE 2013*, Cachan, 2013.

[11] Mikael Foissac. "Black box" EMC model for power electronics converter. *Energy Conversion Congress and Expo*, pages 3609–3615, 2009.

[12] Jerome Genoulaz, Chaiyan Jettanasen, François Costa, and Christian Vollaire. Modeling of common mode conducted noise emissions in PWM inverter-fed AC motor drive systems. *Power Electronics and ...*, 2007.

[13] Chaiyan Jettanasen. *Modélisation par approche quadripolaire des courants de mode commun dans les associations convertisseurs-machines en aéronautique ; optimisation du filtrage*. PhD thesis, 2008.

[14] Hemant Bishnoi, Andrew Carson Baisden, Paolo Mattevelli, and Dushan Boroyevich. EMI modeling of half-bridge inverter using a generalized terminal model. *2011 Twenty-Sixth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, pages 468–474, March 2011.

[15] Satoshi Ogasawara, Hideki Ayano, and Hirofumi Akagi. An active circuit for cancellation of common-mode voltage generated by a PWM inverter. *Power Electronics Specialists . . .*, pages 1547–1553, 1997.

[16] Marwan Ali and Eric Labouré. Design of a Hybrid Integrated EMC Filter for a DC ?DC Power Converter. *IEEE Transactions on Power Electronics*, 27(11) :4380–4390, 2012.

[17] Juergen Biela and Alexander Wirthmueller. Passive and active hybrid integrated EMI filters. *IEEE Transactions on Power Electronics*, 24(5) :1340–1349, 2009.

[18] Remi Robutel. *Etude des composants passifs pour l'électroniques de puissance à "haute température" : application au filtre CEM d'entrée.* PhD thesis, 2011.

[19] Janet Ho and Richard Jow. Capacité HT.pdf. Technical report, Army Research Laboratory, Adephi, 2009.

[20] An Zhou. *Modèles de composants passifs et couplages électromagétique pour filtres HF de puissance - Optimisation du placement.* PhD thesis, 2012.

[21] Thomas De Oliveira. *Optimisation du routage d ?un filtre CEM.* PhD thesis, Université de Grenoble, 2012.

[22] Sylvain Mandray, Guichon Jean-Michel, Jean-Luc Schanen, and Adrian Perregaux. Reduction of conducted EMC using busbar stray elements. *2009 Twenty-Fourth Annual IEEE Applied Power Electronics Conference and Exposition*, pages 2028–2033, February 2009.

[23] Baïdy Birame Touré. *Modélisation Haute Fréquence des variateurs de vitesse pour Aéronefs : Contribution au Dimensionnement et à l'Optimisation de Filtres CEM.* PhD thesis, 2012.

[24] François Costa and Didier Magnon. Graphical Analysis of the Spectra of EMI Sources in Power Electronics. *IEEE Transactions on Power Electronics*, 20(6) :1491–1498, November 2005.

« Une personne qui n'a jamais commis d'erreurs n'a jamais tenté d'innover. »

Albert Einstein

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

I- TABLE DES MATIERES

II- INTRODUCTION

Mon sujet de thèse porte sur la conception et le dimensionnement d'un convertisseur statique dans un environnement sévère. Il s'inscrit dans le cadre d'un projet pour l'Agence Nationale de la Recherche (ANR) financée par la Fondation de la Recherche en Aéronautique et Espace (FRAE). Ce document présente les premiers éléments des recherches effectuées lors de ma première année de thèse dans le cadre du projet Actionneur électrique Compact avec Convertisseur Intégré pour Températures Extrêmes (ACCITE), sur la partie WP2-TACHE2 « Intégration de l'électronique de puissance».

La thèse de doctorat que j'effectue au laboratoire est placée sous la direction du M. Christian Vollaire, Professeur des Universités à l'Ecole Centrale de Lyon (ECL), et l'encadrement de M. Cyril Buttay, chargé de recherche CNRS, et de M. Dominique Bergogne, Maître de Conférences à l'Université Claude Bernard Lyon I (UCBL). Les travaux s'opèrent au sein du laboratoire AMPERE UMR CNRS 5005 sur le site de l'Ecole Centrale de Lyon.

Dans ce mémoire nous introduirons dans un premier temps, le contexte de l'étude, l'électronique de puissance dans l'avion plus électrique.

Puis, une partie sera consacrée à l'étude du convertisseur de puissance et à l'état de l'art actuel.

La troisième partie nous permettra de présenter les principaux travaux effectués précédemment au laboratoire sur ce sujet.

La quatrième partie aura pour sujet la méthode de conception que nous avons définie.

Dans la cinquième partie, nous présenterons la manipulation mise en place parallèlement aux travaux bibliographiques et aux simulations.

Dans la dernière partie, nous ferons une conclusion sur les études bibliographiques et les manipulations réalisées avant de présenter les perspectives des travaux à effectuer pendant les deux dernières années de thèse.

Cette thèse de doctorat s'effectue actuellement à l'École Centrale de Lyon concerne le domaine de l'électronique de puissance dans le milieu aéronautique. Les principaux verrous technologiques auxquels nous avons du et allons devoir faire face durant ces 3 années de recherche sont les améliorations pour l'avion « plus électrique » et l'intégration de l'électronique de puissance en vue de l'alimentation des moteurs dans un milieu où les températures sont extrêmes.

Actuellement, les convertisseurs statiques d'électronique de puissance sont déportés dans un environnement « froid » (environ 100°C), ce qui n'amène aucun problème de vieillissement ou de déficience. Cependant les problèmes de cyclage thermique restent présents.

Le projet ANR ACCITE, figure 1, a pour but de réaliser un « smart moteur » dans lequel le convertisseur sera implanté directement sur le flasque arrière d'une machine synchrone à aimants permanents. C'est dans ce contexte que le milieu de température extrême dans l'aéronautique est abordé dans les pages suivantes.

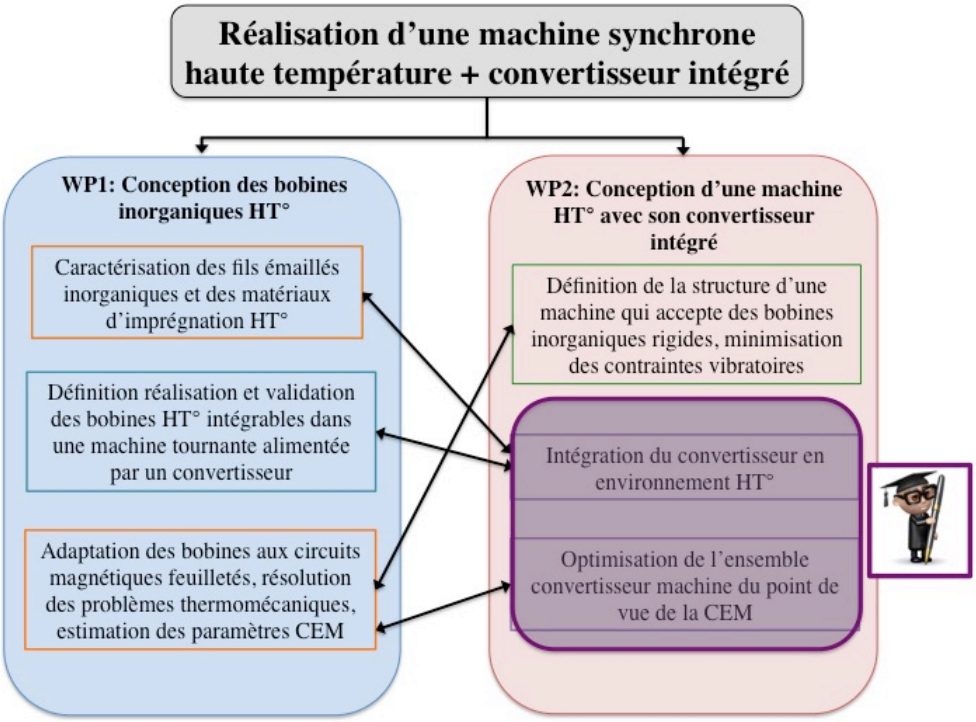


Figure 1 : Structure collaborative du projet et situation de la thèse

Les tâches des différents laboratoires acteurs du projet sont représentées à la figure 1. Le laboratoire LSEE de l'université d'Artois est symbolisé en orange, le laboratoire GREEN de l'Ecole Nationale Supérieure d'Electronique et Mécanique (ENSEM) en vert, le laboratoire LAPLACE de l'Institut National Polytechnique de Toulouse (INPT) en turquoise et enfin le laboratoire AMPERE en violet.

Les matériaux inorganiques constituent l''autre grande catégorie des isolants électriques. Des fils isolés par des fines couches inorganiques (céramique ou matériaux composites à base de mica par exemple) ont été développés pour des applications à faibles contraintes. Ces fils supportent des températures extrêmes de l''ordre de 1000°C mais leurs propriétés électriques et mécaniques sont très inférieures à celles des fils émaillés organiques classiques, ils ne conviennent pas pour bobiner les machines électriques actuelles.

A) L'AVION PLUS « ELECTRIQUE »

La réduction des coûts d'acquisition et d'exploitation ainsi que le bon fonctionnement et la sécurité des systèmes embarqués gouvernent les avancées technologiques dans le domaine de l'aéronautique. L'électrification des avions s'impose progressivement.

Au sein de l'avion « plus électrique », certaines fonctions, jusqu'alors régies par pression hydraulique ou pneumatique deviennent électriques. On peut notamment noter parmi celles-ci l'inverseur de poussée et le freinage des trains d'atterrissage. Ces fonctions sont réalisées *via* des machines électriques, elles-mêmes commandées par des convertisseurs.

Concernant le *TRL* (*Technology Readiness Level*) [1], il sera de niveau 3, c'est-à-dire qu' « une analyse et une expérimentation de la fonction critique » est attendue. Une recherche et un développement actifs sont initiés. Ceci inclut des études analytiques et des études en laboratoire afin de valider physiquement les prévisions analytiques des éléments séparés de la technologie.

B) ENVIRONNEMENT HAUTES TEMPERATURES ET AERONAUTIQUE

Les systèmes d'électronique de puissance ont une plage de fonctionnement située entre -55°C et 85°C (voire 125°C). Les applications « haute température » concernent un fonctionnement transitoire ou continu au-delà de 125°C et dans notre cas supérieur à 200°C. Lorsqu'il s'agit d'environnement sévère, les températures ambiantes et de jonction des puces de puissance sont élevées. La température ambiante correspond à la température de l'environnement. *A contrario*, la température de jonction est celle qui régne au sein des composants semi-conducteurs. Elles sont reliées par l'équation (1).

$$T_j = T_a + P.R_{thja} \quad (1)$$

avec T_j, la température de jonction, T_a la température ambiante, P la puissance dissipée par l'élément semi-conducteur et R_{thja} la résistance thermique jonction-ambiante.

Contrairement aux travaux effectués par M. Rémi Robutel [2] qui ne s'était concentré uniquement sur l'étude des composants magnétiques utilisés dans le filtre d'entrée d'un convertisseur statique de puissance haute température, cette problématique nécessite une nouvelle approche systémique. À savoir, les aspects packaging, compatibilité électromagnétique, comportement dans un environnement sévère, et test sur une machine synchrone à aimants permanents seront réalisés au cours de ces trois années d'étude.

IV- PERIMETRE DU CONVERTISSEUR

Dans le cadre du projet ACCITE, le laboratoire AMPÈRE à travers mes travaux de thèse de doctorat se charge de réaliser un convertisseur statique en environnement extrême (Température ambiante : 300°C) pour l'alimentation d'une machine synchrone à aimants permanents d'une puissance d'environ 5 kW.

A) ACTIONNEUR

Un onduleur est un dispositif d'électronique de puissance permettant de délivrer des tensions et des courants alternatifs à partir d'une source d'énergie continue. Il présente la fonction inverse d'un redresseur. Les onduleurs sont basés sur une structure de pont en H, constituée le plus souvent d'interrupteurs électroniques tels que les IGBTs, MOSFETs, pour les transistors de puissance ou de thyristors. Par un jeu de commutations commandées de manière appropriée, la source continue est modulée afin d'obtenir un signal alternatif d'amplitude et de fréquence désirées. Dans le cadre de ce projet, le convertisseur va être directement implanté sur une machine synchrone à aimants permanents d'environ 5kW.

B) TOPOLOGIE

Le point novateur de ce projet au niveau électronique de puissance est l'intégration du convertisseur dans un environnement sévère ($T_{ambiante}$>125°C) au plus près du moteur. Il sera intégré directement sur le flasque de la machine synchrone à aimants permanents au niveau de la sortie de l'arbre, figure 1.

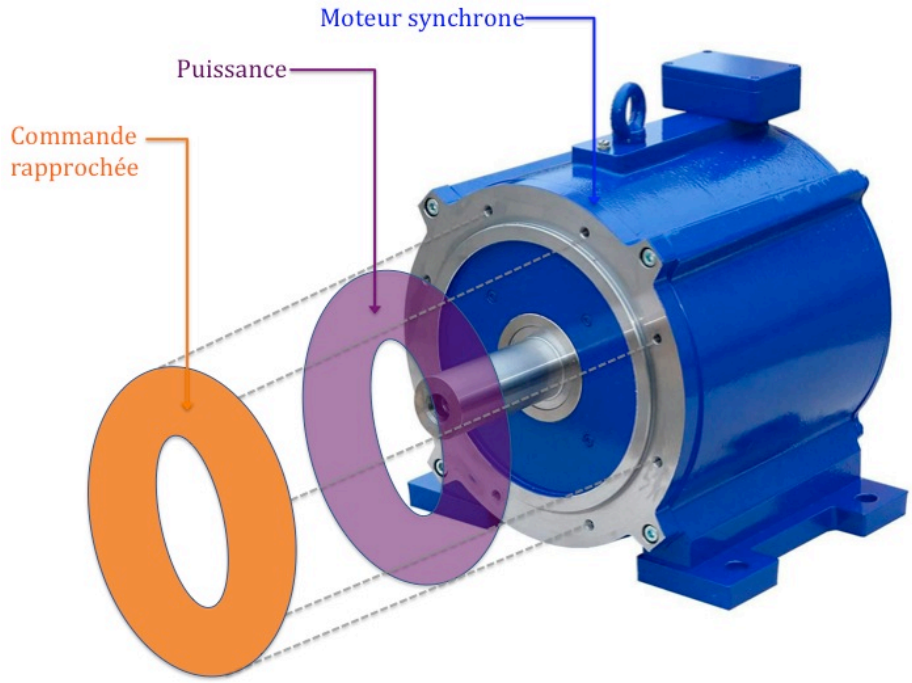


Figure 1. Schéma du système global

La partie d'électronique de puissance (disque violet), sera composée de trois bras d'onduleur constituant un onduleur triphasé. Le convertisseur possèdera un filtre d'entrée et un filtre de sortie. Le filtre d'entrée pourra par exemple être composé de capacités de découplage qui filtrera les courants de mode commun et de mode différentiel en entrée du système.

Le filtre CEM d'entrée doit être optimisé afin de limiter le poids du système, ce qui se révèle important dans le domaine de l'aéronautique. Le filtre CEM d'entrée va protéger le réseau électrique général de l'avion contre les perturbations de mode commun et de mode différentiel générées par le convertisseur et transmises

à travers le câble et la machine. L'onduleur étant monté directement sur l'actionneur, il n'y a en théorie pas besoin de filtre de sortie d'un point de vue de la normalisation CEM. L'absence de câbles longs à la sortie permet d'obtenir des commutations rapides.

Au niveau de la compatibilité électromagnétique, le courant et la tension de mode différentiel sont contenus dans la boucle d'alimentation et dépendent fortement de la fréquence des commutations. Le courant de mode différentiel est filtré par l'inductance de ligne et par la capacité de découplage du filtre d'entrée. Le courant de mode commun circule depuis les lignes de puissance *via* des impédances parasites jusqu'à la masse de l'aéronef. Le courant de mode commun est le principal responsable des perturbations conduites à hautes fréquences. Les condensateurs de filtrage de mode commun seront directement implantés au sein du module de puissance, ceci permettra donc d'avoir des commutations rapides des transistors et de contenir au sein du convertisseur les perturbations CEM de mode commun à hautes fréquences.

C) COMPOSANTS DE PUISSANCE & CHOIX DES TRANSISTORS

La méthode d'analyse présentée ici a été appliquée sur un bras d'onduleur mais peut être étendue à un onduleur triphasé complet. Le module doit offrir les caractéristiques suivantes : une résistance aux hautes températures, peu de réaction aux effets parasites, de par la température de jonction élevée et la vitesse de commutation rapide des interrupteurs de puissance. La gestion thermique et le packaging de ce module constituent les points clés de l'étude [4].

Depuis le début du XXI^{ème} siècle, la technologie des composants électroniques de puissance a évolué vers les températures élevées. Les puces en Carbure de Silicium (SiC) disponibles sur le marché sont capables de commuter des tensions supérieures à 3 kV en quelques dizaines de nano secondes, et à des températures de jonction élevées (jusqu'à 315°C) ce qui conduit à une large amélioration du rendement. On classe le SiC dans les semi-conducteurs à large bande d'énergie interdite. La réduction du temps de commutation permet d'augmenter la fréquence de fonctionnement et ainsi de réduire la taille des systèmes. Les composants en SiC sont généralement utilisés pour des applications à haute température, haute fréquence, haute tension et répondent à des contraintes de poids, d'encombrement et de volume exigeantes [7].

Le SiC est un matériau possédant plusieurs avantages notoires. Les composants en SiC ont une tenue en tension élevée, due à un champ électrique critique important. Son grand gap permet de fabriquer des composants avec des courants de fuite plus faibles et assure ainsi un bon fonctionnement à haute température (200°C). D'autre part, pour les applications haute fréquence, les composants en SiC sont plus efficaces que ceux en Si. Il est expliqué dans [8] que la vitesse de saturation des porteurs du SiC est élevée, et sa permittivité faible. Dans le cadre de notre application, ces propriétés permettent une réduction des pertes et de volume. En effet, sa faible résistance à l'état passant et la réduction de la charge stockée, apportent une diminution des pertes en conduction et en commutation. Les valeurs des propriétés énoncées précédemment sont données dans le tableau 1.

Propriétés	SiC	Si
Gap (eV)	3,2	1,1
Champ électrique (10^6V/cm)	3	0,3
Vitesse de saturation des porteurs (10^{-6}cm/s)	22	10

Tableau 1. Propriétés du Si vs SiC

Une liste des différents composants en SiC pouvant satisfaire le cahier des charges de ce projet a été réalisée :

- L'IGBT (*Insulated Gate Bipolar Transistor*) présente une structure 4 couches. C'est un interrupteur unidirectionnel en courant, il est généralement asymétrique en tension. La structure de l'IGBT

permet de résoudre le problème de la forte valeur de la résistance à l'état passant que présente par exemple les MOSFETs (*Metal Oxide Silicon Fiel Effect Transistor*) à haute tension (>plusieurs kV). D'autre part, les faibles chutes de tension à l'état passant de l'IGBT autorisent un fonctionnement à densité de courant plus élevée que celle des transistors bipolaires. Cependant, leur fonctionnement à haute température sans interruption de fonctionnement normal lors des commutations reste un problème dans le cadre du projet ACCITE, étant donné que ce type de composants n'est pas commercialisé en SiC. De plus, les commutations d'un IGBT sont moins rapides que celles d'autres composants, notamment les transistors à effet de champ, [9]. L'IGBT SiC est réservé tensions supérieures à 10kV et n'est pas disponible industriellement.

- Le transistor MOSFET est caractérisé par la charge de ses porteurs majoritaires qui détermine s'il est de type *P* ou *N*. Ce composant a beaucoup été étudié ces dernières années. Le problème de ce composant est sa faible résistance de grille, et son vieillissement accéléré lors de la montée en température. Certains résultats [10], [11] montrent une nette amélioration des performances des composants (résistance à l'état passant R_{dson} = 16 mΩ · cm2 pour une tension de claquage V_{br} = 1 400 V).
- Le JFET SiC demeure le composant le plus mature et le plus robuste sur le marché actuel. Les structures les plus appropriées pour la haute température semblent être les JFETs et les BJTs de par leur relative simplicité et leur robustesse. On utilisera des JFETs pour l'onduleur de puissance, conformément au cahier des charges qui requière un utilisation en HT° (fournisseur : INFINEON). La vitesse de commutation des JFETs SiC est 2 à 10 fois plus importante que celle des IGBT classiques, ce qui implique une hausse des dv/dt et di/dt.

Un interrupteur fonctionne en « *normally on* » lorsqu'il laisse passer le courant en l'absence d'une polarisation de son électrode de commande. Il est « *normally off* » dans le cas contraire. Dans le cadre de ce projet, nous souhaitons utiliser des interrupteurs « *normally off* ». Cependant, parmi les JFET disponibles sur le marché ou bien susceptibles d'être fabriqués durant cette étude, seuls des « *normally on* » répondant au cahier des charges sont disponibles.

Les puces nues que nous avons choisi pour cette applications sont des JFETs « *normally on* » fabriqués par le laboratoire de recherche et développement SICED intégré à INFINEON. Ces composants possèdent un calibre en tension de 1,2 kV, un calibre en courant d'environ 20 A et une résistance à l'état passant ($R_{DS,on}$) de 80 mΩ à température ambiante, [13].

D) TECHNIQUE DE PACKAGING RETENUE

Le reste de la structure et du packaging (boitier et gel) ne tiennent pas en température. Leur point de dégradation est inférieur aux conditions de températures requises lors d'un vol aérien (250°C).

Par exemple, certaines diodes Schottky, admettent une tension de 1000V et leur température maximale en mode non-dégradé est de 500°C. Par contre, l'habillage de cette diode ne résiste pas aux hautes températures. Ainsi, elles ne peuvent être utilisées qu'à une température maximale de 200°C. On a ici à faire à un problème mécanique et de choix de matériaux d'enrobage.

Les changements de température répétés entraînent la brisure des composants (passer de -50°C à 200°C dans le domaine aérien). En effet, les contraintes mécaniques causées par la dilatation thermique entraînent des fissures au niveau du boîtier entourant les composants électroniques.

Ils seront montés sur un substrat par nos soins afin d'éviter les problèmes de *packaging*.

E) CONTRAINTES DE DIMENSIONNEMENT DE L'ONDULEUR

Dans l'environnement sévère qui fait partie du cahier de charges de ce projet, la température estimée au cœur des bobines de la machine est de 400°C voire 450°C. Le but final de ce projet est de réaliser un « smart moteur », c'est-à-dire un moteur avec son convertisseur intégré [14]. Actuellement, seule l'Université de Shieffield a initié des travaux sur un moteur fonctionnant à haute température en étroite collaboration avec le fabricant ROLLS ROYCE.

Le verrou technologique principal est la température. La densité de courant extérieure en régime permanent J_{ex} ne doit pas trop être importante. Les composants unipolaires tels que les JFETs que nous allons utiliser autorisent des vitesses de commutations largement supérieures à celles des IGBTs. Les *dv/dt maximum* mesurés sur un JFET en carbure de Silicium peuvent atteindre 60kV/s.

L'existence quasi systématique d'un filtre à l'entrée (requis à cause des normes CEM en aéronautiques) atténue les perturbations conduites, ainsi que les perturbations rayonnées par les câbles d'alimentation. Le soin apporté à la réalisation de l'onduleur doit permettre de réduire la génération de perturbations. La prise en compte des composants utilisés, de la technique d'intégration de ces derniers et de la qualité des liaisons électriques constitue un facteur déterminant en vue de la conception de l'onduleur dont nous disposons.

Concernant les degrés de liberté, il faut prendre en compte le filtrage, le blindage, la commande, le rendement, le routage mais aussi les différentes technologies pour l'élaboration des composants. Par habitude, lorsque l'on aborde la CEM en électronique de puissance, on s'intéresse aux aspects normatifs. Ce qui n'est pas seulement le cas dans notre contexte. En effet, il faut respecter l'intégrité de la machine, à savoir éviter une dégradation du Système d'Isolation Electrique (SIE) car les variations trop rapides de la tension peuvent en effet endommager ce SIE. D'autre part, le concepteur d'un convertisseur statique est soumis au dilemme perturbations-pertes : la minimisation des pertes par commutation conduit à faire commuter les interrupteurs très rapidement, mais cela accroît les *dv/dt* et *di/dt*, donc les perturbations électromagnétiques et les contraintes sur le SIE. Un compromis doit être trouver.

Comme le mentionne Jérome Genoulaz dans sa thèse de doctorat [7], la norme aéronautique la plus appliquée est la norme D0-160F, [15]. Elle précise les conditions environnementales et les procédures de test pour les équipements aéronautiques. Cette norme a été rédigée par la Commission Technique pour l'Aéronautique (RTCA : *Radio Technical Commission for Aeronautics*). Elle spécifie les caractéristiques des équipements au sein d'un aéronef ainsi que l'environnement dans lequel ils doivent fonctionner. La prise en compte de la source et de la charge en vue d'une optimisation et de la réalisation du filtre d'entrée sont donc primordiales.

En définitive, si le volume du dissipateur diminue, celui des filtres augmente. Un compromis optimal concernant le dimensionnement du système doit être trouvé. Il doit intégrer la technologie des semi-conducteurs et des composants passifs mais également d'autres paramètres qui influent sur ce compromis, comme par exemple la fréquence de découpage. La constitution technologique de la cellule de commutation possède donc un rôle important quant aux perturbations qu'elle est amenée à générer.

Les filtres du commerce sont caractérisés avec une source d'impédance interne 50 Ω et une 50 Ω. Or le trio convertisseur/réseau/charge ne présente pas une impédance équivalente de 50 Ω tout comme le réseau de l'avion. Il sera nécessaire de connaître le niveau de tension de perturbations en mode commun et en mode différentiel généré par le convertisseur afin de dimensionner le filtre (en entrée ou en sortie, si besoin est).

F) FONCTIONS NON INTEGREES EN ZONE CHAUDE

Les composants semi conducteurs de puissance requièrent une commande spécifique que l'on nomme « commande rapprochée » pour permettre leur commutation.

