Université de
Bourgogne
Dijon

Laboratoire Électronique
Informatique et Image
UMR 5158

Département
de
Mathématiques

HABILITATION À DIRIGER DES RECHERCHES

Présentée par

Eric Busvelle

SUR LES OBSERVATEURS DES SYSTÈMES NON-LINÉAIRES

soutenue le 29 Juin 2004 devant le Jury composé de

MM	Guy Bornard, Professeur	Rapporteur
	Michel Fliess, Directeur de Recherche	Rapporteur
	Jean-Paul Gauthier, Professeur	
	Rémi Langevin, Professeur	
	Michel Paindavoine, Professeur	
	Gauthier Sallet, Professeur	Rapporteur

Je dédie ce travail à mon regretté ami, Daniel Rakotopara

Avant tout, je tiens à remercier tous les membres du jury pour avoir accepté de consacrer une partie de leur temps à la lecture de ce rapport. Merci donc à Guy Bornard, Michel Fliess, Jean-Paul Gauthier, Rémi Langevin, Michel Paindavoine et Gauthier Sallet.

Je remercie évidemment plus particulièrement Jean-Paul Gauthier qui m'a fait l'honneur de me diriger pendant l'essentiel des travaux que je présente ici.

Je tiens à remercier toutes les personnes avec lesquelles j'ai travaillé dans les différents laboratoires qui m'ont accueilli et permis d'obtenir ces résultats. Je remercie ainsi tous les membres du laboratoire de mathématiques de l'Université de Rouen et de l'INSA de Rouen, en particulier Denis de Brucq, Jean-Marie Strelcyn, Rachid Kharab et José de Sam Lazaro.

Je remercie aussi tous mes anciens collègues de Shell Research, parmi lesquels Frédéric Viel, François Deza, Denis Bossanne, Pierre Sommelet, Leen Roozemond, Marjoleine van Dootingh, Jérôme de Morant, Philippe Langonnet et Daniel Rakotopara.

A l'Université de Bourgogne, je remercie très sincèrement tous les membres du Laboratoire de Topologie qui m'ont aidé à m'intégrer, et si je ne cite personne nommément, c'est pour ne par heurter ceux que je pourrai oublier.

Je remercie tous les membres du Laboratoire d'Analyse Appliquée et d'Optimisation qui m'ont très chaleureusement accueillis. C'est dans ce laboratoire que j'ai réalisé l'essentiel du travail présenté ici. Je remercie plus spécifiquement Bernard Bonnard pour les nombreuses discussions – enrichissantes pour moi – que nous avons eues ensemble. Mais je remercie aussi Christian Michelot, Abderrahim Jourani, Abdessamad Barbara et tous les autes membres du laboratoire AAO.

Je fais maintenant partie du Laboratoire d'Électronique, Informatique et Image de l'Université de Bourgogne. Je remercie tous les membres de ce laboratoire pour leur accueil dont, j'espère, je saurai être digne.

Je remercie Alexandre Choux qui a bien voulu faire sa thèse sous mon co-encadrement, ainsi que le département micro-cibles du CEA dans lequel se déroule cette thèse.

Merci aussi à Bronislaw Jakubczyk, Vladimir Zakalyukin, Alain Rapaport, Jean-Luc Gouzé, Jean-Baptiste Pomet, Abderahmane Hajraoui, Andrzej Maciejewski pour leurs discussions et/ou leurs invitations à des séminaires.

Je remercie Madame Alexandre pour son extrême disponibilité et sa gentillesse. Je remercie plus généralement toutes les secrétaires des laboratoires dans lesquels j'ai travaillé, et qui font un travail difficile avec une grande efficacité.

Je remercie deux mathématiciens qui, s'ils exercent dans des spécialités un peu éloignée de la mienne, n'en restent pas moins des amis, Lucian Roiu et Abdelaziz El Matouat.

Enfin, merci à ma femme, Mbola, à mes enfants, Paul et Dylan, de m'avoir supporté pendant les étapes difficiles de ce travail, en particulier ces derniers mois.

Table des matières

1. Présentation générale

1.1. Curriculum Vitae.

Nom patronymique :	BUSVELLE
Prénom :	Eric
Date de Naissance :	12 Février 1964
Grade :	MCF 5$^{\text{ième}}$ échelon
Section CNU :	26
Tél. :	03-80-39-58-38
Fax :	03-80-39-59-10
mèl :	busvelle@u-bourgogne.fr
URL :	http ://www.u-bourgogne.fr/monge/e.busvelle/accueil.php
Adresse professionnelle :	Université de Bourgogne, LE2I, UMR CNRS 5158 BP 47870, 21078 Dijon Cedex
Adresse personnelle :	8 rue du Levant, 21110 Genlis

Mon directeur de recherche est le **Professeur Jean-Paul Gauthier**, du laboratoire LE2I de l'Université de Bourgogne.

J'ai préparé ma thèse conjointement dans le laboratoire d'Analyse et Modèles Stochastiques (LAMS), et le laboratoire d'Automatique, Capteurs, Instrumentation et Systèmes (LACIS). Le travail, que j'ai effectué sous la direction du Professeur Denis de Brucq, a porté sur le filtrage non linéaire. J'ai soutenu ma **Thèse de doctorat de l'Université de Rouen** le *21 Octobre 1991* devant le jury composé de

José de Sam Lazaro, Professeur	Président
Denis de Brucq, Professeur	Rapporteur
Etienne Pardoux, Professeur	Rapporteur
Claude Dellacherie, Directeur de recherche au CNRS	Examinateur
Jean-Paul Gauthier, Professeur	Examinateur
Daniel Rakotopara, Ingénieur de recherche Shell	Examinateur

Le titre de la thèse est "*Immersions et filtres de dimension finie ; filtrage particulaire*"

> J'ai établi des conditions d'existence de filtre de dimension finie, aussi bien en temps discret qu'en temps continu. Ces résultats présentent l'originalité d'être explicites et constructifs. Ces travaux ont été présentés dans une conférence internationale (CDC) J'ai aussi démontré un théorème d'approximation de la loi conditionnelle en filtrage non linéaire par une méthode particulaire. Ce résultat a été ensuite appliqué au sein de la société Shell.

J'ai été embauché le *5 Novembre 1990* en tant qu'Ingénieur de recherche au **Centre de Recherche de la Société Shell** à Grand-Couronne. J'ai démissionné en *Novembre 1995*, avant la fermeture du centre, pour pouvoir continuer mes recherches.

> J'ai mené des recherches en théorie du contrôle et plus particulièrement sur la théorie des observateurs. J'ai publié plusieurs articles sur ce sujet pendant cette période. J'ai parallèlement continué à travailler en collaboration avec des membres de l'Université de Rouen. Un article paru dans Applicae Mathematicae a plus particulièrement résulté de ces activités.

Depuis *1*$^{\text{er}}$ *Octobre 1995*, je suis **Maître de Conférences** à l'Université de Bourgogne. J'effectue actuellement mon travail de recherche au sein du laboratoire Electronique, Informatique et Image (LE2I).

> Mes activités de recherche sont détaillées au paragraphe 3.1.

La suite de ce dossier présente très brièvement mes activités de recherche (page ii), d'enseignement et d'encadrements (page v) et administratives (page vi). L'essentiel du dossier consiste en une présentation plus approfondie de mes résultats (page 1) suivie d'une sélection de cinq articles représentatifs :

- [3] High-Gain and Non-High-Gain Observers for Nonlinear Systems, paru en 2002 dans Contemporary Trends in Nonlinear Geometric Control Theory and its Applications, *World Scientific*

- [2] On determining unknown functions in differential systems, with an application to biological reactors, paru en 2003 dans *ESAIM : COCV*
- [1] Observation and Identification Tools for Nonlinear systems. Application to a Fluid Catalytic Cracker. Soumis pour publication en 2003 à *International Journal of Control*
- [4] Geometric optimal control of the atmospheric arc for a space shuttle, paru en 2002 dans Contemporary Trends in Nonlinear Geometric Control Theory and its Applications, *World Scientific*
- [7] Numerical Integration of Differential Equations in Presence of First Integrals : Observer Method, paru en 1994 dans *Applicationes Mathematicae*

La bibliographie se trouve page 24.

1.2. **Recherche.**

1.2.1. *Synthèse.* Après ma thèse consacrée au filtrage non-linéaire, je me suis orienté vers le problème équivalent mais dans sa version déterministe, les observateurs. Ce changement d'approche était motivé d'une part par les résultats de non existence de filtres finis et d'autre part par la lourdeur des algorithmes nécessaires pour le filtrage non-linéaire (résolution d'une équation aux dérivées partielles stochastique).

Dans le monde de l'ingénieur, l'observateur le plus connu est le filtre de Kalman. Issu de l'approche stochastique, il a aussi de très bonnes propriétés déterministes. Sa version non linéaire est le filtre de Kalman étendu. Malheureusement, le filtre de Kalman étendu – très utilisé en pratique – ne présente pas de bonnes propriétes de convergence. Par contre, l'observateur grand-gain est du point de vue mathématique et déterministe un excellent outil. J'ai dans tout le début de ma carrière tenté de développer cet outil, en cherchant à exploiter au mieux ses propriétés mathématiques pour développer et valider des techniques de contrôle non linéaires les plus générales possibles (comme il en existe dans le cas linéaire). Avec d'autres ingénieurs des procédés, nous avons appliqué l'observateur grand-gain à des colonnes à distiller, des réacteurs catalytiques à lit fluidisé et des réacteurs de polymérisation. Dans certains cas, nous avons couplé l'observateur obtenu avec une loi de commande utilisant toute les propriétés de convergence de l'observateur afin d'obtenir un système stable. Toutes nos applications ont été justifiées théoriquement, *i.e.* la convergence et la stabilité ont été mathématiquement prouvées.

Cependant, les ingénieurs utilisent très souvent le filtre de Kalman étendu car, bien que ne reposant pas sur une solide théorie mathématique (c'est une linéarisation assez mal justifiée du filtre de Kalman linéaire), il possède en pratique de très bonnes propriétés locales qui le rendent robuste au bruit, ce que n'est pas l'observateur grand-gain. Plus récemment, j'ai donc cherché, avec mon directeur de recherches Jean-Paul Gauthier, à utiliser les techniques de démonstration du grand-gain pour justifier *a posteriori* les propriétés d'un filtre de Kalman étendu dans une version modifiée s'inpirant de l'observateur grand-gain. L'article le plus abouti donne finallement une version du filtre de Kalman étendu grand-gain justifiée théoriquement, alliant donc la vitesse de convergence élevée de l'observateur grand-gain et la bonne robustesse au bruit du filtre de Kalman étendu.

L'approche fructueuse que nous avons développé pour mettre au point cet observateur a aussi été utilisée pour développer une stratégie d'identification. L'identification est en quelque sorte une généralisation du problème de l'observateur. Il s'agit cette fois d'estimer, en plus de paramètres, des fonctions inconnues. Les résultats que nous avons publié à ce sujet présentent d'ailleurs une certaine similitude avec les résultats concernant l'observabilité et les observateurs, tout en étant plus complexes

Parallèlement à ces travaux, je me suis intéressé à d'autres problèmes toujours liés au filtrage et aux observateurs. Par exemple, en m'inspirant de certaines propriétés des observateurs, j'ai mis au point avec des co-auteurs de l'Université de Rouen un algorithme d'analyse numérique, permettant de calculer la solution numérique de systèmes d'équations différentielles contenant des intégrales premières, qu'ils soient Hamiltoniens où non, et qu'ils contiennent une ou plusieurs intégrales premières. Pour des systèmes Hamiltoniens, nous avons comparé notre algorithme avec les algorithmes symplectiques et les résultats sont excellents pour notre méthode. Je me suis aussi intéressé à la commande optimale de réacteurs batch et au problème de ré-entrée atmosphérique de la navette spatiale. Enfin, j'ai co-publié un article sur une approche prometteuse d'approximation de la solution du filtrage nonlinéaire par maximum d'entropie.

1.2.2. *Description détaillée.* Au centre de recherche Shell, j'ai travaillé à la fois sur des sujets de recherche et aussi sur des applications de mes travaux de recherche. Mes publications témoignent de ce double intérêt. Parmi les travaux que j'ai effectué, plusieurs n'ont été publiés que sous forme de rapport confidentiels. Les travaux publiés à l'extérieur de la société et correspondant à cette période d'activité sont les références [5, 8–11, 18–21].

> La plupart des travaux de cette période sont consacrés aux observateurs non linéaires et à la stabilisation des systèmes non linéaires à l'aide d'observateurs. Nous avons développé une approche assez générale pour construire des observateurs sur une large classe de systèmes non-linéaires tels qu'ils se présentent dans l'industrie de raffinage, l'exemple typique étant la colonne à distiller. Les observateurs sont de type grand-gain (article de Gauthier-Hammouri-Othmann, A simple observer for nonlinear systems, Application to bioreactors, IEEE Trans on Aut Control, 37, No 6, p 875–880, 1992). Leur convergence est exponentielle et arbitrairement rapide. Grâce à cette dernière propriété, nous avons établi un principe de séparation non linéaire (restreint) utilisant l'observateur grand-gain et des contrôles non linéaires, de type géométrique où de Lyapunov.

J'ai appliqué l'observateur non linéaire résultant de ces travaux de recherche sur une unité de production *(ensemble de deux colonnes binaires d'un gas-plant).* J'ai aussi validé l'approche observateurs sur des simulateurs d'unité de raffinage professionnels.

En dehors de ces travaux, j'ai aussi publié un article ([7]) en collaboration avec des membres du laboratoire de mathématiques de l'université de Rouen, concernant une méthode numérique originale, basée sur la théorie des observateurs, permettant l'intégration précise des systèmes différentiels présentant des intégrales premières connues, tels que les systèmes hamiltoniens (mais pas seulement).

> Dans cet article, nous avons décrit la méthode, démontré ses propriétés mathématiques, et étudié numériquement l'algorithme sur plusieurs systèmes classiques tels que le problème de Kepler plan et spatial, le double oscillateur harmonique, le système de Gavrilov-Shil'nikov, les équations d'Euler sur so(4) et le flot d'Anosov. Ces trois derniers systèmes sont réputés pour faire apparaître du chaos numérique. Notre algorithme diminue de façon drastique les erreurs numériques d'intégration et préservent quasi-idéalement les intégrales premières.

Dès *1995*, j'ai commencé à travailler avec le Professeur Bernard Bonnard de Dijon sur le projet de contrôle optimal des réacteurs chimiques, puis sur un problème de rentrée atmosphérique de la navette spatiale.

> Après avoir étudié la théorie du contrôle optimal, j'ai réalisé un logiciel de simulation de réacteur chimique couplé avec un contrôleur optimal, permettant de vérifier l'intérêt pratique de l'approche contrôle optimal et utilisation des trajectoires singulières. Le projet a débouché sur une thèse en génie chimique présentée à l'Université de Caen, visant à mettre en pratique la théorie développée et à valider les simulations effectuées.

Je me suis naturellement intéressé au couplage d'une loi de commande optimale avec un observateur dans un cadre non-linéaire. J'ai montré dans le cas du réacteur batch que la synthèse était régulière et que le contrôleur couplé avec un observateur à convergence exponentielle était sous-optimal. J'ai présenté ce résultat à l'Institut Henri Poincaré le *2 Mars 1998* ([17]).

> La notion adéquate pour comprendre les phénomènes qui interviennent est celle des synthèses régulières et des solutions au sens de Filippov des équations différentielles avec second membre discontinu (cf. travaux de Filippov, Boltianskii, Brunowsky, Hermès, Sussmann). Le théorème que j'ai établi dit que la cible est atteinte en un temps supérieur au temps optimal d'une quantité ne dépendant que de la vitesse de convergence de l'observateur et qui peut être rendue arbitrairement petite, dès que la condition initiale est à l'intérieur du domaine d'accessibilité. La démonstration ne repose que sur le fait que la synthèse optimale est constituée de trajectoires qui sont des solutions de Filippov de l'équation du système, et que cette synthèse est une stratification qui jouit de propriétés de transversalité (au sens des synthèses régulières de Boltianskii). Les propriétés de convergence exponentielle de l'observateur de Kalman étendu grand-gain, que j'ai contribué à développer ces dernières années, sont aussi indispensables au résultat final.

Parallèlement à ce travail, j'ai continué à m'intéresser au problème de la synthèse optimale proprement dite ([4]), en particulier sur le problème de l'entrée atmosphérique d'une navette spatiale où d'un étage de fusée réutilisable (projet du CNES). Le problème a été abordé en utilisant des outils géométriques qui permettent de mieux comprendre les singularités propres à (presque) tout système non linéaire. Ces singularités essentielles peuvent échapper à une étude purement numérique alors que, lorsque leur existence a été établie, elles conduisent à une synthèse optimale qui peut finallement être calculée numériquement sur des portions de trajectoires pour lesquelles on a une parfaite connaissance qualitative de la synthèse. La particularité du problème d'entrée atmosphérique (par rapport au réacteur chimique par exemple) consiste en la présence de nombreuses contraintes d'état (parmi lesquelles la puissance absorbée par la navette).

> Le modèle d'entrée atmosphérique de la navette est constitué de six équations différentielles non linéaires et fortement couplées. Le problème de contrôle optimal est traité par le principe du maximum de Pontriaguine mais l'existence de contraintes d'état oblige à considérer une version beaucoup moins connue de ce principe (Maurer, 1977). Il est en particulier assez difficile de le résoudre numériquement, en raison de la présence de trajectoires aux frontières des contraintes, dont le comportement est similaire à celui de trajectoires singulières, mais pour lesquelles les conditions de sortie sont moins claires. Pratiquement, il faut utiliser une méthode numérique dédiée à ce type de problème (*multiple-shooting method).*

Plus récemment, je suis revenu à l'étude des observateurs grand-gain et plus particulièrement le filtre de Kalman étendu grand-gain qui allie de très bonnes performances pratiques (le filtre de Kalman étendu est utilisé depuis des décénnies dans le milieu industriel) et des résultats de convergence théorique. J'ai écrit avec Jean-Paul Gauthier deux articles dans ce domaine ([2, 3]). Le premier ([3]) est un résultat théorique issu des problèmes appliqués que se posent les ingénieurs depuis très longtemps. Nous exhibons un algorithme permettant de réaliser un observateur alliant de bonnes performances locales et globales, ce qui était considéré comme contradictoire. Le résultat est évidemment basé sur un théorème et donc justifie théoriquement une approche qui avait déjà été expérimentée par des ingénieurs.

> Dans cet article, nous montrons la convergence exponentielle d'un observateur dont le comportement initial est de type grand gain et dont le comportement asymptotique est celui du filtre de Kalman étendu classique. Ce théorème permet en pratique de profiter de la réponse rapide du grand gain aux erreurs initiales et aux perturbations et de la précision de Kalman étendu au voisinage des trajectoires du système. Une application de ce résultat à une colonne à distiller permet d'en illustrer l'intérêt.

Le second résultat ([2]) concerne l'identifiabilité, qui est un problème lié à l'observabilité mais qui n'avait pas encore été étudié sous cet angle. Partant d'un problème concret – un réacteur biologique – nous étudions la possibilité d'estimer une fonction inconnue des variables d'état sur la base des mesures, conjointement avec la reconstruction classique des variables d'état non mesurées via un observateur. Les résultats que nous avons obtenu ont largement dépassé nos espérances et ont ouvert de nouvelles questions pour lesquelles nous n'avons pas encore de réponses.

> Nous avons montré que l'identifiabilité est une propriété générique si le nombre de sorties est supérieur où égale à 3 et qu'en revanche, cette propriété n'est pas générique dans le cas où le nombre de sorties est inférieur où égale à 2. Dans le cas d'une où de deux sorties, nous avons abouti à une classification des systèmes identifiables. Nous avons appliqué cette approche avec succès à un modèle de bioreacteur.

J'ai été invité au Banach Center à Warsaw, Pologne, pour une conférence intitulée "Geometry in Nonlinear Control" en Juin 2003 ([13]). Au cours de cette conférence, j'ai présenté les résultats que j'avais obtenu sur l'identification.

Je viens de terminer un autre article avec le professeur Jean-Paul Gauthier – [1], soumis à *International Journal of Control* – qui fait une synthèse et généralise certains des résultats des deux articles précédents. Il comble aussi une lacune des deux articles précédents concernant certains systèmes observables pour lesquels nous ne savions pas construire un observateur. Les nouveaux résultats sont appliqués à un réacteur catalytique à lit fluidisé.

Je travaille actuellement, toujours avec mon directeur de recherche Jean-Paul Gauthier, à la généralisation des résultats sur l'identifiabilité en présence d'une variable de contrôle, ce qui permettrait de généraliser le résultat précédent. Il semble par contre difficile d'obtenir une description aussi détaillée dans le cas de plusieurs variables de contrôle que celle publiée sans contrôle, tant la combinatoire de la classification est complexe.

Enfin, la coopération avec le CEA de Valduc nous a permis de mettre au point une méthode originale de validation de la géométrie de microbilles, basée sur la transformée de Fourier-Bessel, qui fera l'objet d'un article dès que le CEA aura implémenté cet algorithme sur des données réelles. Ce travail entre dans le cadre du projet Laser Megajoule dont il sera encore question plus loin dans cette introduction.

1.3. **Enseignement et encadrements.**

1.3.1. *Enseignement.* Lors de ma carrière au centre de recherche Shell, j'étais responsable des stages. J'ai bien sur aussi encadré personellement de nombreux stagiaires – issus d'écoles d'ingénieurs et de diverses Universités d'Europe – à cette même période.

Depuis ma nomination en *1995*, j'ai été chargé du cours d'Analyse Numérique de la licence de Mathématiques ainsi que de plusieurs groupes de TDs. J'ai rédigé un cours polycopié d'Analyse Numérique avec exercices de 120 pages. J'ai aussi assuré des TDs de Probabilités et Statistique, niveau licence de mathématiques, et enseigné une UV d'Algorithmique et le Calcul Formel (cours et TDs) en Maîtrise d'Ingéniérie Mathématiques et Maîtrise de Mathématiques pures (UV commun). Suivant les années, le cours était consacré

- à la théorie de la complexité, avec l'étude des principaux algorithmes célèbres, définitions de la calculabilité, calcul de coûts.
- à l'utilisation d'algorithmes algébriques (modulaires et p-adiques) pour la factorisation des polynômes et la décomposition cylindrique des ensembles semi-algébriques en vu de la résolution automatique de problèmes de contrôle (stabilisabilité, calcul de l'espace d'accessibilité)
- aux séries formelles non-commutatives et leur lien avec les $\mathbb{R}$-automates séries rationnelles, reconnaissables), en vu de la bilinéarisation des systèmes non linéaires, permettant d'appliquer des observateurs par exemple.

J'ai été pendant 3 ans responsable des TPs en license et maitrise de Mathématiques, et administrateur de la salle d'informatique qui comporte 30 PCs en réseau (autour d'un serveur sous Windows NT). J'ai assuré de nombreux TPs, que ce soit des projets d'analyse numérique en licence où des projets directement issus de problématiques industrielles en maitrise d'ingéniérie mathématique. J'ai depuis *1995* régulièrement encadré des stagiaires de la Maîtrise d'Ingéniérie Mathématiques, que ce soit dans le cadre de travaux dans le laboratoire où dans le cadre de stages en entreprise.

Depuis *2002*, j'ai décidé de centrer mon enseignement sur le DEUG, ce qui me laisse plus de temps pour intensifier mon travail de recherche et préparer mon habilitation à diriger des recherches.

1.3.2. *Encadrements.* J'ai co-encadré la thèse de doctorat de Frédéric Viel : ”Stabilité des systèmes nonlinéaires contrôlés par retour d'état estimé, Application aux réacteurs de polymérisation et aux colonnes à distiller”. Cette thèse a été soutenue le 8 Juin 1994 à l'université de Rouen devant le jury composé de

Jacques Descusse, Directeur de Recherche au CNRS	Président
Georges Bastin, Professeur	Rapporteur
Gauthier Sallet, Professeur	Rapporteur
Jean-Paul Gauthier, Professeur	Examinateur
Erick Lenglart, Professeur	Examinateur
Bernard Bonnard, Professeur	Examinateur
Andras Kortbeek, Chef de département à Shell	Examinateur
Eric Busvelle, Ingénieur de recherche à Shell	Examinateur

J'ai co-encadré cet étudiant alors que j'était ingénieur de recherche à Shell dans la laboratoire GCAS et qu'il bénéficiait d'une bourse CIFFRE pour préparer sa thèse au GCAS. Notre travail a donné lieu à trois articles [5, 6, 8], la thèse étant constituée d'une introduction et des articles [5, 6].

Je co-encadre actuellement (avec Jean-Paul Gauthier) Alexandre Choux qui prépare une thèse avec le CEA (Valduc) concernant le contrôle de l'épaisseur de la couche de glace d'un micro–ballon. Cette thèse, en partie financée par le CEA, se déroule conjointement avec le laboratoire le2i de l'Université de Bourgogne et le département des micro-ballons au CEA de Valduc dont le chef de projet est Monsieur Ghislain Pascal.

Les micro-ballons (où micro-cibles) font partie du projet de laser mégajoule (LMJ) visant à réaliser en laboratoire les expériences désormais prohibées par le traité d'interdiction complète des essais nucléaires. Les micro-ballons sont des sphères de rayon environ un millimètre, contenant un mélange deuterium-tritium (DT) à une température inférieure à 20 K. A cette température, le DT est sous forme de glace. L'objet de la thèse est de développer un contrôle en température afin de mettre cette couche de glace en conformité, de sorte que la réaction de fusion puisse avoir lieu. L'enjeu est crucial pour le CEA, le problème est majeur en raison des contraintes opératoires et des spécifications finales demandées.

1.4. Administration et coopérations industrielles.

1.4.1. *Administration.* J'ai été, au centre de recherche Shell de Grand-Couronne, chef du projet *"Conception d'un outil d'aide au développement et à la maintenance des applications de contrôle avancé"*. Cet important projet impliquait plusieurs ingénieurs en raffinerie et dans le département contrôle du siège de la société Shell à La Haye, Pays-Bas.

J'ai fait partie de la commission de spécialistes de rang B en 25–26$^{\text{ième}}$ section à Dijon, entre *1998* et *2001*, date à laquelle j'ai démissioné.

Je suis membre extérieur de l'U.F.R. « Signaux, Systèmes et Traitement » de l'Université de Tétouan, Maroc.

1.4.2. *Coopérations industrielles.*

- **CEA, Valduc.** J'ai réalisé pour le CEA de Valduc un logiciel permettant de valider des spécifications très fines sur des micro-ballons. Ce logiciel est essentiellement une application de traitement d'image pour calculer des mesures avec une précision inférieure au pixel. Une technique originale a été élaborée – en collaboration avec le Professeur Jean-Paul Gauthier – basée sur la transformée de Fourier-Bessel. Cette étude devrait faire l'objet d'une publication conjointe avec le CEA, actuellement en cours de rédaction (nous attendons les tests pratiques de la part du CEA).

  C'est cette coopération fructueuse qui nous a permis de lancer la thèse (en cours) d'Alexandre Choux cofinancée par le CEA
- **Lennox.** J'ai été responsable de deux contrats consécutifs avec Lennox qui est une entreprise internationale ayant une usine à Longvic (siège social à Lyon) fabriquant des climatiseurs indistriels (roof-top). L'étude portait sur la réalisation de modèles de comportement statique de ces climatiseurs.
- **Thomson.** Nous avons placé plusieurs étudiants en stage chez Thomson à Genlis, avec qui nous avons des contacts privilégiés. J'ai plus particulièrement co-encadré un étudiant en DESS à Chambéry en 2002 sur un problème d'optimisation qui comportait une grosse contrainte de temps de calcul. L'algorithme développé est actuellement utilisé sur la chaine de production de tubes cathodiques de l'usine de Genlis.
- **CNES, Paris.** J'ai été en contact avec des responsables du CNES qui nous ont soumis un problème de calcul de trajectoire optimale pour l'entrée atmosphérique d'un engin spatial. Un article a résulté de cette étude qui est toujours en cours actuellement, sous la responsabilité de Bernard Bonnard.
- **Shell.** J'ai gardé de nombreux contacts avec cette multinationale, ce qui m'a permis de placer une étudiante en stage. Je garde des contacts avec les centres de recherche des Pays-Bas (La Haye et Amsterdam) et des USA (Houston, Texas). L'unique centre de recherche Shell en France a été fermé.

Le texte qui va suivre est une présentation synthétique des articles [1] à [12] cités en bibliographie.

2. Observabilité et observateurs

On considère un système d'équations différentielles de la forme

$$\left\{ \begin{array}{lcl} \dfrac{dx}{dt}(t) & = & f(x(t), u(t)) \\ x(0) & = & x_0 \end{array} \right. \tag{1}$$

dans lequel $x(t) \in X$ est l'état du système et $u(t) \in U_{\text{adm}}$ est le contrôle. X est une variété de dimension n et $U_{\text{adm}} \subset \mathbb{R}^p$ est un ensemble de valeurs de contrôles dits admissibles. f est une famille de champs de vecteurs paramétrée par u. Pratiquement, ce système d'équations différentielles est le modèle d'évolution d'un procédé chimique où d'un système physique contrôlé. Ce peut-être par exemple l'évolution de la composition des produits d'un réacteur chimique, le contrôle représentant alors la composition des réactifs alimentant le réacteur.

A ces équations différentielles, on ajoute l'équation suivante

$$y(t) = h(x(t), u(t)) \tag{2}$$

où h est une application de $X \times U_{\text{adm}}$ à valeurs dans $\mathbb{R}^q$, $y(t)$ étant l'observation. Au niveau de cette introduction, on ne précisera pas plus les hypothèses qui varieront en fonction des résultats que l'on énoncera. Le problème que l'on se pose est le suivant : sans connaître l'état initial du système (x_0) mais connaissant le contrôle que l'on applique $\left((u(t))_{t\geq 0}\right)$ et au vu de la trajectoire de l'observation $\left((y(t))_{t\geq 0}\right)$, peut-on reconstruire la trajectoire de l'état $\left((x(t))_{t\geq 0}\right)$? Pratiquement, dans le cas d'un réacteur chimique, on mesure des températures jusqu'à un instant t, éventuellement d'autres quantités, et on souhaite connaître l'état du système à ce même instant t où, ce qui revient au même, connaître x_0.

On distingue deux problèmes, le premier problème est l'**observabilité** : étant donné un système constitué des deux équations (1,2), l'application $x_0 \longrightarrow (y(t))_{t\geq 0}$ est elle injective ? Plus précisemment, soit $(u(t))_{0\leq t<T_u}$, notons Σ_u l'application qui à x_0 associe $(y(t))_{0\leq t<T_{x_0,u}}$ où $T_{x_0,u}$ est le temps maximal pour lequel la solution de (1) est définie. On dira que le système précédent est observable si pour tout $(u(t))_{t\geq 0}$, deux points distincts x_1 et x_2 de X sont toujours distinguables *i.e.*

$$\left\{ 0 \leq t \leq \inf\left(T_{x_1,u}, T_{x_2,u}\right) \overset{\text{déf.}}{=} T_\infty ; \Sigma_u(x_1)|_t \neq \Sigma_u(x_2)|_t \right\}$$

n'est pas de mesure nulle.

Le second problème, pour un système observable, est la synthèse d'un **observateur**. Un observateur est en fait un algorithme qui permet effectivement d'inverser Σ_u. Pour des raisons qui deviendront plus claires après un détour par la version stochastique du problème, l'algorithme que nous chercherons sera de la forme suivante : on appelle observateur exponentiellement convergent la donnée d'une équation différentielle

$$\frac{dZ}{dt}(t) = F(Z(t), u(t), y(t))$$

d'une condition initiale $Z(0) = Z_0$ et d'une équation $z(t) = H(Z(t), u(t), y(t))$ telles que : il existe $\Lambda > 0$, p polynôme tel que que soit x_0, on ait

$$\|x(t) - z(t)\| \leq p(t)\, e^{-\Lambda t} \|x(0) - z(0)\|$$

pour tout $t < T_\infty$. Notons que l'on a fait le choix de ne présenter que les observateurs qui convergent exponentiellement car en pratique, tous les observateurs que l'on a construit étaient à convergence exponentielle.

Notons aussi que cette définition n'a un intérêt clair que si $T_\infty = \infty$.

2.1. **Cadre stochastique : filtre de Kalman-Bucy.** La problématique de l'observateur est très liée à la problématique de l'estimation losque l'on se place dans un cadre stochastique. Les équations d'état et d'observation s'écrivent alors

$$(3) \qquad \left\{ \begin{array}{lcl} dX(t) & = & f(X(t), u(t))\, dt + b(X(t), u(t)) \circ dW(t) \\ dY(t) & = & h(X(t), u(t))\, dt + dV(t) \end{array} \right.$$

où W et V sont deux processus de Wiener indépendants et où $X(0)$ est une variable aléatoire indépendante de W et V. Dans ce cadre, la question est de calculer la loi $X(t)$ conditionnellement à la tribu engendrée par le processus Y entre 0 et t, notée $\mathcal{L}(X(t) \diagup \mathcal{F}_Y^t)$. Cette loi est donnée par une équation aux dérivées partielles stochastique, l'équation de Duncan-Mortensen-Zakaï. Cette équation donne donc une solution théorique au problème d'estimation qui, dans ce cadre très général, s'appelle filtrage. Notons que l'équation de Duncan-Mortensen-Zakaï se présente sous la forme d'une équation qui décrit l'évolution de la loi de $X(t)$ en l'absence d'observation, perturbée par un terme additif qui tient compte de l'observation.

Dans un très petit nombre de cas, l'équation de Duncan-Mortensen-Zakaï peut se ramener à un système d'équation différentielles stochastiques : on dit alors que le système admet un filtre de dimension finie. Ma thèse de l'Université de Rouen était consacrée à l'étude de certains de ces cas, en particulier à une caractérisation explicite de ces filtres de dimension finie, cf [22, 24, 25]. Mais il existe une classe très importante de systèmes (3) admettant un filtre de dimension finie, ce sont les systèmes linéaires. Dans le cadre stochastiques, ces systèmes sont de la forme

$$(4) \qquad \left\{ \begin{array}{lcl} dX(t) & = & A(u(t))\, X(t)\, dt + b(u(t))\, dW(t) \\ dY(t) & = & C(u(t))\, X(t)\, dt + dV(t) \end{array} \right.$$

et admettent toujours un filtre de dimension finie. Si la loi initiale de $X(t)$ est gaussienne, alors l'équation de Duncan-Mortensen-Zakaï se ramène aux célèbres équations du filtre de Kalman ([43]). Dans ce cas, la loi $\mathcal{L}(X(t) \diagup \mathcal{F}_Y^t)$ cherchée est une loi normale $\mathcal{N}(m(t), P(t))$ et $m(t)$ et $P(t)$ sont solutions d'équations différentielles où l'on retrouve une équation de prédiction qui représente l'évolution de l'état du système et un terme correctif (additif) qui dépend de l'observation, plus précisemment qui est proportionnel à la différence entre l'observation et une prédiction de l'observation, différence que l'on appelle innovation. Le facteur de proportionalité est une matrice appellée gain de Kalman.

Les équations de Kalman s'écrivent, en supposant $X(0) \sim \mathcal{N}(m(0), P(0))$:

$$(5) \qquad \left\{ \begin{array}{lcl} dm(t) & = & A(u(t))\, m(t)\, dt + P(t)\, C(u(t))^T R^{-1} (dY(t) - C(u(t))\, m(t)\, dt) \\ \dfrac{dP}{dt}(t) & = & A(u(t))\, P(t) + P(t)\, A(u(t))^T + Q - P(t)\, C(u(t))^T R^{-1} C(u(t))\, P(t) \end{array} \right.$$

où Q et R sont les covariances respectives de W et de V.

Revenons au cas plus général du système (3). Plutôt que de chercher $\mathcal{L}(X(t) \diagup \mathcal{F}_Y^t)$, on peut se contenter de chercher $\hat{X}(t) \overset{\text{déf.}}{=} \mathrm{E}[X(t) \diagup \mathcal{F}_Y^t]$ ce qui, dans le cas linéaire, vaut $m(t)$. Dans le cas non linéaire général (en l'absence de filtre fini), $\hat{X}(t)$ ne s'écrit pas comme solution d'un système d'équations différentielles : il faut résoudre l'équation de Duncan-Mortensen-Zakaï pour calculer exactement $\hat{X}(t)$ qui d'ailleurs n'est pas forcément une bonne estimation de $X(t)$. Pratiquement, on va donc linéariser le sytème (3) afin de pouvoir écrire les équations du filtre de Kalman. Ceci conduira au filtre de Kalman étendu qui est le coeur de ce travail. Disons dès maintenant que le procédé conduit à des équations qui, d'un point de vue stochastique, donnent au mieux une approximation de la loi conditionnelle cherchée. D'un point de vue déterministe, on peut espérer un observateur exponentiellement convergent. Ce que l'approche basée sur une linéarisation nous prédit, c'est que le filtre de Kalman étendu devrait avoir de bonnes propriétés locales. Revenons au cas déterministe.

2.2. **Les équations de Kalman dans le cas déterministe.** Le filtre de Kalman étendu associé au système suivant, où on a considéré $X = \mathbb{R}^n$,

$$(6) \qquad \left\{ \begin{array}{lcl} \dfrac{dx}{dt}(t) & = & f(x(t), u(t)) \\ y(t) & = & h(x(t), u(t)) \end{array} \right.$$

s'écrit

$$(7) \qquad \left\{ \begin{array}{lcl} \dfrac{d\xi}{dt}(t) & = & f(\xi(t), u(t)) + P(t)\, H(t)^T R^{-1} (y(t) - h(\xi(t), u(t))) \\ \dfrac{dP}{dt}(t) & = & F(t)\, P(t) + P(t)\, F(t)^T + Q - P(t)\, H(t)^T R^{-1} H(t)\, P(t) \end{array} \right.$$

où on a noté $F(t) = \frac{df}{dx}(\xi(t), u(t))$ et $H(t) = \frac{dh}{dx}(\xi(t), u(t))$. Ici, Q et R sont des matrices symétriques définies positives qui n'ont plus d'interprétation stochastiques mais qui peuvent être interprétées comme des matrices de coût, en considérant le filtre de Kalman comme un problème d'optimisation quadratique. Ces équations sont celles d'un observateur. Il a été démontré (dans [26] et c'est aussi une conséquence de [3]) que cet observateur est *localement* exponentiellement convergent *i.e.* il converge exponentiellement pour x_0 assez proche de ξ_0 (dans le cadre stochastique, cf [45]). Ce résultat théorique est assez faible. Pratiquement, le filtre de Kalman étendu semble donner de très bons résultats et il est utilisé depuis longtemps par les ingénieurs en charge de la commande de procédés. Pourtant, le caractère purement local du filtre de Kalman étendu le rend impropre aux procédés très nonlinéaires et/où dont on ne connaît pas bien l'état initial. Un autre cas pour lequel le filtre de Kalman étendu ne fonctionnera pas correctement est le cas d'un système nonlinéaire soumis à de grandes perturbations non modélisées et non mesurées. Dans ce cas en effet, l'état réel du système s'éloigne de son estimation et un observateur local est insuffisant pour ramener l'estimation autour de l'état réel du système.

2.3. **Observateur grand-gain.** Il existe une grande classe d'observateurs nonlinéaires qui présentent d'excellentes propriétés globales : ce sont les observateurs grand-gain dont la référence historique est l'article de Gauthier-Hammouri-Othman [31] et pour lesquels la référence la plus complète est [32]. Contrairement au filtre de Kalman étendu, ces observateurs sont des observateurs globaux à convergence exponentielle. En revanche, il ne s'appliquent qu'à une classe restreinte de systèmes non linéaires. Afin de simplifier cette introduction, considérons les seuls systèmes à une observation ($q = 1$), la classe considérée est constituée des systèmes de la forme

$$(8) \qquad \left\{ \begin{array}{rcl} \frac{dx_1}{dt} & = & f_1(x_1, x_2, u) \\ \frac{dx_2}{dt} & = & f_2(x_1, x_2, x_3, u) \\ & \vdots & \\ \frac{dx_{n-1}}{dt} & = & f_{n-1}(x_1, x_2, \ldots, x_n, u) \\ \frac{dx_n}{dt} & = & f_n(x_1, x_2, \ldots, x_n, u) \\ y & = & h(x_1, u) \end{array} \right.$$

où pour tout $1 \leq i < n$, $\frac{\partial f_i}{\partial x_{i+1}}(x_1, \ldots, x_i, x_{i+1}, u) \neq 0$ (*i.e.* ne s'annulle jamais quels que soit les valeurs de x et de u).

Cette forme de système s'appelle la forme canonique d'observabilité. Une remarque triviale est que ces systèmes sont effectivement observables puisque la connaissance de y et de ses $n-1$ premières dérivées donne successivement x_1, x_2, etc jusqu'à x_n. Une remarque moins triviale est que beaucoup de systèmes observables (les systèmes uniformément infinitésimalement observables) peuvent se mettre sous cette forme. Les théorèmes qui caractérisent précisemment les systèmes pouvant se mettre sous forme canonique d'observabilité sont donnés dans le livre de Gauthier–Kupka [32].

Les observateurs grand-gain de type Luenberger s'écrivent sous la forme

$$(9) \qquad \frac{d\xi}{dt}(t) = f(\xi(t), u(t)) + K_\theta(y(t) - h(\xi(t), u(t)))$$

où $K_\theta = \Delta_\theta K$, Δ_θ est une matrice diagonale dont les coefficients diagonaux sont $(\Delta_\theta)_{j,j} = \theta^j$ où θ est un paramètre supposé grand et K est une matrice fixée, catactérisée par une équation de type Lyapunov. Alors

Theorem 1 ([32]). *Pour tout $a > 0$, il existe θ assez grand de sorte que pour tout $x(0) \in X$, pour tout $\xi(0) \in X$, il existe un polynôme de degré, k, tel que*

$$\|x(t) - \xi(t)\| \leq k(a) e^{-at} \|x(0) - \xi(0)\|$$

Ce théorème exprime le fait que les observateurs grand-gain de type Luenberger sont des observateurs exponentiellement convergents, et ce résultat est global si le système s'écrit globalement sous la forme (8). Malheureusement, ce résultat souffre de la nature de l'observateur sur lequel il s'appuie : l'observateur de Luenberger est un observateur de systèmes linéaires – à gain constant – qui perd toute son efficacité lorsque le système est non linéaire. Ici, la présence du paramètre θ garantit la convergence théorique mais au prix de gains si grands que les moindres erreurs (bruits de mesure, erreurs de modèle, perturbations) rendent le système instable.

Au cours de mon travail d'ingénieur de recherche au centre de recherche Shell de Grand-Couronne, nous avons testé à plusieurs reprises et sur divers procédés pétrochimiques l'observateur grand-gain. Pour cela, nous avons utilisé des modèles simples des procédés qui nous avons transformés par un changement de coordonnées pour les mettre sous la forme (8), le plus souvent étendue au cas de plusieurs sorties. Lors de nos tests et dans des conditions proches des conditions habituelles de fonctionnement des unités de productions modélisées, l'observateur de Luenberger n'a jamais fonctionné aussi bien que le filtre de Kalman étendu construit sur le modèle initial du procédé (sans être mis sous forme canonique d'observabilité). En revanche, nous avons noté que lors de perturbations assez importantes, l'observateur grand-gain avait toujours tendance à ramener l'état estimé vers la bonne région alors que le filtre de Kalman étendu ne semblait plus corriger l'innovation.

Suivant une démarche d'ingénieurs, nous avons donc testé le filtre de Kalman étendu sur le système dans sa forme (8) : cette approche s'est toujours révélée la plus efficace. D'un point de vue pratique, nous conservions l'efficacité connue par tous les ingénieurs du filtre de Kalman étendu. Du poins de vue théorique cependant, cette combinaison ne reposait sur aucun théorème. Même si le filtre de Kalman est un bon observateur local, on sait que ce n'est pas un onservateur global (nombreux contre-exemples, cf [45]). Cependant, la nature non intrinsèque du filtre de Kalman étendu peut nous laisser supposer que lorsqu'il est appliqué à un système mis sous une forme canonique d'observabilité alors il acquiert des propriétés de convergence exponentielle globales, à l'instar de l'observateur de Luenberger grand-gain. L'essentiel de la contribution du travail présenté ici est basé sur cette remarque.

2.4. **Synthèse : observateur de Kalman étendu grand-gain.** Nous allons maintenant brièvement décrire les résultats obtenus dans [8–12]. Bien que ces 5 articles constituent une partie importante du travail présenté ici, ils sont devenus quelque peu obsolètes à la suite des articles publiés ultérieurement et qui sont joint à ce texte.

Nous considérons toujours un système de la forme générale (1,2), avec une seule observation ($y \in \mathbb{R}$) mais nous allons supposer que ce système est mis sous la forme suivante

$$\begin{aligned} \frac{dx}{dt} &= A + a(x) + b(x)u \\ y &= Cx = x_1 \end{aligned} \tag{10}$$

où A est une matrice de la forme

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & & \cdots & 0 \end{pmatrix}$$

et que les fonctions a et b sont respectivement de la forme

$$\begin{aligned} a(x) &= (a_1(x_1), a_2(x_1, x_2), \ldots, a_n(x_1, x_2, \ldots, x_n)) \\ \text{et } b(x) &= (b_1(x_1), b_2(x_1, x_2), \ldots, b_n(x_1, x_2, \ldots, x_n)) \end{aligned}$$

On rappelle que cette forme de système est une forme canonique d'observabilité : tout système fortement observable ([31]) dont l'équation d'état est affine en l'entrée (*i.e.* s'écrit $\dot{x} = f(x) + g(x)u$) peut s'écrire localement sous cette forme. Le changement de variable qui permet cette transformation est

$$\Phi(x) = \left(h(x), L_f h(x), \ldots, L_f^{n-1} h(x)\right)$$

où $L_f h$ est la dérivée de Lie de h dans la direction f et $L_f^k h$ est l'itérée k-fois de cette dérivée de Lie. Afin de rendre le résultat global, nous supposons de plus que Φ est un difféomorphisme.

Dans le théorème suivant, nous allons supposer que plusieurs fonctions sont Lipschitz et que U_{adm} est compact. Notons que ces hypothèses très fortes sont en pratiques peu restrictives car l'espace d'état est usuellement borné. En conséquence, il suffit de prolonger les fonctions en dehors du domaine physique en veillant à ce que le prolongement soit Lipschitz. Cela est nécessaire en pratique car même si l'état reste dans le domaine physique, il n'en est pas nécessairement de même de l'état de l'observateur.

Soit $r > 0$. Soit Q symétrique définie positive. On pose $\Delta = \text{diag}(1, \frac{1}{\theta}, ..., (\frac{1}{\theta})^{n-1})$ et $Q_\theta = \theta^2 \Delta^{-1} Q \Delta^{-1}$. Pour toute fonction $f(x)$, on note $f^*(x)$ la matrice Jacobienne de $f(x)$.

Theorem 2 ([11,12]). *Supposons que a, b, Φ et Φ^{-1} soient globalement Lipschitz et que u reste borné. Soit θ un paramètre positif. Alors pour θ suffisamment grand,*

$$\begin{aligned}\frac{dz}{dt} &= Az + a(z) + b(z)u + PC'r^{-1}(y - Cz) \\ \frac{dP}{dt} &= Q_\theta + P(A + a^*(z) + b^*(z)u)' \\ &\quad + (A + a^*(z) + b^*(z)u)P - PC'r^{-1}CP\end{aligned}$$

est un observateur à convergence exponentielle.

Ce théorème est un premier résultat justifiant l'utilisation du filtre de Kalman étendu sur un système mis sous forme canonique d'observabilité. Son intérêt pratique est évident (et nous y reviendrons) : il allie le filtre de Kalman étendu à l'observateur grand-gain.

Nous avons ausi démontré la conséquence suivante. D'un point de vue pratique, les mesures auxquelles on peut accéder sur un procédé sont de natures discrètes : y n'est connue qu'aux instants d'échantillonnage que nons noterons $k\tau$, τ étant la période d'échantillonnage. Cette restriction nous interdit l'utilisation de l'observateur précédent. Considérons donc l'observateur suivant :

$$\begin{aligned}\frac{dz}{dt} &= Az + a(z) + b(z)u \\ \frac{dP}{dt} &= Q_\theta + P(A + a^*(z) + b^*(z)u)' + (A + a^*(z) + b^*(z)u)P\end{aligned}$$

et aux instants d'échantillonnage $t = k\tau$

$$\begin{aligned}z(t) &= z(t^-) + P(t)C'r^{-1}\tau\left(y_k - Cz(t^-)\right) \\ P(t) &= P(t^-) - P(t^-)C'\left(CP(t^-)C' + r\right)^{-1}CP(t^-)\end{aligned}$$

Ces équations sont celles du filtre de Kalman étendu dans sa version "continu/discret" ([42]) : l'évolution de l'état reste continue, les observations sont discrètes.

Theorem 3 ([11,12]). *Sous les mêmes hypothèses que dans le théorème précédent, pour τ assez petit, il existe un intervalle $[\theta_0, \theta_1]$ tel que pour tout $\theta \in [\theta_0, \theta_1]$, l'observateur continu-discret est exponentiellement convergent.*

Ce théorème est clairement ce que l'on pouvait attendre de mieux comme conséquence immédiate du théorème précédent. D'autres résultats ont été établis depuis sur ce même sujet, cf. [44].