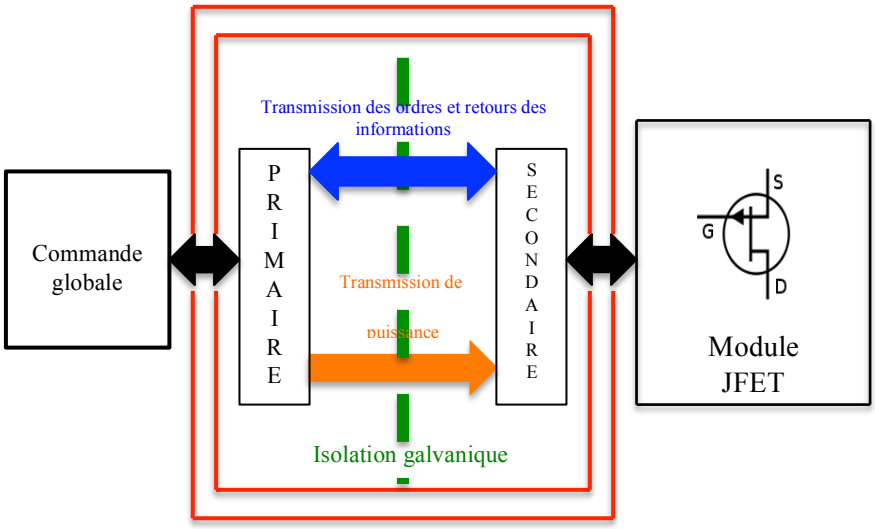
La commande rapprochée a pour but de :

- Mettre en forme le signal arrivant de la commande, ceci nécessite donc un apport d'une puissance, et par conséquent d'une alimentation spécifique.
- Isoler les parties commande et puissance.
- Gérer les temps morts.
- Réaliser le rôle de protection contre les éventuelles perturbations.

Certaines des parties du dispositif de commande rapprochée ne peuvent que difficilement être intégrées en environnement sévère dans le cadre de notre étude. La commande rapprochée dispose d'une alimentation auxiliaire pour fournir l'énergie à la commande du transistor de puissance. L'ordre de commande est transmis par une voie secondaire *via* un dispositif quelconque d'isolation galvanique.

Deux choix peuvent être faits concernant l'isolation commande/puissance:

- Le transformateur d'impulsion : l'isolation galvanique entre l'entrée et la sortie est assurée grâce à un transformateur, fonctionnant à la fréquence de commutations ce qui permet ainsi la réduction de sa taille, [5]. La difficulté dans le cadre de ce projet, est de réaliser un transformateur d'isolement qui puisse résister aux hautes températures.
- L'opto-coupleur : il transmet des signaux de faible niveau, [6]. Ce type de dispositif n'est pas adapté à un fonctionnement en hautes températures.



Commande rapprochée

Les problèmes sur ce type de structure sont les mêmes que pour le transformateur d'impulsion (transformateur fonctionnant en environnement sévère) mais également la proximité des signaux de commande et de puissance qui peuvent engendrer des problèmes de CEM.

V- TRAVAUX ANTERIEURS REALISES DU LABORATOIRE AMPERE

A) CONVERTISSEUR REALISE DANS LA THESE DE REMI ROBUTEL

Dans [3], la capacité de mode commun est directement implantée au sein du module de puissance cela permet de réduire les perturbations CEM de mode commun en HF, comme présenté sur la figure 2. Nous appliquerons ce principe durant notre étude. L'architecture du module de puissance intégrant le filtrage des perturbations électromagnétiques est présentée figure 3.

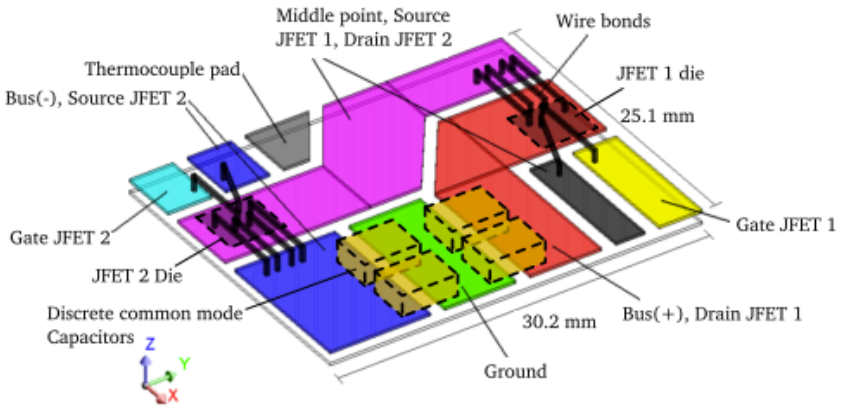


Figure 2. Exemple de routage d'un bras d'onduleur [2]

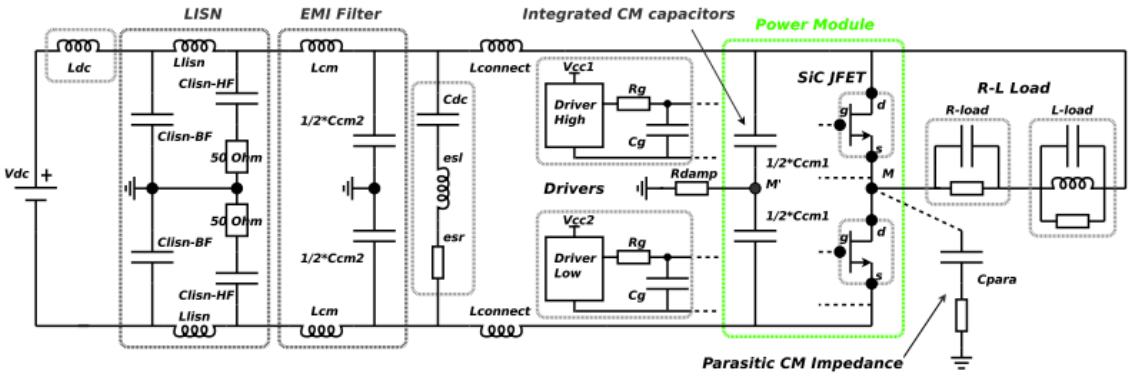


Figure 3. Schéma électrique global de la structure

Au niveau packaging, il faut faire attention aux différentes températures : 350°C pour la température de jonction du JFET SiC, 200°C pour les capacités de filtrage de mode commun car ils partagent le même substrat.

Le *direct bonding copper* (DBC) est en céramique (635 micromètres) avec une couche de cuivre (200 micromètres) du dessous qui constitue le plan de masse, et une autre qui forme les pistes au dessus (200 micromètres). Le contact entre la grille et la source du JFET est assuré *via* des fils de (wire) bonding en aluminium. Des connexions en cuivre assurent les liaisons externes. Un gel de silicone, qui constitue l'encapsulant, a été appliqué pour l'isolation électrique. À la fin du procédé, le module a été relié à un dissipateur thermique *via* son plan de masse, qui possède une résistance thermique faible.

Comme les JFETs choisis pour l'étude du bras d'onduleur présentent une diode de roue libre interne et intrinsèque, il a été convenu que l'utilisation de diodes externes en antiparallèle des transistors ne serait pas nécessaire.

Le refroidissement a été réalisé par conduction *via* la couche de dessous du PCB. Les pertes dans la couche de cuivre sont de 0,1W/cm^2. Cette valeur est négligeable vis à vis des pertes dans les JFETs qui sont comprises entre 100 et 300 W/cm^2. On conclue que les pertes par effet Joule de ce plan de masse sont négligeables. La largeur importante des pistes et le plan de masse permettent de diminuer les inductances dans le circuit. Il faut cependant faire attention à avoir une distance minimale entre la boucle d'alimentation et la boucle de commande. De ce fait, seul un fil de bonding est utilisé pour faire la liaison tension entre grille et source. Au sein de la boucle d'alimentation, quatre fils de bonding sont utilisés pour relier la source du JFET au bus négatif. La température dans le module est mesurée *via* un thermocouple.

La capacité de filtrage de mode commun doit être reliée à la masse. Une piste a donc été prévue sur le PCB à cet effet. En comparaison à un module de puissance classique, l'ajout de ces capacités de filtrage de mode commun accroît la taille du module de 30%. Les JFETs utilisés possèdent les caractéristiques suivantes : puce 4×4 mm^2, tenue en tension de 1.2kV, courant de drain nominal de 15A, résistance à l'état passant 80mΩ à 25°C. A l'état bloqué, V_{gs} doit être < V_p (tension de pincement = -18,5V) et V_{gs}> à V_{pt} (tension d'avalanche=-27V). Pour satisfaire ces contraintes, V_{gs} a été fixée à -23,5V en ajoutant une capacité 2,2nF (C_g) entre grille et source. Dans le but de limiter les impulsions parasites de courant au niveau de la grille, une résistance de 12 Ω a également été ajoutée. L'isolation galvanique est réalisée par deux convertisseurs DC/DC (TRACO power) choisis pour leur faible capacité parasite (13pF max).

En vue de tests pour ce bras, la charge utilisée a été choisie et possède les caractéristiques suivantes: 8 Ω et 2,3mH. L'inductance n'est pas constituée de matériaux magnétiques et a une capacité parasite faible (17pF).

Les inductances équivalentes des fils d'alimentations et des connexions ont été représentées par : $L_{connect}$ = 100nH, figure 3. Les capacités de découplage (figure 2 et 3) ont quant à elles été placées le plus près possible du module pour limiter les inductances parasites durant les commutations. Pour réduire le courant résiduel, une inductance de ligne L_{dc} a été insérée entre le bras d'onduleur et l'alimentation (L_{dc}=500 µH pour un courant de 5A). Au niveau simulation CEM, dans un contexte « classique », la référence est imposée par le RSIL et le plan de masse. Le RSIL (réseau stabilisateur d'impédance de ligne) est un dispositif alimenté par le réseau semblable à un filtre en π utilisé pour mesurer les émissions de courant harmoniques d'un convertisseur connecté à sa sortie. Le RSIL est conforme aux spécifications de la DO 160F. Pour la capacité de mode commun intégrée, les investigations ont été menées entre 1 et 10nF. Des valeurs inférieures à 1nF n'auraient pas eu d'incidence sur la CEM. Dans le cadre de cette étude, ces capacités, doivent filtrer le bruit HF. Pour éviter des oscillations de courant trop importantes, une résistance d'amortissement a été ajoutée à la connexion à la terre. La dissipation due à cet amortissement est considérée comme négligeable. Le spectre CEM est typique pour un onduleur, figure 4.

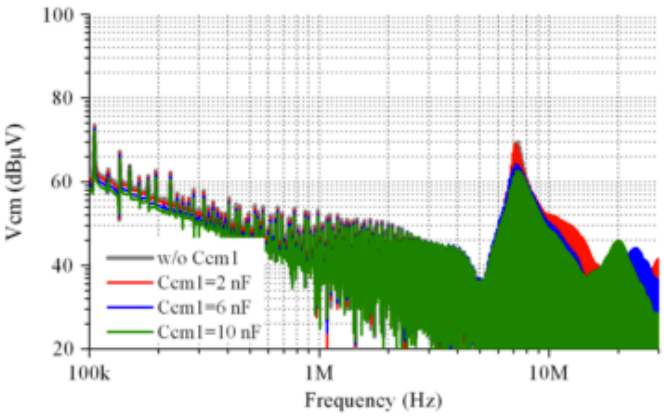


Figure 4. Spectre de mode commun [1]

Le bruit de mode commun est bien filtré jusqu'à quelques MHz, au delà, du fait de la présence d'éléments parasites, le bruit est présent. Les analyses de spectres montrent que le condensateur de 6nF présente le meilleur compromis. L'impédance parasite de mode commun produit des oscillations de courant lors de la

fermeture du JFET. Celles ci sont réduites par la capacité de filtrage. À haute fréquence (7MHz), une atténuation de 6dBµV pour la tension de mode commun est quantifiée. Le filtre CEM d'entrée est donc peu efficace à haute fréquence.

En reprenant ce qui a été montré dans les simulations, l'étude sur un banc d'essai a été menée, avec une capacité de mode commun de 6nF. Plusieurs capacités ont été assemblées pour obtenir 6nF.

B) BRAS D'ONDULEUR HAUTE TEMPERATURE, PROJET THOR

Le projet THOR [18], initié entre le laboratoire AMPERE et l'entreprise SAFRAN a permis de réaliser un bras d'onduleur avec certaines des fonctionnalités des drivers intégrés dans un environnement haute température.

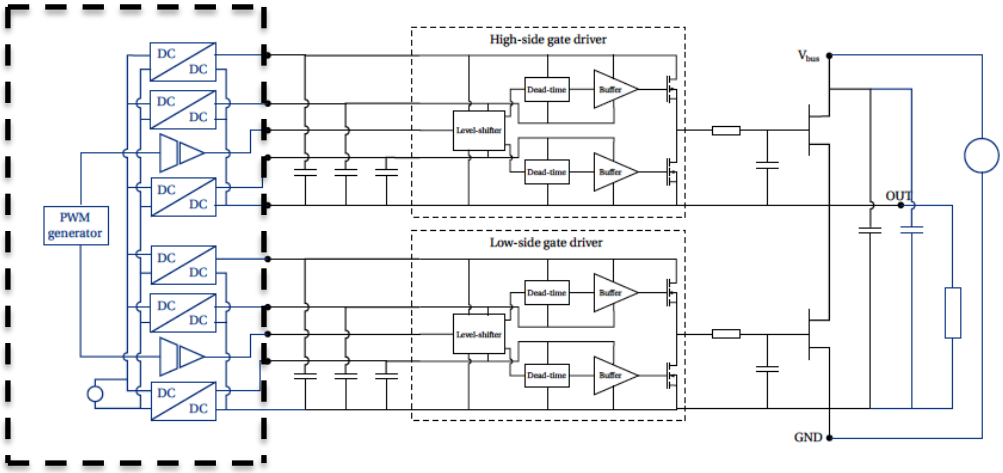


Figure 5. Schéma électrique du module, alimentations externes et charge utilisés.

Sur la figure 5, la partie en pointillés épais est celle qui demeure en environnement à température ambiante, tandis que le reste est en environnement chaud. Les essais réalisés sur ce module de puissance concluent à un fonctionnement satisfaisant jusqu'à 310°C.

En somme, les études menées jusqu'ici ont montré concernant le filtre CEM d'entrée du convertisseur au sein du laboratoire AMPERE,

- Avec une modélisation adaptée, il est possible de prédire les niveaux de perturbation émis et l'efficacité d'un filtre CEM.
- Le dimensionnement d'un filtre CEM d'entrée pour un onduleur de tension d'environ 2 kW a montré qu'avec un choix de condensateurs et de matériaux magnétiques appropriés, l'impact de la température ambiante entre 25 ◦C et 200 ◦C sur le filtrage des perturbations conduites est modéré.
- Une part importante du courant de mode commun est contenue dans le convertisseur avec les choix réalisés (ajout de capacités de mode commun au plus près de la cellule de commutation par exemple).
- La réalisation d'un bras d'onduleur avec les drivers intégrés (hormis leur alimentation) en environnement chaud, fonctionnant jusqu'à 310°C a été menée.

Les travaux réalisés au laboratoire AMPERE sur le filtre CEM concernent l'identification des perturbations conduite ainsi que le découplage des modes commun et différentiels à partir des normes. En fonction de la nature et de l'intensité des perturbations, un filtre peut éventuellement être choisi pour chaque mode. Chaque élément du filtre doit ensuite être dimensionné en prenant en compte les différentes contraintes électriques. L'étude des perturbations de compatibilité électromagnétique conduites permet ainsi de structurer le filtre puis de dimensionner les composants, à savoir leurs valeurs et leurs natures. Ensuite, vient la partie du routage et de l'assemblage. Le tout est enfin testé afin de vérifier le respect du cahier des charges, [15], [16].

VI- METHODE DE CONCEPTION

A) PRESENTATION DES OUTILS DE CONCEPTION

Notre démarche de conception repose sur des outils informatiques. Pour valider la méthode de conception et les outils associés, nous allons étudier le module de puissance du projet THOR [18], pour lequel nous disposons de tous les éléments, ce qui va permettre une comparaison rapide simulations/expériences. Le module d'électronique de puissance étudié est composé d'un substrat, des composants passifs, et des composants actifs.

Le schéma électrique est présenté sur la figure 7.

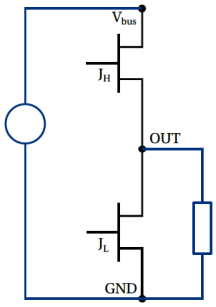


Figure 7. Schéma électrique du module de puissance étudié

La première étape consiste en une simulation du comportement électrique du circuit. Celle-ci est réalisée *via* le logiciel SABER**.** Il permet de simuler le comportement électrique du système en intégrant différents niveaux de complexité (semi-conducteurs, éléments parasites...).

Le schéma électrique du système « parfait » est représentée sur la figure 8.

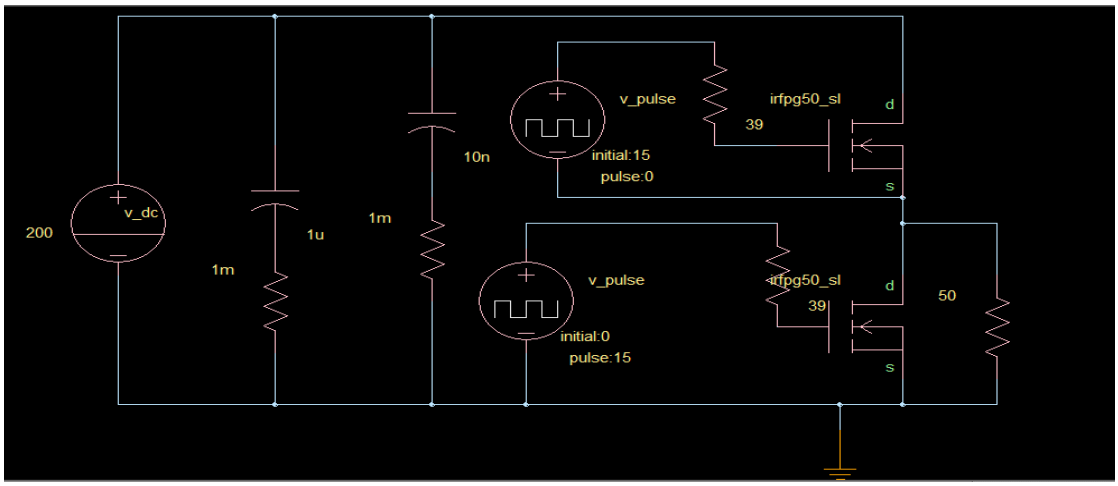


Figure 8. Bras d'onduleur sur charge R = 50 Ω.

Les transistors représentés sur la figure sont des MOSFETs supposés parfaits. Leur résistance à l'état passant est peu élevée. Ils sont commandés par des drivers également supposés parfaits.

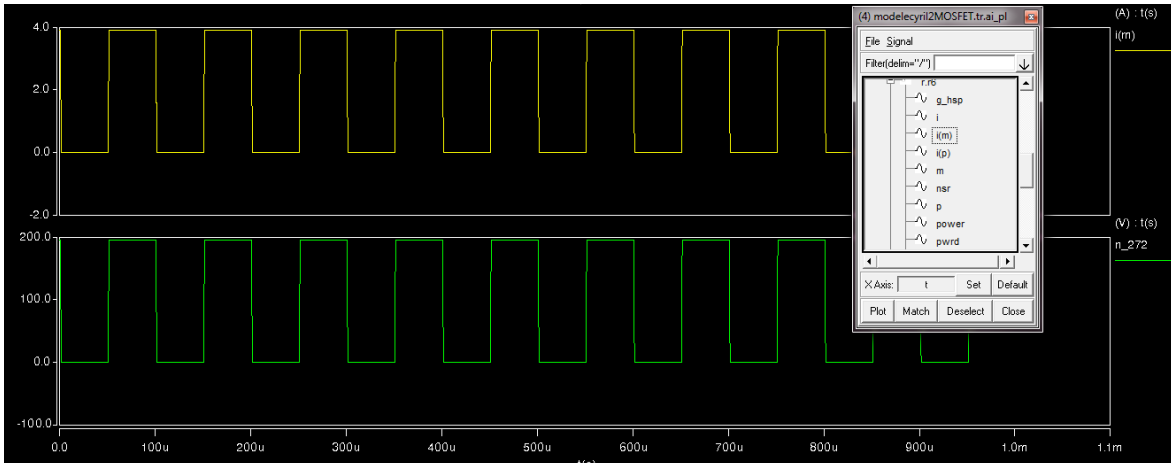


Figure 9. Formes d'ondes en sortie du bras

La première étape a été de prendre en compte les imperfections liées au routage du circuit (inductances, résistances et capacités parasites des pistes). Le logiciel Agilent Design System (ADS) a été utilisé pour extraire les matrices d'impédances des pistes.

B) AUGMENTATION DE LA PRECISION DU MODELE ELECTRIQUE

ELEMENTS PASSIFS

Le processus de calibration des outils de conception s'effectue en deux étapes. Les valeurs physiques des impédances des différents composants passifs ont été évaluées grâce à l'impédancemètre (Agilent 4924A). Les passifs étudiés sont la résistance de grille des JFETs, la capacité de grille, et la capacité de découplage. Une confrontation des modèles implantés dans le logiciel SABER sera effectuée dans la suite des travaux, entre 40 Hz et 110 MHz. Les résultats sont donnés sur les figures 10 à 12.

Comme il a été fait dans [19], les valeurs mesurées physiquement *via* l'impédancemètre sont ajustées avec les valeurs calculées à partir du modèle élaboré. Les résultats présentés dans la suite de cette partie ont été obtenus grâce à un algorithme d'optimisation développé avec le logiciel MATLAB®. La figure 10 présente les résultats du comportement de la capacité de découplage. Sa valeur est de 10nF. Le modèle programmé choisi est un circuit RLC série.

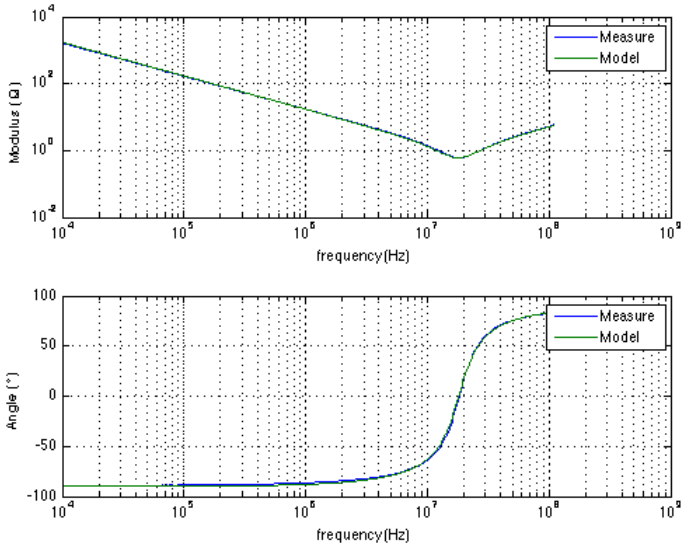


Figure 10. Comportement de la capacité de découplage

Une capacité, que l'on nomme capacité de grille est connectée entre la source et la grille pour chacun des JFETs. Cette capacité est rajoutée en parallèle de la capacité grille-source intrinsèque du JFET pour des raisons de stabilisation de la commande.

Le comportement physique de cette capacité est présenté à la figure 11.

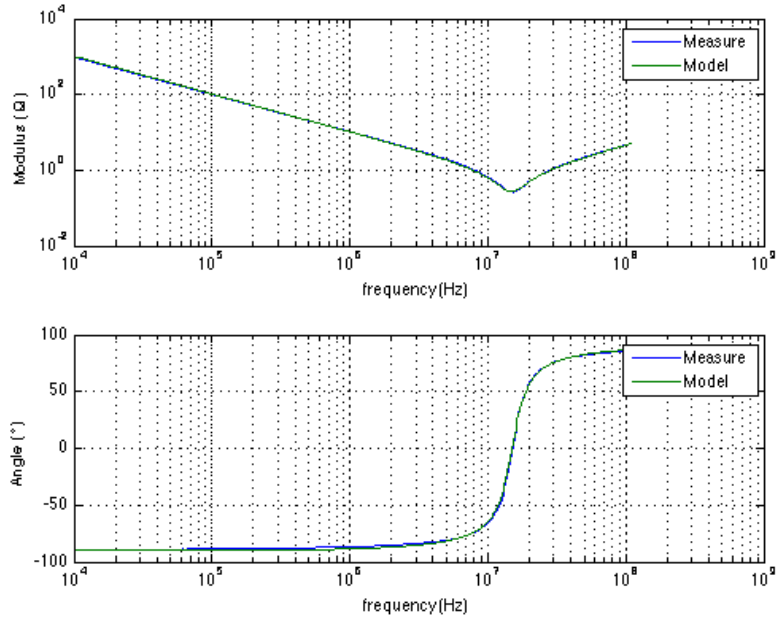


Figure 11. Comportement de la capacité de grille

La résistance de grille (1206 High Precision Wraparound - High Temperature (230 °C) Thin Film Chip Resistors) a également été étudiée. Les résultats sont présentés sur la figure 12.

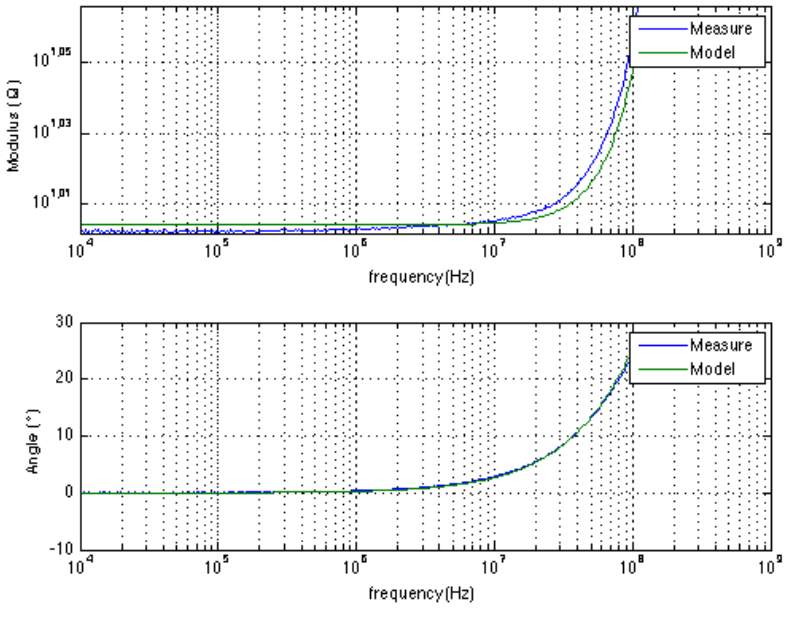


Figure 12. Comportement de la résistance de grille

En conclusion, les éléments passifs présentent un comportement classique sur la plage de fréquence étudiée.