Ces deux théorèmes, dans des versions plus générales, ont été appliqués avec succès à un dépropaniseur ([5,10]), un réacteur de polymérisation ([6]) et un réacteur catalytique à lit fluidisé ([8,19]). Dans les deux premiers cas, cet observateur a été couplé à un contrôleur, nous reviendrons à cet aspect important du travail dans le paragraphe 4 consacré au principe de séparation nonlinéaire.

De tous les modèles d'unités que nous avons testé à des fins d'estimation de l'état, le modèle le plus intéressant est celui de la colonne à distiller binaire, dont le dépropaniseur est un exemple. En effet, la structure "par plateaux" du procédé donne un modèle basé sur des bilans matière sur chaque plateaux qui est très proche de la forme canonique d'observabilité. Nous allons décrire ce modèle et montrer comment l'utiliser dans le chapitre suivant. Il est aussi largement commenté dans [3]. Le réacteur catalytique à lit fluidisé est aussi une unité très nonlinéaire sur lequel nous avons construit un filtre de Kalman étendu grand-gain, que nous allons maintenant décrire.

2.5. Observateur de Kalman étendu grand-gain, asymptotiquement petit gain ([3]).

2.5.1. *Introduction.* Dans un cadre déterministe, le filtre de Kalman étendu est un observateur non linéaire très utilisé en pratique par les ingénieurs. Il s'obtient en appliquant le classique filtre de Kalman linéaire (optimal) au système linéarisé autour de la trajectoire estimée. Cette construction empirique permet d'obtenir un observateur ayant de bonnes propriétés locales, *i.e.* lorsque l'erreur initiale est petite et lorsque le système non linéaire lui même est bien approché par le système linéarisé. De plus il a été montré depuis (cf. [32] par exemple) que sous des conditions d'observabilité uniforme, le système peut être mis sous une forme canonique d'observabilité et que sous cette forme, un observateur de type Kalman étendu grand gain peut être construit. Il est démontré que cet observateur converge arbitrairement vite vers l'état du système (voir [32] pour les hypothèses exactes).

Il n'existe pas de preuves de la convergence du filtre de Kalman étendu, mais en revanche de nombreux exemples de systèmes non linéaires ou le filtre de Kalman étendu ne converge pas (cf [45]) : cela est du au fait que si la trajectoire réelle du système est éloignée de la trajectoire estimée, les non linéarités peuvent être impossibles à compenser par un observateur à petit gain. Il parait donc naturel d'appliquer un observateur de type Kalman étendu grand gain lorsque l'on a très peu d'information sur l'état réel du système (par exemple au début de la trajectoire) puis d'appliquer le filtre de Kalman étendu classique lorsque les deux trajectoires sont suffisamment proches. Des constructions reposant sur cette remarque ont déjà été décrites par des ingénieurs.

Dans le paragraphe suivant, nous énoncerons le théorème qui établit la convergence de cet observateur. Puis nous expliquerons comment utiliser ce théorème pour construire un observateur robuste aux perturbations mais néanmoins proche du filtre de Kalman étendu classique en l'absence de grandes perturbations. Enfin, nous présenterons quelques résultats de simulation.

2.5.2. *Le théorème de convergence.* On considère un système non linéaire de la forme suivante :

$$\left\{\begin{array}{rcl} \dfrac{dx}{dt} & = & A(u)x + b(x,u) \\ y & = & C(u)x \end{array}\right. \tag{11}$$

où $x\left(t\right) \in \mathbb{R}^n$ et $u\left(t\right) \in \mathcal{U} \subset \mathbb{R}^d$, $x\left(0\right)$ étant fixé mais inconnu. Pour simplifier les notations, nous supposerons $y\left(t\right) \in \mathbb{R}$ mais ce que nous allons dire dans la suite peut être aisément adapté à des systèmes avec plusieurs sorties (ce sera d'ailleurs le cas de l'exemple donné en dernière section). On suppose que ce système est sous forme canonique d'observabilité, c'est à dire que les matrices $A\left(u\right)$ et $C\left(u\right)$ sont de la forme

$$C(u) = (a_1(u), 0, ..., 0),$$

$$A(u) = \begin{pmatrix} 0 & a_2\left(u\right) & 0 & \cdots & 0 \\ & & a_3\left(u\right) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ & & & & a_n\left(u\right) \\ 0 & & \cdots & & 0 \end{pmatrix}.$$

où les fonctions $u \mapsto a_i\left(u\right)$, $i = 1, \ldots, n$, sont supposées différentiables et bornées

$$0 < a_m \leq a_i(u) \leq a_M \qquad \forall u \in \mathcal{U}$$

et où $b(x,u)$ est un champ de vecteur à support compact, fonction différentiable de $u \in U$ et ayant la structure triangulaire en x suivante :

$$b\left(x,u\right) = b_1(x_1,u)\frac{\partial}{\partial x_1} + b_2(x_1,x_2,u)\frac{\partial}{\partial x_2} + ... + b_n(x_1,...,x_n,u)\frac{\partial}{\partial x_n}.$$

Cette forme canonique d'observabilité (ou sa généralisation pour des systèmes avec plusieurs sorties) apparaît naturellement sous des hypothèses d'observabilité uniforme, cf [32].

Nous allons considérer l'observateur suivant :

$$\left\{\begin{array}{rcl} \frac{dz}{dt} & = & A(u)z + b(z,u) - S(t)^{-1}C\left(u\right)' r^{-1}(C\left(u\right)z - y(t)) \\ \frac{dS}{dt} & = & -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) \\ & & \qquad +C\left(u\right)' r^{-1} C\left(u\right) - SQ_\theta S \\ \frac{d\theta}{dt} & = & \lambda(1-\theta) \end{array}\right. \tag{12}$$

où $\Delta = diag(1, \frac{1}{\theta}, ..., (\frac{1}{\theta})^{n-1})$ et $Q_\theta = \theta^2\Delta^{-1}Q\Delta^{-1}$ en notant $b^*(z,u)$ la matrice Jacobienne de $b(z,u)$ par rapport à z. Q et r sont les paramètres classiques (covariances des bruits) du filtre de Kalman étendu (Q est supposée symétrique définie positive et $r > 0$) alors que θ est la paramètre habituel de l'observateur grand gain, à ceci près qu'il est ici dépendant du temps et plus précisément décroissant exponentiellement à partir de $\theta\left(0\right) > 1$ jusqu'à 1 et à une vitesse dépendant de $\lambda > 0$ fixé. Alors :

Theorem 4. *Il existe $\lambda_0 > 0$ tel que pour tout $0 \leq \lambda \leq \lambda_0$, il existe θ_0 dépendant de λ tel que pour tout $\theta\left(0\right) > \theta_0$, pour tout $S\left(0\right) \geq c\ Id$, pour tout compact $K \subset \mathbb{R}^n$, pour tout $z\left(0\right) \in K$ alors si on pose $\varepsilon\left(t\right) = z\left(t\right) - x\left(t\right)$, on a pour tout $t \geq 0$*

$$||\varepsilon(t)||^2 \leq R(\lambda, c)e^{-at}\Lambda(\theta\left(0\right), t, \lambda)||\varepsilon\left(0\right)||^2$$

où

$$\Lambda(\theta\left(0\right), t, \lambda) = \theta\left(0\right)^{2(n-1)+\frac{a}{\lambda}} e^{-\frac{a}{\lambda}\theta(0)(1-e^{-\lambda t})}$$

et où a est un réel positif alors que $R(\lambda, c)$ est une fonction décroissante de c.

Il est important de noter que $\Lambda(\theta\left(0\right), t, \lambda)$ est une fonction décroissante de t qui peut être rendue arbitrairement petite à partir de n'importe quel temps $T_0 > 0$ en augmentant $\theta\left(0\right)$. Si $\lambda = 0$, nous retrouvons le filtre de Kalman étendu grand gain tel que présenté dans [32]. Pour la preuve de ce théorème dans le cas $\lambda > 0$, cf [3].

2.5.3. *Remarque importante.* La forme canonique d'observabilité décrite précédemment n'inclus pas les systèmes de la forme

$$\text{(13)} \qquad \left\{\begin{array}{rcll} \dot{x}_1 & = & F_1\left(x_1, x_2, u\right) & \frac{\partial F_1}{\partial x_2} \neq 0 \\ \dot{x}_2 & = & F_2\left(x_1, x_2, x_3, u\right) & \frac{\partial F_2}{\partial x_3} \neq 0 \\ & \vdots & & \\ \dot{x}_n & = & F_n\left(x, u\right) & \end{array}\right.$$

qui constituent pourtant une classe importante de systèmes observables non affines en u. En effet, Gauthier et Kupka montrent dans [32] que tout système infinitésimalement observable peut s'écrire sous cette forme. Nous ne savons pas si le filtre de Kalman étendu grand-gain converge lorsqu'il est appliqué directement à un tel système. Néanmoins, en s'inspirant de [33], et en supposant que u est une fonction C^1 de t, on peut effectuer le changement de variable suivant :

$$\text{(14)} \qquad \left\{\begin{array}{rcl} \xi_1 & = & y = x_1 \\ \xi_2 & = & F_1(x_1, x_2, u) \\ \xi_3 & = & \dfrac{\partial F_1}{\partial x_2} F_2\left(x_1, x_2, u\right) \\ & \vdots & \\ \xi_{i+1} & = & \dfrac{\partial F_1}{\partial x_2} \cdots \dfrac{\partial F_{i-1}}{\partial x_i} F_i\left(x_1, \ldots, x_{i+1}, u\right) \end{array}\right.$$

qui conduit au système

$$\left\{\begin{array}{rcl} \dot{\xi}_1 & = & \xi_2 \\ \dot{\xi}_2 & = & \xi_3 + \frac{\partial F_1}{\partial x_1}\dot{x}_1 + \frac{\partial F_1}{\partial u}\dot{u} \\ & \vdots & \\ \dot{\xi}_n & = & G\left(x, u, \dot{u}\right) \end{array}\right.$$

sur lequel nous pouvons appliquer un observateur grand gain de type Kalman étendu puisque ce système est bien sous la forme (11). Nous reviendrons à ce changement de variable dans la description d'une application de l'identification à un réacteur catalytique à lit fluidisé, cf paragraphe 3.4.

2.5.4. *Application à une colonne à distiller binaire.* Tel quel, l'observateur exponentiellement convergent que nous avons décrit dans la section précédente peut être utilisé pour estimer la trajectoire d'un système dont on ne connaît pas l'état initial mais dont on connaît parfaitement les équations d'évolution, et c'est bien le rôle d'un observateur. En pratique, on souhaite avoir un observateur capable de ré-estimer l'état du système même en cas de grande perturbation (non modélisée) venant modifier la trajectoire de l'état. Un moyen simple d'y parvenir est d'utiliser en parallèle plusieurs observateurs du type 12 que l'on notera $(\mathcal{O})$ en faisant en sorte qu'à chaque instant cohabitent au moins un observateur de type grand gain et un observateur de type Kalman étendu classique (i.e. un observateur du type précédent mais tel que le paramètre $\theta\left(t\right)$ soit assez proche de 1).

Remarquons que cette construction rejoint celle faite par certains spécialistes du filtrage non-linéaire lorsqu'ils cherchent à approcher les solutions de l'équation de Duncan-Mortensen-Zakaï par un filtre particulaire ([22,41,49]). Cependant, alors qu'ils doivent utiliser un grand nombre de copies (particules) pour estimer la loi conditionnelle, nous nous satisfaisons ici d'un nombre très restreint de filtres de Kalman étendu pour estimer uniquement $x\left(t\right)$.

On considère donc une famille d'observateurs $\left(\mathcal{O}_i\right)_{i=1,\ldots,N}$, tous de la même forme $(\mathcal{O})$ mais tels qu'à chaque instant de la forme $\left(i + \frac{j-1}{N}\right)T$, $j = 1, \ldots N$, $i \in \mathbb{N}$, le $j^{\text{ième}}$ observateur soit réinitialisé en $\left(z\left(0\right), S\left(0\right), \theta\left(0\right)\right)$. Ainsi, chaque observateur à une "durée de vie" de T et a tout instant, il existe un observateur qui a été réinitialisé depuis un intervalle de temps n'excédent pas $\frac{T}{N}$. En choisissant T et N, on peut donc être certain d'avoir a tout instant

- au moins un observateur de type Kalman étendu classique (T assez grand)
- au moins un observateur de type Kalman étendu grand gain (N assez grand)

et donc au moins un observateur adapté a la situation, cf Fig. 1.

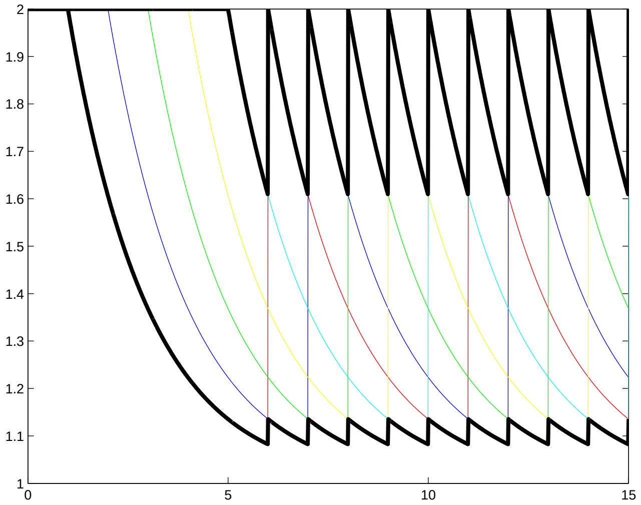


FIG. 1. Valeurs de θ et enveloppes supérieure (grand gain) et inférieure (petit gain)

Il reste à choisir le meilleur des N observateurs. Le critère choisi est assez naturellement celui qui minimise l'innovation, qui est donc celui qui prédit le mieux le comportement du système réel. Ce choix sera justifié en pratique dans l'exemple que nous allons décrire ci-dessous.

2.5.5. *Résultats de simulation.* Pour illustrer la construction précédente, nous allons considérer l'exemple d'un modèle de colonne à distiller à 5 plateaux théoriques[1] et dont les équations sont les suivantes :

$$\begin{aligned}
H_1\frac{dx_1}{dt} &= V(y_2 - x_1)\\
H_2\frac{dx_2}{dt} &= L(x_1 - x_2) + V(y_3 - y_2)\\
H_3\frac{dx_3}{dt} &= F(Z_F - x_3) + L(x_2 - x_3) + V(y_4 - y_3)\\
H_4\frac{dx_4}{dt} &= (L+F)(x_3 - x_4) + V(y_5 - y_4)\\
H_5\frac{dx_5}{dt} &= (L+F)(x_4 - x_5) + V(x_5 - y_5)
\end{aligned}$$

avec

$$y_i = k\left(x_i\right),\ i = 1, \ldots, 5$$
$$\text{où } k\left(x\right) = \frac{\alpha x}{1 + \left(\alpha - 1\right)x}$$

et où F est le débit d'alimentation de la colonne, L est le débit de reflux, V est le débit de rebouillage (L et V sont des variables de contrôle), les H_i sont les rétentions sur chaque plateau $1, \ldots, 5$ et où l'état est constitué des 5 variables $(x_i)_{i=1,\ldots,5}$ représentant les concentrations liquides de l'élément le plus léger du mélange binaire dont la volatilité relative est $\alpha > 1$ ainsi que de Z_F représentant la concentration de la charge, supposée inconnue. Une description justifiée et détaillée de ce modèle peut être trouvée dans [48].

On suppose que les variables observées sont les concentrations en tête et en fond de colonne qui sont respectivement x_1 et x_5. On suppose en outre les contrôles bornés

[1] les plateaux théoriques sont des plateaux physiques regroupés dans le modèle en une seule équation différentielle

$$0 < L_m \leq L(t) \leq L_M \text{ et } 0 < V_m \leq V(t) \leq V_M$$

Introduisons le changement de variable suivant : $\xi = (x_1, k(x_2), x_3, x_4, x_5, Z_F)$, alors le système peut être réécrit sous la forme canonique d'observabilité suivante :

$$\begin{aligned} \frac{d\xi_t}{dt} &= A(L,V)\,\xi_t + \widetilde{b}(L,V,\xi_t) \\ y &= C\xi_t \end{aligned}$$

avec

$$A(L,V) = \begin{pmatrix} 0 & \frac{V}{H_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{F}{H_3} \\ 0 & 0 & \frac{L+F}{H_4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{L+F}{H_5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

et

$$\widetilde{b}(L,V,\xi) = \begin{pmatrix} -\frac{V}{H_1}\xi_1 \\ k'\left(k^{-1}(\xi_2)\right)\left(L(\xi_1 - k^{-1}(\xi_2)) + V(k(\xi_3) - \xi_2)\right) \diagup H_2 \\ \left(-F\xi_3 + L(k^{-1}(\xi_2) - \xi_3) + V(k(\xi_4) - k(\xi_3))\right) \diagup H_3 \\ \left(-(L+F)\xi_4 + V(k(\xi_5) - k(\xi_4))\right) \diagup H_4 \\ \left(-(L+F)\xi_5 + V(\xi_5 - k(\xi_5))\right) \diagup H_5 \\ 0 \end{pmatrix}$$

et enfin

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Du fait que k est un difféomorphisme de $\left]-\frac{1}{\alpha-1}, +\infty\right[$ dans $\left]-\infty, \frac{\alpha}{\alpha-1}\right[$ et parce que le système laisse $[0,1]^6$ positivement invariant, on peut remplacer $\widetilde{b}(L,V,\xi)$ par $b(L,V,\xi) = \widetilde{b}(L,V,\Phi(\xi))$ où $\Phi(\xi_1,\ldots,\xi_6) = (\varphi(\xi_1),\ldots,\varphi(\xi_6))$ et $\varphi(\xi)$ est une fonction C^∞ de $\mathbb{R}$ dans $[0,1]$ qui vaut 1 dans $[0,1]$ et 0 en dehors de $\left]-\frac{1}{\alpha-\frac{1}{2}}, \frac{\alpha}{\alpha-\frac{1}{2}}\right[$, ceci sans modifier les trajectoires physiques du système. Cette modification nous permet d'être dans les hypothèses du théorème et donc d'appliquer l'observateur décrit dans la section précédente avec la garantie de la convergence. Notons cependant que nous sommes dans le cas à plusieurs sorties et donc qu'il est nécessaire d'adapter l'observateur à ce cas pour conserver le résultat de convergence, ce qui peut être fait facilement en adaptant la preuve de [3].

Finalement, chaque observateur s'écrit :

$$\left\{ \begin{array}{lcl} \frac{dz}{dt} & = & A(u)\,z + b(u,z) - S(t)^{-1} C^T R_\theta^{-1}(Cz - y(t)) \\ \frac{dS}{dt} & = & -(A(u) + b^*(z,u))' S - S(A(u) + b^*(z,u)) + C' R_\theta^{-1} C - S Q_\theta S \\ \frac{d\theta}{dt} & = & \lambda(1-\theta) \end{array} \right.$$

où $u = (L,V)$ .On a posé ici

$$\Delta = \mathrm{diag}\left(\frac{1}{\theta^2}, \frac{1}{\theta^3}, \frac{1}{\theta^2}, \frac{1}{\theta}, 1, \frac{1}{\theta^3}\right)$$

avec $Q_\theta = \theta^2 \Delta^{-1} Q \Delta^{-1}$ et $R_\theta = \left(C\Delta^{-1}C'\right) R \left(C\Delta^{-1}C'\right)$, Q et R étant les deux matrices symétriques définies positives du filtre de Kalman étendu classique.

Pour les simulations, nous n'avons pas cherché à calculer les constantes λ_0 et θ_0 mais nous avons choisi λ et θ de façon à assurer des performances correctes de l'observateur. Nous avons aussi choisi arbitrairement $N = 5$, $\theta(0) = 10$, $T = 50\,\mathrm{mn}$ et $\lambda = \frac{1}{600}\,\mathrm{s}^{-1}$ de sorte qu'à tout moment, il existe un observateur avec $\theta > 4.3$ et un autre avec $\theta < 1.2$. Afin d'assurer un comportement stationnaire de l'observateur au temps $t = 0$, nous avons initialisés les N observateurs avec $\theta_i = 1 + e^{-\lambda\frac{(i-1)T}{N}}(\theta(0) - 1)$ au début de la simulation.

Pour illustrer le comportement de l'observateur, nous avons simulé deux sortes de perturbations non prises en compte dans le modèle :

- au temps $t = 66\,\mathrm{mn}\,40\,\mathrm{s}$, nous avons simulé un brusque changement de la concentration de la charge qui passe de 0.45 à 0.60 puis,

- au temps $t = 116\,\text{mn}\,40\,\text{s}$, nous avons simulé une oscillation de la concentration de la charge d'amplitude 0.01.

Nous avons aussi introduit un bruit de mesure et nous avons simulé un contrôle rudimentaire permettant de réguler les deux sorties représentées avec leur bruit de mesure sur la Figure 2, courbes du haut et courbe du bas.

Les résultats obtenus sont tout a fait conformes à ce que la théorie nous prédisait. L'estimation de la qualité de la charge fournie par l'observateur ayant la plus petite innovation est montré sur la Figure 2 en tiretés, la courbe étant en général complètement superposée avec la courbe représentant la composition réelle de la charge.

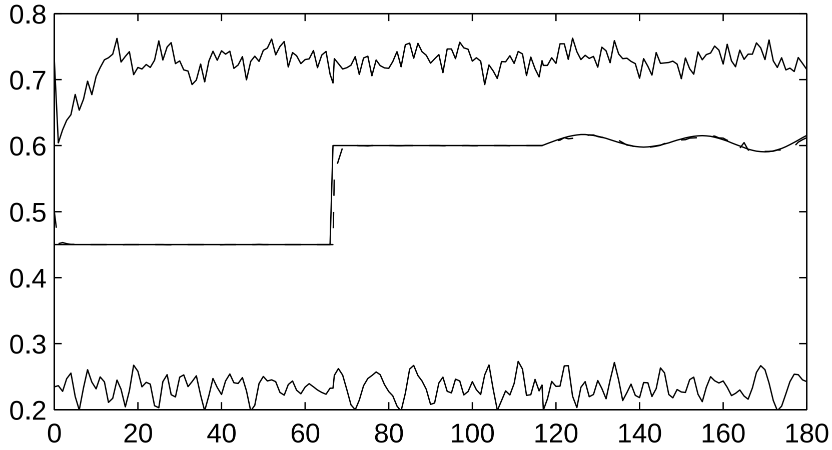


Fig. 2. Sorties mesurées (bruitées) et estimation de la charge.

Dans la figure suivante (Fig. 3), nous avons représenté l'évolution de θ pour les cinq observateurs en pointillé. On voit clairement les différentes réinitialisations et la durée de vie de chaque observateur. On a aussi représenté en trait plein la valeur de θ qui donne l'innovation la plus faible et donc qui fourni l'estimation de la Figure 2. Ainsi, en l'absence de perturbation et après compensation de l'erreur initiale, le meilleur observateur est le filtre de Kalman étendu. Lors de la première grosse perturbation au temps $t = 66\,\text{mn}\,40\,\text{s}$, l'observateur ayant le plus grand gain devient presque immédiatement le meilleur puis l'observateur reprend son cours normal. En présence de perturbations persistantes mais d'amplitude moyenne, un compromis est trouvé entre les plus grandes et les plus petites valeurs de θ.

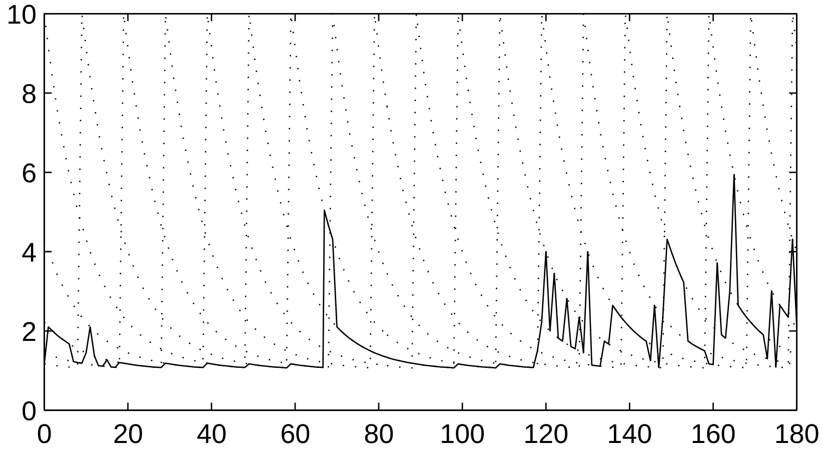


Fig. 3. Valeurs de θ en pointillé et celle correspondant au meilleur observateur en trait continu.

Pour conclure sur la partie simulation, il faut noter que

(1) Le choix des paramètres est plus simple que pour le filtre de Kalman étendu : en effet, ce dernier nécessite choisir Q et R de façon à trouver un compromis entre un comportement efficace en l'absence de perturbations et en leur présence. Ici, l'introduction de nouveaux paramètres permet de choisir Q et R de façon à avoir de bonnes performances en l'absence de perturbations, les autres paramètres garantissant la bonne performance lors des perturbations.