Nous avons implanté la géométrie du routage sous le logiciel ADS.

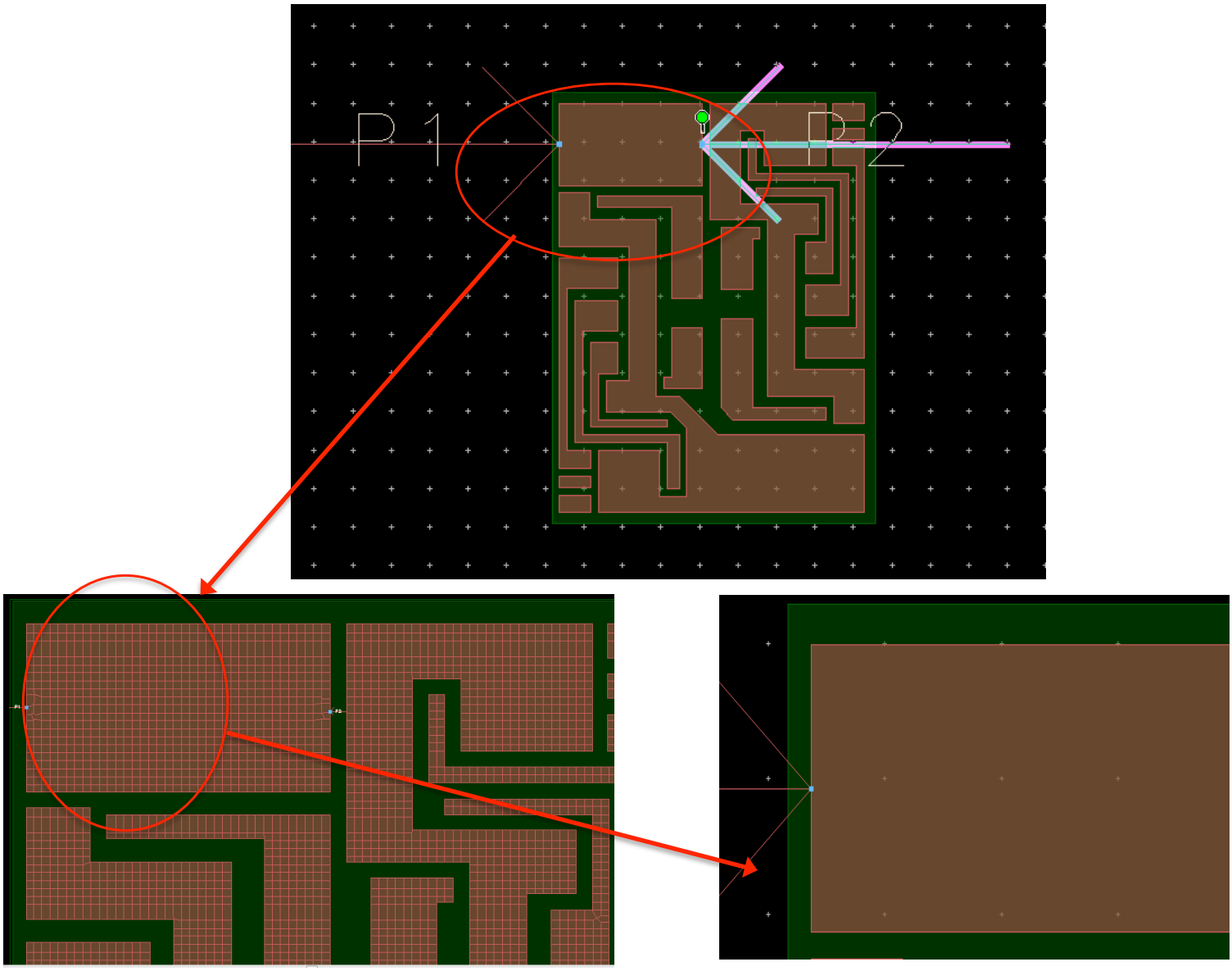


Figure 13 : Détail du routage réalisé sous le logiciel ADS®

Notre approche consiste à extraire des valeurs de capacités entre pistes et plan de masse, et de la comparer à des valeurs obtenues par des approches analytiques et expérimentales. La partie du circuit que nous avons étudié est représentée sur la figure 13.

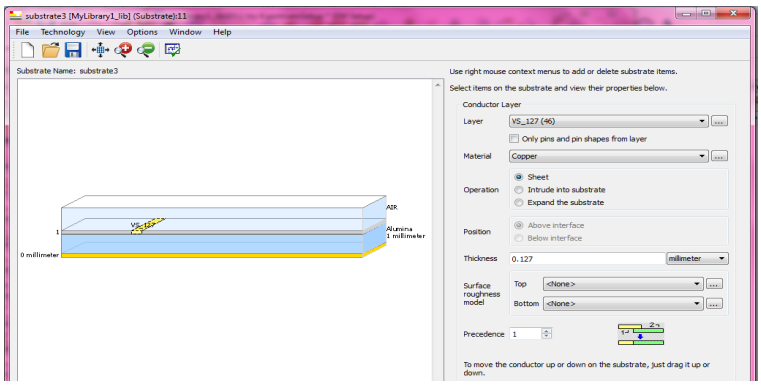


Figure 14. Constitution du substrat

La deuxième étape concernait le choix de la géométrie du substrat à étudier (piste, diélectrique, plan de masse) et de ses différentes propriétés (épaisseur, matériau, permittivité…).

Les simulations *via* ADS nous permettent également de trouver la capacité parasite. La simulation électromagnétique *via* Momentum nous permet d'extraire la matrice d'impédances des paramètres S. Puis grâce à l'utilisation de schematic, la matrice est implémentée dans un bloc, figure 15.

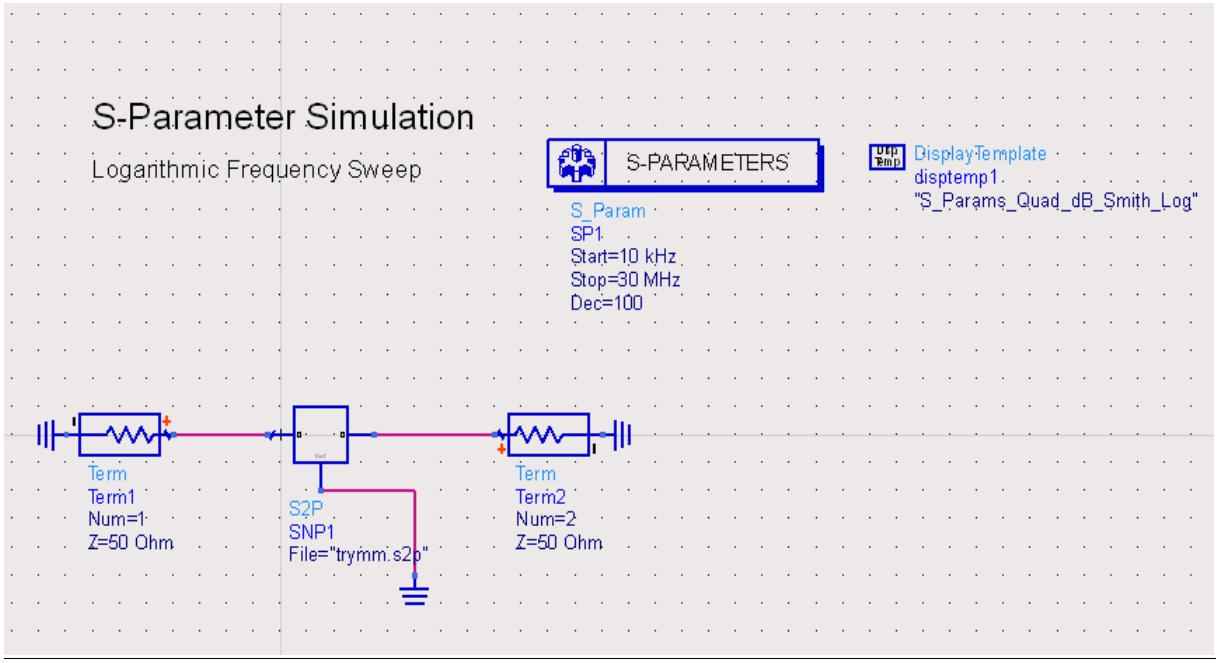


Figure 15. Schéma bloc utilisé pour passer des paramètres S à Z

Les connexions sont ensuite réalisées avec différents blocs, représentant chaque port, figure 15.

La simulation nous permet de retrouver les diagrammes d'amplitude et de phase de l'ensemble, figure 16.

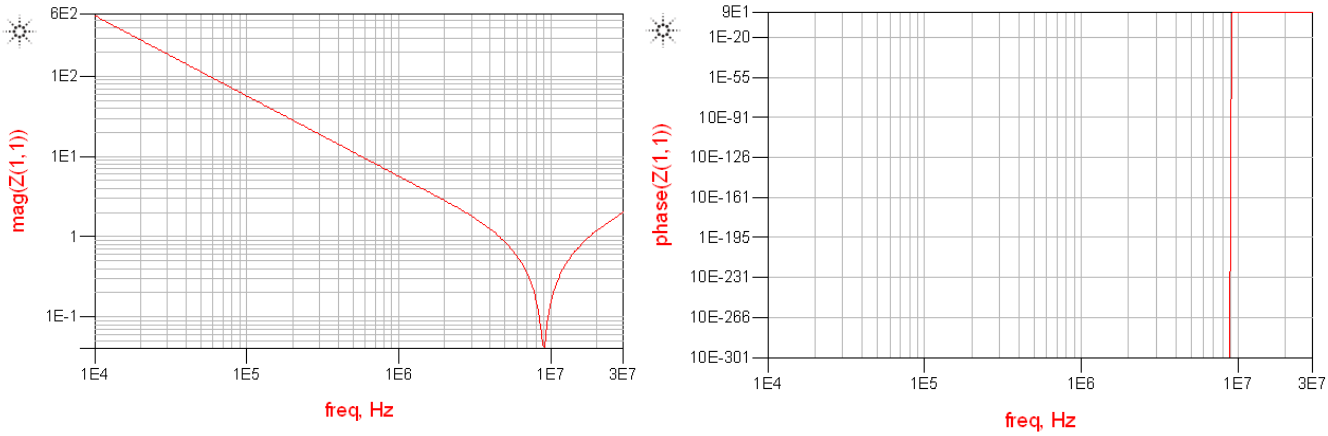


Figure 16. Diagrammes logarithmiques d'amplitude et de phase pour la partie du routage étudiée

Ici, la capacité qui nous intéresse correspond à Z_{11}.

À la fréquence de 1,019 MHz, l'amplitude est |Z|=84809,298Ω. On a ici à faire à un comportement capacitif que l'on peut déterminer grâce à l'allure du diagramme qui présente une pente en $\frac{1}{jC\omega}$.

Donc, $Z = \frac{1}{jC\omega}$ => $|Z| = \frac{1}{C\omega}$ => $C = \frac{1}{|Z|\omega} = \frac{1}{|Z|2\pi f}$ (5)

A.N. C=1.84e-12F.

Nous avons également déterminé cette capacité expérimentalement grâce à la méthode utilisée avec l'impédancemètre présentée à la section III.B).

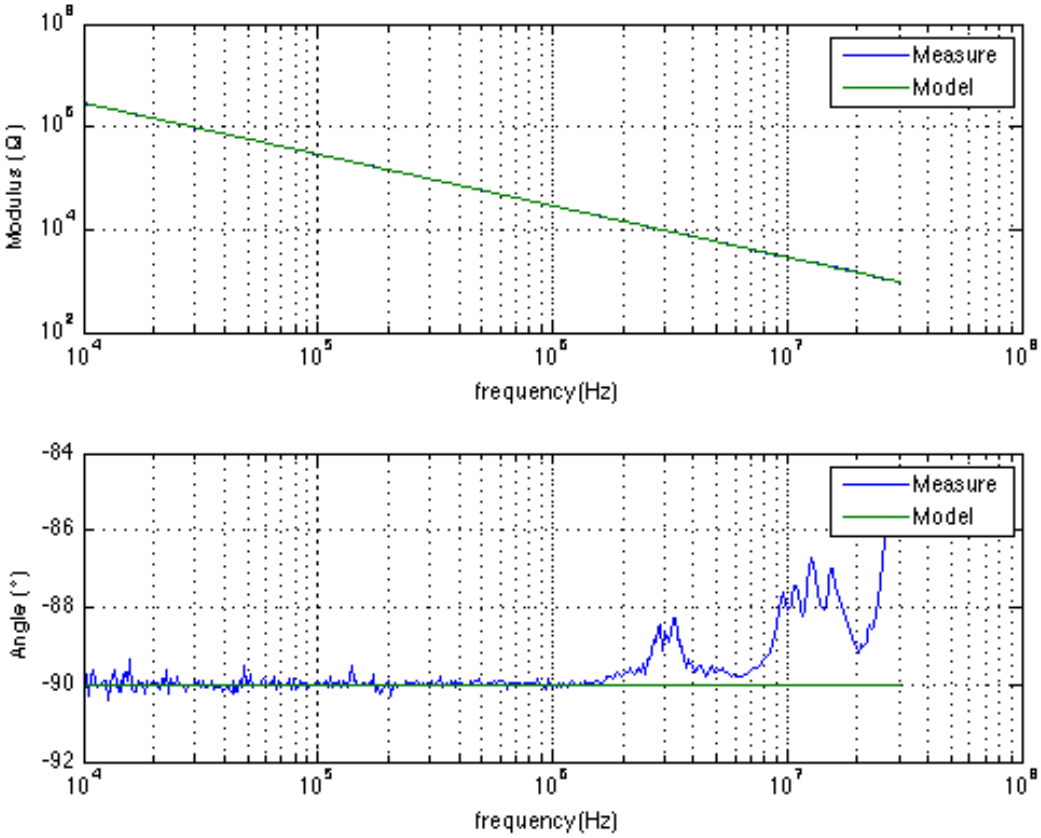


Figure 17. Diagramme de magnitude et de phase de la capacité obtenue expérimentalement

La valeur de la capacité déterminée expérimentalement est C=5,45e-12F.

La matrice d'impédances que nous avons obtenue est une matrice (34×34) puisque 17 pistes sont présentes sur le substrat. Cette matrice sera implantée sous le logiciel SABER® afin de réaliser une simulation circuit du système complet qui tient compte de l'influence du routage sur une plage de fréquence allant de 10kHz à 30MHz.

Les valeurs utilisées pour le calcul analytique de la capacité parasite entre le plan de masse et une piste sont données ci-dessous (cela correspond à la partie entourée dans la figure 13) :

W=5,3mm largeur de la piste

L=9,23mm longueur de la piste

h=1,00mm épaisseur du diélectrique

t=0,127mm épaisseur de la piste

ε_0=8,85714e-12F/m permittivité relative du vide

ε_r=9,6 permittivité du diélectrique

Plusieurs méthodes analytiques sont présentées pour calculer la capacité parasite étudiée.

La première est la relation donnée entre deux plaques conductrices planes séparées par un diélectrique de permittivité ε connue et d'épaisseur h.

$$C = \varepsilon_o \times \varepsilon_r \times \frac{W \times L}{h} \tag{2}$$

L'application numérique donne C= 4.16e-12F

La deuxième est la méthode des transformations conformes. Elle peut être notamment utilisée pour des géométries complexes.

$$C = \varepsilon_0 \times \varepsilon_r \times \left[\frac{W}{h} + \frac{2}{\pi} \times \left(1 + \ln\left(\frac{\pi \times W}{2 \times h} + 1\right)\right)\right] \times L \tag{3}$$

La dernière est une méthode empirique et est connue sous le nom de formule de Sakurai et Tamaru, [20]. Elle possède la particularité de prendre en compte les effets de bords. Ceux-ci sont non négligeables lorsque la largeur de la piste est faible devant celle du plan de masse, ce qui est le cas ici. C'est donc celle que nous utiliserons.

$$C = \varepsilon_0 \times \varepsilon_r \times \left[1.15 \times \frac{W}{h} + 2.8 \times \left(\frac{t}{h}\right)^{0.222}\right] \times L \tag{4}$$

VII- ASPECTS THERMIQUES

Le but de la manipulation présentée ici est d'évaluer les pertes dans un bras d'onduleur réalisé avec des transistors JFETs en SiC, [21]. Cette manipulation a été engagée afin de se familiariser avec les éléments d'électronique de puissance qui vont être étudiés dans le cadre du projet ACCITE. Une étude de manière thermique se révèle être un bon moyen pour évaluer précisément les pertes dans un module d'électronique de puissance.

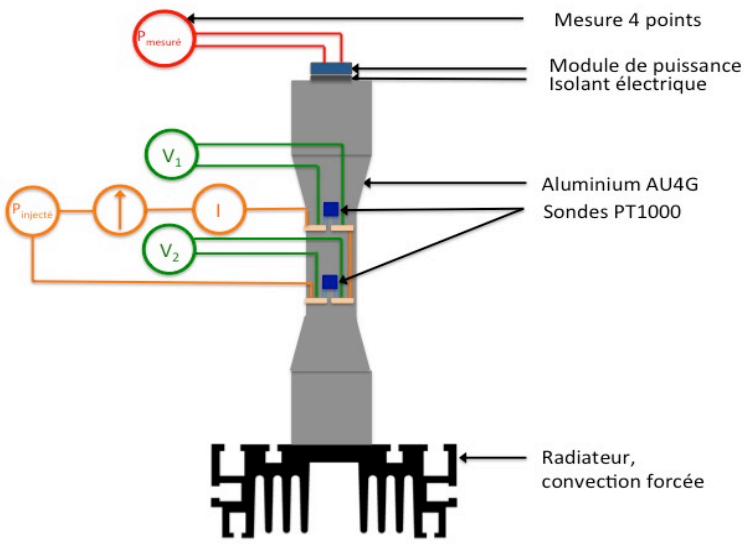


Figure 18 : schéma global du principe de mise en œuvre

Il existe plusieurs méthodes pour déterminer les pertes dans un module d'électronique de puissance tel qu'un convertisseur. L'évaluation de ces pertes dans un bras d'onduleur en mesurant le courant dans le drain et la tension drain-source constitue une tâche difficile due à la vitesse de commutation élevée des transistors, [22]. D'autre part, si le rendement du convertisseur atteint les 98%, la méthode de mesure de pertes, qui repose sur l'évaluation de la différence entre la puissance en sortie et la puissance en entrée devient inappropriée de part les incertitudes sur la mesure électrique. Plusieurs méthodes pour l'évaluation des pertes par calorimétrie ont déjà été mises au point. Cependant, elles ne garantissent pas une précision optimale des mesures. Dans [23], pour une puissance injectée de 30W, la précision est de 2% sur le dispositif de convertisseur DC-DC étudié. De plus, ces méthodes de calorimétrie, [24] [25] [26], se révèlent difficiles à mettre en œuvre (conception d'une chambre adiabatique).

A) PRINCIPE

DIMENSIONNEMENT DE LA PARTIE THERMIQUE

Dans cette optique, nous avons engagé des travaux sur l'étude des pertes dans un module de puissance par une voie thermique. Afin de déterminer le flux thermique à travers un circuit, une étude de la résistance thermique de ce dernier doit être réalisée, Figure 18.

Le dimensionnement de la colonne en aluminium a été réalisé utilisant une résistance thermique adaptée (environ 1°C/W pour faciliter les calculs). Cette dernière a été calculée sur le cylindre grâce à l'équation (6).

$$R_{th} = \frac{e}{\lambda S} \tag{6}$$

La conductivité de l'alliage de l'aluminium λ =181W.m^{-1}.K^{-1} étant connue, la surface a été fixée avec R, le rayon du cylindre, R=15,01mm pour permettre de mettre les composants de puissance au dessus de la colonne. La distance entre les deux sondes est e=10,083cm, la résistante thermique peut être calculée, R_{th}=0,7870°C/W.

Théoriquement, pour une puissance injectée, lorsque le courant et la tension continus aux bornes du composant (dans un premier temps une résistance) sont connus, les pertes peuvent être déterminées par la formule de la puissance électrique.

APPAREILS DE MESURES UTILISES

Les températures ont été mesurées par le biais de deux sondes de température (Labfacility, Platinum Sensing Resistors Thin Film Pt1000). Dans un premier temps, la source chaude est constituée par une résistance (Vishay, Power Resistor Thick Film Technology, LTO 100) qui a été placée au dessus de la colonne. Toute la puissance dissipée par les composants sous test est évacuée par conduction vers le radiateur, qui constitue ici la source froide. Le système est isolé thermiquement par de la vermiculite. La mesure des tensions V_1 et V_2 (figure 18) ainsi que le courant I (figure 18) a été réalisée *via* des multimètres (Keithley, 6 ½ Digit USB Digital Multimeter, 2100).

B) EVALUATION DES PERTES

MESURE, PRINCIPE DE L'UTILISATION DE LA DIFFERENCE DE TEMPERATURE

La résistance des sondes à 0°C est de 1000 Ω et l'intervalle fondamental entre 0°C et 100°C est de 385 Ω. Sur la plage de fonctionnement de notre expérience, on peut linéariser avec une précision de 0,15%, selon la norme IEC 751 [27]. L'expression de la température est donnée par l'évolution de la résistance (7).

$$T = \frac{R-1000}{3{,}85} \tag{7}$$

Les résistances des deux sondes PT1000 sont calculées en mesurant les tensions et courant dans le système, quand ce dernier a atteint son fonctionnement en régime établi. On obtient donc la différence de température sur la hauteur de la colonne. La résistance thermique est définie par l'équation (8).

$$R_{th} = \frac{\Delta T}{\phi} \tag{8}$$

ϕ correspond au flux thermique dans la colonne, i.e. les pertes.

CALIBRAGE ET INCERTITUDES

Dans la section VII.B), il a été noté que la différence de température pouvait être évaluée seulement une fois le régime établi atteint. Selon les mesures (Figure 19), la réponse temporelle du système à un échelon de puissance injectée correspond à un système du premier ordre. Le temps de réponse à 95% a été calculé par la méthode des moindres au carrés et est égal à 8 min. Les mesures ont donc été effectuées 30 min après la mise sous tension une fois le régime établi atteint.

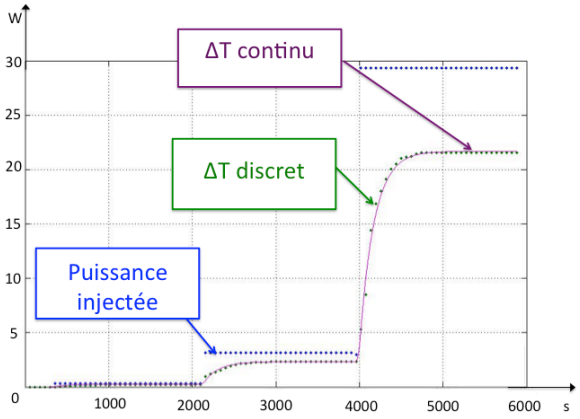


Figure 19. Réponse temporelle du système pour une gamme de puissance entre 0,3W et 30W

Toute la puissance dissipée par le composant est transférée par conduction dans la colonne. Celle-ci a été placée dans une boîte en plexiglas et remplie d'un isolant thermique (la vermiculite). Sa faible conductivité (λ=0,0694 $W.m^{-1}.K^{-1}$) nous permet de négliger les pertes par convection et par rayonnement. La figure 20 illustre une image du système sans l'isolation. Deux colonnes en aluminium sont placées côte à côte afin de pouvoir recevoir plusieurs composants ou des composants plus grands le cas échéant.

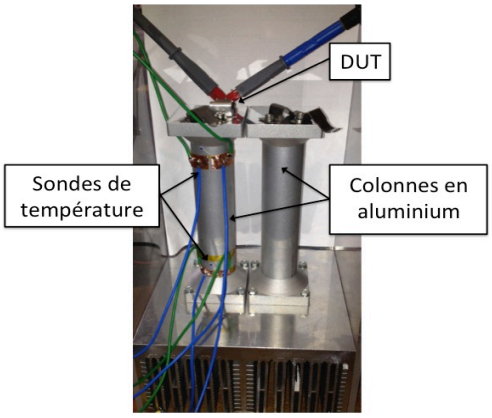


Figure 20. Système sans isolation thermique

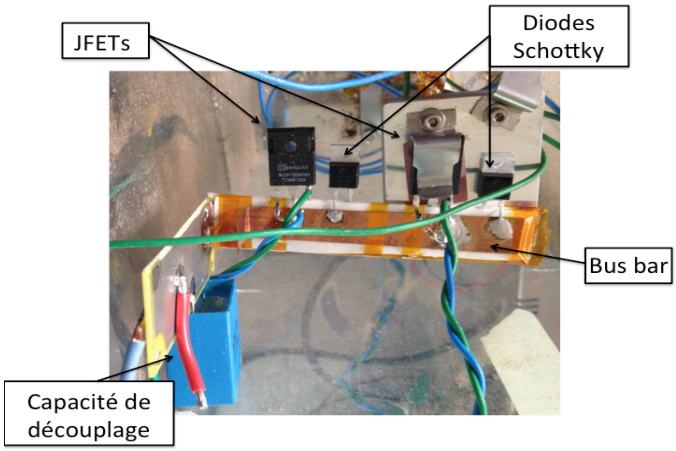


Figure 21. Montage du bras d'onduleur

La résistance est une fonction de la température des sondes. En mettant les deux sondes PT1000 en série, on impose une tension et l'on mesure le courant dans le circuit ainsi que la tension aux bornes de chaque sonde.

Les pertes (*i.e.* qui correspondent au flux thermique) sont déterminées expérimentalement avec la méthode qui vient d'être expliquée via l'équation (9).

$$P = \frac{\Delta T}{R_{th}} \tag{9}$$

Le tableau 2 récapitule les résultats.