(2) L'observateur présenté nécessite l'intégration de nombreuses équations différentielles mais il est par nature très facilement parallélisable (les observateurs étant indépendant les uns des autres).

(3) Enfin, la multiplicité des observateurs fourni un outil de diagnostic performant : il suffit de regarder la Figure 3 pour comprendre à quels moments sont apparues les perturbations non mesurées.

3. Identifiabilité et identification ([1, 2])

3.1. **Introduction.** La problématique de l'observateur est l'estimation de variables d'état inconnues car non mesurées. Nous allons ajouter à ce problème celui de l'estimation de fonctions de l'état elles-même inconnues. Ce problème est en fait un problème d'identification en ce sens que nous allons supposer qu'une partie du modèle est inconnue.

Notre but est donc d'estimer à la fois l'état x du système et une fonction inconnue $\varphi : X \longrightarrow I$, I étant un intervalle compact de $\mathbb{R}$. Sans φ, le problème est un problème d'observation tel que nous l'avons exposé dans les paragraphes précédents de cette introduction. Avec φ à estimer uniquement, lorsque l'état x est supposé intégralement observé, ce problème est un problème d'inversion (estimation de l'entrée), cf [39, 40, 47]

3.2. **Principaux résultats.** On considère le système analytique suivant

$$\Sigma : \begin{cases} \frac{dx}{dt} &= f(x, \varphi \circ \pi(x(t))) \\ y &= h(x, \varphi \circ \pi(x(t))) \end{cases} \tag{15}$$

où la variable d'état $x = x(t)$ appartient à une variété analytique[2]X de dimension n, $x(0) = x_0$, la variable observée y est dans $\mathbb{R}^{d_y}$, et f, h sont respectivement un champ de vecteurs analytique paramétré et une fonction analytique. La fonction φ est une fonction inconnue d'une partie de l'état $\pi(x)$

$$\begin{array}{cccccc} \varphi : & X & \to & Z & \to & I \subset \mathbf{R} \\ & x & \to & z = \pi(x) & \to & \varphi(\pi(x)) \end{array}$$

Toutes les trajectoires solutions du système (15) seront supposées définies dans des intervalles $[0, T_{x_0,\varphi}[$.

Dans ([2]), nous donnons des définitions de l'identifiablilté pour des systèmes avec contrôle. Cependant, la majorité des résultats que nous avons obtenus concernent des systèmes sans contrôle, donc nous donnerons ici des définitions adaptées à (15).

Soit $\Omega = \{(x_0, \hat{\varphi}(\cdot)) \in X \times L^\infty[I]\,;\ \exists \varphi \in L^\infty[Z \to I]$ t.q.. $\hat{\varphi}(t) = \varphi(x(t))$ pour $x(t)$ trajectoire de $\Sigma_{x_0,\varphi}\}$. On défini l'application entrées/sorties par

$$\begin{array}{cccc} P_\Sigma : & \Omega & \longrightarrow & L^\infty\left[\mathbb{R}^{d_y}\right] \\ & (x_0, \hat{\varphi}(\cdot)) & \longrightarrow & y(\cdot) \end{array}$$

La définition naturelle de l'identifiabilité est la suivante :

Definition 5. *Σ est dit identifiable si P_Σ est injective.*

De même que pour l'observabilité, on défini une version infinitésimale de l'identifiabilité. Considérons la variation d'ordre 1 du système $\Sigma_{x_0,\varphi}$ (où $\hat{\varphi} = \varphi \circ x$) :

$$T\Sigma_{x_0,\hat{\varphi},\xi_0,\eta} \begin{cases} \dfrac{dx}{dt} &= f(x, \hat{\varphi}) \\ \dfrac{d\xi}{dt} &= T_x f(x, \hat{\varphi})\,\xi + d_\varphi f(x, \hat{\varphi})\,\eta \\ \hat{y} &= d_x h(x, \hat{\varphi})\,\xi + d_\varphi h(x, \hat{\varphi})\,\eta \end{cases} \tag{16}$$

[2] dans ce chapitre, une variété analytique sous-entend une variété analytique connexe, paracompacte et Hausdorf.

et l'application entrées/sorties de $T\Sigma$

$$\begin{array}{rccc} P_{T\Sigma,x_0,\hat{\varphi}}: & T_{x_0}X \times L^\infty[\mathbb{R}] & \longrightarrow & L^\infty\left[\mathbb{R}^{d_y}\right] \\ & (\xi_0, \eta(\cdot)) & \longrightarrow & \hat{y}(\cdot) \end{array}$$

Definition 6. *Σ est dit infinitésimallement identifiable si $P_{T\Sigma,,x_0,\hat{\varphi}}$ est injective pour tout $(x_0, \hat{\varphi}(\cdot)) \in \Omega$ i.e.* $\ker(P_{T\Sigma,x_0,\hat{\varphi}}) = \{0\}$ *pour tout $(x_0, \hat{\varphi}(\cdot))$.*

Enfin et toujours en analogie avec l'observabilité, nous allons maintenant définir l'identifiablilité différentielle.

Soit $D_k\Phi = X \times (U \times \mathbb{R}^{(k-1)d_u}) \times (I \times \mathbb{R}^{k-1})$ l'espace des jets d'ordre k du système Σ.

On notera $j^k(u) = \left(u(0), u'(0), \ldots, u^{(k-1)}(0)\right)$, on pose

$$\begin{array}{rccc} \Phi_k^\Sigma: & D_k\Phi & \rightarrow & \mathbb{R}^{kd_y} \\ & (x_0, j^k(u), j^k(\hat{\varphi})) & \rightarrow & j^k(y) \end{array}$$

$$\begin{array}{rccc} \Phi_{k,2}^{\Sigma,*}: & D_k\Phi \times D_k\Phi & \rightarrow & \mathbb{R}^{kd_y} \times \mathbb{R}^{kd_y} \\ & (z_1, z_2) & \rightarrow & (\Phi_k^\Sigma(z_1), \Phi_k^\Sigma(z_2)) \end{array}$$

Definition 7. *Σ est différentiellement identifiable d'ordre k si et seulement si $\Phi_{k,2}^{\Sigma,*}(z_1, z_2) \in \Delta_k \Rightarrow (x_1, \hat{\varphi}_1(0)) = (x_2, \hat{\varphi}_2(0))$*

La proposition suivante est claire :

Proposition 8. *L'identifiabilité différentielle entraine l'identifiabilité*

Le résultat suivant est beaucoup moins évident et il est actuellement en cours de rédaction : il assure l'équivalence entre "identifiabilité" et "identifiabilité pour des fonctions C^∞" :

Theorem 9. *Si Σ est infinitésimallement identifiable dans la classe des fonctions analytiques alors il est infinitésimallement identifiable dans la classe des fonctions L^∞.*

Plus explicitement, si Σ n'est pas infinitésimallement identifiable en raison de l'existence de $(x_0, \hat{\varphi}(\cdot)) \in \Omega$ tel que $(\xi_0, \eta) \in \ker(P_{T\Sigma,x_0,\hat{\varphi}})$, $(\xi_0, \eta) \neq 0$, alors il existe $\left(\tilde{x}_0, \widetilde{\hat{\varphi}}(\cdot)\right) \in \Omega$ où φ (dans $\widetilde{\hat{\varphi}} = \varphi(x)$) est analytique et $\left(\tilde{\xi}_0, \tilde{\eta}\right) \in \ker(P_{T\Sigma,x_0,\hat{\varphi}})$, $\left(\tilde{\xi}_0, \tilde{\eta}\right) \neq 0$, avec $\tilde{\eta}$ analytique.

La preuve de ce lemme est basée sur le lemme suivant qui est une adaptation d'un lemme de [32] et de certains résultats de [38] :

Lemma 10. *Soit $\dfrac{dx}{dt} = f(x, u)$ un système analytique $\bar{\Sigma}$ défini sur $X \times U$ où X est une variété analytique et U un compact sous-analytique de $\mathbb{R}^d$. Soit S un sous-ensemble sous-analytique de $X \times U$. Soit $(x(t), u(t))_{t\in[0,T[}$ une trajectoire L^∞ de $\bar{\Sigma}$ telle que $E = \{t \in [0, T[;\quad (x(t), u(t)) \in S\}$ soit de mesure de Lebesgue non nulle. Alors il existe une trajectoire analytique $(\tilde{x}(t), \tilde{u}(t))_{t\in[0,\tilde{T}[}$ de $\bar{\Sigma}$ telle que $(\tilde{x}(t), \tilde{u}(t))$ soit dans S pour tout $t \in \left[0, \tilde{T}\right[$.*

Nous avons aussi le résultat suivant

Theorem 11. *Si Σ est identifiable dans la classe des fonctions analytiques alors il est identifiable dans la classe des fonctions L^∞.*

Ces deux résultats justifient nos définitions de l'identifiabilité qui pourraient paraître trop restrictives sans ces théorèmes.

L'un des résultats principaux concernant l'identifiabilité est le suivant :

Theorem 12. *Si $d_y \geq 3$, l'identifiabilité differentielle d'ordre $2n + 1$ **est générique** dans la classe des systèmes C^∞.*

Idée de la preuve (nous ne traitons que le cas difficile, parmi les cas qui apparaissent) : Posons $Z_i = \left(x_i, \varphi_i, \varphi_i', \ldots, \varphi_i^k, j_\Sigma^k(x_i, \varphi_i)\right)$, $i = 1, 2$, $Z = (Z_1, Z_2)$, $\Phi(Z) = \Phi_k^\Sigma(Z_1) - \Phi_k^\Sigma(Z_2) \in R^{k\,d_y}$, où $k = 2n + 1$. Admettons que Φ soit une submersion donc $\text{codim}\Phi^{-1}(0) = k\,d_y$

Soit $\Pi\Phi^{-1}(0) = \left\{\left(x_i, \varphi_i, j_\Sigma^k(x_i, \varphi_i)\right)_{i=1,2}\right\}$:

$$\begin{aligned} \text{codim}\Pi\Phi^{-1}(0) &\geq k\,d_y - 2\,(k-1) = k\,(d_y - 2) + 2 \\ &\geq k + 2 \geq 2n + 3 \end{aligned}$$

Définissons

$$\begin{array}{cccc} \rho_\Sigma : & (X \times I)^2 \setminus \Delta & \rightarrow & \left(J_\Sigma^k\right)^2 \\ & (x_1, \varphi_1, x_2, \varphi_2) & \rightarrow & \left(x_i, \varphi_i, j_\Sigma^k(x_i, \varphi_i)\right)_{i=1,2} \end{array}$$

Le théorème de transversalité multijet nous affirme que l'ensemble des Σ tels que ρ_Σ est transverse à $\Pi\Phi^{-1}(0)$ est résiduel. Mais puisque $\dim (X \times I)^2 \setminus \Delta = 2n + 2$, alors il signifie que génériquement, ρ_Σ ne rencontre pas $\Pi\Phi^{-1}(0)$

Theorem 13. *Si $d_y < 3$, l'identifiabilité differentielle* ***n'est pas*** *générique.*

Si d_y est égale à 1 où 2, l'identifiabilité est donc une hypothèse très restrictive (de codimension infinie). Dans [2], nous décrivons la classification complète des systèmes infinitésimallement identifiables. Nous donnons pour cela des propriétés géométriques intrinsèques, qui sont localement équivalentes aux formes normales que nous présentons dans les Théorèmes 14 et 19 ci-dessous.

Theorem 14 ([2])**.** *($d_y = 1$) Si Σ est uniformément infinitésimallement identifiable alors il existe un sous-ensemble sous-analytique fermé Z de X, de codimension 1 au moins, tel que pour tout $x_0 \in X \backslash Z$, il existe un système de coordonnées $(x_1, \ldots, x_n,\ V_{x_0})$, $V_{x_0} \subset X \backslash Z$ dans lequel Σ (restreint à V_{x_0}) peut s'écrire :*

$$(17) \qquad \Sigma_1 \left\{ \begin{array}{lcl} \dot{x}_1 & = & x_2 \\ & \vdots & \\ \dot{x}_{n-1} & = & x_n \\ \dot{x}_n & = & \psi(x, \varphi) \\ y & = & x_1 \end{array} \right. \quad \textit{et } \frac{\partial}{\partial \varphi}\psi(x, \varphi) \neq 0$$

Le théorème précédent admet la pseudo-réciproque suivante

Theorem 15. *Si Σ satisfait les conditions précédentes,*

- $\frac{\partial}{\partial \varphi}\left\{h, L_{f_\varphi}h, \ldots, (L_{f_\varphi})^{n-1}h\right\} \equiv 0$
- $\frac{\partial}{\partial \varphi}L_{f_\varphi}^n h \neq 0$
- $d_x h \wedge \ldots \wedge d_x L_{f_\varphi}^{n-1} h \neq 0$,

alors Σ est localement identifiable, localement uniformément infinitésimalement identifiable, et localement différentiellement identifiable d'ordre $n + 1$.

Idée de la preuve : Soit $k < n$ le plus petit k tel que $d_\varphi L_f^k h \not\equiv 0$:

$$\Sigma \quad \left\{ \begin{array}{lcl} y & = & x_1 \\ \dot{x}_1 & = & x_2 \cdots \\ \dot{x}_{k-1} & = & x_k \\ \dot{x}_k & = & L_f^k(x, \varphi) = f_k(x, \varphi) \cdots \\ \dot{x}_n & = & f_n(x, \varphi) \end{array} \right.$$

$$T\Sigma \quad \left\{ \begin{array}{lcl} \dot{x} & = & f(x, \varphi) \\ \widehat{y} & = & \xi_1 \\ \dot{\xi}_1 & = & \xi_2 \cdots \\ \dot{\xi}_{k-1} & = & \xi_k \\ \dot{\xi}_k & = & d_x f_k(x, \varphi)\,\xi + d_\varphi f_k(x, \varphi)\,\eta \end{array} \right.$$

Le retour d'état $\eta = -\dfrac{d_x f_k\left(x,\varphi_0\right)\xi}{d_\varphi f_k\left(x,\varphi_0\right)}$ en un φ_0 tel que $d_\varphi f_k\left(x,\varphi_0\right) \neq 0$ donne $\frac{d\xi_k}{dt} = 0$ donc η rend le système inobservable.

Supposons maintenant $\frac{\partial}{\partial\varphi} L^n_{f_\varphi} h = 0$ en (x,φ)

$$\begin{array}{rcl} X \times I & \supset & E = \left\{(x,\varphi)\,;\, d_\varphi L^n_f h = 0\right\} \\ & & \downarrow \Pi \\ X & \supset & \Pi E \end{array}$$

Le théorème de Hardt ([34]) assure l'existence de $\widehat{\varphi}$ tel que

$$\left\{\begin{array}{lllllllll} y & = & x_1, & \dot{x}_1 & = & x_2, & \ldots & \dot{x}_n & = \psi\left(x,\widehat{\varphi}\left(x\right)\right) \\ \widehat{y} & = & \xi_1, & \dot{\xi}_1 & = & \xi_2, & \ldots & \dot{\xi}_n & = d_x\psi\left(x,\widehat{\varphi}\left(x\right)\right) + 0 \end{array}\right.$$

Avant de donner le théorème concernant les systèmes avec deux sorties, donnons quelques définitions supplémentaires :

Posons $E_l = \left\{d_x h_i, d_x L_{f_\varphi} h_i, \ldots, d_x L^{l-1}_{f_\varphi} h_i\ , \, i = 1,2\right\}$ et $N\left(l\right) = \mathrm{rang}\left(E_l\right)$ en un point générique :

On défini k et m par $N\left(0\right) = 0$, $N\left(1\right) = 2, \ldots\ N\left(k\right) = 2k$, $N\left(k+1\right) = 2k+1, \ldots\ N\left(k+m\right) = 2k+m$, puis $N\left(j\right) = 2k+m$ pour $j \geq k+m$. k est donc l'indice où se produit une chute du rang de E_k.

Definition 16. *On appelle ordre du système le plus petit entier r tel que $d_\varphi L^r_{f_\varphi}\left(h_1,h_2\right) \not\equiv 0$.*

Lemma 17. *Si Σ est uniformément infinitésimalement identifiable alors (1) $2k+m = n$*
(2) $r \leq k+m$

Démonstration. (1) $\varphi = \varphi_0$ =cte $\left\{\begin{array}{lll} \dot{x} & = & f\left(x,\varphi_0\right) \\ \dot{\xi} & = & g\left(x,\xi,\varphi_0\right) \\ y & = & h\left(x,\varphi_0\right) \end{array}\right.$ contredit l'observabilité (2) $\left\{\begin{array}{lll} \dot{x} & = & f\left(x\right) \\ y & = & h\left(x\right) \end{array}\right.$ contredit l'identifiabilité □

Definition 18. *Un système qui vérifie (1) et (2) est dit régulier.*

Theorem 19 ([2]). *($d_y = 2$) Si Σ est uniformément infinitésimallement identifiable, alors il existe un semi-analytique ouvert dense $\tilde{U} \subset X \times I$, tel que tout point $\left(x_0,\varphi_0\right)$ de $\tilde{U}$, possède un voisinage $V_{x_0} \times I_{\varphi_0}$,et des coordonnées x dans V_{x_0} telles que le système Σ restreint à $V_{x_0} \times I_{\varphi_0}$, noté $\Sigma_{|V_{x_0}\times I_{\varphi_0}}$, admette l'une des formes normales suivantes :*

– ***forme normale de type 1 :*** *($r > k$)*

$$(18) \qquad \Sigma_{2,1}\left\{\begin{array}{llllll} y_1 & = & x_1 & y_2 & = & x_2 \\ \dot{x}_1 & = & x_3 & \dot{x}_2 & = & x_4 \\ & \vdots & & & \vdots & \\ \dot{x}_{2k-3} & = & x_{2k-1} & \dot{x}_{2k-2} & = & x_{2k} \\ \dot{x}_{2k-1} & = & f_{2k-1}(x_1,...,x_{2k+1}) & & & \\ \dot{x}_{2k} & = & x_{2k+1} & & & \\ & \vdots & & & & \\ \dot{x}_{n-1} & = & x_n & & & \\ \dot{x}_n & = & f_n(x,\varphi) & & & \end{array}\right.$$

avec $\frac{\partial f_n}{\partial\varphi} \neq 0$.

– ***forme normale de type 2 :*** *($r < k$)*

$$(19) \qquad \Sigma_{2,2}\left\{\begin{array}{llllll} y_1 & = & x_1 & y_2 & = & x_2 \\ \dot{x}_1 & = & x_3 & \dot{x}_2 & = & x_4 \\ & \vdots & & & \vdots & \\ \dot{x}_{2r-3} & = & x_{2r-1} & \dot{x}_{2r-2} & = & x_{2r} \\ \dot{x}_{2r-1} & = & \psi(x,\varphi) & \dot{x}_{2r} & = & F_{2r}(x_1,\ldots,x_{2r+1},\psi(x,\varphi)) \\ & & & \dot{x}_{2r+1} & = & F_{2r+1}(x_1,\ldots,x_{2r+2},\psi(x,\varphi)) \\ & & & & \vdots & \\ & & & \dot{x}_{n-1} & = & F_{n-1}(x,\psi(x,\varphi)) \\ & & & \dot{x}_n & = & F_n(x,\varphi) \end{array}\right.$$

avec $\frac{\partial\psi}{\partial\varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0, ...,\ \frac{\partial F_{n-1}}{\partial x_n} \neq 0$

– ***forme normale de type 3 :*** *(*$r = k$ *et* $n = 2k$*)*

$$(20) \qquad \Sigma_{2,3} \left\{ \begin{array}{lllllll} y_1 & = & x_1 & & y_2 & = & x_2 \\ \dot{x}_1 & = & x_3 & & \dot{x}_2 & = & x_4 \\ & \vdots & & & & \vdots & \\ \dot{x}_{n-3} & = & x_{n-1} & & \dot{x}_{n-2} & = & x_n \\ \dot{x}_{n-1} & = & f_{n-1}(x,\varphi) & & \dot{x}_n & = & f_n(x,\varphi) \end{array} \right.$$

avec $\frac{\partial}{\partial\varphi}(f_{n-1}, f_n) \neq 0$

Si $r = k$ *et* $2k < n$ *i.e. si la chute de rang coïncide avec l'ordre et ceci avant la dernière dérivation, alors on a* $d_x h_1 \wedge \cdots \wedge d_x L_{f_\varphi}^{k-1} h_1 \wedge d_x L_{f_\varphi}^{k-1} h_2 \wedge d_x L_{f_\varphi}^{k} h_2 \not\equiv 0$. *Si* $d_\varphi L_{f_\varphi}^{k} h_1 \neq 0$, *on obtient* φ *par* y_1 *et* $x_{2k}, \ldots, x_n$ *par* y_2, *et le système se met sous forme normale de type 2. Si* $d_\varphi L_{f_\varphi}^{k} h_1 \equiv 0$, *on trouve* φ *par* y_2 *et le système se met sous forme normale de type 1.*

Dans ([2]), nous n'avons pas donné de forme normale d'identifiabilité pour les systèmes identifiables pour lesquels $d_y \geq 3$. Nous pouvons néanmoins donner une forme normale d'identifiablilté pour un système identifiable générique en un point générique.

Considérons un système analytique générique $\Sigma_{x_0,\varphi}$ avec au moins 3 observations ($d_y \geq 3$) en un point générique x_0 tel que $\left.\frac{\partial h}{\partial\varphi}\right|_{x=x_0} \neq 0$. On peut écrire le système sous la forme

$$\check{\Sigma} \left\{ \begin{array}{lll} \frac{dx}{dt} & = & f(x,\varphi) \\ y & = & h(x,\varphi) \\ \check{y} & = & \check{h}(x,\varphi) \end{array} \right.$$

où $y \in \mathbb{R}^{d_y - 1}$ et $\check{y} \in \mathbb{R}$ et tel que $\left.\frac{\partial \check{h}}{\partial\varphi}\right|_{x=x_0} \neq 0$. Le théorème des fonctions implicites nous permet d'écrire $\varphi = \Phi(x, \check{y})$ de sorte que le système $\check{\Sigma}$ s'écrive

$$\Sigma \left\{ \begin{array}{lll} \frac{dx}{dt} & = & f(x,\check{y}) \\ y & = & h(x,\check{y}) \end{array} \right.$$

où $\check{y}$ est considéré comme une entrée. On peut donc appliquer la théorie de Gauthier–Kupka ([32]) pour les systèmes analytiques comportant plus de sorties que d'entrées (puisque $d_y - 1 \geq 2$) : on peut donc en déduire une forme normale pour les systèmes génériques identifiables aux pouts génériques ($\frac{\partial h}{\partial\varphi} \neq 0$) basée sur la représentation dans l'espace des phases des systèmes observables. On peut en fait montrer le théorème suivant (la densité des systèmes est définie dans [32]) :

Theorem 20. *(*$d_y \geq 3$*) Si* Σ *est un système infinitésimallement observable générique, alors il existe un sous-espace fermé sous-analytique* Z *de* X, *de codimension* 2 *au moins, tel que pour tout* $x_0 \in X \backslash Z$, *il existe une fonction* C^∞ F *et un plongement* $\Psi_{\check{y},\ldots,\check{y}^{(2n)}}(x)$, *dépendant de* $\left(\check{y}, \check{y}', \ldots, \check{y}^{(2n)}\right)$ *tel que en dehors de* Z, *les trajectoires de* $\Sigma_{x_0,\varphi}$ *soient envoyées par* $\Psi_{\check{y},\ldots,\check{y}^{(2n)}}$ *sur les trajectoires du système suivant :*

$$\Sigma_{3+} \left\{ \begin{array}{lll} \frac{dz_1}{dt} & = & z_2, \quad \frac{dz_2}{dt} = z_3, \quad \ldots \quad, \frac{dz_{2n}}{dt} = z_{2n+1} \\ \frac{dz_{2n+1}}{dt} & = & F\left(z_1, z_2, \ldots, z_{2n+1}, \check{y}, \check{y}', \ldots, \check{y}^{(2n+1)}\right) \\ \bar{y} & = & z_1 \end{array} \right.$$

où z_i, $i = 1, \ldots, 2n+1$ *est un vecteur de dimension* $d_y - 1$, *et où*

$$(2) \qquad \left\{ \begin{array}{lll} x & = & \Psi^{-1}_{\check{y},\ldots,\check{y}^{(2n)}}(z) \\ \varphi & = & \Phi(x, \check{y}) \end{array} \right.$$

Pratiquement, plusieurs constructions explicites d'observateurs grand-gain peuvent être appliquées à Σ_{3+} afin d'estimer $z(t)$ en observant y, $\check{y}$ et ses dérivées ([2,32]). Ensuite, x et φ sont estimés grâce à (2). Dans [2], nous montrons comment appliquer ce type d'observateurs pour des systèmes non-génériques de la forme Σ_1 et $\Sigma_{2,i}$.

3.3. **Application à un réacteur biologique.** A titre d'exemple, nous avons étudié le cas d'un reacteur biologique (en réalité, c'est l'étude de cet exemple qui a motivé l'étude de l'identifiabilité). Le modèle de bio-réacteur suivant est très classique ([27])

$$\left\{\begin{array}{rcl}\dfrac{ds\left(t\right)}{dt} & = & -\mu\left(s\left(t\right)\right)x\left(t\right)+D(t)(S_{in}-s\left(t\right)) \\ \dfrac{dx\left(t\right)}{dt} & = & (\mu\left(s\left(t\right)\right)-D(t))x\left(t\right)\end{array}\right.$$

où les variables désignent

$$\begin{array}{lcl} s(t) & : & \text{substrat} \\ x(t) & : & \text{biomasse} \\ D(t) & : & \text{débit} \\ S_{in} & : & \text{substrat en entrée}\end{array}$$

Dans ce modèle la fonction μ représente la fonction de croissance. C'est une fonction positive vérifiant $\mu\left(0\right)=0$. Typiquement, elle peut avoir les formes suivantes :

- loi de Monod $\mu(s)=\frac{\mu_0 s}{k_m+s}$
- loi de Haldane $\mu(s)=\frac{\mu_0 s}{k_m+s+\frac{s^2}{k_i}}$

mais en fait un très grand nombre d'autres formes possibles. Cette fonction est en général très mal connue, c'est cette dernière que nous chercherons à identifier. Nous supposerons que nous mesurons $s\left(t\right)$.

Le système n'est pas identifiable. Posons en effet

$$\begin{array}{rcl} X & = & x+s \\ \widetilde{D}\left(t\right) & = & \int_0^t D\left(\tau\right)d\tau \\ \Lambda\left(t\right) & = & e^{D(t)}(s-S_{in})+S_{in}\end{array}$$

alors

$$\dot{\Lambda}=-e^{\tilde{D}(t)}(X-s)\mu\left(s\right)=(\Lambda-X_0)\mu\left(s\right)$$

avec $\Lambda\left(0\right)=s\left(0\right)$. Si $s\left(t_0\right)=s\left(t_1\right)$, $t_0<t_1$ alors

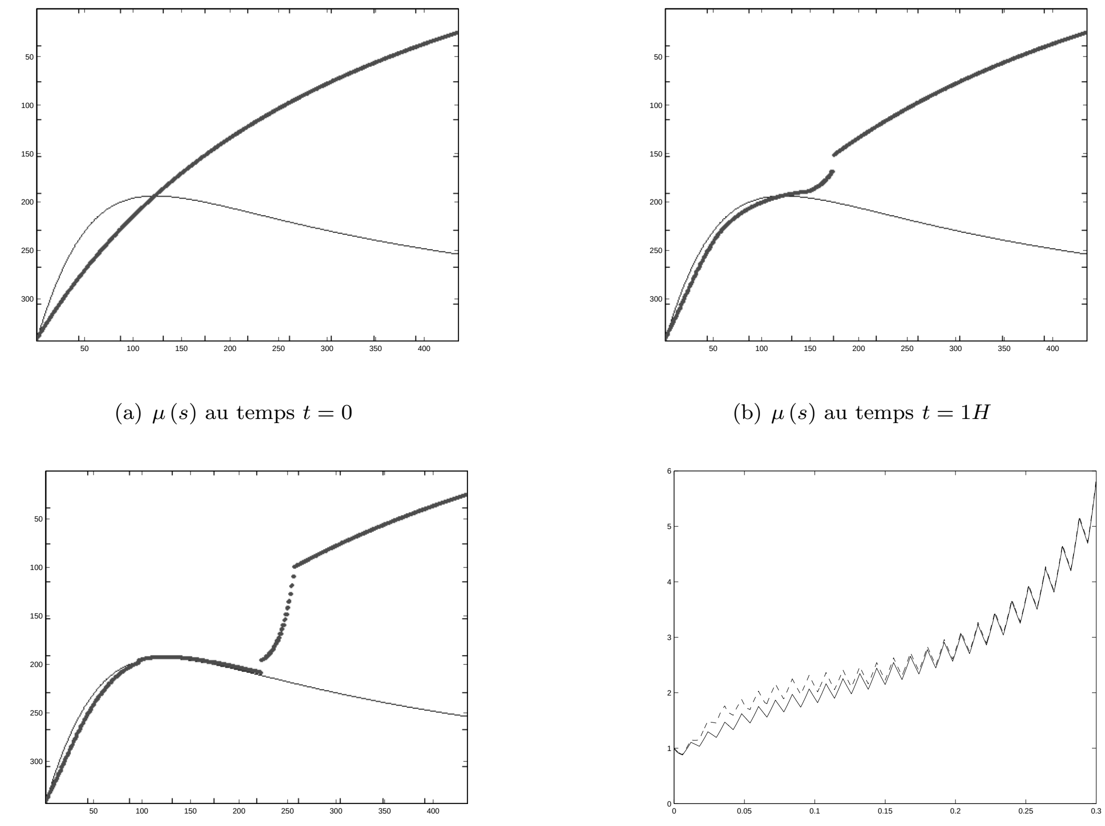
$$\frac{\dot{\Lambda}(t_0)}{\Lambda(t_0)-X_0}=\mu(s(t_0))=\mu(s(t_1))=\frac{\dot{\Lambda}(t_1)}{\Lambda(t_1)-X_0}$$

donne X_0 donc $\mu(s(t))=\frac{\dot{\Lambda}(t)}{\Lambda(t)-X_0}$. Ainsi, $\mu\left(s\right)$ est identifiable si et seulement si $s\left(t\right)$ visite deux fois une valeur identique.

Afin d'illustrer les résultats précédents, nous avons effectué une simulation sous Matlab/Simulink. La Figure 4 montre les résultats numériques de la simulation telle qu'elle est expliquée dans [2]. Le premier tracé 4(a) montre la fonction inconnue $\mu\left(s\right)$ qui est de type Haldane (elle n'est pas monotone) et l'estimation initiale $\widehat{\mu}_0\left(s\right)$ que l'on a choisi de type Monod. Les deux figures suivantes 4(b) et 4(c) montrent l'estimation de $\mu\left(s\right)$ à différents moments de la simulation. Ces estimations de μ servent à estimer la variable d'état inconnue $x\left(t\right)$. Cette estimation est représentée sur le dernier tracé de la Figure 4(d). Nous avons simulé un contrôle très oscillant afin de rendre le système identifiable. Dans la première moitié de la simulation, $x\left(t\right)$ est estimé avec un biais du à l'erreur de modélisation de la fonction de croissance. Au cours de la simulation, la fonction de croissance est identifiée et donc le biais disparaît lorsque l'état retourne en des points où μ a été correctement identifiée.

3.4. **Application à un réacteur catalytique à lit fluidisé ([1]).** Le reacteur catalytique à lit fluidisé (FCC) est une unité de traitement cruciale dans une raffinerie car elle apporte une très forte valeur ajoutée à ses produits. C'est un réacteur qui craque des molécules d'hydrocarbone en molécules plus petites par l'action conjuguée de la chaleur et d'un catalyseur. A la suite de cette réaction endothermique, le catalyseur est envoyé dans un régénérateur où le coke qui s'est formé sur le catalyseur est brulé en présence d'air que l'on insuffle dans le régénérateur (c'est un des contrôles). Cette réaction est exothermique, le catalyseur est ensuite renvoyé dans le réacteur où la réaction continue. Il y a donc recirculation du catalyseur et de l'énergie. Cette unité est très nonlinéaire et peut-être instable hors de sa plage de fonctionnement nominale.

Le modèle du FCC est basé sur des bilans matière (catalyseur) et énergétique au sein du réacteur et du régénérateur. La partie la moins connue du modèle est la combustion du coke déposé sur le catalyseur dans le régénérateur. Nous utilisons donc l'approche que nous venons de décrire pour identifier cette partie du modèle.



(a) $\mu(s)$ au temps $t = 0$

(b) $\mu(s)$ au temps $t = 1H$

(c) $\mu(s)$ au temps $t = 2H$

(d) $x(t)$

FIG. 4. Simulation d'un bioréacteur

Le modèle ne se met pas directement sous la forme canonique d'observabilité des systèmes affines en le contrôle (11) sur laquelle on peut directement appliquer un filtre de Kalman étendu grand-gain. En effet, la forme canonique d'identifiabilité du FCC est celle de type 2 donnée en (19). Mais cette forme est un cas particulier de (13) donc, quitte à dériver une entrée, on peut effectivement mettre le modèle du FCC sous une forme propice à l'application du filtre de Kalman étendu grand-gain. Le modèle est un peu trop compliqué pour être décrit ici mais on peut le mettre sous la forme suivante :

$$
\begin{array}{rclrcl}
y_1 & = & x_1 & y_2 & = & x_5 \\
\dot{x}_1 & = & a_2 x_2 + g_1(x_1, x_5) & \dot{x}_5 & = & \alpha_1 x_5 + \alpha_2 x_6 + \alpha_3 y_1 + \beta \\
\dot{x}_2 & = & a_3 x_3 + g_2(x_1, x_2, x_5, x_6) & \dot{x}_6 & = & x_7 \\
\dot{x}_3 & = & a_4 x_4 + g_3(x_1, x_2, x_3, x_5) & \dot{x}_7 & = & x_8 \\
\dot{x}_4 & = & g_4(x) & \dot{x}_8 & = & 0
\end{array}
\tag{21}
$$

où les fonctions g_1, g_2, g_3, g_4, α_1, α_2, α_3 et β dépendent aussi du contrôle u et de sa dérivée $\dot{u}$.

Nous avons supposé que la fonction à identifier satisfaisait localement à un modèle du second ordre $\varphi = at^2 + bt + c$ qui correspond dans (21) à ($x_6 = \varphi$), $\dot{x}_6 = x_7$, $\dot{x}_7 = x_8$, $\dot{x}_8 = 0$. Le reste du modèle provient d'un changement de variable du type (14)

Le résultat est consititué de deux systèmes ayant presque une structure triangulaire : le système à droite ne dépend pas du système à gauche excepté par une sortie mesurée. On applique donc un filtre de Kalman linéaire au système de droite, qui nous donne une estimation de la fonction inconnue φ, quel l'on utilise ensuite dans le système de gauche sur lequel on applique un filtre de Kalman étendu grand gain asymptotiquement petit gain. Le résultat est brièvement montré sur les Figures 5.

Cette application est donc une synthèse des résultats récents en terme d'observateurs et d'identification.

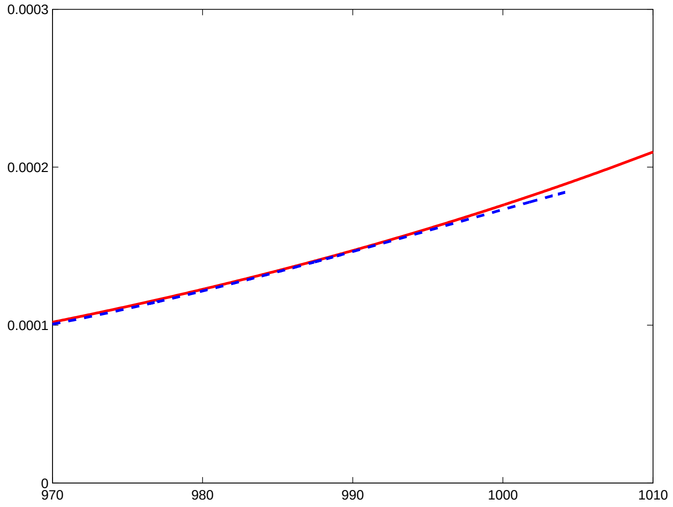


Fig. 5. Fonction inconnue (trait plein) et fonction estimée (pointillés)

4. Principe de séparation nonlinéaire ([5, 6])

Un observateur non linéaire, basé sur le modème physico-chimique de procédé, donne des informations utiles pour l'opération manuelle d'une unité, pour la maintenance, la sécurité, le maintien des spécifications, l'estimation du rendement, etc, et ceci pour un prix dérisoire par rapport au prix d'un capteur fournissant une mesure de qualité d'un produit. Mais l'utilisation principale d'un observateur non linéaire est son utilisation en vue du contrôle d'un procédé. La propriété de convergence exponentielle de l'observateur grand-gain, en particulier avec une vitesse de convergence arbitraire, en fait même un outil de prédilection. Cette propriété nous a permis d'élaborer un principe de séparation non linéaire. Ce dernier n'est pas aussi puissant que dans le cas linéaire où la donnée d'un observateur optimal d'une part et d'un contrôleur optimal d'autre part donne une boucle observateur + contrôleur qui est elle-même optimale. Néanmoins, nous construisons effectivement une boucle de contrôle basée sur un observateur grand-gain d'une part et un contrôleur non linéaire basé sur un retour d'état d'autre part, et nous montrons la stabilité asymptotique global de l'ensemble lorsque le retour d'état est remplacé par un retour d'état estimé. Ce travail a fait l'objet de la thèse de Frédéric Viel [50] regroupant les articles [5, 6].

Considérons donc un système que l'on suppose sous une forme canonique d'observabilité

$$\left\{\begin{array}{rcl} \dfrac{dx}{dt} & = & Ax + b\left(x, u\right) \\ y & = & Cx \end{array}\right. \tag{22}$$

pour lequel on dispose

- d'un observateur grand-gain exponentiellement convergent, par exemple du type Kalman étendu grand-gain

$$\left\{\begin{array}{rcl} \dfrac{dz}{dt} & = & Az + b(z, u) - S(t)^{-1}C\left(u\right)' r^{-1}(C\left(u\right)z - y(t)) \\ \dfrac{dS}{dt} & = & -(A + b^{*}(z, u))'S - S(A + b^{*}(z, u)) + C'r^{-1}C - SQ_{\theta}S \end{array}\right. \tag{23}$$

- d'un contrôleur par retour d'état globalement asymptotiquement stable autour d'un point x^{*}

On considère alors le système bouclé constitué des équations (22,23) :

$$\begin{array}{rcl} \dfrac{dx}{dt} & = & Ax + b(x, u\left(z\right)) \\ \dfrac{dz}{dt} & = & Az + b(z, u\left(z\right)) - S(t)^{-1}C'r^{-1}(Cz - y(t)) \\ \dfrac{dS}{dt} & = & -(A + b^{*}(z, u\left(z\right)))'S - S(A + b^{*}(z, u\left(z\right))) + C'r^{-1}C - SQ_{\theta}S \end{array} \tag{24}$$

Theorem 21 ([5]). *Soit S^* l'unique solution définie positive de l'équation de Riccati algébrique en z^**

$$-(A+b^*(z^*,u^*))'S - S(A+b^*(z^*,u^*)) + C'r^{-1}C - SQ_\theta S = 0$$

en notant $u^* = u(z^*)$. Le point (z^*,z^*,S^*) est un point d'équilibre localement asymptotiquement stable de (24).

La preuve de ce résultat repose simplement sur la structure triangulaire de (24) et de la convergence exponentielle de l'observateur. Il faut ajouter une hypothèse pour atteindre la stabilité asymptotique globale :

Theorem 22 ([5]). *Si la demi-trajectoire $(x(t))_{t>0}$ extraite de la solution du système (24) est bornée, alors la solution complète de (24) est entièrement contenue dans le bassin d'attraction du point d'équilibre (x^*,x^*,S^*).*

La preuve de ce théorème repose elle-aussi sur la convergence exponentielle de l'observateur : la bornitude de la solution en x entraine la bornitude des trajectoires positives en z et en S. L'étude de l'ensemble ω–limite des trajectoires positives permet de conclure.

Dans les cas pratiques que nous avons étudiés, les modèles des procédés admettaient un compact positivement invariant si bien que l'hypothèse de bornitude de la demi-trajectoire en x était automatiquement satisfaite. Dans ce cas (courant), on peut donc conclure immédiatement à la stabilité asymptotique globale du système bouclé.

5. Contrôle optimal ([4, 17])

Une partie de mon travail ces dernières anneés a consisté à étudier le contrôle optimal et certains des aspects du couplage contrôleur optimal– observateur. Une partie de ce travail a été publiée [4] et présentée [17].

5.1. **Contrôle optimal de la navette spatiale ([4]).** L'article [4] est une étude sur le contrôle optimal de la rentrée atmosphérique d'une navette spatiale. Il s'agit d'une étude préliminaire qui a depuis été complétée par de nombreux travaux beaucoup plus précis.

Le modèle utilisé est un modèle simplifié de l'entrée atmosphérique basé sur les équations de la mécanique. Certains termes de Coriolis en $O(\Omega)$ ont été simplifiés, le modèle résultant étant constitué de trois équations différentielles dans lesquelles on a effectué une reparamétrisation du temps afin de transformer la minimisation du flux thermique en un problème temps minimal :

$$(25) \qquad \left\{ \begin{array}{lcl} \frac{dr}{dt} & = & \psi(r,v)\, v \sin(\gamma) \overset{\text{def.}}{=} F_r(r,v,\gamma) \\ \frac{dv}{dt} & = & \psi(r,v)\left(-g\sin(\gamma) - \frac{1}{2}\rho(r)\frac{S\,C_D}{m}v^2\right) \overset{\text{def.}}{=} F_v(r,v,\gamma) \\ \frac{d\gamma}{dt} & = & \psi(r,v)\left(\cos(\gamma)\left(-\frac{g}{v}+\frac{v}{r}\right) + \frac{1}{2}\rho(r)\frac{S\,C_L}{m}vu\right) \overset{\text{def.}}{=} F_\gamma(r,v,\gamma,u) \end{array} \right.$$

où

$$\frac{1}{\psi(r,v)} = C_q\sqrt{\rho(r)}v^3$$

et on considère le système adjoint

$$(26) \qquad \left\{ \begin{array}{lcl} \frac{dp_r}{dt} & = & -\frac{\partial H}{\partial r} \\ \frac{dp_v}{dt} & = & -\frac{\partial H}{\partial v} \\ \frac{dp_\gamma}{dt} & = & -\frac{\partial H}{\partial \gamma} \end{array} \right.$$

où

$$\begin{aligned} H(r,v,\gamma,p_r,p_v,p_\gamma,u) &= p_rF_r(r,v,\gamma) + p_vF_v(r,v,\gamma) + p_\gamma F_\gamma(r,v,\gamma,u) \\ &= G(r,v,\gamma,p_r,p_v,p_\gamma) + p_\gamma\left(\frac{1}{2}\rho(r)\frac{S\,C_L}{m}v\psi(r,v)\right)u \end{aligned}$$

Dans [4], nous avons explicité géométriquement les conséquences du principe du maximum sur ce système en dimension 3. Nous avons borné (localement) le nombre de commutations possibles et caractérisé les cas où le système est susceptible d'admettre des trajectoires singulières. Les difficultés viennent essentiellement des contraintes sur l'état, difficiles à prendre en compte dans le principe du maximum sous sa forme habituelle.

Nous avons résolu (25,26) et l'équation de reparamétrisation (27) numériquement

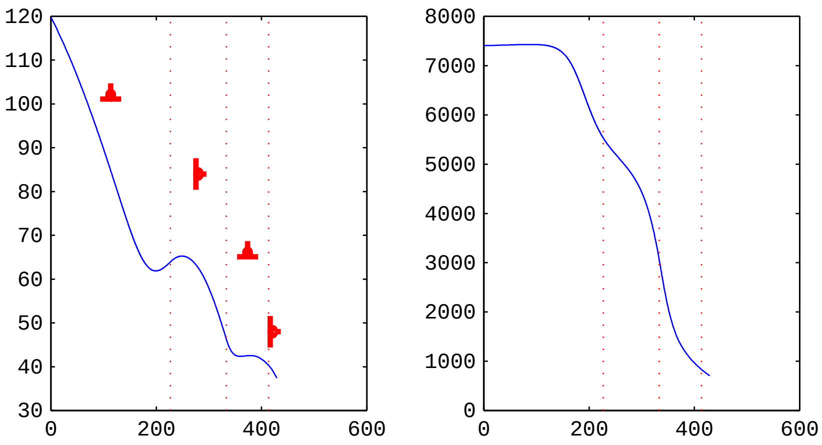


FIG. 6. Altitude h (in km) and relative speed v (in m / s) versus time

$$(27)\qquad \frac{d\tau}{dt} = \psi\left(r, v\right)$$

en choisissant à chaque instant le contrôle $u \in U = \{u \diagup 0 \leq u\left(t\right) \leq 1\}$ minimisant l'Hamiltonien *i.e.*

$$u\left(t\right) = \left\{ \begin{array}{ll} 0 & \text{si } p_\gamma\left(t\right) > 0 \\ 1 & \text{sinon} \end{array} \right.$$

Ainsi, nous avons calculé la solution donnée par le principe du minimum où le coût est le flux thermque. Les équations ont été implémentées en FORTRAN avec l'intégrateur LSODAR (ref. [37]) qui permet la détection des surfaces de commutation. Cette optionalité nous a permis de calculer les commutations ($p_\gamma = 0$) et les violations des contraintes sur l'état.

Les résultats de simulation n'ayant pas été publiés dans [4], nous allons donner ici une de ces simulations qui n'a été publiée que sous forme de prépublication du laboratoire d'Analyse Appliquée et Optimisation de l'Université de Bourgogne.

- Valeurs initiales : $r\left(0\right) = 6\,497\,960$ m, $v\left(0\right) = 7\,404.95$ m/s, $\gamma\left(0\right) = -0.032$ rad.
- Modèle atmosphérique : $\rho\left(r\right) = \rho_0 \exp\left(-\frac{r-r_0}{h_s}\right)$ où $\rho_0 = 1.22499$ kg/m^3, $r_0 = 6\,378\,140$ m, $h_s = 7\,143$ m.
- Constantes : $g = 9.78033$ m/s^2, $m = 7\,169.6$ kg, $S = 100$ m^2, $C_D = 0.5$, $C_L = 0.1$, $C_q = 1.705\,10^{-5}$.

Nous avons effectué des simulations avec plusieurs valeurs initiales du vecteur adjoint (p_r, p_v, p_γ). Les courbes données sur la Figure 6 montrent la trajectoire optimale correspondant au choix (arbitraire) $p_r = 1$, $p_v = 1$ and $p_\gamma = -1$. Les lignes verticales en pointillé représentent les instants de commutation. La valeur initiale du contrôle est $+1$. Le premier instant de commutation intervient au temps 227 s, un second au temps 333.3 s puis un dernier au temps 413.7 s, la valeur finale du contrôle étant 0. Cependant, une contrainte d'état est violée au temps 70.8 s.

5.2. **Couplage contrôle optimal–observateur ([17]).** La référence [17] n'a pas fait l'objet d'une publicaition écrite, je vais donc résumer très bièvement le contenu de cette communication.

Dans son article [36] sur les champs de vecteurs discontinus, Hermes défini la notion de stabilité par rapport aux mesures, en soulignant le fait qu'elle n'est pas équivalente à la stabilité usuelle lorsque le champ de vecteur est discontinu. Il montre aussi que l'existence d'une solution au sens de Filippov est une condition nécessaire à la stabilité par rapport aux mesures. Brunovsky ([29]) montre d'autre part que tout système linéaire possède une synthèse optimale régulière, et dans leur article, Brunovsky-Mirica ([30]) ont montré que dans le cas des systèmes linéaires, les synthèses régulières sont solutions au sens de Filippov du problème en temps optimal. Plus récemment, Benedetto Piccoli et Hector Sussmann ont introduit une notion différente de synthèses régulières, cf [46] pour une référence récente.

Nous avons donc étudié les propriétés de certaines synthèses, régulières en un sens qui fait l'objet de la Définition 23.

On s'interesse à des systèmes de la forme

$$\begin{cases} \dfrac{dx(t)}{dt} & = & f\left(x\left(t\right),u\left(t\right)\right) \\ y(t) & = & h(x(t)) \end{cases}$$

x évolue sur une variété analytique M et $x(0) = x_0 \in M$. Le champ de vecteur paramétré f est supposé analytique sur M. La classe des contrôles admissibles est l'ensemble U_{adm} des fonctions $u(t)$ continues par morceaux et à valeurs dans $[-1, 1]^p$. La fonction h de M dans $\mathbb{R}^q$ est elle aussi supposée analytique. Pour tout contrôle u fixé, le système est défini sur un intervalle $[0, T(x_0, u)[$.