$P_{mesurée}$	0,292W	2,744W	7,535W	13,21W	24,50W
Δ T	0,228°C	2,057°C	6,016°C	10,39°C	19,99°C
$P_{calculée}$	0,290W	2,616W	7,651W	13,22W	25,42W

Tableau 2. Résultats des différentes puissances injectées, mesurées et des différences de température

Les incertitudes sur le calcul des pertes, quelle que soit la puissance injectée peuvent être calculées suivant les équations (10) à (12).

$$\frac{\Delta\phi}{\phi} = \sqrt{(\frac{\Delta T}{T})^2 + (\frac{\Delta R_{th}}{R_{th}})^2} \tag{10}$$

avec,

$$\frac{\Delta T}{T} = \sqrt{(\frac{\Delta V_1}{V_1})^2 + (\frac{\Delta V_2}{V_2})^2 + (\frac{\Delta I}{I})^2} \tag{11}$$

et,

$$\frac{\Delta R_{th}}{R_{th}} = \sqrt{(\frac{\Delta e}{e})^2 + (\frac{\Delta \lambda}{\lambda})^2 + (\frac{\Delta S}{S})^2} \tag{12}$$

Les valeurs des différentes incertitudes sont données dans le tableau 3.

Incertitude	$\frac{\Delta e}{e}$	$\frac{\Delta \lambda}{\lambda}$	$\frac{\Delta S}{S}$	$\frac{\Delta V_1}{V_1}$	$\frac{\Delta V_2}{V_2}$	$\frac{\Delta I}{I}$
Valeur	$9{,}92e^{-3}$%	0,50%	0,067%	0,055%	0,055%	0,055%

Tableau 3. Valeurs des différentes incertitudes

C) CONCLUSION ET PERSPECTIVES DE LA PARTIE THERMIQUE

Nous avons présenté une nouvelle approche calorimétrique pour l'évaluation des pertes dans un module de puissance. Cette méthode garantit une bonne précision à plus ou moins 0,6%, et se révèle très stable dans le temps, une fois le régime permanent atteint. La moyenne des pertes par conduction et par commutation dans un bras d'onduleur (Figure 19) sur une période de temps sera évaluée. Ce bras d'onduleur sera constitué de deux JFETs verticaux et deux diodes Schottky. Une comparaison des pertes entre un bras d'onduleur avec et un bras sans diode Schottky sera présentée dans la suite des travaux, il permettra de compléter les travaux initiés au laboratoire AMPERE sur les JFETs verticaux, [28].

VIII- CONCLUSION ET PERSPECTIVES

Dans ce mémoire de première année, une étude bibliographique sur le convertisseur statique de puissance dans un environnement sévère vous a été présentée.

La méthode de travail que nous avons mise au point concernant la prise en compte du routage futur du convertisseur a constitué l'avant dernière section. Les composants passifs (capacité de découplage, résistance de grille, capacité de grille) ont été modélisés à partir des mesures réalisées à l'impédancemètre. Le comportement de ces composants passifs et d'une piste du routage ont été étudiés entre 40Hz et 10MHz. Les résultats obtenus physiquement ont été comparés aux simulations sous ADS® et aux calculs analytiques.

Parallèlement, dans la dernière partie, les travaux de recherche sur les pertes de puissance dans un bras d'onduleur constitué de diodes Schottky et de JFETs verticaux ont été présentés. La démarche scientifique et l'explication de la manipulation ont été expliquées.

En termes de perspectives, la première des choses à réaliser et la finalisation de la manipulation présentée dans la dernière section sur les aspects thermiques. L'estimation des pertes de puissance dans un bras d'onduleur avec et sans diodes de roue libre sera réalisée.

Parallèlement la modélisation du bras d'onduleur étudié dans la section *VI* d'un point de vue de la CEM sera effectuée. La matrice d'impédances du routage, obtenue grâce au logiciel ADS sera implantée sous SABER afin d'établir le comportement du système sur une plage de fréquence donnée. L'étude des composants semi-conducteurs constituant le bras et leur comportement sur la même plage de fréquence sera également faite. Une comparaison entre les résultats analytiques, logiciels, et physiques sera menée.

Le bras d'onduleur haute température réalisé via le projet THOR sera testé en vue d'une alimentation pour trois bobines haute température faites entre le laboratoire GREEN (ENSEM, Nancy) et le laboratoire LSEE (Université d'Artois, Béthune). Le tout sera testé au GREEN dans un environnement haute température. L'ensemble bras d'onduleur-bobine(s) sera mis dans un four afin de tester son fonctionnement à 300°C. Ceci sera effectué au GREEN courant avril 2014.

La réalisation d'un nouveau bras d'onduleur que nous modéliserons et dimensionnerons en nous basant sur les travaux réalisés au laboratoire AMPERE constituera la suite du projet. Il sera fait en prenant en compte les contraintes de *dv/dt* imposés par la constitution de la machine synchrone (bobines).

Enfin, la réalisation d'un convertisseur statique de puissance qui fonctionnera à haute température sera menée dans la dernière partie du doctorat afin de pouvoir l'implanter sur le flasque arrière de la machine synchrone à aimants permanents réalisée par les laboratoires partenaires du projet ACCITE.

L'ensemble des tâches à effectuer est visible sur le tableur excel en annexe.

IX- REFERENCES

[1] J. C. Mankins, "TECHNOLOGY READINESS LEVELS," pp. 4–8, 1995.

[2] R. Robutel, "Étude des composants passifs pour l'électronique de puissance à 'haute température' : application au filtre CEM d'entrée," *INSA Lyon*, 2011.

[3] R. Robutel, C. Martin, H. Morel, C. Buttay, P. Mattavelli, D. Boroyevich, and R. Meuret, "Design and Implementation of Integrated Common Mode Capacitors for SiC JFET Inverters," *IEEE Transactions Power Electronics*, vol. X, no. X, pp. 1–1, 2013.

[4] C. Buttay, and D. Bergogne, "Thermal Stability of Silicon Carbide Power JFETs Thermal Stability of Silicon Carbide Power JFETs."

[5] T. Friedli, S. Round, D. Hassler, and J. W. Kolar, "Design and Performance of a 200-kHz All-SiC JFET Current DC-Link Back-to-Back Converter," *IEEE Trans. Ind. Appl.*, vol. 45, no. 5, pp. 1868–1878, 2009.

[6] M. Metz, "Alimentations à découpage Le transformateur," in *Techniques de l'ingénieur*, vol. 33, no. 0, 1991, pp. 0–12.

[7] F. Costa, "CEM en électronique de puissance Sources de perturbations, couplages, SEM," in *Techniques de l'ingénieur*, vol. 33, no. 0, 2013.

[8] J. Rabkowski, D. Peftitsis, and H. Nee, "Silicon Carbide Power Transistors: A New Era in Power Electronics Is Initiated," *IEEE Industrial Electronics Magazine*, vol. 6, no. 2, pp. 17–26, Jun. 2012.

[9] G. Jérome, "Contribution a l'Etude du Rayonnement des Câbles Soumis aux Signaux de l' Electronique de Puissance dans un Environnement Aéronautique," *Université des sciences et technologies de Lille*, 2008.

[10] Z. Xu, F. Xu, P. Ning, and F. Wang, "Development of a 30 kW Si IGBT based three-phase converter for operation at 200 °C with high temperature coolant in hybrid electric vehicle applications," in *2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, pp. 3027–3033, 2013.

[11] TAN (J.), COOPER Jr. (J.A.) et MELLOCH (M.R.), "Highvoltage accumulation-layer UMOSFET.s in 4H-SiC. *IEEE Electron Device Letters*, 487-489, 1998.

[12] J. D. Scofield, H. Kosai, B. Jordan, S. H. Ryu, S. Krishnaswami, F. Husna, and A. Agarwal, "High Temperature DC-DC Converter Performance Comparison Using SiC JFETs, BJTs and Si MOSFETs," *Material Science Forum*, vol. 556–557, pp. 991–994, 2007.

[13] R. Burgos, Z. Chen, D. Boroyevich, and F.Wang. "Design considerations of a fast 0-ohm gate-drive circuit for 1.2 kv sic jfet devices in phase-leg configuration," IEEE Energy Conversion Congress and Exposition (ECCE) ,2009.

[14] D. Malec, G. Vélu, S. Duchesne, B. Nahidmobarakeh, C. Buttay, and C. Vollaire, "Canevas de projet ACCITE." pp. 1–30, 2012.

[15] RTCA, "DO-160F." p. 460, 2007.

[16] Center For Power Electronics Systems. Safran final report - discrete passive emi filter design. Livrable du Projet CPES-Safran Phase 1 (2008).

[17] Y.Maillet, R. Lai, S. Wang, F. Wang, R. Burgos, and D. Boroyevich. High- density emi filter design for dc-fed motor drives. IEEE Transactions on Power Electronics, Volume 25, Numéro 5, p1163-1172 (2010).

[18] C. Butttay, P. Bevilacqua, K. E L FAlahi, S. H AscoE, L. Phung, D. Turnier, B. Allard, D. Panson, I. Lyon, and B. Lyon, "High temperature , Smart Power Module for aircraft actuators Power devices," in *HiTEN'13, Oxford : United Kingdom*, 2013.

[19] E. Rondon, F. Morel, C. Vollaire, and J.-L. Schanen, "Impact of SiC components on the EMC behaviour of a power electronics converter," in *2012 IEEE Energy Conversion Congress and Exposition (ECCE)*, pp. 4411–4417, 2012.

[20] T. Sakurai and K. Tamaru, "Simple formulas for two- and three-dimensional capacitances," *IEEE Transactions Electron Devices*, vol. 30, no. 2, pp. 183–185, 1983.

[21] P. G. Neudeck and R. S. Okojie, "High-temperature electronics - a role for wide bandgap semiconductors?," *IEEE Proceedings*, vol. 90, no. 6, pp. 1065–1076, 2002.

[22] C. Cai, W. Zhou, and K. Sheng, "Characteristics and Application of Normally-Off SiC-JFETs in Converters Without Antiparallel Diodes," *IEEE Transactions Power Electronics*, vol. 28, no. 10, pp. 4850–4860, 2013.

[23] L. Hoffmann, C. Gautier, S. Lefebvre, and F. Costa, "Thermal measurement of losses of GaN power transistors for optimization of their drive," in *EPE Power Electronics and Applications, 2013 European Conference on*, pp. 1–8, 2013.

[24] D. Christen, U. Badstuebner, J. Biela, and J. W. Kolar, "Calorimetric Power Loss Measurement for Highly Efficient Converters," *2010 International Power Electronics Conference - ECCE ASIA -*, pp. 1438–1445, Jun. 2010.

[25] K. Bradley, W. Cao, J. Clare, and P. Wheeler, "Predicting Inverter-Induced Harmonic Loss by Improved Harmonic Injection," *IEEE Transactions Power Electronics*, vol. 23, no. 5, pp. 2619–2624, Sep. 2008.

[26] K. J. Bradley and A. Ferrah, "Development of a High-Precision Calorimeter for Measuring Power Loss in Electrical Machines," *IEEE Transactions Instrumental Measurement*, vol. 58, no. 3, pp. 570–577, Mar. 2009.

[27] "DIN IEC 751 Temperature / Resistance Table for Platinum Sensors."

[28] R. Ouaida, X. Fonteneau, F. Dubois, D. Bergogne, F. Morel, H. Morel, and S. Oge, "SiC vertical JFET pure diode-less inverter leg," in *2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2013, pp. 512–517.

Ecole Centrale de Lyon - INSA de Lyon – Université Claude Bernard Lyon 1

Laboratoire Ampère

Unité Mixte de Recherche du CNRS - UMR 5005

Génie Electrique, Electromagnétisme, Automatique, Microbiologie environnementale et Applications

Mémoire doctorant 1ère année 2012 -2013

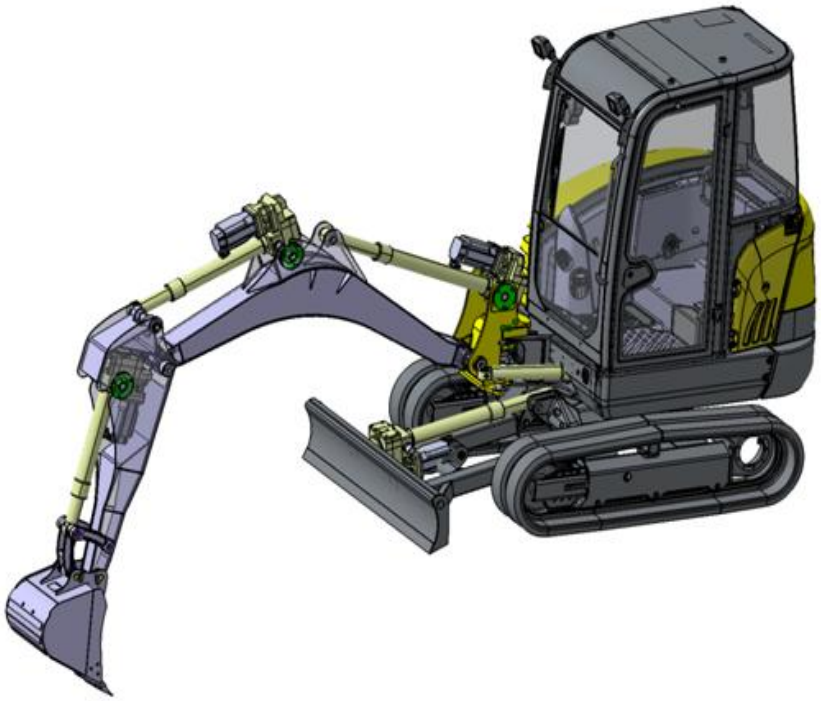
Nom - Prénom	**Than Tien Thinh**
Titre de la thèse	Actuation architecture evaluation for an electric mini-excavator
Directeur de thèse	**Prof. Eric Bideaux**
Co- encadrants	**Dr. Mohamed Smaoui**
Dpt. de rattachement	**Méthode pour l'Ingénierie Système**
Date début des travaux	**11/09/2012**
Type de financement	**Bourse CNRS**

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France

Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

INSA de Lyon

Actuation architecture evaluation for an electric mini-excavator



Supervisors: Prof. Eric Bideaux; Dr. Mohamed Smaoui

Student: Than Tien Thinh

FUI ELEXC** project is financed by **OSEO** and **FEDER

Project Partners: Volvo Technology, Volvo Construction Equipment, EFS, ELBI, Bonfiglioli, ProLion, SymbioFCell

Table of content

Abstract

The requirement of the market as well as environmental concerns have created opportunities for electrically actuated construction equipments. Among these, mini-excavators have the potential to become the new kind of electric machine. A project has been launched to manufacture a prototype for electric excavator. Within this project, there rises the necessity to model and reconstruct the operation of the electric as well as hydraulic excavators using simulation tools, which are the tasks for my part in the project. After conducting a literature review, I have modelled and simulated the machine's operation kinematically and dynamically. For kinematic modelling, I have utilised the Denavit-Hartenberg method as well as conventional geometry. For dynamic modelling, I have utilised the Lagrangian method, or more exactly, derived the Lagrangian equation of the second kind. For calculation and simulation, I have used MATLAB and AMESim, dedicated softwares for these tasks. I have also designed preliminary PID controllers for the excavator model, to be able to reconstruct the digging cycle for the excavator, as data regarding operator's activities have not been provided.

For the task of reconstructing the reaction forces acting on electro-mechanical actuators, I have studied a test bench with a hydraulic actuator to reconstruct the force and an electro-mechanical actuator to follow a course. I have controlled this test bench model with PID control and have also studied the mathematical model to prepare for non-linear model-based control. The sliding mode control is available but yet to be validated, so only the PID control is presented.

The advance so far is good for the project as there is now a model ready to be implemented with control strategies. This model will form the foundation for later works of global control, comparison and optimisation.

1 Introduction

As the price of fossil fuels is increasing almost everyday, there is currently a growing trend to move away from vehicles, machines, and equipments using fossil fuels to ones using electricity. While it can be seen that there are a lot of achievements in machines and transportational vehicles, there has not been such a level of advancement in the field of construction equipments.

For this sub-set of equipments, earth-moving machines have another unique character. Nowadays, they are actuated mainly by hydraulic systems, and although this type of actuation has many advantages, it still poses disadvantages that can discourage its use in some circumstances.

Firstly the hydraulic actuators are very noisy. This creates acoustic pollutants, especially when they are operated in dense urban area. Imagine if earth-moving equipments are quieter, this may allow inner city construction sites to be worked day and night, without the complains from the people living nearby having their sleeps disturbed, and strongly increases the speed of the construction works. To this aspect, electric equipments are usually much less noisy. Anyone who is reading this paragraph can remember the "not too bad" noise from electric home appliances. It is clearly advantages of electric equipments over hydraulic ones.

Secondly, the hydraulic actuators produce a lot of polluting oils. In this time of "green" economy, this is truly a big disadvantage. Of course there may be some efforts going on to develop "greener" oil, but a system that produces no oil at all, like an electric one, would have clear advantage.

Thirdly, the current state of energy management for electric systems are quite good. As electric motors can also work as generators at time, this allows energy management methods to be applied to have good energy performance for the equipments.

Among the earth-moving equipments, the excavator is one of the most widely used. It can range from mini-excavator with about half-ton earth-moving capacity used in small construction sites to giant heavy super excavator that can move close to 100 tons of soil to be used in large open-pit mines. The excavator, then, has been chosen as the object for this research because of its popularity among earth-moving equipments.

Figure 1.1. A comparison of a giant excavator and a normal one (Source: Internet).

Therefore, the requirement of the market as well as the environmental conscience lead to the necessity for novel types of earth-moving equipments that are less noisy, less polluting and have better energy efficiency. The hydraulic excavators, although popularly used at the moment, use a large quantity of polluting oil and generate acoustic "pollutants" as well. Therefore, together with the current state of mechatronic technology, it may give rise to new opportunities to improve the performance of excavators energetically and acoustically.

An idea of building an electrically-actuated excavator was then proposed to explore the abilities of current available technologies to design and manufacture a mini excavator prototype to survey the potential of such a product. The context part for this project will be conducted in Laboratoire Ampère, Lyon.

The state of France, with their commitments to support pioneering and innovative research in high technology toward a green economy, has always been side by side with research institutions

by offering assistances in policies as well as finance, is a major stakeholder in the existence development of many scientific projects, among them is this one. Such commitment can be considered invaluable, especially against many challenges and adversaries of our time.

The research is conducted at INSA de Lyon within Université de Lyon, one of the top academic and research institution in Europe. Among many laboratories affiliated with our institution, Laboratoire Ampère is a dynamic and productive laboratory. With many experts present at this laboratory in many research domains including modelling, control, biomechanics, … the project can be seen to have got a very capable academic partner.

Volvo Corporation, which is a big player in the multinational market of cars, trucks, and construction equipments, is the project's main industrial partner. As proved by many years of providing excellent products and services, no one can doubt their expertise and ability in being a very pro-active and state-of-the-art partner.

Together with Volvo, there are many other partners, namely EFS, ELBI, Bonfiglioli, ProLion, SymbioFCell. Together, we form a good chain of cooperation to develop this very interesting project.

Returning to the context of the project, it is framed by the hypothesis that excavators based on a complete electric architecture can now operate within a small range with light duty. This hypothesis argues that according to the current state of the electrical technology it is still more cost-effective to design small electric excavators to work with light-duty tasks like road repair or landscaping, rather than in large building sites or open-pit mines. Designing electric excavators for heavier-duty tasks would lead to very high cost and heavy electric specifications, which may not be justified. Since the performance of the machine is light-duty, its specification can differ greatly from heavier duty excavators.

There are three PhD students currently working in this project. Among the three, my part in the project is to handle the mechanical part of the project. Utilising calculation and simulation tools, the task is to evaluate differences that may exist between the three architectures: electric only,

electro-hydraulic, and the currently in used hydraulic architecture. This task forms the foundation for the modelling of the whole excavator, as the other two parts, control using power electronics and energy management will be mounted with the respective architectures.

There have not been many researches regarding the field of research and development for electric excavators. However, the hydraulic excavator is a very popular object for scientists and engineers alike. The state of the art for this project will mainly consist of works about hydraulic excavators as well as electric machines and control strategies.

As stated in the title, this report is the summary of my first year progress, basically a study of the modelling of the proposed electric mini-excavator, typical of those common in road repair work. The broad objective of the work done is to understand how the machine is used and how it can be modelled kinematically and dynamically.

The modelling stage gives way to the control stage of the excavator model. Although another person is mainly involved with this stage, close cooperation will be a necessity in proposing control strategies for the global excavator model. Novelties in methodology, strategy or modelling will be proposed and evaluated in this stage, though the author of this report can only claim partial credits on them.

As the main difference between the hydraulic excavator and the electric one is the use of electro-mechanical actuators in place of hydraulic excavators, an evaluation step must be taken for the electro-mechanical actuator. The main task for this step should be identifying various forms of mechanical energy loss as well as evaluating their effects. This task will be accomplished with the assistance of a test bench. The work surrounding this test bench, like modelling, control designing, result evaluating will also be a source for novelties, especially in the control and evaluation aspects. It should be noted that this test bench would have two parts to serve two purpose. The electro-mechanical part is an electro-mechanical actuator, it will be modelled then controlled to follow a defined course to study the character of its performance, while the hydraulic part is a hydraulic actuator, it will be modelled then controlled to follow a defined force to reconstruct the soil force effect acting on the electro-mechanical actuator.

After the validation of the modelling step, which means that the results from the simulation on the excavator virtual model can be seen as reliable, the comparison will be made between the hydraulic excavator and the electric excavator. The comparison will focus mainly on the energetical performance of the two systems. All effects resulted from differences between the two systems will be evaluated, including architecture, energy flow, weight and load, operational differences, etc… Until now, little is known about the performance duty-tradeoff for electric excavators, although there is expectation that such a trade-off exists. One of the contributions of the project will also be an insight into elements of the tradeoff with the specific machine under study - the VOLVO EC27C-model electric excavator.

With the modelling and computing capability present, it is possible to participate in the sub-field of proposing optimal trajectories for earth-moving processes. This is a very interesting research domain where new and newer mathematical models, theories, methods are proposed. This project will try to have an interesting voice in this domain, while trying to make the most from what can be available, whether theoretical or experimental.

The final stage of this project will be the implementation of an on-board real-time system capable of monitoring excavator activity and teaching various operational strategies for machine operators that can optimize the performance and working cost of the excavator. Although this kind of system has been available for several years with some other types of manned machine, to the knowledge of the author, this is the very first presence of such a system for the electric excavator.

In conclusion, my work for the first year is around two main tasks: One, is the kinematic and dynamic modelling of the excavator, and simple controllers to be in the place of the operator for the reconstruction purpose. Two, is the modelling and control of the test bench of two parts, electro-mechanical part and hydraulic part. The first task can be considered more important, however, the second task is very important, in the sense that the hydraulic actuator must be able to reconstruct the reaction force acting against the electro-mechanical actuator, while the electro-mechanical actuator must be able to follow a pre-defined trajectory.

1.1 State of the art

For the modelling of multi-link mechanical systems like excavators, there have been a lot of books, but one of the most interesting ones is the work by Dombre and Khalil [1]. Their study investigate both kinematic and dynamic aspects, as well as contemporary methodologies and control aspects. Also, this book introduces the popular as well as comprehensive techniques currently under usage in this domain. For the kinematic modelling of excavators, normally one with very strong geometric competence can model the system using high school knowledge, however, a systematic and convenient approach has been designed by Denavit and Hartenberg [2]. Since decades, this method has been used extensively in kinematic modelling of multi-link systems. In this project, this method is also used for kinematic modelling. For dynamic modelling, currently there are two popular methods: Euler-Lagrange's equations or Newton-Euler's equations. As stated by Koivo et al [3] and other sources, the Euler-Lagrange's equations may give very complex expressions, while Newton-Euler's equations provide a recursive calculating procedure, which may be more convenient to compute. However, although very lengthy, if one can cleverly arrange elements in the expression, the Euler-Lagrange's equations are actually manageable. This approach has been taken by the author of this report to derive the equations of motion for the excavator, which can be seen under the matrix form in Appendix C. The equations, although still lengthy, have been shortened by applying trigonometric grouping to produce shorter expressions.

The excavator, after being modelled kinematically and dynamically, will be modelled with AMESim, which is a dedicated software for simulating dynamic systems. The excavator AMESim model will then be tested with control attempts, will be made by the author as well as a project colleague, then, some literature review regarding how to control the excavator and its actuators should be done. Before entering this, it should be made clear the reason for our control attempts. As the data provided by Volvo does not contain any information regarding the operator's actions, controllers should be put into the model so digging cycles can be reconstructed, hence the requirement for controllers. This is very important as one may question the "reason to be" of control schemes introduced in this report.

One of the most popular methods of control for hydraulic excavator is PID control, which is also described by Kelly [4], and will be used in the beginning part of this project. However, as the hydraulic system is highly non-linear, usually, a pure PID control approach is not taken, but instead a modified one. For instance, Morita and Sakawa [5] used PID control with a feedforward controller based on inverse dynamics, Song and Koivo [6] used PID control together with a neural network controller, while Seward Lucie [7] went as far as integrating artificial intelligence for a high-level control system, then (s)he used a PID controller as a low-level control system to serve the high-level one. As stated previously, in the beginning part of this project, a pure PID control approach is taken for the electric excavator model. This is to facilitate other tasks in this project until a more complex control strategy designed by a colleague is ready for the global excavator model.

As the focus of this project is the electric excavator, not the hydraulic excavator, the control of electric machines is of high interest. The type of electric motors utilised in this project is synchronous motor (see Figure 1.2), therefore, the control of this particular type of electric machine will be studied extensively. Due to the popularity of this type of electric machine, many of control techniques ever proposed has its representative with the synchronous motor. For example, field-oriented control by Saleh et al [8], feedback linearisation by Kuroe et al [9], sliding mode by Yang et al [10], Elmas and Ustun [11]. However, in industrial applications, the PI(D) controller is still the most widely used control strategy.

Figure 1.2. A Bonfiglioli synchronous motor, like the ones used in this project (Source: Bonfiglioli).

As the PID controller is by far the most popular control method, it is interesting to study several well-known PID tuning methods, among maybe hundreds of ones ever proposed. Most noteworthy is the method proposed by Ziegler and Nichols [12], as well as the IMC PID-tuning method by Rivera et al [13], and the one by Smith and Corripio [14]. The Ziegler-Nichols [12] tuning can result in a very good disturbance response for integrating processes, but are also known to result in quite aggressive tunings, and give unsatisfied performance for processes with a dominant delay, according to Tyreus and Luyben [15]. On the other hand, the IMC-tunings of Rivera et al., though generally give very good responses for setpoint changes, are known to result in poor disturbance response for integrating processes (Chien and Fruehauf [16]). For PID tuning in this project so far, the Ziegler-Nichols method is preferred. The reason for this needs to be stated again, that for the excavator, the purpose of the controllers is to reconstruct the operator's actions, which have not been provided, so a simple method can be preferred.

One caution should be raised while using this method is that the physical system can enter its saturation mode. However, it should be noted again that the control design process at this stage only serves the reconstruction purpose, therefore, results will not be used unconditionnally without caution.

Though the PID controller can control the global excavator model, as well as both the hydraulic and electric actuators of the test bench, it does not require knowledge about the system to design. Therefore, two model-based control strategies of sliding mode and backstepping are proposed for controlling the hydraulic actuator of the test bench. As stated before, the hydraulic actuator needs to be force-controlled to reconstruct the force against the electro-mechanical actuator.

Regarding model-based control strategies, as stated by Slotine and Li [17], the most useful and general approach for evaluating the stability of controlled non-linear systems is the Lyapunov theory. Generally speaking, with any system, if there exists a function (called Lyapunov function) which is globally strictly positive, and that function's derivation over time is negative definite, then the system is stable. Then, the idea is, if a controlled non-linear system is designed in such a way that a Lyapunov function can be found, the controlled system is stable.

The two control strategies presented next, which, as stated above, would be applied to control the part of the test bench, will utilise Lyapunov theory to have a rigorous theoretical proof that they

can do the job. They both have their own Lyapunov functions, and the advantages as well as the reason why these strategies have been chosen lie in the fact that they can be proved by a theory, while PID control is not always mathematically proved to be able to control a system.

Sliding mode control is a control strategy developed about fifty years ago, starting with the book by Emelyanov [18]. The idea is to create a defined mathematical "surface" and utilising Lyapunov criterion to design a controller to "push" the system toward that "surface", then guaranteeing the stability of the system. Until now, sliding mode control has possessed many applications, to name a few are examples with chemical processes [19] or mechanical systems [20]. Sliding mode control can also be seen as a very useful strategy in Multi-Input Multi-Output control of non-linear systems, as stated in [21] and [22].

Backstepping technique is an interesting control strategy, and also intensively studied in our laboratory. Basically, it provides a recursive control design, starting with the simplest stable subsystem to build a series of virtual feedback controllers in order to stabilize each and every subsystem until the final external control is reached [23], [24]. Due to its ease of implementation, this technique has been applied with success to control purposes in many fields, including hydraulics [25], pneumatics [26], electrics [27], magnetics [28], and robotics [29].

As these control strategies require knowledge about the system, or more exactly the mathematical model of the system, this aspect also forms an important part of the literature review. For the electromechanical actuator, it can be seen that the mathematical model for the synchronous motor has become somewhat classical, as stated in [30]. However, for the part of the roller screw, there are still various developping models.

Roller screw is a device used to transform a rotational motion into a translational motion and sometimes the other way round. Due to their higher efficiency in comparison with lead screws and ball screws, roller screws are used extensively in heavy-load, long-life and heavy-use applications. Figure 1.3 shows a roller screw with its threaded shaft and rollers (hence the name roller screw).

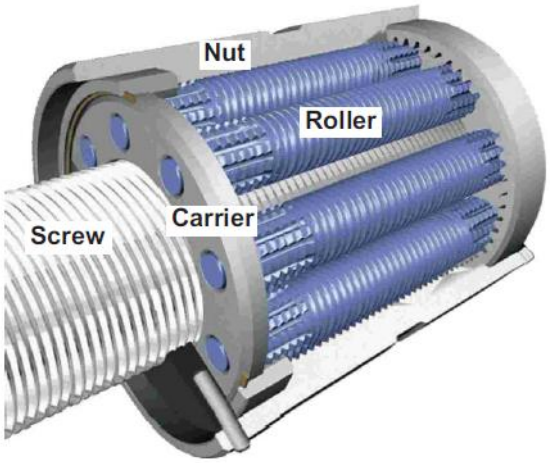


Figure 1.3. Cut view of a roller screw ([31]).

Obviously, the mechanical energy losses at roller screw are due to friction and deformation, and any scientific work must take into account both phenomena [32] [33] [34]. However, different authors usually have different set of assumptions, therefore they will arrive with different conclusions. Therefore, the modelling of the roller screw has been done with the fact that there is no dominant theory regarding this subject. Despite this fact, and taking into account that the author has yet access to a test bench with the roller screw mounted on, the ideas from Velinsky [31], in which the motion of the components are meticulously calculated and simple, basic friction expressions are applied, shared the author's agreement. As some other theories focus on deformation [33] [34], when the test bench is available, deformation can be checked to see whether they really count, or some kinematic study, as in [31] can be enough.

For the hydraulic actuator, a very good book about modelling and controlling of this kind system is the one by Merritt [35], which presents almost all aspects surrounding the hydraulic actuator, with various levels of modelling, from very detailed to some levels of model reduction. The mathematical model presented in this report is also adapted from this work. Although there have been many more updated additions to the hydraulic actuator's mathematical model, such as ones by Dransfield [36] or Gotz [37], the basic form of the model is similar to Merritt [35], as presented by Bishop et al [38] . As the hydraulic actuator modelling is very important, literature review will continue to see whether there are important dynamic phenomena that have not been taken into account by Merritt [35].

Return to the global model of the excavator, beside Dombre and Khalil, Koivo is also an established researcher in the field of excavator modelling ([39] and [3]), who investigated the kinematic and dynamic aspects of excavators, and also worked with control of this kind of machine. His research group did extensive work in modelling construction equipments, as well as stating the importance of surface material science in studying about this kind of equipments and other kinds alike. The global modelling in this project, so far, has inherited the notations (how to name points, how to define angles) mainly from these two research groups.

As the excavator is a earth-moving equipment, the study of machine-soil interaction is of fundamental importance. In this domain the work by Alekseeva et al. is considered both pioneering and classical [40]. In their work various mathematical models for soil force have been proposed, as well as the effects of soil force on the excavator bucket. These models will play a crucial role in the future optimisation of the performance of the excavator.

Among these models, our current focus is the Fundamental Equation of Earthmoving (FEE) by Reece [41]. This equation takes into account many aspects of a digging process, among them soil parameters like density and cohesion; geometric parameters like tool depth and tool width; and tool parameters like tool profile and material.

Utilising various earthmoving models, the parametric identification and optimisation of digging process is in the interest of many researchers. Among many studies in this sub-field, we can note the ones by Hall and McAree, [42], Chang and Lee [43], Luengo et al [44], Vahed et al [45], and Kim et al [46], who examined a number of algorithms to track pre-defined or online-defined trajectories for the excavator bucket to serve the purpose of optimisation. This work will also be done by the team at Ampère Laboratory – our team - to try to define an optimal trajectory and follow that trajectory by utilising controllers.

In conclusion, the literature review, though not able to cover the whole scope of related fields, like very large fields of non-linear control, soil mechanics or excavator researches, tries to cover the classical, popular and closed-to-the-topic studies. For small sub-field like roller screws, the

author has tried to note the more cited literatures, as there are not so many of them. This literature review, therefore, has been useful for the development of the works done in the first year.

1.2 The VOLVO EC27C-model electric excavator

The VOLVO EC27C-MODEL excavator to serve as the equipment reference for this project, in a face shovel configuration, is shown in Figure 1.4. It is primarily used for excavation and loading. While it can propel to transport material short distances, it is used almost exclusively to load truck vehicles by slewing from the face to a dump area. The maximum propel speed is moderate at around 5 km/hour but the excavator has good positional capability and can work in closed quarters allowing it to dig selectively.

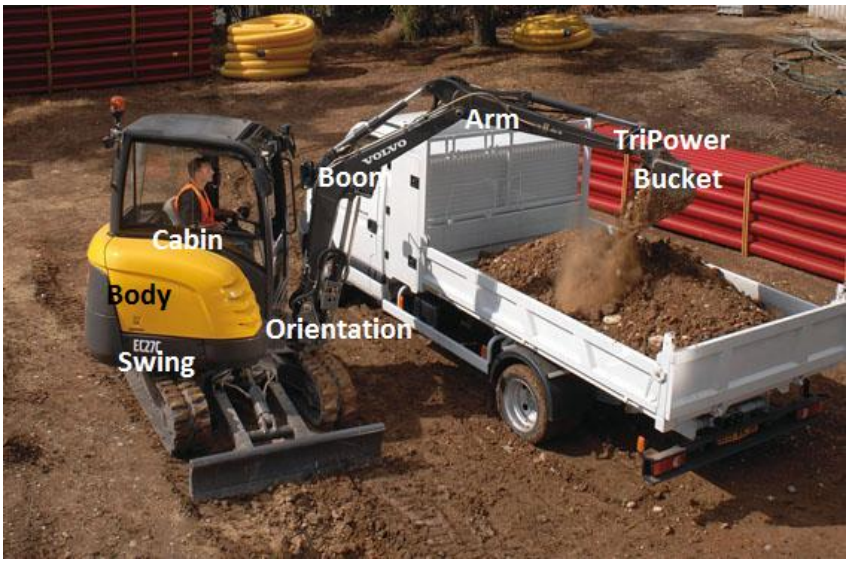


Figure 1.4: Schematic of the VOLVO EC27C excavator (source: VOLVO Construction Equipment)

The following terminology is used when referring to the excavator,

Excavator body: The main structure of the excavator, housing the operator, the engine, fuel, electric pumps, and supporting the digging arm equipment.

Operator cabin: Usually located on the left side looking outward along the arm equipment. The operator therefore has good visibility while digging and quite moderate visibility when dumping, particularly to large trucks.

Boom: Proximal link of excavation arm whose main function is to raise and lower the bucket through actuation of the boom cylinder. The boom connects to the excavator body at the boom root through a rotary joint about which it pivots.

Arm: Medial link of the excavation arm equipment whose main function is to extend the arm horizontally under actuation of the arm cylinder. The arm is connected to the boom by a rotary joint about which it pivots.

Bucket: The digging tool which is located as the distal link on the excavation arm equipment.

TriPower: Arrangement linking the boom and bucket roll cylinders via a triangular rocker that pivots relative to the boom. This arrangement gives a constant carry angle of the bucket under hoist and crowd motions and is claimed (by the long history of excavators) to improve digging performance.

Stunga: The term taken for the solid link connecting to the triangular rocker to the excavator body.

Boom cylinder: Cylinder placed on the boom connecting the excavator body to the triangular rocker. This cylinder raises and lowers the boom.

Arm cylinder: Cylinder on the arm that connects between the boom and arm. This cylinder extends and retracts the arm.

Bucket-roll cylinder: This cylinder is connecting the triangular rocker to the bucket. This cylinder enables the rotation of the bucket.

Swing system: The motors and mechanical drive transmission mounted in the excavator body enabling the excavator to pivot about a vertical slew axis.

Orientation system: The cylinder mounted in the excavator body enabling the arm equipment to pivot about a vertical axis.

Propel system: The motors and mechanical components driving crawler tracks that propel the excavator.

All the hydraulic actuators are replaced by electric ones in this project. The cylinders are replaced by electromechanical actuators, while hydraulic motors are replaced by electric motors.

1.3 Basic excavator operation

The basic machine operating cycle consists of a digging pass through the bank, a loaded swing (or carry) to the dump position, the dump into a haul truck, empty swing back to the digging face and repositioning or spotting of the bucket at the face.

The machine is controlled by dual, two degrees of freedom joysticks that enable the operator to control the velocity of the boom, arm, and bucket cylinders and to the swinging system. In the case of a hydraulic system, the joystick angle drives the distributor opening and therefore the flow rate into the cylinder. The arm and bucket cylinders are controlled by one joystick; the boom-cylinder and swinging system by the other.

Vertical (hoist) and horizontal (crowd) motion of the bucket are achieved by actuating the boom and arm cylinders, facilitating motion of the bucket through the bank.

The orientation of the bucket is controlled by the bucket cylinder. (The TriPower arrangement decouples bucket rotations from hoist and crowd motions.) The bucket is usually rotated as it moves through the bank so that it cradles collected material [47].

1.4 Summary of field data generating trials

The work and conclusions of this project are based on data collected during data generating trials conducted by Volvo Construction Equipment at their factory in Belley. The trials involved the fitting of a set of sensors to the excavator. A total of ten cycles of operational and performance data was recorded. The excavator dug in the same type of earth for the duration of the trials. The data generating cycles are summarized in Table 1.

Cycle	Description of task	Percentage of operation time
A1	Trenching (trenching angle less than 30°)	25%
A2	Loading trucks (raising angle less than 45°)	15%
A4	Like A1 but without orientation, only with deportation	3%
A5	Digging along a wall	3%

B1	Bottom digging	8%
C1	Scouring	8%
D1	Levelling	5%
E1	Backfilling	2%
F	Translation with blade on soil	19%
G	Translation with elevated blade	5%

Table 1: Cycle periods.

The following data was collected on the hydraulic excavator.

- Arm, boom, and bucket cylinder displacement.
- Pressures in each chamber of the arm, boom, and bucket cylinder. Measurements were made using pressure sensors placed at each cylinder chamber supply core. The measurement of the pressures in each chamber enables to determine the hydraulic force applied to the piston.

1.5 Report structure

The remainder of **Section 1** has given relevant background material in preparation for the main works of this project so far. This includes an overview of the VOLVO EC27C-model electric excavator and a summary of the data gathering.

Section 2 analyses this overview data in order to show the preliminary part of modelling. Also, this section presents how the electro-mechanical actuator is modelled and controlled.

Section 3 describes the kinematic model developed as part of this work that links the bucket trajectory to the cylinder displacements. Section 3 also applies this algorithm to track the bucket trajectories kinematically for each of the three jacks that worked on the excavator during one of the working cycles. Cycle A1, which is a popular and distinct digging modes has been chosen. It is shown that the three jacks under PID controllers follow very well the trajectory. Also in this section describes an algorithm to identify the external forces acting on the bucket by the soil utilising the excavator's dynamics. With the force data provided by Volvo Construction

Equipment on three jacks, the result of simulation verifies the correctness of the proposed algorithm.

Section 3 is the centre of this report as it covers the kinematic modelling of the arm equipment, and for the dynamic modelling, it is expected to cover the whole excavator. However, until now, only the dynamic modelling of the arm equipment is ready. The dynamic modelling of other movements of the excavator is pending to wait for more parametric data.

Section 4 describes attempts to model and control the hydraulic and electro-mechanical actuator on a test bench in development at Ampère.

Section 5 summarizes the work on the project so far and. Suggestions are also made to future steps towards the vision of developing an electric excavator with optimal energy performance.

2 Modelling of the electro-mechanical actuator and initial modelling for the excavator

The software tool that was selected for simulation during this project is AMESim, which is a very strong dedicated software for energy-based modelling. MATLAB-Simulink will also be used extensively in this project for calculation and control modelling.

The objective for this section is to prepare for the next section, which means, to prepare these inputs for the AMESim models: the arm equipment of the excavator for a digging cycle, then electro-mechanical actuators to actuate that arm equipment, and then controllers to control those electro-mechanical actuators. This means that the section prepares elements for the modelling, calculation and simulation to be presented in Section 3.

2.1 Modelling of the arm equipment of the excavator

AMESim provides very intuitive libraries for modelling systems. Firstly, elements were taken from the software's library to represent the boom, arm, bucket, … bodies. These elements were then connected by revolutional joints. Jacks were then mounted and connected with electromechanical actuators.

Then, the electric excavator's arm equipment model is built in AMESim as in Figure 2.1 and Figure 2.2. The meaning of icons are detailed in Appendix A.

To be clear, one should look at Figure 2.1, the eight degrees of freedom for the excavator are presented, where there are four rotations in the xOy plane: boom, arm, bucket and blade; there are two rotations in the xOz plane: orientation and rotation; and there are two translations in the xOz plane for the two chain wheels.

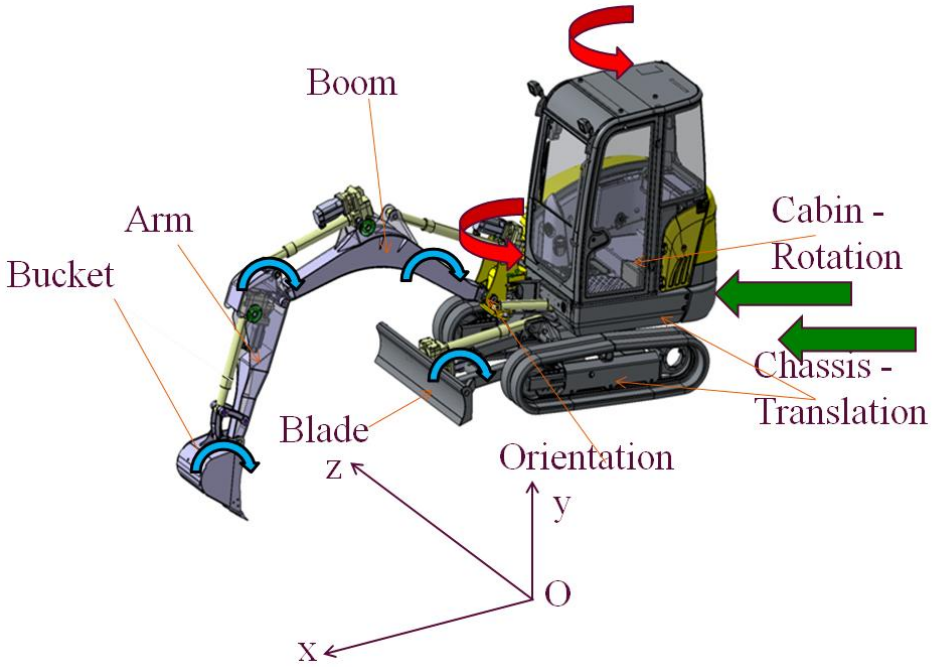


Figure 2.1. Electric excavator model in AMESim – Assembly CAD image for the excavator (Adapted from one provided by Volvo).

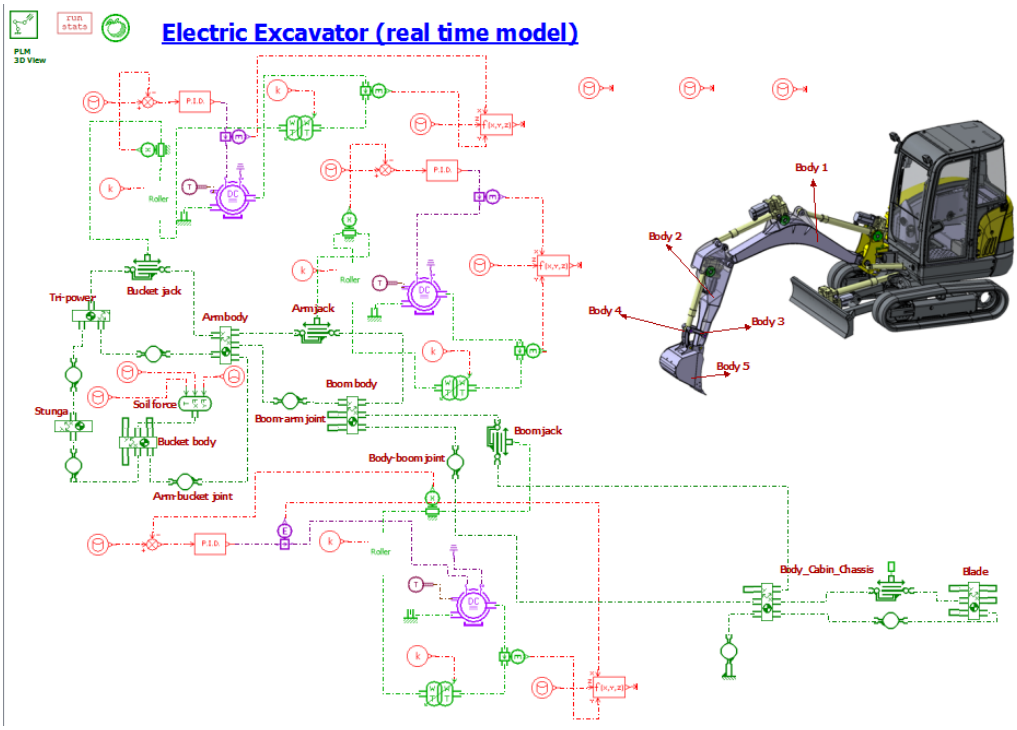


Figure 2.2. Electric excavator arm equipment model in AMESim – Mechanical model for the excavator.

The advantage of AMESim is that the icons used are very intuitive. In Figure 2.2, maybe the only symbols that are difficult to understand are the ones representing the excavator's elements. There are five elements being modelled, respectively the boom, the arm, the tri-powers, and the bucket for the arm equipment of the excavator.

2.2 Modelling of the electromechanical actuator

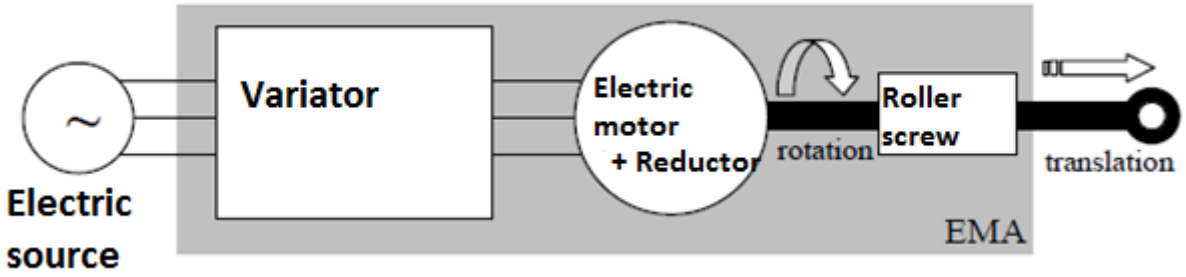


Figure 2.3. Block diagram of the electro-mechanical actuator (EMA) (Adapted from [48]).

The electromechanical actuator (Figure 2.3) is constituted of the roller screw, the velocity reductor and the electric motor. The roller screw is modelled by a dedicated part of the software packet of AMESim called AMESet, which allows to develop specific components based on a mathematical model to be used in AMESim.

The velocity reductor is taken from AMESim library. Basically, it is just a velocity-torque transformer with a ratio k, which is 6.9 for the bucket, arm, and boom.

The electric motor is a three-phase synchronous motor, and it should be modelled as such. However, at an energetical level it can be approximated by a DC motor. In order to facilitate the evaluation stage later in this project, the author has specified the DC motor's parameters to be as equivalent as possible to the synchronous motor. Naturally, in the later stage of this project, the three-phase synchronous motor with its controller will replace the DC motor in modelling.

As there has not been many literatures about the roller screw, then so far in this project, the base for modelling the roller screw is the works by Velinsky et al ([31], [49]) and Karam, [48].

The roller screw is modelled energetically as a transformer that transforms mechanical energy in rotational form into mechanical energy in translational form. This transformation involves many kinds of energy loss, mainly to translational friction and rolling friction.

Karam [48] modelled the roller screw with varying efficiencies depending on the direction of the rotating shaft, while Velinsky et al [31] modelled the roller screw with constant efficiencies depending on many parameters. As the test bench for characterising the roller screw has not been available, the Velinsky model has been chosen. The idea is, as stated previously in the literature review, that the kinematics of the roller screw, or more precisely, the relative velocities of roller screw's elements are meticulously calculated. Then, friction is calculated according to simple Coulomb friction. Presently, this approach considers a constant efficiency which is simpler to model in AMESim. But experimental works are planned to characterise precisely the losses of the roller screw in the electromechanical actuator.

The Velinsky model presents:

$$v \,.\, F = \omega \,.\, T \,.\, \eta \qquad (2.1)$$

with

ω: Rotational velocity; T: Input torque; η: Efficiency; v : Translational velocity; F: Force

Where

$$\eta = \frac{r_S \ tan\,\alpha_S (cos\,\alpha_S \ cos\,\rho \ \sin\beta - \sin\,\alpha_S \ \sin\rho - f_k\,(\sin\,\alpha_S \ \cos\rho + \cos\,\alpha_S \ \sin\rho \ \sin\beta))}{r_S(\cos\rho \ \sin\,\alpha_S \ \sin\beta + \sin\rho \ \cos\,\alpha_S) + \ r_{RP} \ (1-\cos\rho)(\sin\rho + \cos\rho)\cos\,\alpha_S \ \cos\beta} \qquad (2.2)$$

With

f_k : Coulomb kinetic coefficient of friction; f_r : rolling coefficient of friction; r_{RP} : radius of curvature of the roller side profile; r_S : screw radius; β : contact angle; α_S : screw helix angle; ρ : friction angle.

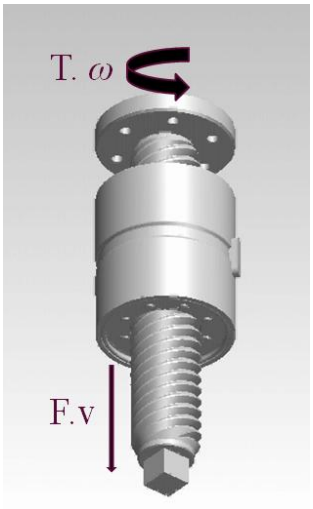


Figure 2.4. Roller screw with torque and force.

This expression (2.2) is based on kinematic calculation, with the hypothesis that there is no slip between the rollers and the screw in the rotational direction, also the nut is prevented from rotating. Also, the assumptions including isotropic linear elastic material properties, no surface tractions, smooth and continuous surfaces are hypothesised. Although these hypotheses may not always hold true during the operation of the electromechanical actuator, for the current stage in the project, it can be considered sufficiently correct for the modelling [31].