Definition 23. *Une synthèse u^* est appellée synthèse régulière de Filippov si et seulement si les solutions $x\left(t\right)$ de*

$$\frac{dx(t)}{dt} = f\left(x\left(t\right),u^*\left(x\left(t\right)\right)\right)$$
$$x\left(0\right) = \eta$$

au sens de Caratheodory (i.e. une fonction absolument coninue pour presque tout t) coïncident avec les solutions au sens de Filippov i.e. pour presque tout t

$$\frac{dx(t)}{dt} \in \bigcap_{\delta>0} \bigcap_{\lambda N=0} \overline{\mathrm{conv}}\, f\left(B\left(x\left(t\right),\delta\right) \setminus N\right)$$

Bien que disposant d'un début de théorie plus générale, nous allons exposer nos résultats dans le cas d'un réacteur chimique de type batch. On considère une réaction consécutive $A \to B \to C$. Ce type de réaction est très courant et se rencontre par exemple lors des oxydations partielles et chlorinations partielles. Le modèle cinétique correspondant s'écrit

$$\begin{cases} \dfrac{d}{dt}\left[A\right]_t & = & -k_1\left(T_t\right)\left[A\right]_t \\ \dfrac{d}{dt}\left[B\right]_t & = & k_1\left(T_t\right)\left[A\right]_t - k_2\left(T_t\right)\left[B\right]_t \end{cases}$$

où $\left[A\right]_t$ et $\left[B\right]_t$ sont les concentrations en espèce A et B respectivement au temps t, et où k_i, $i = 1, 2$ sont les constantes cinétiques données par les lois d'Arrhénius

$$k_i\left(T_t\right) = A_i \mathrm{e}^{-\frac{E_i}{RT_t}}$$

A_i et E_i sont des constantes liées aux produits et R est une constante universelle. On suppose que $\left[A\right]_0 = 1$ et $\left[B\right]_0 = 0$, et notre objectif est d'atteindre en temps minimal la surface $\frac{[B]}{[A]} = d$ où $\frac{[B]}{[A]}$ est usuellement appellée selectivité. Nous noterons N cette courbe dans $\mathbb{R}^2$. Le contrôle est la puissance de chauffe Q_t qui est lié à la température par une équation dynamique

$$\frac{dT_t}{dt} = -h\left(T_t\right)\,Q_t$$

Le système est bien de la forme

$$\frac{dx(t)}{dt} = f\left(x\left(t\right),u\left(t\right)\right) \tag{28}$$

où lorsqu'il est bouclé avec la synthèse optimale

$$\frac{dx(t)}{dt} = f\left(x\left(t\right),u^*\left(x\left(t\right)\right)\right) \tag{29}$$

La synthèse optimale a été obtenue dans [28].

Lemma 24. *La synthèse en temps minimal du réacteur batch est une synthèse régulière au sens de Filippov.*

> Le preuve est élémentaire : elle consiste à montrer que la trajectoire singulière est une trajectoire au sens de Filippov et qu'il n'existe pas de solutions de Filippov autres que les solutions au sens de Caratheodory.

On considère un observateur à convergence exponentielle tel que ceux décrits précedemment (très facile à construire sur ce procédé)

$$\frac{dz(t)}{dt} = f\left(z\left(t\right),u\left(t\right)\right) + K_\theta\left(t, z\left(t\right)\right)\left(y\left(t\right) - h\left(x\left(t\right)\right)\right) \tag{30}$$

Theorem 25. *Considérons le système constitué du réacteur de polymérisation bouclé avec le contrôle calculé sur la base de la synthèse optimale i.e.*

$$\left\{\begin{array}{lll}\dfrac{dx(t)}{dt} & = & f\left(x\left(t\right),u^{*}\left(z\left(t\right)\right)\right)\\ \dfrac{dz(t)}{dt} & = & f\left(z\left(t\right),u^{*}\left(z\left(t\right)\right)\right)+K_{\theta}\left(t,z\left(t\right)\right)\left(y\left(t\right)-h\left(x\left(t\right)\right)\right)\end{array}\right. \tag{31}$$

alors pour tout $\varepsilon>0$, il existe θ assez grand de sorte qu'il existe T tel que

$$\left|T-T^{*}\right|<\varepsilon$$

$$\min_{\xi\in N}\left\Vert x\left(T\right)-\xi\right\Vert <\varepsilon$$

où T^{*} est le temps optimal donné par la synthèse.

Le théorème exprime le fait qu'un voisinage arbitrairement petit de la cible N est atteint en un temps arbitrairement proche du temps optimal. On montre plus précisemment que la trajectoire de $x\left(t\right)$ dans (31) est une trajectoire qui s'approche exponentiellement de la trajectoire optimale solution de (29).

6. Intégration numérique : la méthode de l'observateur ([7])

Le travail que je présente dans ce paragraphe est le plus ancien. C'est une application très inhabituelle des observateurs qui, à l'origine, était uniquement faite pour résoudre un problème de calcul numérique posé par un astronome mais qui a donné de tels résultats que nous avons approfondi la méthode et publié un long article dans *Applicationes Mathematicae.*

On considère le problème de Cauchy suivant, constitué d'un système d'équations différentielles écrit sous forme vectorielle et d'une condition initiale donnée :

$$\left\{\begin{array}{lll}\frac{dx}{dt} & = & F(x)\\ x(t_{0}) & = & x_{0}\end{array}\right. \tag{32}$$

avec $F\in C^{1}(U,\mathbb{R}^{p})$ lipschitz et où U est un ouvert de $\mathbb{R}^{p}$ contenant x_{0}, fixé.

On suppose que $H_{1},\ldots,H_{q}\in C^{1}(U,\mathbb{R})$, $1\leq q<p$, sont des intégrales premières fonctionnellement indépendantes associées au problème (32). On note

$$H=\left(\begin{array}{c}H_{1}\\ \vdots\\ H_{q}\end{array}\right)\in C^{1}(U,\mathbb{R}^{q}) \tag{33}$$

et on note D_{H} la matrice Jacobienne de H. On notera $H_{0}=H(x_{0})$. On a donc la relation

$$D_{H}F=0 \tag{34}$$

Soit Γ la surface de niveau dans laquelle se trouve la trajectoire solution de (32), $\Gamma=\{x\in U;\quad H(x)=H_{0}\}$. Générallement, Γ est une réunion de plusieurs composantes connexes et pas nécessairement compactes.

Hypothèse A *On suppose Γ compact.*

Le résultat que l'on veut établir concerne la résolution numérique du problème $(\mathcal{P})$ sur des grands temps d'intégration. Pratiquement, ce type de problème se rencontre essentiellement pour des problèmes $(\mathcal{P})$ dont les orbites sont bornées *a priori.* C'est cette hypothèse qui est cruciale pour la méthode exposée. Usuellement, Γ étant elle-même bornée, nous faisons donc cette hypothèse supplémentaire.

Hypothèse B *On suppose que pour tout $x\in\Gamma$,* $\operatorname{rang}\left(D_{H}\left(x\right)\right)=q$*.*

Sous cette hypothèse, Γ est une sous-variété lisse de U. Il est à noter que l'Hypothèse B n'est pas facilement vérifiable *a priori* pour un problème (32) donné, mais qu'elle peut-être vérifiée numériquement pendant l'intégration. Là encore, l'essentiel du point de vue pratique est que les orbites ne traversent pas un point pour lequel $D_{H}\left(x\right)$ admette une chute de rang.

On note $r\left(x,\Gamma\right)\overset{\text{déf.}}{=}\left\Vert H_{0}-H\left(x\right)\right\Vert$ et on défini $\tilde{\Gamma}_{\varepsilon}=\{x\in U;\quad r\left(x,\Gamma\right)\leq\varepsilon\}$ ($\left\Vert\cdot\right\Vert$ désigne la norme Euclidienne sur $\mathbb{R}^{p}$). Le théorème des fonctions implicites nous dit qu'il existe ε_{0} assez petit tel que pour tout $0<\varepsilon\leq\varepsilon_{0}$, la composante connexe Γ_{ε} de $\tilde{\Gamma}_{\varepsilon}$ contenant Γ est aussi un compact de U pour lequel l'Hypothèse B est vérifiée.

En raison de l'Hypothèse A, le problème (32) admet une solution sur $\mathbb{R}$ que l'on notera $x(t)$ (cf [35] par exemple). Pour tout $t \in \mathbb{R}$, $H(x(t)) = H_0$.

On considère maintenant le système

$$\left\{ \begin{array}{lcl} \frac{d\xi}{dt} & = & \tilde{F}(\xi) \\ \xi(t_0) & = & \xi_0 \end{array} \right. \tag{35}$$

où $\tilde{F}(\xi) = F(\xi) + D_H^T(\xi)\left(D_H(\xi)\,D_H^T(\xi)\right)^{-1} K(\xi)(H_0 - H(\xi))$ et $K(\xi)$ est une matrice diagonale $K(\xi) = \operatorname{diag}(\theta_1(\xi), \ldots, \theta_q(\xi))$ dont les éléments diagonaux sont des fonctions de $\xi \in \Gamma_\varepsilon$ telles que

$$\eta \overset{\text{déf.}}{=} \min_{1\leq i\leq q} \min_{\xi\in\Gamma_\varepsilon} \theta_i(\xi) > 0 \tag{36}$$

La matrice $D_H^T(\xi)\left(D_H(\xi)\,D_H^T(\xi)\right)^{-1}$ est l'inverse généralisée de la matrice $D_H(\xi)$ au sens de Moore-Penrose. La matrice $D_H(\xi)\,D_H^T(\xi)$ est inversible sur Γ sous l'Hypothèse B. Notons que

$$D_H\tilde{F} = K(\xi)(H_0 - H(\xi)) \tag{37}$$

Il est clair que les orbites solutions de (32) sont aussi des solutions de (35), puisqu'elles restent sur Γ. Néanmoins, le lemme suivant montre que pour (35), la surface Γ est un attracteur exponentiel.

Lemma 26 ([7]). *Soit $\xi_0 \in \Gamma_\varepsilon$. L'unique solution de (35) est définie pour tout $t \geq 0$ et vérifie $r(\xi(t), \Gamma) \leq r(\xi_0, \Gamma)\,\mathrm{e}^{-\eta t}$.*

La surface Γ étant attractive pour les orbites de (35), on peut s'attendre à ce que les trajectoires calculées numériquement, par n'importe quelle méthode d'intégration numérique stable, restent dans un voisinage de Γ. La matrice K est une matrice de poids qui permet de pondérer les contributions respectives des erreurs sur les intégrales premières ainsi que la correction globale appliquée. Suivant les problèmes considérés, K pourra être une matrice diagonale constante, $\theta_i(\xi) = k_i$, où on pourra choisir $\theta_i(\xi)$ de la forme $k_i \left\| D_H^T(\xi)\left(D_H(\xi)\,D_H^T(\xi)\right)^{-1} \right\|$ de façon à tenter de contrôler l'amplitude des corrections suivant les positions.

Considérons donc une méthode d'intégration numérique explicite à un pas, stable et consistante

$$\left\{ \begin{array}{lcl} \xi_{j+1} & = & \xi_j + h\Phi(\xi_j, h) \\ \eta_0 & = & \xi_0 \end{array} \right. \tag{38}$$

telle que, par exemple, la méthode de Runge–Kutta–Fehlberg.

La méthode étant consistante, pour tout ξ_0, il existe $C(\xi_0, t)$ telle que

$$\|\xi(t+h) - \xi(t) - h\Phi(\xi(t), h)\| \leq C(\xi_0, t)\,h$$

On suppose que pour tout compact $Q \subset U$:

$$\sup_{\xi\in Q}\ \sup_{0\leq h\leq 1} \|\Phi(\xi, h)\| < \infty$$
$$\text{et } \sup_{\xi_0\in Q}\ \sup_{0\leq h\leq 1} \|C(\xi_0, h)\| < \infty$$

Alors le théorème suivant montre la préservation effective des intégrales premières par la méthode numérique de l'observateur :

Theorem 27 ([7]). *Considérons le problème (35). Pour tout $\varepsilon > 0$,il existe un pas h, $0 < h \leq 1$, assez petit, et η défini par 36 assez grand de telle sorte que si $\xi_0 \in \Gamma$ alors $\xi_j \in \Gamma_\varepsilon$ pour tout $j \geq 0$.*

Nous avons testé et comparé cette méthode très simple à mettre en oeuvre sur le problème de Kepler plan et spatial, l'oscillateur harmonique double, le système de Gavrilov–Shil'nikov, le système d'Euler sur $so(4)$ et le système d'Anosov. Les résultats constituent l'essentiel de [7].

7. Perspectives de recherche

Les questions liées à l'identification restent nombreuses, tant théoriques que pratiques.

Du point de vue théorique, la question la plus difficile est probablement celle de la recherche de formes normales en présence d'une variable de contrôle. Aux deux entiers invariants permettant de caractériser les trois formes normales dans le cas sans contrôle s'ajoute le degré relatif du contrôle et sa position par rapport aux deux premiers.

Une autre question liée au filtre de Kalman étendu grand-gain est celle de sa convergence pour un système avec contrôle mis sous forme canonique d'observabilité. Au prix d'une dérivation de la variable de contrôle, on sait construire un observateur de Kalman étendu grand-gain exponentiellement convergent (en utilisant la transformation (14)). On ne sait pas si ce même observateur est exponentiellement convergent pour la forme canonique d'observabilité générale.

Du point de vue plus appliqué, nous aimerions beaucoup tester notre approche de l'identification sur des systèmes physiques, autrement qu'en simulation. Le fait d'avoir intégré le laboratoire LE2I devrait nous permettre, en plus d'avoir trouvé d'excellentes conditions de travail, de pouvoir tester nos algorithmes sur des systèmes électroniques réels.

Articles

[1] *E. Busvelle, J.-P. Gauthier* Observation and Identification Tools for Nonlinear systems. Application to a Fluid Catalytic Cracker. Soumis pour publication à *International Journal of Control*, 2003

[2] *E. Busvelle, J.-P. Gauthier* On determining unknown functions in differential systems, with an application to biological reactors, *ESAIM : COCV* **9**, p. 509–552 *(2003)*

[3] *E. Busvelle, J.-P. Gauthier* High-Gain and Non-High-Gain Observers for Nonlinear Systems, Contemporary Trends in Nonlinear Geometric Control Theory and its Applications, *World Scientific*, p. 233–256 *(2002)*

[4] *B. Bonnard, E. Busvelle, G. Launay* Geometric optimal control of the atmospheric arc for a space shuttle, Contemporary Trends in Nonlinear Geometric Control Theory and its Applications, *World Scientific*, p. 257–286 *(2002)*

[5] *F. Viel, E. Busvelle, J.-P. Gauthier* A stable control structure for binary distillation columns, *Int. J. of Control,* **67**, No 4, pp 475-505, *(1997)*

[6] *F. Viel, E. Busvelle, J.-P. Gauthier* Stability of a non-linear controller using a high-gain observer for polymerization reactor, *Automatica* **31**, pp 971-984, *(1995)*

[7] *E. Busvelle, R. Kharab, A. J. Maciejewski, J.-M. Strelcyn,* Numerical Integration of Differential Equations in Presence of First Integrals : Observer Method, *Applicationes Mathematicae* **22**, No 3, pp 373–418 *(1994)*

[8] *F. Viel, E. Busvelle, J.-P. Gauthier* A new method for instrument fault detection, *Revue européenne diagnostic et sureté de fonctionnement*, **4**, No 3, Article sélectionné du *TOOLDIAG'93*, Toulouse, France *(1994)*

[9] *F. Deza, E. Busvelle, J.-P. Gauthier, D. Rakotopara* Exponential observers for nonlinear systems, *IEEE Transactions on Automatic Control* **38**, No 3, *(1993)*

[10] *F. Deza, E. Busvelle, J.-P. Gauthier* Exponentially Converging Observers for Distillation Columns and Internal Stability of the Dynamic Output Feedback, *Chemical Engineering Science,* **47**, No 15/16, *(1992)*

[11] *F. Deza, E. Busvelle, J.-P. Gauthier, D. Rakotopara* High Gain Estimation for Nonlinear Systems, *Systems & Control Letters*, **18** *(1992)*

[12] *F. Deza, E. Busvelle, J.-P. Gauthier, D. Rakotopara* A stability result on the continuous-continuous and continuous-discret extended Kalman filters, *Compte-Rendus de l'Académie des Sciences de Paris*, **314**, *Février 1992*

Conférences

[13] *E. Busvelle* (présentation d'un travail commun avec J.-P. Gauthier) Observation and identification for nonlinear systems, *Banach Center Workshop,* Geometry in Nonlinear Control, *16–20 June 2003, Warsaw, Poland*

[14] *E. Busvelle* (présentation d'un travail commun avec J.-P. Gauthier) On determining unknown functions in differential systems, with an application to biological reactors, Colloque Commande des systèmes non-linéaires, *19-21 Juin 2002, Metz*

[15] *E. Busvelle* (présentation d'un travail commun avec J.-P. Gauthier) Observateur grand-gain et observateur de Kalman étendu, *CIRO'02, Marrakesh 2002*

[16] *E. Busvelle*, (présentation d'un travail commun avec J.-P. Gauthier), Sur la convergence d'un filtre de Kalman étendu asymptotique, *Journées Nationales d'Automatique, GdR–CNRS, 31 Janvier au 2 Février 2001 à Autrans, France*

[17] *E. Busvelle*, Observateurs et synthèse optimale de Filippov. Application à un réacteur chimique, *Séminaire dans le cadre du trimestre* "théorie du contrôle et applications" *à l'I.H.P., Paris, Mars 1998*

[18] *E. Busvelle, Y. Le Rouzic, D. Bossanne, F. Viel* Dynamic Quality Estimators, *4th International Symposium on Analytical Techniques for Industrial Process Control*, Anatech *1994*, Mandelieu-la-Napoule, France

[19] *E. Busvelle, F. Deza, F. Mokrani, S. Hapiak* Application of a nonlinear observer to a fluid catalytic craker, *NOLCOS'92, Juin 1992* Bordeaux, France

[20] *F. Viel, E. Busvelle, J.-P. Gauthier* A new approach for nonlinear state estimation applied to free radical polymerization, *Annual AIChE meeting 1992*, Miami, USA

[21] *E. Busvelle, F. Deza, J.-P. Gauthier* A Stability Result for the Extended Kalman Filter. Application to Industrial Processes, *SIAM Conference on Control and Its Applications*, *September 1992*, Mineapolis, USA

[22] *E. Busvelle*, Immersions et filtres de dimension finie ; filtrage particulaire, *Thèse de l'université de Rouen*, soutenue en *Octobre 1991*

[23] *D. de Brucq, P. Courtellemont, E. Busvelle* Evolution d'état non-linéaire et filtrage approché par maximum d'entropie, *GRETSI*, *Septembre 1991*, Juan-les-Pins, France

[24] *E. Busvelle, D. Rakotopara* A necessary condition for the existence of finite dimensional filters for discrete-time nonlinear systems, *IEEE* 29th *Conference on Control Applications*, *Decembre 1990*, Honolulu, Hawai

[25] *E. Busvelle, D. Rakotopara* Immersion in conditionally gaussian systems and finite dimensional filters, *IEEE* 29th *Conference on Control Applications*, *Decembre 1990*, Honolulu, Hawai

Bibliographie annexe

[26] *J. S. Baras, A. Bensoussan, M. R. James,* Dynamic observers as asymptotic limits of recursive flters : special cases, *SIAM J. Appl. Math.,* **48,** 1147–1158, *(1988)*

[27] *G. Bastin, D. Dochain,* Adaptive control of bioreactors, Elsevier, *(1990)*

[28] *B. Bonnard, J. de Morant,* Towards a geometric theory in the time optimal control of chemical batch reactors, *SIAM J. Control*, **33** (5) pp 1279–1311, *(1995)*

[29] *P. Brunovsky,* Existence of regular synthesis for general control problems, *J. Differential Equations* **38** no. 3, 317–343, *(1980)*

[30] *P. Brunovsky, S. Mirica,* Classical and Filippov solutions of the differential equation defined by the time-optimal feedback control. *Rev. Roumaine Math. Pures* Appl. **20** no. 8, 873–883, *(1975)*

[31] *J.-P. Gauthier, H. Hammouri, S. Othman,* A simple observer for nonlinear systems, *IEEE Trans. Aut. Control,* **37**, pp. 875–880, *(1992)*

[32] *J.-P. Gauthier, I. Kupka,* Deterministic observation theory and applications, Cambridge University Press, 2001

[33] *H. Hammouri, M. Farza,* Nonlinear observers for locally uniformly observable systems, *COCV*, **9**, 343-352, *(2003)*

[34] *R. Hardt*, Stratification of real analytic mappings and images, *Invent. Math.* **28**, pp. 193-208, *(1975)*

[35] *P. Hartman*, *Ordinary Differential Equations*, Wiley, New York, *(1964)*

[36] *H. Hermes*, Discontinuous vector fields and feedback control. In J. Hale andJ. Lasalle, editors, *Differential Equations and Dynamical Systems*, Academic Press, pp 155–165, NY, *(1967)*

[37] *A. C. Hindmarsh*, *Odepack, a systematized collection of ode solvers*, in scientific computing, r. s. Stepleman et al. (eds.), North-Holland, Amsterdam, pp. 55-64, *(1983)*

[38] *H. Hironaka,* Subanalytic sets, in Number Theory, Algebraic Geometry and Commutative Algebra, Volume in honor of Y. Akizuki ; Kinokuniya, Tokyo, pp. 453-493. *(1973)*

[39] *R.M. Hirschorn*, Invertibility of control systems on Lie Groups, *SIAM Journal on Control and Opt.*, **15**, No 6, pp 1034-1049, *(1977)*

[40] *R.M. Hirschorn*, Invertibility of Nonlinear Control Systems, *SIAM J. Control and Opt.*, **17**, No 2, 1979, pp. 289-297.

[41] *T. Huillet, G. Salut,* Interprétation des équations du filtrage non-linéaire, séances du GdR Automatique du C. N. R. S. (pôle non-linéaire), Paris, *(1989)*

[42] *A. Jazwinski*, Stochastic Processes and Filtering Theory, New York Academic, *(1970)*

[43] *R. E. Kalman*, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, **82**, (series D), pp. 35–45, *(1960)*

[44] *M. Pengov, E. Richard, J.-C. Vivalda,* On global stabilization of nonlinear systems with continuous-discrete observers, European Journal of Control, à paraître

[45] *J. Picard,* Efficiency of the extended Kalman filter for nonlinear systems with small noise, *SIAM J. Appl. Math.,* **51,** 3, 843–885, *(1991)*

[46] *B. Piccoli, H. Sussmann,* Regular synthesis and sufficiency conditions for optimality, *SIAM J. Control Optim.* **39** , no. 2, pp. 359–410 *(2000)*

[47] *W. Respondek*, Right and Left Invertibility of Nonlinear Control Systems, in Nonlinear Controllability and Optimal Control, H.J. Sussmann ed., Marcel Dekker, New-York, pp. 133-176, *(1990)*

[48] *P. Rouchon,* Simulation dynamique et commande non linéaire des colonnes à distiller, Thèse de l'école des mines de Paris, *(1990)*

[49] *M. van Doothing, F. Viel, D. Rakotopara, J.-P. Gauthier,* Coupling of non-linear control with a stochastic filter for state estimation : application on a continuous free radical polymerization reactor, *I. F. A. C. International Symposium* ADCHEM'91, Toulouse, France, *(1991)*

[50] *F. Viel*, Stabilité des systèmes nonlinéaires contrôlés par retour d'état estimé. Application aux réacteurs de polymérisation et aux colonnes à distiller. Thèse de l'Université de Rouen, *(1994)*

HIGH-GAIN AND NON-HIGH-GAIN OBSERVERS FOR NONLINEAR SYSTEMS.

E. BUSVELLE, J.P.GAUTHIER

Dedicated to Velimir Jurdjevic

Abstract. In this paper, following ideas already developped in [10], we construct an observer for nonlinear systems that looks like the extended Kalman filter. In fact, it is asymptotically (in time) exactly the deterministic version of the extended Kalman filter, and when the "innovation" is large, it is an high gain observer. In the context of the theory developped in [10], we show that it works for "all observable systems". In the paper, we prove convergence of the estimation error, we give several estimates of this error, and we show a convincing illustrative application (a distillation column).

1. Introduction, systems under consideration

1.1. **Systems under consideration.** We consider nonlinear systems of the following form (1.1), on $\mathbb{R}^n$. The control space U, is a closed subset of $\mathbb{R}^d$. **Only for simplicity of the exposition of the proof of the main result**, the observation is taken to be single-valued: it is a $u-$ dependant linear form on $\mathbb{R}^n$.

$$\begin{aligned} \frac{dx}{d\tau} &= A(u)x + b(x,u), \\ y &= C(u)x. \end{aligned} \tag{1.1}$$

$A(u)$, $C(u)$ are matrices:

$$C(u) = (a_1(u), 0, ..., 0),$$

$$A(u) \begin{pmatrix} 0, a_2(u), 0, ..., 0 \\ 0, 0, a_3(u), 0, ..., 0 \\ . \\ . \\ 0, ..., 0, a_n(u) \\ 0, ..., 0 \end{pmatrix}.$$

where $a_i(.)$, $i = 1, ..., n$, are positive smooth functions, bounded from above and from below:

$$0 < a_m \leq a_i(u) \leq a_M.$$

Also, $b(x,u)$is a smooth, $u-$dependant vector field, depending triangularly on x and compactly supported:

$$b = b_1(x_1, u)\frac{\partial}{\partial x_1} + b_2(x_1, x_2, u)\frac{\partial}{\partial x_2} + ... + b_n(x_1, ..., x_n, u)\frac{\partial}{\partial x_n}.$$

These assumptions look very strong. In fact, under either genericity hypotheses or observability hypotheses, **for the purpose of synthesis of observers, it is sufficient to restrict to these systems, under the normal form** (1.1) **(or similar multi-output normal forms), and meeting these assumptions**. This will be discussed in the next section 2.