2.3 PID controller design for the electromechanical actuator

With the DC motor, the mathematical model can be expressed as

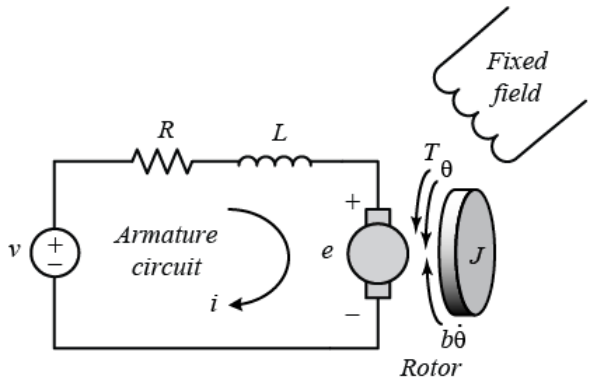


Figure 2.5. DC motor scheme and circuit. (Source: Internet – University of Michigan database)

$$T = K_T . i \qquad (2.3)$$

With T as output torque, i as current and K_T is a constant parameter called torque constant.

The back electromotive force (back emf), *e*, is proportional to the angular velocity of the shaft, $\dot{\theta}$, by a constant factor K_e.

$$e = K_e.\dot{\theta} \qquad (2.4)$$

As it is assumed that the electrical energy loss from the circuit (e.i) is equal to the mechanical energy gain at the shaft (T. $\dot{\theta}$), then $K_T = K_e = K$.

Then, we can derive the following equations based on Newton's second law and Kirchhoff's voltage law.

$$J\ddot{\theta} + b\dot{\theta} = K.i \qquad (2.5)$$

$$L\frac{di}{dt} + Ri = V - K\dot{\theta} \qquad (2.6)$$

With: J: moment of inertia of the rotor system (including the reductor and roller screw)
b: motor viscous friction constant; L : electric inductance; R : electric resistance; V : electric voltage

Applying the Laplace transform, the above equations can be expressed in terms of the Laplace variable *s*.

$$s(Js + b)\theta(s) = KI(s) \qquad (2.7)$$

$$(Ls + R)I(s) = V(s) - Ks\theta(s) \qquad (2.8)$$

We arrive at the following transfer function between the rotational angle and the voltage:

$$\frac{\theta(s)}{V(s)} = \frac{K}{s(Js+b)(Ls+R)+sK^2} \qquad (2.9)$$

Note that the rotational angle of the shaft and the translational displacement of the jack are mechanically linked to each other by a ratio p (p = reductor ratio x roller screw ratio for our system), the transfer function between the jack displacement and the voltage is:

$$\frac{X(s)}{V(s)} = \frac{Kp}{s(Js+b)(Ls+R)+sK^2} \qquad (2.10)$$

This is a third-order system, theoretically, there is no analytical way to establish a PID controller for such a system, therefore, the popular Ziegler-Nichols method [12] is chosen to design the PID controller with the help of Matlab-Simulink.

Although the detailed calculation to get the P-I-D gains is not presented here, the result of this PID controller will be presented in the next sections.

3 Kinematic and dynamic modelling

In the previous sections we presented data collected in the normal operation of the excavator. To understand the machine more precisely, there rises the need to model the machine in operation as a mechanical system. Then, the most important cycle, cycle A1, with 25% of operational time, has been chosen to be presented here kinematically and dynamically. However, it should be noted that once a cycle is modelled successfully, all the other cycles can also be modelled and tracked using the same methods and algorithms, as there is no theoretical nor structural differences among the cycles. The results of other cycles are under construction and will be presented in future reports. For this digging cycle, the excavator base does not move, so the results are presented only for the digging arm equipment. It should be noted that in the Appendix the dynamic model of the whole eight degrees of freedom is presented.

3.1 Kinematic modelling

This section presents a kinematic model that allows the angles of the three parts of the equipment: bucket, arm, and boom to be determined from knowledge of the cylinder displacements. This calculation is based on kinematic geometry. Please see Appendix B for more details.

The equipment linkage is shown in Fig. 3.1. The three primary links, namely the boom, arm, and bucket, are connected by rotary joints. Actuation is by extension and retraction of three electro-mechanical actuators, commanded independently by operator joysticks. The three degrees of freedom of the excavation equipment are augmented by swing motion about the axis *y* passing through the excavator body.

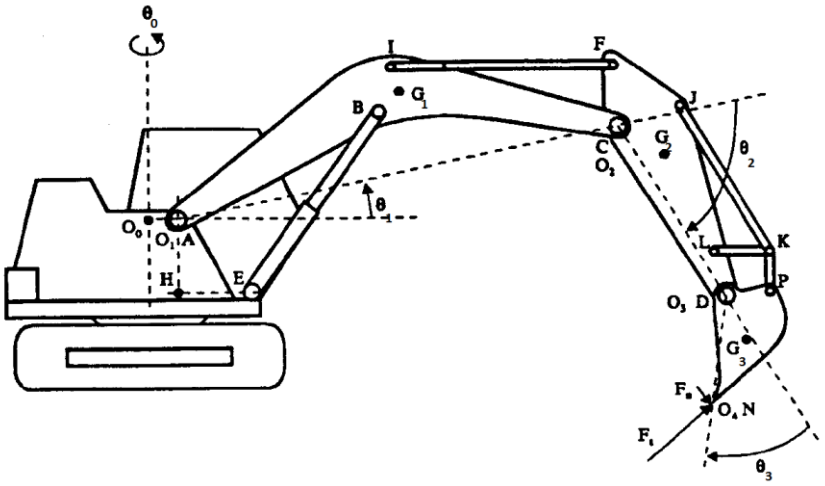


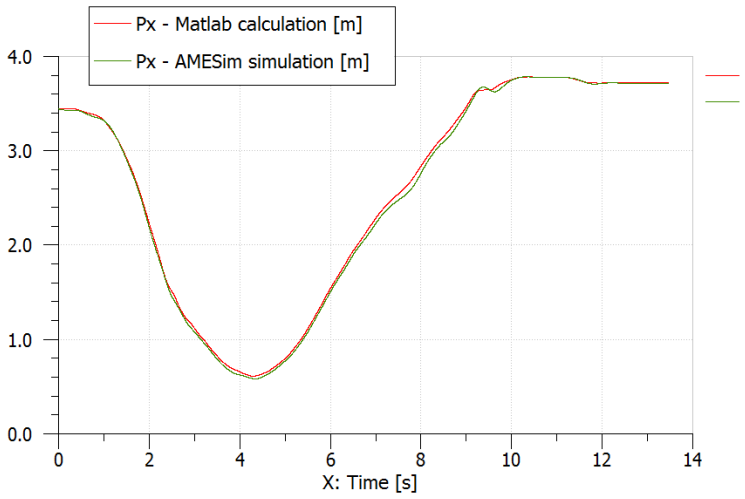
Figure 3.1: Kinematic layout of the excavation arm equipment. The arm equipment is actuated by electric actuators. Solid circles indicate revolute joints [50].

The model described in this section gives a method for modelling the angles of the boom, arm and bucket links relative to a global reference frame. Having determined these angles, and with knowledge of the pivot angle, the spatial position of the bucket tip can be readily determined by applying the transform method of Denavit and Hartenberg ([2], [51]).

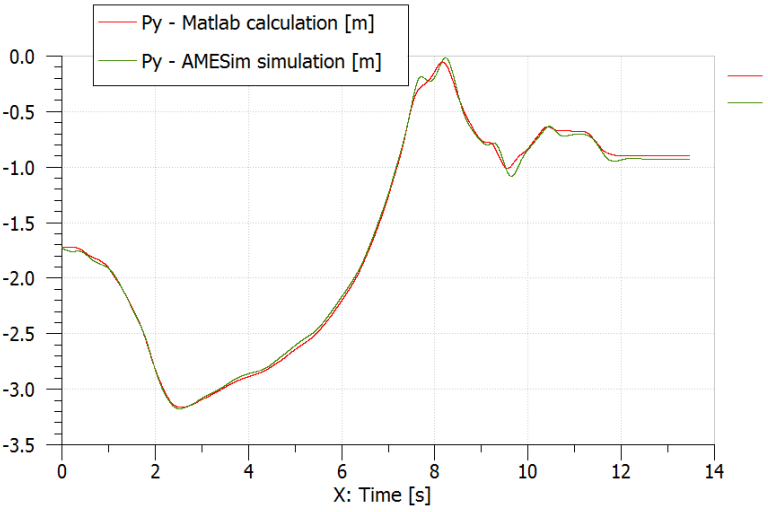
Basically, there are two algorithms, one to transform the design data and extensions data into rotating angles $\theta_{1,2,3}$, and the other one to transform the design data and angles data $\theta_{1,2,3}$ into extensions and global (O_0) coordinates of all points on the arm equipment. The latter algorithm is very important for dynamic modelling, as all coordinates and angles for the arm equipment are calculated from this algorithm, and will give much shorter expressions in the dynamic formulae.

Then, the AMESim model is designed with PID controllers as stated in the previous section to track the kinematics of cycle A1. It should be noted that if the tracking is successful, the direct and indirect kinematic algorithms has been validated.

Figure 3.2 shows the successful modelling of cycle A1. Figure 3.2.a shows the calculation and simulation result of global coordinates of the bucket tip P in x, while Figure 3.2.b shows the calculation and simulation result of P in y. It should be noted that the MATLAB calculation is using data from Volvo, and should be seen as a reference.



a.



b.

Figure 3.2. Kinematic modelling result, a. Result for Px, b. Result for Py.

3.2 Dynamic modelling

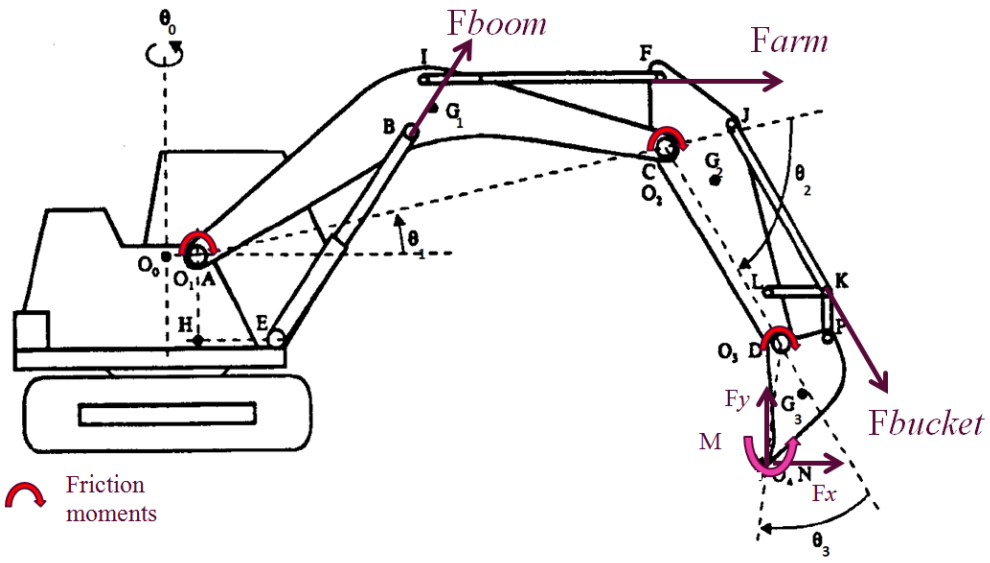


Figure 3.3. Dynamic modelling – the excavator arm equipment under external forces and moments.

For dynamic modelling of the electric excavator, or more exactly the reconstruction of dynamic forces acting on the arm equipment, the algorithm is a little more complicated, as the forces from soil has to be identified.

We note that the whole excavator has eight degrees of freedom, three for the arm equipment, one for orientation, one for rotation, one for blade, and two for translation. As one and two degrees of freedom systems have simpler dynamics, the focus will be on the arm equipment, which has very complicated dynamics. Readers can see Appendix C for dynamics of the arm equipment of the excavator.

In the previous section there is possibility to calculate rotating angles from jack extension data, then, these angles are numerically derivated using the three-point formula approximation to get the rotational velocities and accelerations in preparation for the coming dynamic calculation.

Naturally, the force from soil while digging acts in a broad surface with multiple directions. The modelling software does not permit such area-based modelling, and it is also impossible to identify such force profile, therefore, this force profile can only be “moved” to the bucket tip.

According to the classical law of "force moving", this force profile will be replaced by a force F and a moment M, acting at the bucket tip.

As the Volvo data supplied three forces: boom force, arm force, and bucket force, respectively the actuating forces at the three jacks, it is mathematically possible to calculate the two elements of the force F: Fx and Fy, together with the moment M.

The dynamic modelling has been done by the author, which some suggestions from works by Koivo et al [3]. Basically, it utilises the classical Lagrangian equations, by calculating the kinetic energy of the arm equipment, and the generalized external forces.

The Lagrange equation of the second kind takes the form:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta_n}}\right) - \frac{\partial T}{\partial \theta_n} = -\frac{\partial \mathrm{V}}{\partial \theta_n} + Q_n \qquad (\mathrm{n} = 1, 2, 3) \qquad (3.1)$$

to be processed and rearranged in the matrix form as

$$D(\theta).\ddot{\theta} + C\left(\theta, \dot{\theta}\right).\dot{\theta} + G(\theta) + B\left(\dot{\theta}\right) = L(\theta).F - F_{soil} \qquad (3.2)$$

Where:

T: kinetic energy of the arm equipment

θ_n: respectively boom, arm, and bucket angles according to Figure 3.1.

V: potential energy of the arm equipment

Q_n: respectively external generalized forces acting on boom, arm, and bucket coordinates. These forces are calculated based on virtual work principle.

θ: angle matrix of boom, arm, and bucket angles representing three generalized coordinates for the three degrees of freedom

$D(\theta)$: inertial matrix ; $C\left(\theta, \dot{\theta}\right)$: Coriolis matrix ; $G(\theta)$: gravity matrix

$B\left(\dot{\theta}\right)$: viscous friction matrix ; $L(\theta)$: geometrical matrix

F: force matrix of boom, arm, and bucket forces

F_{soil}: soil force matrix, which is actually moments at boom, arm, and bucket by soil force profile

The processes and expressions for all these terms above are very lengthy so they are included in the Appendix C. This model comes with these hypotheses:

- The jacks are considered absolute solid bodies, their masses are included in the masses of the excavator's elements, and the effects of their movements are totally neglected. Also, the masses of the tri-power and stunga are neglected due to their small values.
- The dry friction forces at the jacks are considered to be already included in the efficiencies of the electro-mechanical actuators, and are not presented in the model.
- The friction at the joints are considered to be purely viscous.

The most important note is that it is possible to calculate Fx, Fy, and M from F_{soil}, with

$$F_{soil} = L(\theta).F - (D(\theta).\ddot{\theta} + C(\theta,\dot{\theta}).\dot{\theta} + G(\theta) + B(\dot{\theta})) \qquad (3.3)$$

Then, Fx, Fy, and M are input into the AMESim model, to recalculate the forces by actuators at boom, arm, and bucket. The modelling setup for actuators is presented in Figure 3.4 and the result is presented in Figure 3.5.

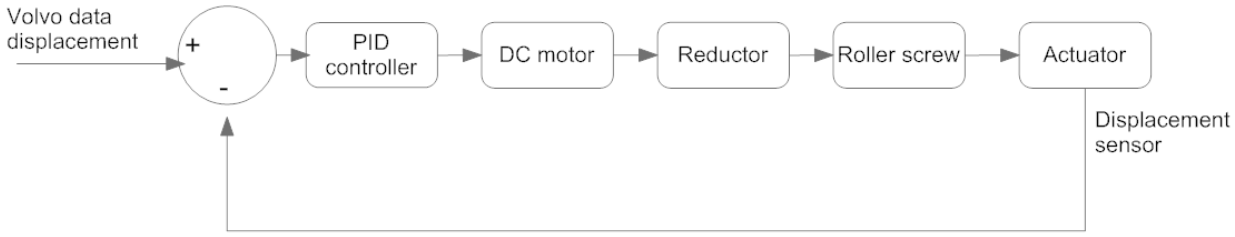
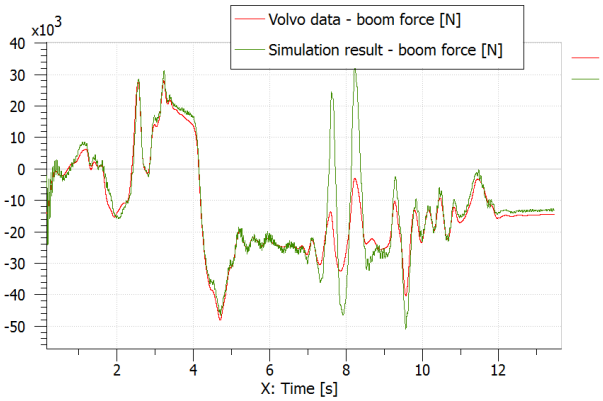
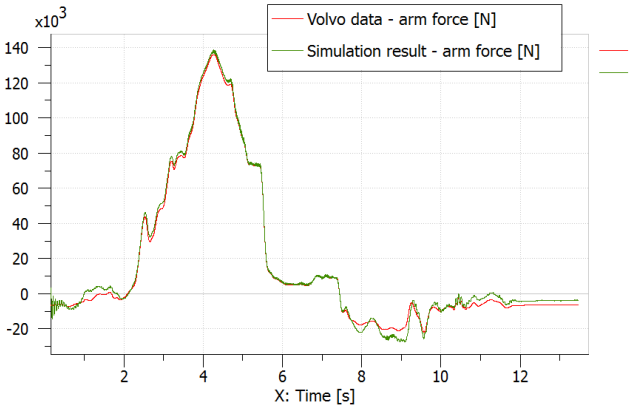


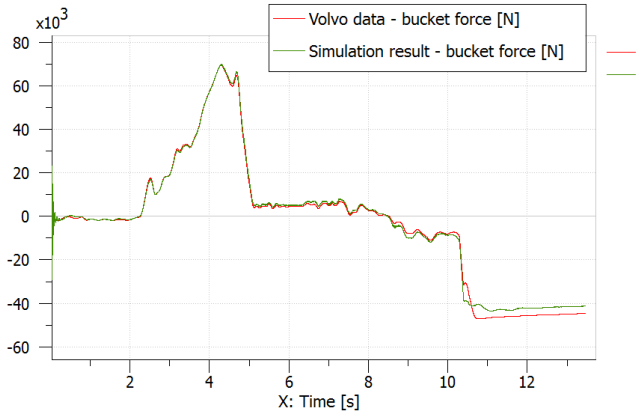
Figure 3.4. Electric excavator model in AMESim – Electro-mechanical model for the actuator.



a.



b.



c.

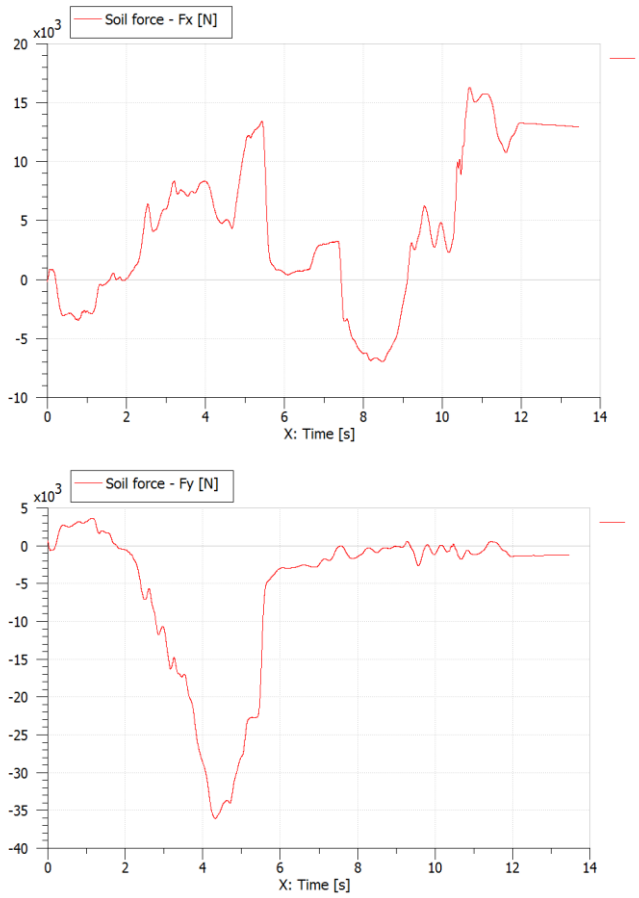
Figure 3.5. Dynamic modelling of the excavator, a. boom, b. arm, c. bucket.

Looking at Figure 3.5, it can be seen that because of the PID controllers, there are still small differences in tracking. However, there is reason to believe that if the controllers worked absolutely perfectly, the simulation results would match the force data from Volvo.

Please note that the correct masses for the electromechanical actuators are not yet available, so this dynamic model will be modified with new parameters in the coming future.

3.3 Discussion

Figures 3.2 and 3.5 have shown the successful kinematic and dynamic modelling of the excavator. This also validated the identification algorithms. Figure 3.6 shows the Fx, Fy and M of the soil force. It can be seen easily that there are four phases in cycle A1.



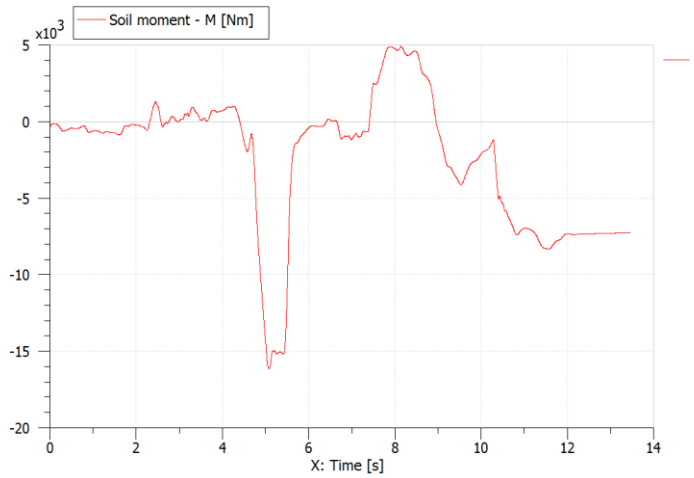


Figure 3.6. Soil Fx, Fy and M.

Looking at Figure 3.6, from 0 to 2 second, the soil force may be the result of scratching effect as the excavator moves to the digging position while holding the bucket very close to the ground. To the end of the cycle, Fx and M do not approach zero, which also may be the result of mechanical singularity in the bucket jack.

This force and moment profile will serve as the base for calculating the soil force for next steps in this project. Also, a soil force model is being studied and will be applied in the future stages.

4 Test bench control

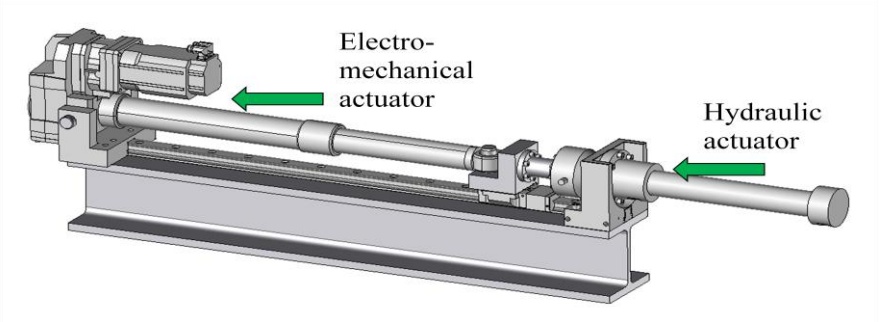


Figure 4.1. CAD model for the test bench.

In order to develop a more accurate dynamic model for the electric actuator, a test bench is currently under construction for this project (Figure 4.1). Beside this purpose, this test bench can also represent the digging activity by demanding the hydraulic actuator to reconstruct the reaction force and the electro-mechanical actuator to reconstruct the jack displacement. Basically, then, this test bench has two parts. The hydraulic part is a hydraulic actuator to represent the force, and the electric part is an electro-mechanical actuator to follow a pre-defined trajectory. The AMESim model for this test bench has been developed, and has also been controlled successfully by PID controllers. More sophisticated control methods are being developed for this test bench.

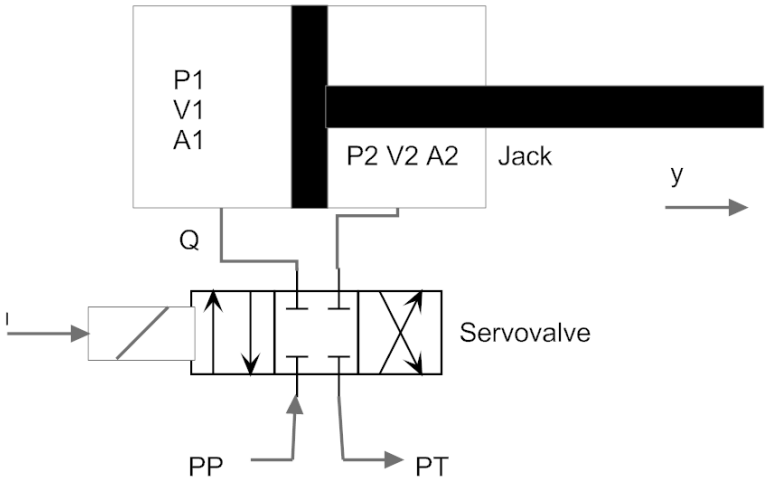


Figure 4.2. Block diagram of the hydraulic part of the test bench.