We stress again that in all the paper, **the single output assumption can be removed everywhere,** and we leave this to the reader, but, in Section 4, we will deal with a $2-$ outputs system, in a similar normal form.

Date: December 2000.

Key words and phrases. Control Theory, Observability, Observation Theory, Nonlinear Systems.

1.2. **Presentation of the paper.** Our purpose herein is **to construct observers**, for the observable systems described above.

In fact, for these systems, several types of nonlinear observers can be constructed. We will focus on two types of construction that both turn around the "extended Kalman filter", in either its deterministic or its stochastic form:

1. **First construction**: The Extended Kalman Filter itself,
2. **Second construction:** The High Gain Extended Kalman Filter,
3. **Our construction in this paper:** a mixing of 1. and 2.

Let us just give some details now, to explain where we want to go.

1. The extended Kalman Filter.

For long, the engineers introduced and successfully used the extended Kalman filter (EKF), either in its stochastic or its deterministic form. The EKF is just the standard Kalman filter for linear time-dependant systems, applied to the **linearized system along the estimate trajectory**. We will give precise equations later on.

It is easy to see that it is a non-intrinsic object (depending on coordinates). It would be intrinsic if it was dealing with the linearized along the real trajectory, but this trajectory is unknown.

It is known that, **under observability conditions**, the Extended Kalman filter, has good properties:

(i) In its deterministic form, it is a local observer in the following sense. For sufficiently small initial error on the estimate of the state, the estimation error converges exponentially to zero. The prototype of these results can be found in [2] for instance.

For our systems (1.1), with the assumptions of Section 1.1, it is not hard to check that the linearized systems along any trajectory are uniformly observable, (in the classical sense of the linear theory, and with uniform bounds on the Gramm observability matrices). Hence, this result applies.

(ii) In its stochastic form, except for the linear case, where the EKF is the "optimal" filter, there is no general theoretical result that applies. Even for good observable systems in our normal form 1.1, for small noise, small initial variance and dimension 1: there is a counterexample of such a system, in [16] for instance, where the EKF doesn't work at all.

Nevertheless, despite the lack of these theoretical justifications, people use it in practice for nonlinear filtering and it may give very good results (even for systems that have much weaker observability properties than those considered here).

In the application of our techniques, presented in section 4 below, we will show a (family of) practical examples which is very interesting because, it seems that, the results of [16] on the EKF for small noise, apply in general, and that the "small parameter" has a physical interpretation.

We will not say more about that because this is beyond the scope of this paper. But it is one more justification of the use of our method developed here to this application.

2. The High Gain Extended Kalman Filter.

The following results have been proved in [4], [5], [10].

We consider the equations of the extended Kalman filter, in which the "covariance matrix Q" depends on a real parameter θ, $\theta \geq 1$, in the following way:

$$Q_{ij} = \theta^{i+j+1} Q^0_{i,j}.$$

For $\theta = 1$, it is exactly the EKF. For θ large enough, it is what we call here the "High gain extended Kalman filter" (HGEKF).

(i) In the deterministic setting, the estimation error has **arbitrarily large exponential decay** (depending on θ). ([10], for instance). This holds **whatever the initial error is, (that is, this is a global result)**.

(ii) In the stochastic setting, it is a nonlinear filter with "bounded variance" (the variance is bounded in θ^n, which is not that good, but it is bounded anyway). ([4], for instance).

3. What we want to do in this paper.

The idea in this paper is the **very simple** following one: we give the parameter θ in the HGEKF an exponential decay from θ_0 large, to 1.

What is expected, (and what happens) is the following:

(i) The beginning of the transient of the estimation error is the one of the high gain extended Kalman observer: there is an exponential decay that can be made arbitrarily large.

(ii) There is a global exponential decay of the estimation error (but, of course, it cannot be controlled).

(iii) The asymptotic behavior is the one of the standard "extended Kalman filter", (that people like in practice, as stated above).

Our main result, Theorem 1 in Section 3 proves (i) and (ii). The proof is more or less an improvement of the proof of convergence of the high gain Kalman observer, as given in [10].

Of course, this construction has a terminal defect: it is time dependant. In deterministic terms, it will work for large initial estimation errors, but not for big "jumps" of the state at intermediate times. In the section 3.3, we propose a very simple practical way to make the observer "recursive".

In the section 4, we show the application of this procedure to a binary distillation column in which the "quality of the feed" is unknown, an subject to large changes. It was already noticed in the book [10] that this application is a nontrivial nice application of the observability theory, and of high gain observers.

Here, it is even much more convincing: when the feed changes, (a big "state jump"), the behavior of the observer is the one of a high-gain observer: recovering arbitrarily fast the quality of the feed, and when the feed does not move, the asymptotic behavior of the observer is the one of the extended Kalman filter, almost optimal with respect to small noise in that case (but we do not prove anything about this optimality in this paper).

For first applications of "high gain observers" to distillation columns, see [19], [20].

2. Justification of the assumptions and observability

2.1. **Justification of the normal form.**

Let us recall here the main results of an observability theory summarized in [10], and developed in [6], [7], [8], [9], [13]. This theory leads to the consideration of systems under the normal form (1.1), or similar multi-output normal forms such as (2.1) below, that meet the assumptions of the section 1.1. Here, by "observability", we mean "observability for every fixed input function $u(t)$". For details, see [10].

The main results of this theory are as follows. They concern general nonlinear smooth systems of the form:

$$\begin{aligned}\frac{dx}{dt} &= f(x,u),\\ y &= h(x,u),\end{aligned}$$

on a smooth manifold X, n dimensional, $y \in \mathbb{R}^p$, $u \in U$, subset of $\mathbb{R}^d$.

Basically, there are two cases.

Case 1. $p \leq d$. In that case, observability is a non generic property. It is even a property of infinite codimension, at the level of germs of systems. This high degeneracy leads to the fact that, in the control affine case, all observable systems can be put under normal forms similar to (1.1). (moreover, one can take $a_i = 1$, $i = 1, ..., n$).

This result is only a local result in the state space, but it is a global result with respect to the control variable. Moreover, in most of the practical cases we know, it is also global in x. In particular, it will be global in the application of Section 4.

In the non control affine case, there is another result, that we don't want to recall here. It leads naturally to high gain observers of another type ("Luenberger type"). Let us just say that the results herein can be easily generalized to this normal form and these observers.

Case 2. $p > d$. In that case, the situation is completely opposite. Observability becomes a **generic property,** and generically, a system can be put **globally** under a normal form similar to (1.1), but the dimension of the state in the normal form is bigger than the dimension of the state of the original system: it is at most double plus one. Also, the control in the normal form contains a certain number of derivatives of the control of the initial system. But this is more or less unimportant for observation problems, where the control, and hence its derivatives, are known.

In fact, generically, the systems can be put in a form which is a very special case of the form (1.1), called the "phase variable representation":

$$y^{(N)} = \varphi(y, \dot{y}, ..., y^{(N-1)}, u, \dot{u}, ..., u^{(N-1)}), \quad N \leq 2n+1. \tag{2.1}$$

Other cases: there are also other (nongeneric) interesting cases where the original system can be put under the "phase variable representation" (2.1). For instance, systems **without control** that are such that the mapping:

$$initial - state \rightarrow derivatives\ of\ y:$$

$$x_0 \rightarrow (y, \dot{y}, ..., y^{(M)}),$$

has "finite multiplicity" for a certain integer M. (See [10], and originally [13]).

Note 1. The reasons for which we make the matrix $A(u)$ depend on u in the normal form (1.1) may look not clear, because, in all the cases described above, it doesn't.

In fact, the only reason to consider this dependance is the following: the formal computations we do in the proof of our main result, work for that type of systems. Moreover, in the application we describe in Section 4, the matrix A actually does depend on u.

Note 2. In that case were a_i depends on u, the following should also be noticed: even the high gain version of the extended Kalman filter is much better in practice than the "high gain Luenberger observer" mentioned above: the high gain observers both kill the nonlinearities contained in the vector field b. But the extended Kalman filter takes into account the variations of u, through the matrix $A(u)$. The standard high gain observers in Luenberger form don't do this. This is the case in the application, Section 4 below.

2.2. Justification of the technical assumptions.

Let us consider successively the two technical assumptions we made in the section 1.1:

A. $0 < a_m \leq a_i(u) \leq a_M$, $i = 1, ..., n$,

B. The functions b_i are compactly supported.

In fact, the assumption A is always satisfied in the cases 1., 2. of the previous section 2.1: the a_i are constant and equal to 1. In the application of section 4, this assumption is also satisfied, as we shall see.

Let us just notice the following.

A1. The Assumption $a_i(u) \neq 0$ just implies observability of systems in the normal form (1.1):

- If the output $y(t)$ is known, the input being also known, the fact that $a_1(u)$ is nonzero implies that we can compute $x_1(t)$ from $y(t)$.

- The fact that $a_2(u) \neq 0$ implies that we can compute $x_2(t)$ from the knowledge of $x_1(t)$, and by induction, we can reconstruct the whole state $x(t)$ from the knowledge of $y(t)$.

Modulo a trivial change of variables, the condition $a_i(u) \neq 0$ is equivalent to $a_i(u) > 0$.

A2. The a_i being smooth, restricting to a compact subset of the set of values of control implies that we can find the a_m, a_M, of assumption A.

The assumption B above can be trivially realized, by multiplying by a cut-off function, compactly supported, leaving the original **vector field b unchanged on an arbitrarily large compact subset of $\mathbb{R}^n$**.

We cannot expect more than that. As explained in the book [10], the problem of synthesis of observers is an ill-posed problem outside compact sets of the state space. This is easily understandable: on noncompact sets, it can happen that the estimation error goes to zero for certain metrics, but to infinity for others. So that, reasonable observers work only as long as the state trajectory $x(t)$ of the system remains in a given compact set, or they work for semi trajectories $\{x(t), t \geq 0\}$ that are entirely contained in a given compact set.

To finish, let us mention that this restriction to compact sets (unavoidable in a general observation theory), has not so important consequences: for instance, the high gain observers can be used in general for **global** dynamic output stabilization (again, see [10]).

3. Statement and proof of the theoretical result

The observer we propose, is based upon the High gain extended Kalman filter, proposed in [10], [4], [5]. For computational details about the Riccati matrix equation, we refer to [3], or [10].

3.1. **The observer and the statement of the theorem.** The equation of the observer is:

$$\left\{\begin{array}{c} (i)\ \frac{dz}{d\tau} = A(u)z + b(z,u) - S(t)^{-1}C'r^{-1}(Cz - y(t)), \\ (ii)\frac{dS}{d\tau} = -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + \\ C'r^{-1}C - SQ_\theta S, \\ \frac{d\theta}{d\tau} = \lambda(1-\theta), \end{array}\right. \tag{3.1}$$

where $C = (a_1(u), 0, ..., 0)$, $Q_\theta = \theta^2\Delta^{-1}Q\Delta^{-1}$, $\Delta = diag(1, \frac{1}{\theta}, ..., (\frac{1}{\theta})^{n-1})$. Here, $b^*(z,u)$ denotes the Jacobian matrix of $b(z,u)$ w.r.t. z, and r, λ are positive scalars. Q is a symmetric positive definite matrix.

Comments:

1. Q, r, in the stochastic context, are the covariances of the state noise and output noise respectively.

2. If $\lambda = 0$ and $\theta_0 = 1$, or if $\lambda > 0$, but t is large, this is exactly the (deterministic version of) the extended Kalman filter.

3. If θ_0 is large, and if $\tau \leq T$, then, this equation is almost the equation of the high gain extended Kalman filter with gain $\theta(T)$. Hence, for $\tau \leq T$, setting $\varepsilon(\tau) = z(\tau) - x(\tau)$, ($\varepsilon$ is the **estimation error**), we can expect the following, for θ_0 large enough in front of T:

$$||\varepsilon(\tau)||^2 \leq \theta(\tau)^{2(n-1)}H(c)e^{-(a_1\theta(T)-a_2)\tau}||\varepsilon(0)||^2. \tag{3.2}$$

Here, a_1, a_2 are positive constants, $H(c)$ is a decreasing positive function of c, where $S(0) \geq c\ Id$. Also, $\theta(T) = 1 + (\theta_0 - 1)e^{-\lambda T}$.

In particular, this implies that the error $\varepsilon(t)$ **can be made arbitrarily small, in arbitrarily short time, increasing** θ_0. For θ constant, this is the behavior of the "high gain extended Kalman filter. In that case (θ constant), this estimate follows from [10], [5]. We will prove it below for θ nonconstant.

Our main result herein will be the following:

Theorem 1. *1. For all $0 \leq \lambda \leq \lambda_0$, ($\lambda_0 = \frac{Q_m\alpha}{4(n-2)}$, where $Q \geq Q_mId$ and α comes from Lemma 1 below), for all θ_0 large enough, depending on λ, for all $S_0 \geq c\ Id$, for all $K \subset \mathbb{R}^n$, K a compact subset, for all $\varepsilon_0 = z_0 - x_0$, $\varepsilon_0 \in K$, the following estimation holds, for all $\tau \geq 0$:*

$$\begin{aligned} ||\varepsilon(\tau)||^2 &\leq R(\lambda,c)e^{-a\ \tau}||\varepsilon_0||^2\Lambda(\theta_0,\tau,\lambda), \\ \Lambda(\theta_0,\tau,\lambda), &= {\theta_0}^{2(n-1)+\frac{a}{\lambda}}e^{-\frac{a}{\lambda}\theta_0(1-e^{-\lambda\tau})}, \end{aligned} \tag{3.3}$$

where $a > 0$. $R(\lambda,c)$ is a decreasing function of c.

2. Moreover the short term estimate (3.2) holds for all $T > 0$, $\tau \leq T$, for all $\theta_0 \geq \bar{\theta}_0$, $\bar{\theta}_0 = e^{\lambda T}(\frac{L'}{Q_m\alpha} - 1) + 1$, where $L'$is the sup of the partial derivatives of b w.r.t. x.

Comments.

a. Note that the function $\Lambda(\theta_0,\tau,\lambda)$ is a decreasing function of τ, and that, for all $\tau > 0$, $\lambda > 0$, $\Lambda(\theta_0,\tau,\lambda)$ can be made arbitrarily small, increasing θ_0.

b. This means that, provided that λ is smaller than a certain constant λ_0, and θ_0 is large in front of λ, the estimation error goes exponentially to zero, and can be made arbitrarily small in arbitrary short time.

c. The asymptotic behavior of the observer is the one of the extended Kalman filter,

d. The "short term behavior" is the one of the "high gain extended Kalman filter".

3.2. **Proof of Theorem 1.**

3.2.1. *Preparation for the proof.* Let us recall that:

$$\theta(\tau) = 1 + (\theta_0 - 1)e^{-\lambda\tau}, \tag{3.4}$$

and let us set $F = diag(0, 1, 2, ..., n-1)$. Then:

$$\begin{aligned} \frac{d(\frac{1}{\theta})}{d\tau} &= -\frac{\lambda(1-\theta)}{\theta^2}, \\ \frac{d\Delta}{d\tau} &= -F\Delta\frac{\lambda(1-\theta)}{\theta}, \\ \frac{d\Delta^{-1}}{d\tau} &= F\Delta^{-1}\frac{\lambda(1-\theta)}{\theta}. \end{aligned} \tag{3.5}$$

The equations under consideration are:

$$\begin{gathered} (i)\ \tfrac{d\varepsilon}{d\tau} = A(u)\varepsilon + b(z,u) - b(x,u) - S(t)^{-1}C'r^{-1}C\varepsilon, \\ (ii)\tfrac{dS}{d\tau} = -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'r^{-1}C - SQ_\theta S, \\ (iii)\ \tfrac{d\theta}{d\tau} = \lambda(1-\theta). \end{gathered} \tag{3.6}$$

We make the following changes of variables, with $P = S^{-1}$:

$$\begin{aligned} &\tilde{x} = \Delta x, \tilde{z} = \Delta z, \varepsilon = z - x, \tilde{\varepsilon} = \Delta\varepsilon, \tilde{S} = \theta\Delta^{-1}S\Delta^{-1}, \\ &\tilde{P} = \tilde{S}^{-1} = \frac{1}{\theta}\Delta P\Delta, \tilde{b}(z) = \Delta b(\Delta^{-1}z), \tilde{b}^*(z) = \Delta b^*(\Delta^{-1}z)\Delta^{-1}. \end{aligned} \tag{3.7}$$

Remark : It should be noted that the Lipschitz constant of $\tilde{b}$ is the same as the one of b, and the maximum of $||\tilde{b}^*||$ is the same as the one of $||b^*||$ (recall that the component b_i of b is compactly supported with respect to all of its arguments $(x_1, ..., x_i, u)$, and that $\theta \geq 1$).

An obvious computation gives:

$$\frac{d}{d\tau}(\tilde{\varepsilon}) = \theta[(A - \tilde{P}C'r^{-1}C)\tilde{\varepsilon} + \frac{1}{\theta}(\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x})) - \frac{\lambda(1-\theta)}{\theta^2}F\tilde{\varepsilon}], \tag{3.8}$$

$$\begin{aligned} \frac{d}{d\tau}(\tilde{S}) &= \theta[-(A + \frac{1}{\theta}\tilde{b}^*(\tilde{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\theta)}{\theta^2})'\tilde{S} \\ &\quad - \tilde{S}(A + \frac{1}{\theta}\tilde{b}^*(\tilde{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\theta)}{\theta^2}) + C'r^{-1}C - \tilde{S}Q\tilde{S}], \\ \frac{d\theta}{d\tau} &= \lambda(1-\theta). \end{aligned} \tag{3.9}$$

Important comment. At this place, we used the observability properties: the normal form (1.1) is crucial in the computation above.

Now, we can make a time rescaling. We set:

$$\begin{gathered} dt = \theta(\tau)d\tau, \text{ or } t = \int_0^\tau \theta(v)dv, \\ \tilde{\varepsilon}(\tau) = \bar{\varepsilon}(t), \tilde{S}(\tau) = \bar{S}(t), \tilde{P}(\tau) = \bar{P}(t), \theta(\tau) = \bar{\theta}(t), \end{gathered}$$

to get the final set of equations:

$$\begin{aligned} &(i)\ \frac{d}{dt}(\bar{\varepsilon}) = [(A - \bar{P}C'r^{-1}C)\bar{\varepsilon} + \frac{1}{\bar{\theta}}(\tilde{b}(\bar{z}) - \tilde{b}(\bar{x})) - \frac{\lambda(1-\bar{\theta})}{\bar{\theta}^2}F\bar{\varepsilon}], \\ &(ii)\ \frac{d}{dt}(\bar{S}) = [-(A + \frac{1}{\bar{\theta}}\tilde{b}^*(\bar{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\bar{\theta})}{\bar{\theta}^2})'\bar{S} \\ &\qquad - \bar{S}(A + \frac{1}{\bar{\theta}}\tilde{b}^*(\bar{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\bar{\theta})}{\bar{\theta}^2}) + C'r^{-1}C - \bar{S}Q\bar{S}], \\ &(iii)\ \frac{d\bar{\theta}}{dt} = \lambda\frac{(1-\bar{\theta})}{\bar{\theta}}. \end{aligned} \tag{3.10}$$

First, there are some classical results allowing to bound the solutions of the Ricatti equation (3.10), (ii), for $\theta_0 > 1$, and $\lambda < 1$. To apply these results, one has to notice that the linear time dependant systems:

$$\begin{aligned} \frac{dx}{dt} &= (A(u(t)) + \frac{1}{\bar{\theta}}\tilde{b}^*(\bar{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\bar{\theta})}{\bar{\theta}^2})x(t), \\ y &= C(u(t))x(t), \end{aligned}$$

are uniformly observable (in the sense of linear systems), for all bounded measurable functions $a_i(u(t))$, $\tilde{b}^*_{i,j}(\bar{z}(t)), \bar{\theta}(t)$, with $a_M \geq a_i \geq a_m > 0$. Precisely, we have:

Lemma 1. *If the functions $a_i(u(t))$, $|\tilde{b}^*_{i,j}(\bar{z}(t))|$, $\bar{\theta}(t)$, are all smaller than $a_M > 0$, and if $a_i(u(t)) > a_m > 0$, (which is the case by our assumptions), if $0 \leq \lambda \leq 1$, and $1 < \bar{\theta}(t)$ then, the solution of the Ricatti equation 3.10, (ii), satisfies the following inequality,*

$$\alpha\ Id \leq S(t) \leq \beta\ Id,$$

for all $T_0 > 0$, for all $t \geq T_0$, where α and β depend on T_0, a_m, a_M ***(but do not depend on** c, $\bar{S}_0 \geq c$ Id !)*

This result is more or less classical. It is contained in [3] for instance. A detailed proof is given in [10], because there are several mistakes in many textbooks. The key point for a **simple** proof is the precompactness of the weak-* topology on $L^\infty[0,T]$, and the continuity of the input-state mapping of a control-affine system, for the weak-* topology on controls, and the uniform topology on trajectories $x(t)$, $t \in [0,T]$.

Straightforward computations with (3.10) give:

$$\begin{aligned} \frac{d}{dt}(\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t)) \leq& -Q_m\ \bar{\varepsilon}'\bar{S}(t)^2\bar{\varepsilon} + 2\bar{\varepsilon}'\bar{S}(t)(\frac{1}{\bar{\theta}}(\tilde{b}(\bar{z}) - \tilde{b}(\bar{x}) - \tilde{b}^*(\bar{z})\bar{\varepsilon})) \\ &+ \frac{\lambda(1-\bar{\theta})}{\bar{\theta}^2}\bar{\varepsilon}'\bar{S}(t)\bar{\varepsilon}, \end{aligned} \tag{3.11}$$

where $Q \geq Q_m\ Id$.

In particular, if $t \geq T_0$, with α given by Lemma 1, this gives:

$$\begin{aligned} \frac{d}{dt}(\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t)) \leq& -(Q_m\alpha + \frac{\lambda(\bar{\theta}-1)}{\bar{\theta}^2})\ \bar{\varepsilon}'\bar{S}(t)\bar{\varepsilon} + \\ & 2\bar{\varepsilon}'\bar{S}(t)(\frac{1}{\bar{\theta}}(\tilde{b}(\bar{z}) - \tilde{b}(\bar{x}) - \tilde{b}^*(\bar{z})\bar{\varepsilon})). \end{aligned} \tag{3.12}$$

Using this equation, and again Lemma 1, we will now prove the theorem.

3.2.2. *Proof of the short term estimation 3.2.* This proof is in two steps. We will first prove an estimation for $T \geq t \geq T_0 > 0$, and after for $t \leq T_0$. Gluing them together, we get the short term estimation 3.2. This is the standard high gain reasoning, and it is done in details in [10] for θ constant. We omit the computational details.

Step 1, $T \geq t \geq T_0$.

Straightforward computations using (3.12), Lemma 1 and the remark in Section 3.2.1 give:

$$\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t) \leq \bar{\varepsilon}(T_0)'\bar{S}(T_0)\bar{\varepsilon}(T_0)e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})(t-T_0)}. \tag{3.13}$$

Therefore $\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t) \leq \beta||\bar{\varepsilon}(T_0)||^2e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})(t-T_0)}$, and finally:

$$\begin{gathered} T \geq t \geq T_0 : \\ ||\bar{\varepsilon}(t)||^2 \leq \frac{\beta}{\alpha}e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})(t-T_0)}||\bar{\varepsilon}(T_0)||^2. \end{gathered} \tag{3.14}$$

Step 2, $t \leq T_0$.

We need a more straightforward estimation here. A very rough one is obtained just using Gronwall's identity. For certain $s, k > 0$, we have:

$$||\bar{P}(t)|| \leq (||\bar{P}(0)|| + k)e^{sT_0}. \tag{3.15}$$

We assume that $S(0) = S_0$ lies in the compact set: $c\ Id \leq S_0 \leq d\ Id$. As a consequence, $P(0) \leq \frac{1}{c}Id$.