PID controllers design for the test bench: The mathematical model for the electrical part of the test bench as well as the PID design have been presented in Section 2. For the hydraulic actuator, according to [52], a simplified mathematical model can be written as

$$\begin{cases} \frac{dP_1}{dt} = -\frac{\beta}{V_1(y)}.A_1.v + \frac{\beta}{V_1(y)}\varphi_1\left(P_1, P_P, P_T, sign(I)\right).I \\ \frac{dP_2}{dt} = \frac{\beta}{V_2(y)}.A_2.v + \frac{\beta}{V_2(y)}\varphi_2\left(P_2, P_P, P_T, sign(I)\right).(-I) \end{cases} \quad (4.1)$$

Where $\varphi_1 = \frac{Q_n}{I_n}\sqrt{\frac{|P_P - P_1|}{\Delta P_n}}.\frac{sign(I)+1}{2} + \frac{Q_n}{I_n}\sqrt{\frac{|P_1 - P_T|}{\Delta P_n}}.\frac{1-sign(I)}{2}$ (4.2)

$$\varphi_2 = \frac{Q_n}{I_n}\sqrt{\frac{|P_P - P_2|}{\Delta P_n}}.\frac{1-sign(I)}{2} + \frac{Q_n}{I_n}\sqrt{\frac{|P_2 - P_T|}{\Delta P_n}}.\frac{sign(I)+1}{2} \quad (4.3)$$

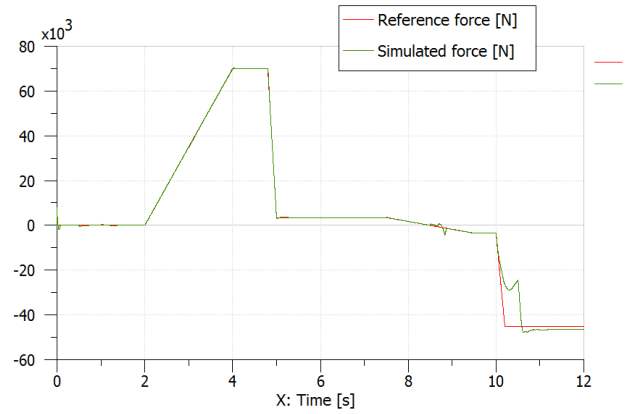
With: *y: piston displacement; v: velocity; A_1: cross-sectional area of the piston, chamber 1 side; A_2: cross-sectional area of the piston, chamber 2 side; M: mass of the piston; P_1: pressure at port 1; P_2: pressure at port 2; b: viscous friction coefficient; F_{fs}: dry friction; β: bulk modulus; V_1: volume of chamber 1; V_2: volume of chamber 2; P_P: pressure at port P; P_T: pressure at port T; I: hydraulic servovalve control current; Q: flow rate; Q_n: flow rate at maximum valve opening; I_n: valve rated current; ΔP_n: maximum pressure drop.*

This mathematical model goes with a hypothesis that the dynamic between control current I and flow rate Q is very very fast, which means, for example, if $I = I_n$ which is the rated current, Q is immediately equal to Q_n, which is the flow rate at maximum valve opening. In reality, this may or may not hold true, depending on the characteristics of the device. Therefore, when the test bench is available, this hypothesis needs to be verified, otherwise, a more accurate mathematical model needs to be derived.

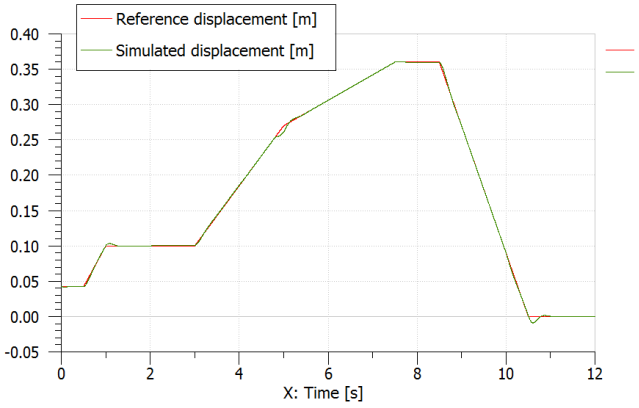
As this is a highly non-linear system, therefore, there is also no analytical method to design a PID controller for such a system. Although several linearisation approaches have been undertaken by Wang [53], Karpenko and Shapehri [54], or Pommier et al [55], linearisation is not expected to be conducted in this project as they are still quite rough so with the utilisation of AMESim, a Ziegler-Nichols approach is taken to tune the PID controller for the hydraulic part. Non-linear control strategies will be conducted in later stages of this project.

As another purpose of this test bench is to reproduce the soil force. The hydraulic part of this test bench must be able to track a pre-defined profile of force as the soil force in reality can come in many different forms. The result in Figure 4.3.a is just one example of the soil force profile that the people involved in this project assumed.

Figure 4.3 shows the very successful tracking of force and the successful tracking of course in the AMESim model of this test bench.



a. Force tracking



b. Course tracking

Figure 4.3. Tracking of force and course for the test bench – Simulation result in AMESim.

The sliding mode and backstepping controls for this test bench is under construction and will be presented in a separate report.

5 Conclusion

5.1 Summary of results

So far, the electric actuator has been modelled with significant success, as its model can work without any problem both in the global excavator model as well as the test bench model. The kinematic and dynamic modellings for the global excavator arm equipment model also work well, which means the identification modelling are accurate.

This means that the results from the excavator model can start being useful in studying all aspects of operation for the excavator, especially during the digging cycle of the arm equipment. It also means that there is a foundation available for control and energy management algorithms developed by other doctorants also working in this project. Therefore, the major contribution of this work to the project so far has been to provide a preliminary characterization and modelling of the operation of a specific electric excavator (an electric VOLVO EC27C-MODEL) from data obtained by monitoring the excavator during a normal operation period.

5.2 Recommendations for future investigation steps

The next steps will take advantage of the availability of the dynamic model as well as the AMESim model. They will be used intensively to study more about the performance of the electric excavator during its operation.

The first thing first will be the characterisation and re-modelling of the electric actuator and working on the test bench, as currently some theoretical models are being used. For the other next steps, as agreed upon between the professors and the companies, they will of course follow the Gantt chart of this project, as shown in Figure 5.1.

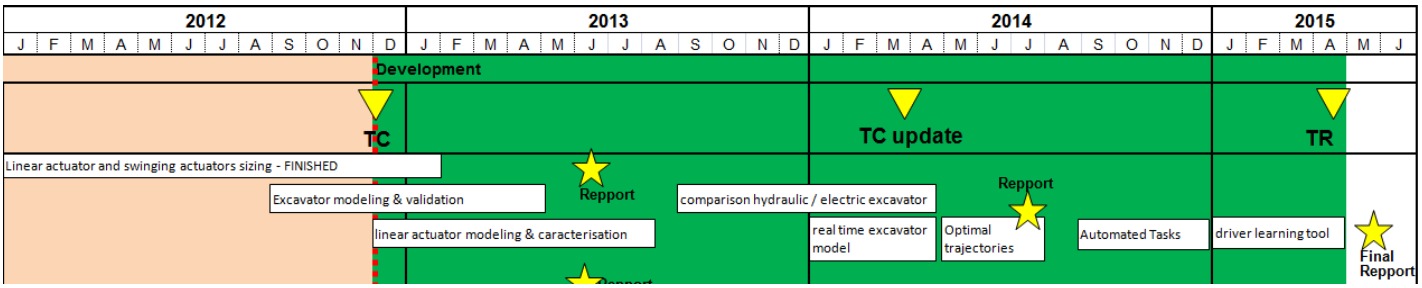


Figure 5.1. Gantt chart for the project.

There is now the ability to compare between the different architectures. One of the most important new features of the electric excavator to the hydraulic excavator is that the electromechanical actuators are way heavier than the hydraulic ones. This fact will have a lot of impacts on the operation as well as the energetical performance of the excavator. Also, the structure of the excavator body may have to be modified to cope with the new dynamic demand. This will be studied in the next step of this project, utilising the current dynamic and AMESim model of the electric excavator.

This dynamic model will also be tested with real time. It should be strongly noted that the full dynamic model (as in Appendix C) may have to be reduced by applying some assumptions as the model can be too long for real time calculation. These various assumptions, if there must be some, will be evaluated in the next step of the project.

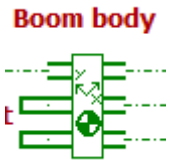
Regarding the test bench, beside the PID results, there is some good news regarding the model-based non-linear control of the test bench. The sliding mode control has got some interesting results, but as the control design is quite complicated, it needs to be checked carefully with the control experts available in the laboratory to get approvement and validation. Dr. M. Smaoui, who is an expert in sliding mode control, will take an active role in this step.

The optimisation of digging trajectory will be conducted after the thorough study of the control design and energy management of the excavator. This will apply knowledges about optimisation, as well as soil mechanics and energy management. The author of this report then will have to spend time to study and explore the new knowledge, models, as well as opportunities for the finding of optimal digging trajectories.

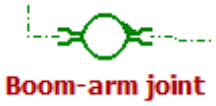
In a nutshell, the next steps for this project will be both challenging and interesting. With the progress so far, it should be trusted that the work done can form a good foundation on which the works that will be done in the next steps can lie with stability.

Appendix A

The assembly of any mechanical structure in AMESim is very intuitive. Basically they contain only rigid bodies and joints. For the arm equipment of the excavator, all joints are revolutional.

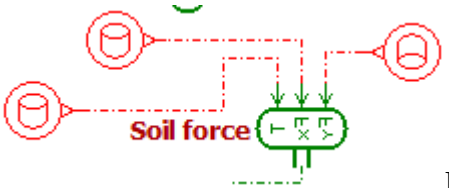


There are many kinds of bodies, ranging from 2 to 6 ports. Usually complex bodies with many joints should be modelled with many ports. Number of ports may be bigger than number of joints for visualisation purpose.



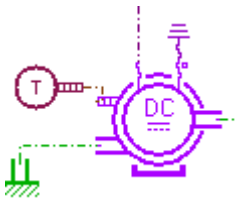
Revolutional joints to connect the bodies.

General jacks (which is just mathematically a jack) to form the structure for **any** kind of actuator, whether hydraulic, electric or pneumatic…

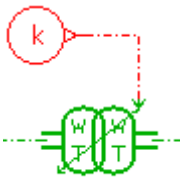


Force and moment generator, serves the purpose to convert signal from .data files into force or moment.

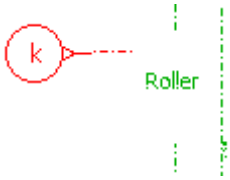
For the excavator body, the same elements and method are used.



The DC motor is the permanent magnet DC motor. The control of synchronous motor is quite complex to be integrated in AMESim alone, so cosimulation with Simulink is being conducted. For the architecture starting point, a DC motor is used.



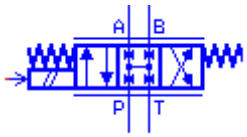
The reductor is a simple rotary-rotary transformer.



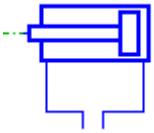
The roller screw is not available in AMESim library, therefore, it has to be created using AMESet, another software developped by LMS. There are two models of roller screw to choose from, one after the theory by Velinsky et al [31] and the other after the work by Karam [48].

PID controller is chosen for modelling for the performance quality of this simple kind of controller.


Pressure source and sink



Hydraulic servovalve


Hydraulic piston

Appendix B

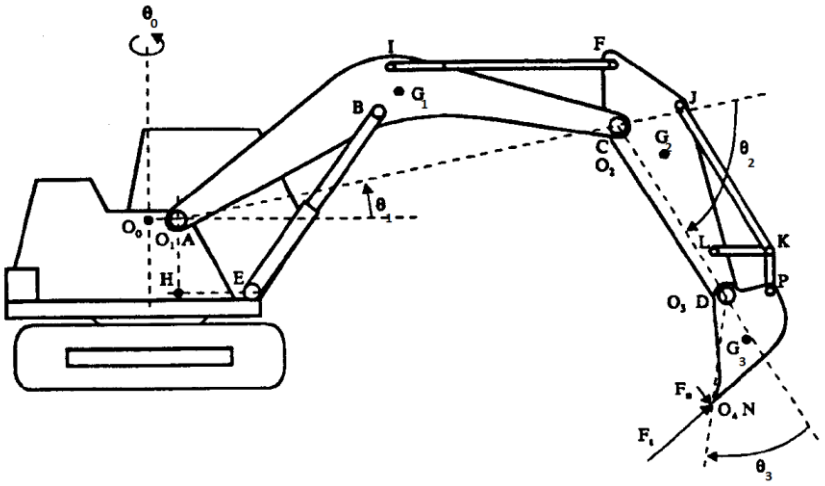


Figure B.1. Notation for Appendix B.

1. **Forward kinematics – from angles to coordinates and displacements**

   For the given $[\theta_1 \ \ \theta_2 \ \ \theta_3]$, find the coordinates of an arbitrary point P.

To determine the positions of the points on the excavator in the base Cartesian coordinate frame, the relations between the fixed coordinate system and other coordinate systems is necessary. Therefore, the transformation matrix relating two adjacent coordinate frames was studied by Koivo [39] and Denavit-Hartenberg [2] as follows:

$$A_{i-1}^{i} = \begin{bmatrix} \cos\theta_i & -\cos\alpha_i . \sin\theta_i & \sin\alpha_i . \sin\theta_i & a_i \cos\theta_i \\ \sin\theta_i & \cos\alpha_i . \cos\theta_i & -\sin\alpha_i . \cos\theta_i & a_i \sin\theta_i \\ 0 & \sin\alpha_i & \cos\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Where α_i is the twist angle of link i, a_i is the length of link i, and d_i is the offset distance in link i. i = 1, 2, 3.

Then, with an arbitrary point P in link i, absolute coordinates of P will be:

$$P^0 = A_0^1 . A_1^2 \dots A_{i-1}^i . P^i$$

With this formula, the absolute coordinates of all points from A to N in Figure B.1 can be calculated. This is very important to calculate the dynamic of the arm equipment.

As absolute coordinates for points E, B; I, F; J, K are known, it is therefore very easy to calculate distances EB, IF, JK and jack displacements using Pythagorean equation. All angles in Figure B.1 will be calculated using the cosine theorem. In the dynamic modelling, it is assumed that as long as θ_1 , θ_2 , θ_3 are known, all other coordinates, lengths, angles on the arm equipment are known.

2. Inverse kinematics – from displacements to angles

The inverse kinematics from displacements to angles will be solved by conventional geometrics. For example, to find θ_2, we look at joint O_2, it can be seen:

$2\pi = \angle ACI + \angle ICF + \angle FCD + \angle DCA$ while

∠ACI can be calculated as AC, CI, IA are known using cosine theorem. ∠ICF can be calculated as IC, CF, FI are known using cosine theorem. ∠FCD can be calculated as FC, CD, DF are known using cosine theorem. Note that

∠DCA = $\pi + \theta_2$ then θ_2 can be found easily.

Appendix C

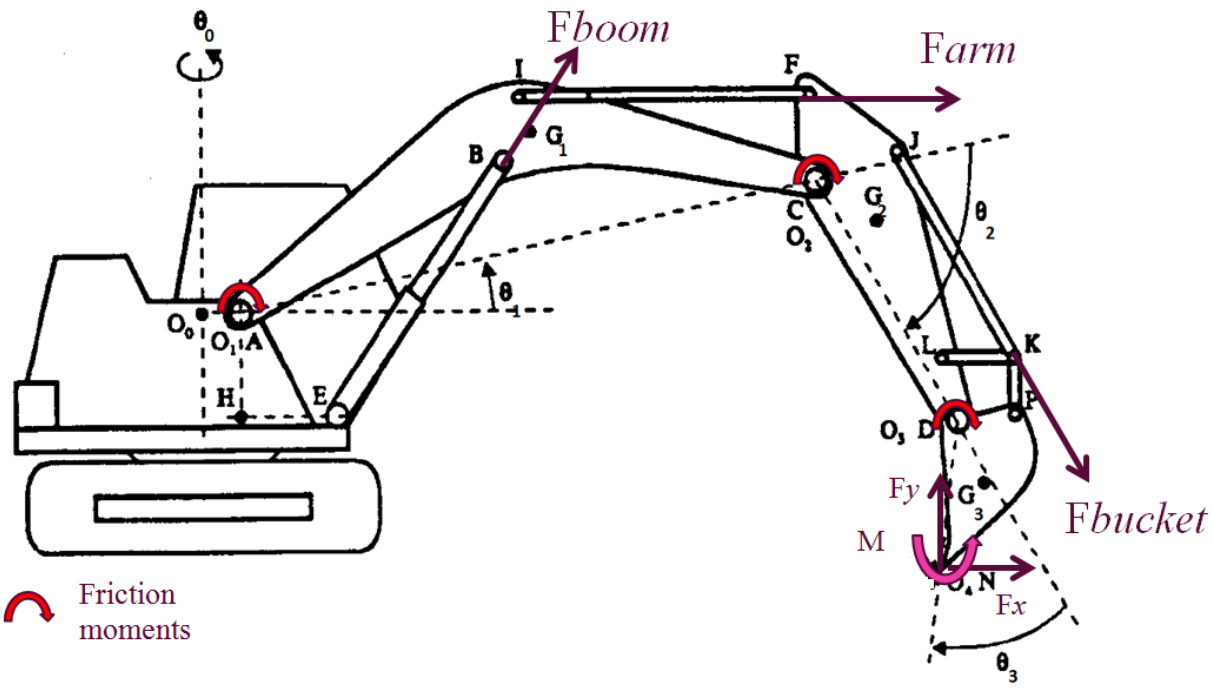


Figure C.1. Notations for Appendix C.

Before dynamic modelling, it should be stated again that with known θ_1 , θ_2 , θ_3, all coordinates, lengths, angles of and created from points in Figure C.1 are known from the kinematic modelling. The dynamic modelling, therefore, focuses only on finding motion equations.

The Lagrange equation of the second kind takes the form:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta_n}}\right) - \frac{\partial T}{\partial \theta_n} = -\frac{\partial V}{\partial \theta_n} + Q_n \qquad \text{(n = 1, 2, 3)}$$

The kinetic energy T of the arm equipment is calculated as

$$T = \frac{1}{2}\sum_{n=1}^{3} m_n.(\dot{x_{Gn}}^2 + \dot{y_{Gn}}^2) + \frac{1}{2}\sum_{n=1}^{3} I_n.\dot{\theta_{1-n}}^2$$

Where

m_n: mass of link n

x_{Gn} and y_{Gn}: absolute coordinates of mass centres G1, G2, G3 for each link

I_n: moment of inertia around mass centre Gn in rotating plane

θ_{1-n}: absolute rotating angles for link n. Note that $\theta_{12} = \theta_1 + \theta_2$ and $\theta_{123} = \theta_1 + \theta_2 + \theta_3$

Q_n: total external generalized forces acting on link n

The potential energy V of the arm equipment is calculated as

V = $m_1(O1G1sin(\theta_1 + \alpha_1)) + m_2((a12.\sin(\theta_1) + O2G2.\sin(\theta_{12} + \alpha_2)) + m_3((a12.\sin(\theta_1) + a23.\sin(\theta_{12}) + O3G3.\sin(\theta_{123} + \alpha_3))$

= $m_1 y_{G1} + m_2 y_{G2} + m_3 y_{G3}$

Then, the kinetic energy is the sum of 6 elements, 3 translational elements and 3 rotational elements for each of the 3 degrees of freedom.

$T = \sum_{k=1}^{6} T_k$ with

$$T_1 = \frac{1}{2} m_1((O1G1.\dot{\theta}_1.sin(\theta_1 + \alpha_1))^2 + (O1G1.\dot{\theta}_1.cos(\theta_1 + \alpha_1))^2)$$

$$T_2 = \frac{1}{2} I_1(\dot{\theta_1})^2$$

$$T_3 = \frac{1}{2} m_2((a12.\sin(\theta_1)\,\dot{\theta}_1 + O2G2.\sin(\theta_{12} + \alpha_2).\dot{\theta}_{12})^2 + (a12.\cos(\theta_1)\,\dot{\theta}_1 + O2G2.\cos(\theta_{12} + \alpha_2).\dot{\theta}_{12})^2)$$

$$T_4 = \frac{1}{2} I_2(\dot{\theta_{12}})^2$$

$$T_5 = \frac{1}{2} m_3((a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3).\dot{\theta_{123}})^2 + (a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3).\dot{\theta_{123}})^2)$$

$$T_6 = \frac{1}{2} I_3(\dot{\theta_{123}})^2$$

Then comes 18 elements of $\frac{\partial T_k}{\partial \dot{\theta_n}}$ (6 elements of T x 3 degrees of freedom)

Element 1-1: $\frac{\partial T_1}{\partial \dot{\theta_1}} = m_1(\left((O1G1.sin(\theta_1 + \alpha_1))^2 \dot{\theta_1}\right) + \left((O1G1.cos(\theta_1 + \alpha_1))^2 \dot{\theta_1}\right) = m_1 O1G1^2 \dot{\theta_1}$

Element 2-1: $\frac{\partial T_2}{\partial \dot{\theta_1}} = I_1 \dot{\theta_1}$

Element 3-1: $\frac{\partial T_3}{\partial \dot{\theta_1}} = m_2((a12.\sin(\theta_1)\,\dot{\theta_1} + O2G2.\sin(\theta_{12} + \alpha_2).\dot{\theta_{12}})(a12.\sin(\theta_1) + O2G2.\sin(\theta_{12} + \alpha_2)) + (a12.\cos(\theta_1)\,\dot{\theta_1} + O2G2.\cos(\theta_{12} + \alpha_2).\dot{\theta_{12}})(a12.\cos(\theta_1) + O2G2.\cos(\theta_{12} + \alpha_2)))$

Element 4-1: $\frac{\partial T_4}{\partial \dot{\theta_1}} = I_2 \dot{\theta_{12}}$

Element 5-1: $\frac{\partial T_5}{\partial \dot{\theta_1}} = m_3(\left(a12.\sin(\theta_1)\,\dot{\theta_1} + a23.\sin(\theta_{12})\,.\,\dot{\theta_{12}} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}\right).\left(a12.\sin(\theta_1) + a23.\sin(\theta_{12}) + O3G3.\sin(\theta_{123} + \alpha_3)\right) + \left(a12.\cos(\theta_1)\,\dot{\theta_1} + a23.\cos(\theta_{12})\,.\,\dot{\theta_{12}} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}\right).((a12.\cos(\theta_1) + a23.\cos(\theta_{12}) + O3G3.\cos(\theta_{123} + \alpha_3))))$

Element 6-1: $\frac{\partial T_6}{\partial \dot{\theta_1}} = I_3 \dot{\theta_{123}}$

Element 1-2: $\frac{\partial T_1}{\partial \dot{\theta_2}} = 0$

Element 2-2: $\frac{\partial T_2}{\partial \dot{\theta_2}} = 0$

Element 3-2: $\frac{\partial T_3}{\partial \dot{\theta_2}} = m_2((a12.\sin(\theta_1)\,\dot{\theta_1} + O2G2.\sin(\theta_{12} + \alpha_2).\dot{\theta_{12}})O2G2.\sin(\theta_{12} + \alpha_2) + (a12.\cos(\theta_1)\,\dot{\theta_1} + O2G2.\cos(\theta_{12} + \alpha_2).\dot{\theta_{12}})O2G2.\cos(\theta_{12} + \alpha_2))$

Element 4-2: $\frac{\partial T_4}{\partial \dot{\theta_2}} = I_2 \dot{\theta_{12}}$

Element 5-2: $\frac{\partial T_5}{\partial \dot{\theta}_2} = m_3((a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}).(a23.\sin(\theta_{12}) + O3G3.\sin(\theta_{123} + \alpha_3)) + (a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}).((a23.\cos(\theta_{12}) + O3G3.\cos(\theta_{123} + \alpha_3))))$

Element 6-2: $\frac{\partial T_6}{\partial \dot{\theta}_2} = I_3\dot{\theta_{123}}$

Element 1-3: $\frac{\partial T_1}{\partial \dot{\theta}_3} = 0$

Element 2-3: $\frac{\partial T_2}{\partial \dot{\theta}_3} = 0$

Element 3-3: $\frac{\partial T_3}{\partial \dot{\theta}_3} = 0$

Element 4-3: $\frac{\partial T_4}{\partial \dot{\theta}_3} = 0$

Element 5-3: $\frac{\partial T_5}{\partial \dot{\theta}_3} = m_3((a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}).(O3G3.\sin(\theta_{123} + \alpha_3)) + (a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta_{123}}).(O3G3.\cos(\theta_{123} + \alpha_3))))$

Element 6-3: $\frac{\partial T_6}{\partial \dot{\theta}_3} = I_3\dot{\theta_{123}}$

After these 18 elements, then comes another 18 elements of $\frac{\partial T}{\partial \theta_n}$ (6 elements of T x 3 degrees of freedom)

Element 1-1a: $\frac{\partial T_1}{\partial \theta_1} = m_1\left(\left(O1G1.\dot{\theta}_1.sin(\theta_1 + \alpha_1)\right)\left(O1G1.\dot{\theta}_1.cos(\theta_1 + \alpha_1)\right)\right) + \left(O1G1.\dot{\theta}_1.cos(\theta_1 + \alpha_1)\right).(O1G1.\dot{\theta}_1.\left(-sin(\theta_1 + \alpha_1\right))) = 0$

Element 1-2a: $\frac{\partial T_1}{\partial \theta_2} = 0$

Element 1-3a: $\frac{\partial T_1}{\partial \theta_3} = 0$

Element 2-1a: $\frac{\partial T_2}{\partial \theta_1} = 0$

Element 2-2a: $\frac{\partial T_2}{\partial \theta_2} = 0$

Element 2-3a: $\frac{\partial T_2}{\partial \theta_3} = 0$

Element 3-1a: $\frac{\partial T_3}{\partial \theta_1} = m_2((a12.\sin(\theta_1)\,\dot{\theta}_1 + O2G2.\sin(\theta_{12} + \alpha_2).\dot{\theta}_{12})(a12.\cos(\theta_1)\,\dot{\theta}_1 + O2G2.\cos(\theta_{12} + \alpha_2).\dot{\theta}_{12}) + (a12.\cos(\theta_1)\,\dot{\theta}_1 + O2G2.\cos(\theta_{12} + \alpha_2).\dot{\theta}_{12})(a12.(-\sin(\theta_1))\dot{\theta}_1 + O2G2.(-\sin(\theta_{12} + \alpha_2)).\dot{\theta}_{12})) = 0$

Element 3-2a:

$$\frac{\partial T_3}{\partial \theta_2} = m_2\Big((a12.\sin(\theta_1)\,\dot{\theta}_1 + O2G2.\sin(\theta_{12} + \alpha_2)\,.\,\dot{\theta}_{12})(O2G2.\cos(\theta_{12} + \alpha_2)\,.\,\dot{\theta}_{12}) + (a12.\cos(\theta_1)\,\dot{\theta}_1 + O2G2.\cos(\theta_{12} + \alpha_2)\,.\,\dot{\theta}_{12})(O2G2.(-\sin(\theta_{12} + \alpha_2)).\dot{\theta}_{12})\Big) = -m_2 a12\; O2G2\dot{\theta}_1.\dot{\theta}_{12}\sin(\theta_2 + \alpha_2)$$

Element 3-3a: $\frac{\partial T_3}{\partial \theta_3} = 0$

Element 4-1a: $\frac{\partial T_4}{\partial \theta_1} = 0$

Element 4-2a: $\frac{\partial T_4}{\partial \theta_2} = 0$

Element 4-3a: $\frac{\partial T_4}{\partial \theta_3} = 0$

Element 5-1a: $\frac{\partial T_5}{\partial \theta_1} = m_3\Big(\big(a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\big(a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big) + \big(a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\big(a12.(-\sin(\theta_1))\dot{\theta}_1 + a23.(-\sin(\theta_{12}))\big).\,\dot{\theta}_{12} + O3G3.(-\sin(\theta_{123} + \alpha_3)).\,\dot{\theta}_{123}\big) = 0$

Element 5-2a: $\frac{\partial T_5}{\partial \theta_2} = m_3\Big(\big(a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\big(a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big) + \big(a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\big((a23.(-\sin(\theta_{12}))\big).\,\dot{\theta}_{12} + O3G3.(-\sin(\theta_{123} + \alpha_3)).\,\dot{\theta}_{123}\big) =$

$m_3\big(-a12a23\dot{\theta}_1\dot{\theta}_{12}\sin(\theta_2) - a12O3G3\dot{\theta}_1.\,\dot{\theta}_{123}\sin(\theta_{23} + \alpha_3) - 0 - a23.O3G3.\dot{\theta}_{12}.\,\dot{\theta}_{123}\sin(\theta_3 + \alpha_3) + a23O3G3.\,\dot{\theta}_{123}.\,\dot{\theta}_{12}\sin(\theta_3 + \alpha_3) - 0\big) =$

$m_3(-a12a23\dot{\theta}_1\dot{\theta}_{12}\sin(\theta_2) - a12O3G3\dot{\theta}_1.\,\dot{\theta}_{123}\sin(\theta_{23} + \alpha_3))$

Element 5-3a: $\frac{\partial T_5}{\partial \theta_3} = m_3\Big(\big(a12.\sin(\theta_1)\,\dot{\theta}_1 + a23.\sin(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\sin(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\big(O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big) + \big(a12.\cos(\theta_1)\,\dot{\theta}_1 + a23.\cos(\theta_{12})\,.\,\dot{\theta}_{12} + O3G3.\cos(\theta_{123} + \alpha_3)\,.\,\dot{\theta}_{123}\big)\Big).\,O3G3.(-\sin(\theta_{123} + \alpha_3)).\,\dot{\theta}_{123}\big) =$

$m_3(-a12.O3G3.\,\dot{\theta}_1.\,\dot{\theta}_{123}\sin(\theta_{23} + \alpha_3) - a23.O3G3.\,\dot{\theta}_{12}.\,\dot{\theta}_{123}\sin(\theta_3 + \alpha_3) + 0)$

Element 6-1a: $\frac{\partial T_6}{\partial \theta_1} = 0$

Element 6-2a: $\frac{\partial T_6}{\partial \theta_2} = 0$

Element 6-3a: $\frac{\partial T_6}{\partial \theta_3} = 0$

Putting these 36 elements into the expression $\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\theta_n}}\right) - \frac{\partial T}{\partial \theta_n}$, then rearrange we have

$$D(\theta).\ddot{\theta} + C(\theta,\dot{\theta}).\dot{\theta}$$

The generalized forces are calculated according to the principle of virtual work, which says for our excavator arm equipment:

$Q_i = \sum_{j=1}^{m}(F_j \frac{\partial v_j}{\partial \dot{\theta}_i} + M_j \frac{\partial \vec{\omega}_j}{\partial \dot{\theta}_i})$, i =1,2,3

With F_j: external force, v_j: velocity, M_j: external moment, $\vec{\omega}_j$: rotational velocity.

As the bodies are considered rigid, and the joints are ideal joints, we then have:

For the $\theta_1$coordinate, there are five external forces, namely boom force, viscous friction moment, two soil forces and one soil moment.

For the $\theta_2$coordinate, there are five external forces, namely arm force, viscous friction moment, two soil forces and one soil moment.

For the $\theta_3$coordinate, there are five external forces, namely bucket force, viscous friction moment, two soil forces and one soil moment.

The external forces Q_i are then calculated using conventional formulae.

Then, rearrange into the form:

$D(\theta).\ddot{\theta} + C(\theta,\dot{\theta}).\dot{\theta} + G(\theta) + B(\dot{\theta}) = L(\theta).F - F_{soil}$

with:

θ: angle matrix of boom, arm, and bucket angles

$D(\theta)$: inertial matrix

$C(\theta,\dot{\theta})$: Coriolis matrix

$G(\theta)$: gravity matrix

$B(\dot{\theta})$: viscous friction matrix

$L(\theta)$: geometrical matrix

F: force matrix of boom, arm, and bucket forces

F_{soil}: soil force matrix, which is actually moments at boom, arm, and bucket by soil force profile

We have in more detail,

G = $[G_1 \ G_2 \ G_3]^T$

B = $[B_1 \ B_2 \ B_3]^T$, the three elements are constants and are currently set as 150, 100, and 50 Nms/rad, these values will be checked with Volvo to see whether they are representative or not, however, the effect of friction is extremely small in comparison with the effects by gravity and reaction force.

F = $[F_{boom} \ F_{arm} \ F_{bucket}]^T$, with notice that the pushing force will be positive, while the pulling force will be negative.

$$F_{soil} = \begin{bmatrix} O_{1y} - N_y & N_x - O_{1x} & 1 \\ O_{2y} - N_y & N_x - O_{2x} & 1 \\ O_{3y} - N_y & N_x - O_{3x} & 1 \end{bmatrix} . \begin{bmatrix} F_x \\ F_y \\ T \end{bmatrix}$$

C = $\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$

D = $\begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{bmatrix}$

L = $\begin{bmatrix} L_{11} & L_{12} & L_{13} \\ 0 & L_{22} & L_{23} \\ 0 & 0 & L_{33} \end{bmatrix}$

In detail,

$G_1 = -m_3 g$ (a_{12} cos(θ_1) + a_{23} cos(θ_{12}) + O3G3cos(θ_{123} + α_3)) - $m_2 g$ (a_{12} cos(θ_1) + O2G2cos(θ_{12} + α_2)) - $m_1 g$O1G1cos(θ_1 + α_1)

G_2 = -$m_3 g$ (a_{23} cos(θ_{12}) + O3G3cos(θ_{123} + α_3)) - $m_2 g$O2G2cos(θ_{12} + α_2)

G_3 =-$m_3 g$O3G3cos(θ_{123} + α_3)

C_{11}=

$-m_2 a12\ O2G2.\dot{\theta}_{12}\sin(\theta_2+\alpha_2) - m_3 a12 a23\dot{\theta}_{12}\sin(\theta_2) - m_3.a12.O3G3.\dot{\theta}_{123}\sin(\theta_{23}+\alpha_3)$

C_{12} =

$-m_2 a12\ O2G2.\dot{\theta}_{12}\sin(\theta_2+\alpha_2) - m_3 a12 a23\dot{\theta}_{12}\sin(\theta_2) - m_3.a12.O3G3.\dot{\theta}_{123}\sin(\theta_{23}+\alpha_3)$

$C_{13}= -m_3.a12.O3G3.\dot{\theta}_{123}\sin(\theta_{23}+\alpha_3)$

$C_{21} = m_3 a12 a23\dot{\theta}_1\sin(\theta_2) + m_2 a12\ O2G2.\dot{\theta}_1\sin(\theta_2+\alpha_2) - m_3 a23.O3G3.\dot{\theta}_{123}\sin(\theta_3+\alpha_3)$

$C_{22} = -m_3 a23.O3G3.\dot{\theta}_{123}\sin(\theta_3+\alpha_3)$

$C_{23} = -m_3 a23.O3G3.\dot{\theta}_{123}\sin(\theta_3+\alpha_3)$

C_{31}=

$m_3 a12.O3G3.\dot{\theta}_1\sin(\theta_{23}+\alpha_3) + m_3 a23.O3G3.\dot{\theta}_1\sin(\theta_3+\alpha_3) + m_3 a23.O3G3.\dot{\theta}_2\sin(\theta_3+\alpha_3)$

$C_{32} = m_3 a23.O3G3.\dot{\theta}_2\sin(\theta_3+\alpha_3)$

$C_{33} = 0$

$D_{11} = I_3 + m_3 O3G3^2 + I_2 + m_2 O2G2^2 + m_3(a_{23}^2 + 2a_{23}O3G3\cos(\theta_3+\alpha_3)) + I_1 + m_1 O1G1^2 + m_2(a_{12}^2 + 2a_{12}O2G2\cos(\theta_2+\alpha_2)) + m_3(a_{12}^2 + 2a12.a_{23}\cos(\theta_2) + 2a_{12}O3G3\cos(\theta_{23}+\alpha_3))$

$D_{21} = D_{12} = I_3 + m_3 O3G3^2 + m_3 a_{23}O3G3\cos(\theta_3+\alpha_3) + m_3 a_{12}O3G3\cos(\theta_{23}+\alpha_3) + I_2 + m_2(O2G2^2 + 2a_{12}O2G2\cos(\theta_2+\alpha_2)) + m_3(a_{23}^2 + a12.a_{23}\cos(\theta_2) + a_{23}O3G3\cos(\theta_3+\alpha_3))$

$D_{22} = I_3 + m_3 O3G3^2 + I_2 + m_2 O2G2^2 + m_3(a_{23}^2 + 2a_{23}O3G3\cos(\theta_3+\alpha_3))$

$D_{31} = D_{13} = I_3 + m_3 O3G3^2 + m_3 a_{23}O3G3\cos(\theta_3+\alpha_3) + m_3 a_{12}O3G3\cos(\theta_{23}+\alpha_3)$

$D_{32} = D_{23} = I_3 + m_3 O3G3^2 + m_3 a_{23}O3G3\cos(\theta_3+\alpha_3)$

$D_{33} = I_3 + m_3 O3G3^2$

Elements of the geometrical L is calculated according to the virtual work principle. As proved by Zatkorsky in [56], the external generalized force for each joint is equal to the external torque. Note that for joint 3, one has to use the velocity relationship for four-bar linkage, as such linkage is formed by the arm, tri-power, stunga and bucket. After checking with [57], therefore we have:

L_{11} = O1Bsin(∠ (O1BE)); L_{12} = 0; L_{13} = 0; L_{22} = O2Fsin(∠ (IFO2)); L_{23} = 0;

L_{33} = $\mathrm{LJsin}\big(\angle(\mathrm{LJK})\big)\frac{O3P\sin(\alpha-\varphi)}{LKsin(\alpha-\sigma)}$;

With: α, φ, σ respectively angles between KP, O3P and LK with positive x_2.

Where:

a12: length from O1 to O2

a23: length from O2 to O3

α_i: angle between OiGi and positive x_i.

$\theta_{23} = \theta_2 + \theta_3$

$\theta_{12} = \theta_1 + \theta_2$

$\theta_{123} = \theta_1 + \theta_2 + \theta_3$

Dynamic model for the excavator body:

The excavator body's dynamic model is made with these assumptions: 1. The excavator is operated absolutely planar, which means, gravity plays exactly no role in the dynamic model; 2. There is no slip in the movement, the translational velocity of the chain track is linked to the rotational velocity of its wheel by a constant.

These assumptions are reasonable, due to the fact that the mini-excavator often moves around on levelled surface or a surface with small slope gradient. Also, whenever the chain track slips, the excavator is no longer considered safe to operate and it is often required to stop operation or change the working condition (put mat, wait for the soil to dry…) to make slippage disappear. Then, this assumption should be hold true during the whole process of modelling.

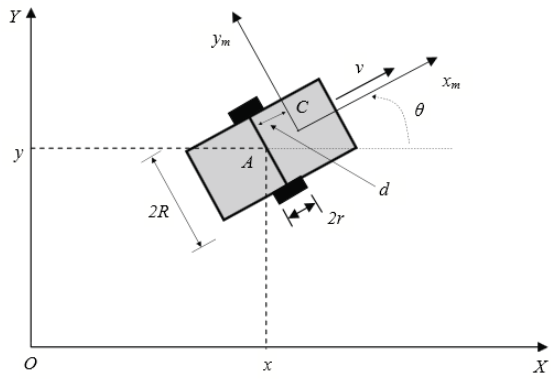


Figure C.2. Top view of the excavator body (from [58] and [59]).

At Figure C.2, A is the geometrical centre of the excavator body, which means, when the rotational velocities of the left and right wheels are equal in value but opposite in direction, the body rotates around A. C is the centre of mass of the excavator body.

By simple geometrical reasoning, we have

$$\dot{x_A} = \frac{r}{2}\left(\dot{\theta_L} + \dot{\theta_R}\right).\cos(\theta)\,;\ \dot{y_A} = \frac{r}{2}\left(\dot{\theta_L} + \dot{\theta_R}\right).\sin(\theta)\,;\ \dot{\theta} = \frac{r}{2R}\left(\dot{\theta_R} - \dot{\theta_L}\right)$$

Since,

$\dot{x_C} = \dot{x_A} - d\dot{\theta}\sin(\theta)$ and $\dot{y_C} = \dot{y_A} + d\dot{\theta}\cos(\theta)$, then

$\dot{x_C} = \frac{r}{2}\left(\dot{\theta_L} + \dot{\theta_R}\right).\cos(\theta) - d\dot{\theta}\sin(\theta)$ and $\dot{y_C} = \frac{r}{2}\left(\dot{\theta_L} + \dot{\theta_R}\right).\sin(\theta) + d\dot{\theta}\cos(\theta)$

The expression for kinematic energy of the excavator body is:

$$T = \frac{1}{2}M\left(\dot{x_C}^2 + \dot{y_C}^2\right) + \frac{1}{2}I_A(\dot{\theta})^2 + \frac{1}{2}I_0(\dot{\theta_R})^2 + \frac{1}{2}I_0(\dot{\theta_L})^2$$

Where M is the mass of the excavator body, I_A is the moment of inertia of the excavator body around point A, I_0 is the moment of inertia of the chain wheel complex, $\dot{\theta_R}$ and $\dot{\theta_L}$ are respectively rotational velocities of right and left chain wheels.

Then:

$$T = \left(\frac{Mr^2}{8} + \frac{(I_A + Md^2)r^2}{8R^2} + \frac{1}{2}I_0\right)(\dot{\theta_R})^2 + \left(\frac{Mr^2}{8} + \frac{(I_A + Md^2)r^2}{8R^2} + \frac{1}{2}I_0\right)(\dot{\theta_L})^2 + \left(\frac{Mr^2}{4} - \frac{(I_A + Md^2)r^2}{4R^2}\right)\dot{\theta_R}\dot{\theta_L}$$

Applying the Lagrange equation of the second kind with θ_R and θ_L

$A\ddot{\theta_R} + B\ddot{\theta_L} = \tau_R - \mu\dot{\theta_R}$

$B\ddot{\theta_R} + A\ddot{\theta_L} = \tau_L - \mu\dot{\theta_L}$

With A = $2\left(\frac{Mr^2}{8} + \frac{(I_A+Md^2)r^2}{8R^2} + \frac{1}{2}I_0\right)$ and B = $\frac{Mr^2}{4} - \frac{(I_A+Md^2)r^2}{4R^2}$; τ_R and τ_L: right and left actuation torques, μ: coefficient of reacted friction seen from the wheel actuation.

For the blade, the motion equation is similar to the boom, however, as the blade is mainly used for levelling, which means that the blade then acts as a part of the excavator body, and the dynamic equation in this case is included above. For other cases, when the blade data is available, it will be processed.

The dynamic model for the remaining two rotation and orientation degrees of freedom is very simple, they both have the same form in horizontal state of the excavator:

For orientation: $T_{ori} - T_{fri} = J_{ori}.\dot{\omega}_{ori}$ with T_{ori}: actuated torque, T_{fri}: friction torque, J_{ori}: moment of inertia of the orientation body to the orientation centre, $\dot{\omega}_{ori}$: orientation acceleration

For rotation: $T_{rot} - T'_{fri} = J_{rot}.\dot{\omega}_{rot}$ with T_{rot}: actuated torque, T'_{fri}: friction torque, J_{rot}: moment of inertia of the rotation body to the rotation centre, $\dot{\omega}_{rot}$: rotation acceleration

Bibliography

[1] Dombre and Khalil, Modeling Identification & control of robots, Kogan Page Science, 2002.

[2] J. Denavit and R. Hartenberg, "A kinematic notation for lower-pair mechanisms based on matrices," *Journal of Applied Mechanics,* vol. 22, pp. 215-221, 1955.

[3] A. Koivo, M. Thoma, E. Kocaoglan and J. Andrade-Cetto, "Modeling and control of excavator dynamics during digging operation," *Journal of Aerospace Engineering,* 1996.

[4] R. Kelly, "A tuning procedure for stable PID control of robot manipulators," *Robotica,* vol. 13, 1995.

[5] T. Morita and Y. Sakawa, "Modelling and control of a power shovel," *J-Soc-Instrum Control Engineers,* vol. 22, no. 1, pp. 69-75, 1986.

[6] B. Song and A. Koivo, "Neural adaptive control of excavators," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, 1995.

[7] S. D. Lucie, "The autonomous robot excavator," *Industrial Robot,* vol. 19, no. 1, pp. 14-8, 1992.

[8] K. Saleh, O. Mohammed and M. Badr, "Field oriented vector control of synchronous motors with additional field winding," *IEEE Transactions on Energy Conversion,* vol. 19, pp. 95-101, 2004.

[9] Y. Kuroe, K. Okamura, H. Nishidai and T. Maruhashi, "Optimal speed control of synchronous motors based on feedback linearization," in *International conference on power electronics and variable-speed drives*, London, UK, 1998.

[10] Z. Yang, M. Wang, C. Liu and Y. Hou, "Variable structure control with sliding mode for self-controlled synchronous motor drive speed regulation," *IEEE International Symposium on Industrial Electronics (ISIE),* vol. 2, pp. 620-624, 1992.

[11] C. Elmas and O. Ustun, "A hybrid controller for the speed control of a permanent magnet synchronous motor drive," *Control Engineering Practice,* vol. 16, no. 3, pp. 260-270, 2008.

[12] J. Ziegler and N. B. Nichols, "Optimum settings for automatic controllers," *Transactions of the ASME,* vol. 64, pp. 759-768, 1942.

[13] D. Rivera, M. Morari and S. Skogestad, "Internal model control. 4. PID controller design," *Ind. Eng. Chem. Res.,* vol. 25, no. 1, pp. 252-265, 1986.

[14] C. Smith and A. Corripio, Principles and Practice of Automatic Process Control, John Wiley & Sons,

1985.

[15] B. Tyreus and W. Luyben, "Tuning PI controllers for integrator/dead time processes," *Ind. Eng. Chem. Res.,* pp. 2628-2631, 1992.

[16] I. Chien and P. Fruehauf, "Consider IMC tuning to improve controller performance," *Chemical Engineering Progress,* pp. 33-41, 1990.

[17] J. E. Slotine and W. P. Li, Applied Non-linear Control, New Jersey: Prentice Hall, 1991.

[18] S. Emelyanov, Variable Structure Control Systems, Moscow: Nauka, 1967.

[19] H. Sira-Ramirez, "Dynamical Sliding Modes Control Strategies in the Regulation of Non-linear Chemical Processes," *International Journal of Control,* 1992.

[20] G. Bartolini, A. Pisano and E. Usai, "A survey applications of second-order sliding mode control to mechanical systems," *International Journal of Control,* vol. 76, no. 9, pp. 875-892, 2003.

[21] J. J. E. Slotine and D. Li, "On Sliding Control for Multi-Input Multi-Output Nonlinear Systems," in *In American Control Conference*, 1987.

[22] F. Fisher and D. Zhou, "MIMO Sliding Mode Control: A Lyapunov Approach," in *In American Control Conference*, 1991.

[23] M. Kristic, I. Kanellakopoulos and P. Kokotovic, Nonlinear and adaptive control design, New York: John Wiley & Sons, 1995.

[24] H. K. Khalil, Nonlinear systems, New Jersey: Prentice-Hall, 2001.

[25] A. Liu and G. Alleyne, "Systematic control of a class of nonlinear systems with application to electrohydraulic cylinder pressure control," *Control Systems Technology, IEEE Transactions on,* vol. 8, no. 4, pp. 623-634, 2000.

[26] M. Smaoui, X. Brun and D. Thomasset, "A study on tracking position control of electropneumatic system using backstepping design," *Control Engineering Practice,* vol. 14, no. 8, pp. 923-933, 2006.

[27] F. J. Lee and C. C. Lin, "Adaptive backstepping control for linear induction motor drive to track periodic references," *IEE Proceedings-Electric Power Applications,* vol. 147, pp. 449-458, 2000.

[28] Z. J. Minashima and M. Yang, "Robust nonlinear control of a feedback linearizable voltage-controlled magnetic levitation system," *Transactions of the Institute of Electrical Engineers of Japan,* vol. 121, no. 7, pp. 1203-1211, 2001.

[29] Z. P. Jiang, E. Jiang and H. Nijmeijer, "Saturated stabilization and track control of a nonholonomic mobile robot," *Systems and Control Letters,* vol. 42, pp. 327-332, 2001.

[30] M. Chami, "The use of the CHDN to model a permanent magnet synchronous motor powered by ultracapacitors," *Vehicle power and propulsion,* 2004.

[31] S. Velinsky, B. Chu and T. Lasky, "Kinematics and Efficiency Analysis of the Planetary Roller Screw Mechanism," *Journal of Mechanical Design,* vol. 131, pp. 11-16, 2009.

[32] P. Lemor, "The roller–screw, an efficient and reliable mechanical component of electro-mechanical actuators," in *Proceedings of the 31st Intersociety*, 1996.

[33] D. Schinstock and T. Haskew, "Dynamic load testing of roller–screw EMAs," in *Proceedings of the 31st Intersociety*, 1996.

[34] A. Tselishchev and I. Zharov, "Elastic elements in roller–screw mechanisms," *Journal of Russian Engineering Research,* vol. 28, no. 11, pp. 1028-1040, 2008.

[35] H. E. Merritt, Hydraulic control systems, New York: John Wiley & Sons, 1967.

[36] P. Dransfield, Hydraulic control systems—design and analysis of their dynamics, Berlin: Springer, 1981.

[37] W. Gotz, Hydraulics. Theory and applications, Ditzingen: Robert Bosch Automation Technology Division Training, 1998.

[38] R. Bishop, Mechatronics Handbook, Florida: CRC Press LLC, 2002.

[39] A. J. Koivo, Fundamentals for control of robotic manipulators, New York: J. Wiley and Sons, 1989.

[40] T. V. Alekseeva, K. A. Artem'ev, A. A. Bromberg, R. I. Voitsekhovskii and N. Ul'yanov, Machines for Earthmoving Work, Theory and Calculations, Rotterdam: A. A. Balkema, 1992.

[41] A. Reece, "The Fundamental Equation of Earthmoving Mechanics," *Proceedings of Institution of Mechanical Engineers,* 1964.

[42] A. Hall and P. McAree, "Robust bucket position tracking for a large hydraulic excavator," *Mechanism and Machine Theory,* vol. 40, pp. 1-16, 2005.

[43] P. H. Chang and S. J. Lee, "A straight-line motion tracking control of hydraulic excavator system," *Mechatronics,* vol. 12, no. 1, pp. 119-138, 2002.

[44] O. Luengo, S. Singh and H. Cannon, "Modeling and Identification of Soil-tool Interaction in Automated Excavation," in *IEEE/RSJ International Conference on Intelligent Robotic Systems*,

Victoria, B.C., Canada, 1998.

[45] S. Vahed, X. Song, J. S. Dai, H. K. Lam, L. Seneviratne and K. Althoefer, "Soil Estimation Based on Dissipation Energy during Autonomous Excavation," in *Proceedings of the 17th World Congress The International Federation of Automatic Control*, Seoul, Korea, 2008.

[46] Y. B. Kim, J. Ha and H. Kang, "Dynamically optimal trajectories for earthmoving excavators," *Automation in Construction,* 2013.

[47] A. Hall, "Characterizing the Operation of a Large Hydraulic Excavator," Master Thesis, University of Queensland, 2003.

[48] W. Karam, "Generators of Static and Dynamic Forces at High Power in Electromechanical Technology," Doctoral Thesis, Université de Toulouse, 2007.

[49] S. Velinsky and M. Jones, "Kinematics of Roller Migration in the Planetary Roller Screw Mechanism," *Journal of Mechanical Design,* vol. 134, 2012.

[50] Y. Liu, M. Hasan and H. Yu, "Modelling and Remote Control of an Excavator," *International Journal of Advanced Mechatronic Systems,* vol. 2, no. 1, 2009.

[51] J. Craig, Introduction to robotics: mechanics and control, Addison-Wesley, 1986.

[52] L. Sidhom, "Sur la Dérivation Numérique : Algorithmes et Applications," PhD thesis, Institut National des Sciences Appliquées de Lyon - INSA Lyon, 2012.

[53] L. Wang, "Force Equalization for Active/Active Redundant Actuation System Involving Servo-hydraulic and Electro-mechanical Technologies," PhD thesis, Institut National des Sciences Appliquées de Toulouse - INSA Toulouse, 2012.

[54] M. Karpenko and N. Shapehri, "Fault – Tolerant control of a servohydraulic positioning system with crossport leakage," *IEEE Trans. on Contr. Syst. Technology,* vol. 13, pp. 155-161, 2005.

[55] V. Pommier, J. Sabatier, P. Lanusse and A. Oustaloup, "Crone control of a nonlinear hydraulic actuator," *Control Engineering Practice,* vol. 1, no. 4, pp. 391-402, 2002.

[56] V. M. Zatsiorsky, Kinetics of Human Motion, Education, 2002.

[57] C. P. Tang, Lagrangian Dynamic Formulation of a Four-Bar Mechanism with Minimal Coordinates, March 2006.

[58] J. Velagic, B. Lacevic and N. Osmic, "Nonlinear Motion Control of Mobile Robot Dynamic Model," *Advanced Robotic System International,* pp. 531-552, 2008.

[59] B. Lacevic, J. Velagic and B. Perunicic, "Reduction of Control Torques of Mobile Robot Using Hybrid Nonlinear Position Controller," in *International Conference on Computer as a Tool*, Belgrade, 2005.

[60] Bonfiglioli, BMD Permanent Magnet AC Synchronous Motors, 2012.