By the equation (3.10), we have, for $t \leq T_0$: $\frac{d}{dt}(\bar{\varepsilon}) = (A - \bar{P}C'r^{-1}C)\bar{\varepsilon} + \frac{1}{\theta}(\tilde{b}(\bar{z}) - \tilde{b}(\bar{x})) - \frac{\lambda(1-\theta)}{\bar{\theta}^2}F\bar{\varepsilon}$, hence:

$$||\bar{\varepsilon}(t)||^2 \leq ||\bar{\varepsilon}(0)||^2 + \int_0^t ||\bar{\varepsilon}(\tau)||^2(2||A|| + 2||C||^2||r^{-1}||\ ||\bar{P}|| + \frac{f}{\bar{\theta}})d\tau,$$

and by 3.15, we know that $||\bar{P}(t)|| \leq \varphi_1(T_0) + ||\bar{P}_0||\varphi_2(T_0)$. Then, since $\bar{P}_0 = \frac{1}{\theta_0}\Delta P_0 \Delta(0)$, $\theta_0 > 1$, $||\bar{P}(t)|| \leq \varphi_1(T_0) + ||P_0||\varphi_2(T_0) \leq \varphi_1(T_0) + \frac{1}{c}\varphi_2(T_0) = \varphi(T_0, c)$.

$$||\bar{\varepsilon}(t)||^2 \leq ||\bar{\varepsilon}(0)||^2 + \bar{\nu}(T_0, c)\int_0^t ||\bar{\varepsilon}(\tau)||^2 d\tau,$$

and $\bar{\nu}(T_0, c)$ is a positive **decreasing** function of c.

Gronwall's inequality implies that:

$$||\bar{\varepsilon}(t)||^2 \leq \Psi(T_0, c)||\bar{\varepsilon}(0)||^2,$$

with: $\Psi(T_0, c) = e^{\bar{\nu}T_0}$, $\Psi(T_0, c)$ is also a decreasing function of c.
In particular, $||\bar{\varepsilon}(T_0)||^2 \leq \Psi(T_0, c)||\bar{\varepsilon}(0)||^2$. Plugging this in (3.14), we get:

$$||\bar{\varepsilon}(t)||^2 \leq \frac{\beta}{\alpha}e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})(t-T_0)}\Psi(T_0, c)||\bar{\varepsilon}(0)||^2, \text{ for } T \geq t \geq T_0. \tag{3.16}$$

Hence, for $T \geq t \geq T_0$,

$$||\bar{\varepsilon}(t)||^2 \leq \frac{\beta}{\alpha}e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})t}e^{Q_m\alpha T_0}\Psi(T_0, c)||\bar{\varepsilon}(0)||^2. \tag{3.17}$$

Going back to $t \leq T_0$, we have:

$$\begin{aligned} ||\bar{\varepsilon}(t)||^2 &\leq \Psi(T_0, c)||\bar{\varepsilon}(0)||^2 \leq \Psi(T_0, c)\frac{\beta}{\alpha}||\bar{\varepsilon}(0)||^2 \\ &\leq \frac{\beta}{\alpha}e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})t}e^{Q_m\alpha T_0}\Psi(T_0, c)||\bar{\varepsilon}(0)||^2, \end{aligned}$$

Hence, in all cases (either $t \leq T_0$ or $T_0 \leq t$), we have:

$$||\bar{\varepsilon}(t)||^2 \leq H(T_0, c)e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})t}||\bar{\varepsilon}(0)||^2,\ 0 \leq t \leq T, \tag{3.18}$$

with $H(T_0, c) = \frac{\beta}{\alpha}\Psi(T_0, c)e^{Q_m\alpha T_0}$, a decreasing function of c. Therefore, going back to the initial time τ, since $t = \int_0^\tau \theta(v)dv$, and $t \leq T$, then, $\tau \leq \tau(T)$, and $t \geq \theta(\tau(T))\tau$:

$$||\tilde{\varepsilon}(\tau)||^2 \leq H(T_0, c)e^{-(Q_m\alpha\theta(\tau(T)) - L')\tau}||\tilde{\varepsilon}(0)||^2, \tau(T) \geq \tau \geq 0,$$

if $\tilde{C} = Q_m\alpha\theta(\tau(T)) - L' > 0$, which is implied by

$$\theta_0 > e^{\lambda\tau(T)}(\frac{L'}{Q_m\alpha} - 1) + 1, \tag{3.19}$$

indeed, if (3.19) holds, since $\theta(\tau(T)) = \bar{\theta}(T) = 1 + (\theta_0 - 1)e^{-\lambda\tau(T)} > \frac{L'}{Q_m\alpha}$.

Since $\varepsilon = \Delta^{-1}\tilde{\varepsilon}$, and $\theta > 1$, $||\varepsilon(\tau)||^2 \leq ||(\Delta^{-1})||^2||\tilde{\varepsilon}(\tau)||^2 \leq \theta^{2(n-1)}||\tilde{\varepsilon}(\tau)||^2$, we get, for all $\tau_0 \geq \tau \geq 0$:

$$\begin{aligned} &||\varepsilon(\tau)||^2 \leq \theta^{2(n-1)}(\tau)H(T_0, c)e^{-(Q_m\alpha\theta(\tau_0) - L')\tau}||\varepsilon(0)||^2, \\ &\text{for } \theta_0 > e^{\lambda\tau_0}(\frac{L'}{Q_m\alpha} - 1) + 1, \\ &\text{or equivalently, } \theta(\tau_0) > \frac{L'}{Q_m\alpha}. \end{aligned}$$

$H(T_0, c)$ is a decreasing function of c.

This is the short term estimation (3.2). If $\lambda = 0$, it gives the standard high gain estimation.

3.2.3. *proof of the long term estimation.* Going back to (3.12), and using Lemma 3, in Section 5, we get, for all λ, $0 \leq \lambda < 1$, $t \geq T_0$,

$$\frac{d}{dt}(\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t)) \leq -k_1\ \bar{\varepsilon}'\bar{S}(t)\bar{\varepsilon} + k_2\ \bar{\theta}(t)^{(n-2)}||\bar{S}||\ ||\bar{\varepsilon}||^3,$$

where $k_1 = Q_m\alpha$, k_2 is a positive constant.

Lemma 1, applied to the Riccati equation in (3.10), implies:

$$\frac{d}{dt}(\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t)) \leq -k_1\ \bar{\varepsilon}'\bar{S}(t)\bar{\varepsilon} + k_2'\ \bar{\theta}^{(n-2)}\ ||\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t)||^{\frac{3}{2}}, \tag{3.20}$$

for another positive constant k_2'.

Now, we apply Lemma 2, in Section 5, to get that, for $t \geq T \geq T_0$:

$$\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t) \leq 4e^{-k_1(t-T)}\bar{\varepsilon}(T)'\bar{S}(T)\bar{\varepsilon}(T), \tag{3.21}$$

as soon as

$$(\mathfrak{P})\ \bar{\varepsilon}(T)'\bar{S}(T)\bar{\varepsilon}(T)\bar{\theta}(T)^{2(n-2)} \leq \frac{(k_1)^2}{4(k_2')^2}.$$

Setting, $q = \bar{\varepsilon}(T)'\bar{S}(T)\bar{\varepsilon}(T)\bar{\theta}(T)^{2(n-2)}$, let us use the short term estimation (3.18). It gives $q \leq \beta H(T_0, c)e^{-(Q_m\alpha - \frac{L'}{\theta(T)})T}||\bar{\varepsilon}(0)||^2\bar{\theta}(T)^{2(n-2)}$,

$$q \leq \beta H(T_0, c)e^{-(Q_m\alpha - \frac{L'}{\bar{\theta}(T)})T}||\bar{\varepsilon}(0)||^2\theta_0^{2(n-2)}.$$

If :

$$\theta_0 \geq e^{\lambda T}(\frac{2L'}{Q_m\alpha} - 1) + 1, \tag{3.22}$$

then $\frac{Q_m\alpha}{L'} - \frac{1}{\bar{\theta}(T)} \geq \frac{Q_m\alpha}{2L'}$. Indeed, in that case, $\bar{\theta}(T) \geq \theta(T) = 1 + (\theta_0 - 1)e^{-\lambda T} \geq \frac{2L'}{Q_m\alpha}$.

Then, let us chose $T = T^* = Log(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{1}{\lambda}} \geq T_0$ (in order to get the equality in (3.22)). This is possible, since we can assume from the very beginning that$\frac{2L'}{Q_m\alpha} - 1 > 0$ (we can increase L' for this) and$\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1} > e^{T_0} > e^{\lambda T_0}$ (we can take θ_0 large enough).

$$\begin{aligned} q &\leq \beta H(T_0, c)(\frac{\frac{2L'}{Q_m\alpha} - 1}{\theta_0 - 1})^{\frac{Q_m\alpha}{2\lambda}}||\bar{\varepsilon}(0)||^2\theta_0^{2(n-2)} \\ &\leq \beta H(T_0, c)||\bar{\varepsilon}(0)||^2(2(\frac{2L'}{Q_m\alpha} - 1))^{\frac{Q_m\alpha}{2\lambda}}\theta_0{}^{2(n-2) - \frac{Q_m\alpha}{2\lambda}}. \end{aligned}$$

Then, if:

$$\lambda < \frac{Q_m\alpha}{4(n-2)}, \tag{3.23}$$

for θ_0 large enough, for $||\varepsilon_0||$ bounded, q is arbitrarily small.

This means that the property $(\mathfrak{P})$ above is met at $T = T^*(\theta_0, \lambda)$, as soon as λ satisfies (3.23) and θ_0 is large enough.

In that case, (3.21) above holds, for $t \geq T^*$ $(\geq T_0)$:

$$\begin{aligned} \bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t) &\leq 4e^{-k_1(t-T^*)}\bar{\varepsilon}(T^*)'\bar{S}(T^*)\bar{\varepsilon}(T^*), \\ &\leq 4e^{-k_1 t}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1}{\lambda}}\bar{\varepsilon}(T^*)'\bar{S}(T^*)\bar{\varepsilon}(T^*). \end{aligned}$$

This implies, with (3.18):

$$\begin{aligned} ||\bar{\varepsilon}(t)||^2 &\leq 4\frac{\beta}{\alpha}e^{-k_1 t}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1}{\lambda}}||\bar{\varepsilon}(T^*)||^2, \\ &\leq 4\frac{\beta}{\alpha}H(T_0, c)e^{L'T^*}e^{-k_1 t}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1}{\lambda}}||\varepsilon_0||^2, \\ &\leq 4\frac{\beta}{\alpha}H(T_0, c)e^{-k_1 t}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1 + L'}{\lambda}}||\varepsilon_0||^2, \end{aligned}$$

for $t \geq T^*$ $(\geq T_0)$.

For $t \leq T^*$, using (3.18), and the fact that $k_1 = Q_m\alpha$:

$$\begin{aligned} ||\bar{\varepsilon}(t)||^2 &\leq H(T_0,c)e^{-k_1 t}e^{L't}||\varepsilon_0||^2, \\ &\leq H(T_0,c)e^{-k_1 t}4\frac{\beta}{\alpha}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2, \end{aligned}$$

because $\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1} > 1$.

Therefore, for all $t \geq 0$:

$$\begin{aligned} ||\bar{\varepsilon}(t)||^2 &\leq 4\frac{\beta}{\alpha}H(T_0,c)e^{-k_1 t}(\frac{\theta_0 - 1}{\frac{2L'}{Q_m\alpha} - 1})^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2, \\ &\leq \tilde{H}(T_0,c,\lambda)e^{-k_1 t}\theta_0^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2, \end{aligned}$$

where $\tilde{H}$ is a decreasing function of c. Hence:

$$||\tilde{\varepsilon}(\tau)||^2 \leq \tilde{H}(T_0,c,\lambda)e^{-k_1 t}\theta_0^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,$$

and, with $t = \tau + \frac{\theta_0-1}{\lambda}(1-e^{-\lambda\tau})$,

$$||\tilde{\varepsilon}(\tau)||^2 \leq \tilde{H}(T_0,c,\lambda)e^{-k_1\tau}\theta_0^{\frac{k_1+L'}{\lambda}}e^{-k_1\frac{\theta_0-1}{\lambda}(1-e^{-\lambda\tau})}||\varepsilon_0||^2.$$

Finally,

$$||\varepsilon(\tau)||^2 \leq \bar{H}(T_0,c,\lambda)e^{-k_1\tau}||\varepsilon_0||^2\theta_0^{\frac{k_1+L'}{\lambda}+2(n-1)}e^{-k_1\frac{\theta_0}{\lambda}(1-e^{-\lambda\tau})},$$

where $\bar{H}(T_0,c,\lambda)$ is a decreasing function of c.

This is the long term estimation. It holds as soon as λ satisfies (3.23), and for θ_0 large, depending on λ.

3.3. **Practical implementation: making the observer "recursive".** We consider a one parameter family $\{O_\tau, \tau \geq 0\}$ of observers of type (3.1), indexed by the time, each of them starting from S_0, θ_0, at the current time τ. In fact, in practice, it will be sufficient to consider, at time τ, a slipping window of time, $[\tau - T, \tau[$, and a finite set of observers $\{O_{t_i}, \tau - T \leq t_i \leq \tau\}$, with $t_i = \tau - i\frac{T}{N}$, $i = 1, ..., N$.

As usual, we call the term $I(\tau) = \hat{y}(\tau) - y(\tau)$, (the difference at time τ between the estimate output and the real output), the **"innovation".** Here, for each observer O_{t_i},we have an innovation $I_{t_i}(\tau)$.

Our suggestion (very natural and very simple), is to take as the estimate of the state, the estimation given by the observer O_{t_i} that minimizes the absolute value of the innovation.

Let us analyze what will be the effect of this procedure in a deterministic setting:

1. Let us assume that there is no "jump" of the state. Then, clearly, the best estimation will be given by the "oldest" observer in the window, O_{t_N}. Then, the error will be given by the "long term" and "short term" estimates at time T :

$$\begin{aligned} ||\varepsilon(\tau+T)||^2 &\leq R(\lambda,c)e^{-a\ T}||\varepsilon(\tau)||^2\Lambda(\theta_0,T,\lambda), \\ ||\varepsilon(\tau+T)||^2 &\leq \theta(T)^{2(n-1)}H(c)e^{-(a_1\theta(T)-a_2)T}||\varepsilon(\tau)||^2. \end{aligned}$$

a. If T is large enough, the asymptotic behavior will be the one of the "extended Kalman filter".

b. At the beginning, the transient is the one of the HGEKF.

c. the error can be made arbitrarily small in arbitrary short time, provided that θ_0 is large enough.

2. If at a certain time we have a "jump" of the state, then, the innovation of the "old observers" will become large. The "youngest" one will be chosen, and the transient will be the same as the one of the HGEKF, first, and of the EKF, after T.

This looks very promising. We show on an example in the next section, that it works very well.

4. Application: observation of a binary distillation column

4.1. **The constant molar overflow model.** The model we consider is the classical "constant molar overflow" (CMO) model. It is one of the most simple distillation models, and it is used by many process engineers (for instance, even in its static form, it is used for simple short-cut distillation calculations).

Since everything here follows from the very special "tridiagonal" structure of this model, and since any reasonable distillation model possesses such a structure, all what we do in this paper can certainly be extended to more precise distillation models.

The equations are based upon:
a. a thermodynamical relation describing the thermal equilibria for each tray.
b. Material balances on each plate.

Thermal balance on each plate is replaced by the "Lewis hypotheses", that lead to the fact that the liquid and vapor flowrates along the column are constant in the "stripping" (above the feed) and "rectification" (below the feed) zones. For justification of these assumptions, see [15].

The equations of this model are:
Total condenser:

$$H_1 \frac{dx_1}{dt} = (V + FV)(y_2 - x_1). \tag{4.1}$$

Rectifying section, $j = 2, \cdots, f-1$:

$$H_j \frac{dx_j}{dt} = L(x_{j-1} - x_j) + (V + FV)(y_{j+1} - y_j). \tag{4.2}$$

Feed tray:

$$\begin{aligned} H_f \frac{dx_f}{dt} &= FL(Z_F - x_f) + FV(k(Z_F) - y_f) \\ &\quad + L(x_{f-1} - x_f) + V(y_{f+1} - y_f). \end{aligned} \tag{4.3}$$

Stripping section, $j = f+1, \cdots, n-1$:

$$H_j \frac{dx_j}{dt} = (L + FL)(x_{j-1} - x_j) + V(y_{j+1} - y_j). \tag{4.4}$$

Bottom of the column:

$$H_n \frac{dx_n}{dt} = (L + FL)(x_{n-1} - x_n) + V(x_n - y_n). \tag{4.5}$$

The parameters have the following physical meaning:

n	number of trays,
f	index of the feed tray,
H_j	liquid hold up on the j^{th} tray (a geometric constant),
x_j	liquid composition on the j^{th} tray,
y_j	vapor composition on the j^{th} tray,
FL, FV, L, V	feed (liquid and vapor), reflux and vapor flow,
Z_F	feed composition (molar fraction of light component in feed).

On each tray the liquid and vapor compositions, x_j and y_j, are linked by the liquid/vapor equilibrium law $y_j = k(x_j)$. We assume that the function k is monotonic, *i.e.* we do not consider azeotropic distillation.

Each of the equations is relative to a tray. It just expresses the accumulation of the liquid on the corresponding tray, and the thermodynamical equilibrium.

The condenser and the bottom of the column are assimilated to tray 1 and tray n respectively. The state of the model is the liquid composition profile of the more volatile component on each tray, denoted by (x_j).

The top and bottom product compositions x_1 and x_n are the two observed variables. In practice, they are also the two variables that one wants to control: they are the "qualities" of the products going out of the column.

The two control variables are the reflux flow-rate L and the vapor flow-rate V.

There are also two disturbances to be counteracted:

a. changes in the feed rate $F = FL + FV$. In general this is a "measured disturbance", (a flowrate measurement),

b. the in-feed composition Z_F. In general, it is unknown, and it is practically very expensive to "observe it". Moreover, it may change brutally. We will consider this feed composition Z_F as an unknown (constant) state variable. When Z_F changes, the consequence is **a jump of the state of the system.**

The qualitative properties of this model are very nice (see [10], [18], [17]):

a. For positive control variables L and V, (negative doesn't physically makes sense), the "physical" domain $D = [O, 1]^n$ is positively invariant under the dynamics. This means that all the state variables x_j remain between 0 and 1.

b. In the domain D, all other variables (than the x_i's and the y_i's) being constant, **there is a single equilibrium, which is globally asymptotically stable.**

c. It has very nice observability properties, as will be discussed later on.

Our goal in this section is to construct an estimator of the state x, and more specifically of the feed composition Z_F, by using the results of the previous sections.

4.2. **Observability of the model and synthesis of the observer.** A complete analysis of observability and observer synthesis has been carried out in [10] in the general case. It happens that, even if the feed is considered as an unknown state variable (meeting the equation $\frac{dZ_F}{dt} = 0$), the model is observable in the strongest possible sense. In particular, as we shall see, it can be put in a normal form similar to (1.1).

Our purpose here is just to apply the observer described in the previous sections. Hence, we will fix a special case of distillation column. But all what we show works in general. We will chose:

- $n = 5$ and $f = 3$,
- The function k is a diffeomorphism from $[0, 1]$ into itself and is given by,
$$k(x) = \frac{\alpha x}{1 + (\alpha - 1) x}.$$
Here α is the "relative volatility" of the mixture. It is a physical parameter larger than 1 (but close to 1). The closer to one, the most difficult distillation. If $\alpha = 1$, the two products are thermodynamically identical, and cannot be distillated (the model is not controllable).
- Let us observe that k is a diffeomorphism from $\left]-\frac{1}{\alpha-1}, +\infty\right[$ to $\left]-\infty, \frac{\alpha}{\alpha-1}\right[$.
- The feed is assumed to enter the column at its "bubble point". As a consequence, $F = FL$.

Let us make the following change of state variables: $\xi_1 = x_1$, $\xi_2 = k(x_2)$, $\xi_3 = x_3$, $\xi_4 = x_4$, $\xi_5 = x_5$ and $\xi_6 = Z_F$.

Then, the system can be rewritten as:

$$\left\{ \begin{array}{lcl} H_1 \frac{d\xi_1}{dt} & = & V(\xi_2 - \xi_1), \\ H_2 \frac{d\xi_2}{dt} & = & k'\left(k^{-1}(\xi_2)\right)\left(L(\xi_1 - k^{-1}(\xi_2)) + V(k(\xi_3) - \xi_2)\right), \\ H_3 \frac{d\xi_3}{dt} & = & F(\xi_6 - \xi_3) + L(k^{-1}(\xi_2) - \xi_3) + V(k(\xi_4) - k(\xi_3)), \\ H_4 \frac{d\xi_4}{dt} & = & (L + F)(\xi_3 - \xi_4) + V(k(\xi_5) - k(\xi_4)), \\ H_5 \frac{d\xi_5}{dt} & = & (L + F)(\xi_4 - \xi_5) + V(\xi_5 - k(\xi_5)), \\ H_6 \frac{d\xi_6}{dt} & = & 0, \end{array} \right. \tag{4.6}$$

or:

$$\frac{d\xi_t}{dt} = A(L, V)\xi_t + \widetilde{b}(L, V, \xi_t), \tag{4.7}$$

where,

$$A(L, V) = \begin{pmatrix} 0 & \frac{V}{H_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{F}{H_3} \\ 0 & 0 & \frac{L+F}{H_4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{L+F}{H_5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and,

$$\widetilde{b}\left(L,V,\xi\right)=\left(\begin{array}{c}-\frac{V}{H_1}\xi_1\\ k'\left(k^{-1}\left(\xi_2\right)\right)\left(L(\xi_1-k^{-1}\left(\xi_2\right))+V(k\left(\xi_3\right)-\xi_2)\right)\diagup H_2\\ \left(-F\xi_3+L(k^{-1}\left(\xi_2\right)-\xi_3)+V(k\left(\xi_4\right)-k\left(\xi_3\right))\right)\diagup H_3\\ \left(-(L+F)\xi_4+V(k\left(\xi_5\right)-k\left(\xi_4\right))\right)\diagup H_4\\ \left(-(L+F)\xi_5+V(\xi_5-k\left(\xi_5\right))\right)\diagup H_5\\ 0\end{array}\right),$$

$$=\left(\begin{array}{c}\widetilde{b}_1\left(V,\xi_1\right)\\ \widetilde{b}_2\left(L,V,\xi_1,\ldots,\xi_5\right)\\ \widetilde{b}_3\left(L,V,\xi_3,\xi_4,\xi_5\right)\\ \widetilde{b}_4\left(L,V;\xi_4,\xi_5\right)\\ \widetilde{b}_5\left(L,V,\xi_5\right)\\ 0\end{array}\right).$$

The observations are then given by

$$y=\left(\begin{array}{cccccc}1&0&0&0&0&0\\ 0&0&0&0&1&0\end{array}\right)\xi=C\xi.$$

Now, since in fact the only pertinent **(and positively invariant)** part of the state space is $D'=[0,1]^6$, we can manage the things for $\tilde{b}$ be compactly supported, as in section 1.1, and unchanged on D'. Let us change $\widetilde{b}\left(L,V,\xi\right)$ in the following way outside $[0,1]^6$: replace $\widetilde{b}\left(L,V,\xi\right)$ by $b\left(L,V,\xi\right)=\widetilde{b}\left(L,V,\Phi\left(\xi\right)\right)$ where $\Phi\left(\xi_1,\ldots,\xi_6\right)=\left(\varphi\left(\xi_1\right),\ldots,\varphi\left(\xi_6\right)\right)$ and $\varphi\left(\xi\right)$ is any C^∞ function from $\mathbb{R}$ to $[0,1]$ equal to one in $[0,1]$ and equal to zero outside $\left]-\frac{1}{\alpha-\frac{1}{2}},\frac{\alpha}{\alpha-\frac{1}{2}}\right[$. This modification does not change the "physical trajectories".

Our system has the property to be *observable for any input*, as soon as the control variables L and V are >0. Here, we assume that L,V are bounded from below (and from above) by >0 constants:

$$L_M\geq L(t)\geq\varepsilon_1>0,\quad V_M\geq V(t)\geq\varepsilon_2>0.$$

This assumption is the analog of the assumption $0<a_m\leq a_i(u)\leq a_M$, in section 1.1. It is a realistic requirement from the physical point of view.

To finish, let us point out the fact that we are in **case 1** of section 2.1 above (i.e. the nongeneric case): The number of observations is equal to the number of control variables (it is 2).

Due to these observability properties, we will be able to apply the observer of the previous section 3.3. In fact, it will be an adaptation of the results of section 3, Theorem 1, to this multi-output case.
We leave the reader to check (this is really straightforward) that all the reasoning in the proof of Theorem 1 can be strictly repeated, and that the statements of this theorem are valid for the distillation column.

Of course, in practice, we didn't compute the theoretical bounds λ_0 and $\theta_0(\lambda)$. We have just got some values for them by experimentation. Also, the number N of "parallel" observers, and the "sampling times" t_i of section 3.3 have been chosen experimentally.

Finally, the state of our observer is the collection of the states of N independent observers $\left(z_i,S_i,\theta_i\right)_{i=1,\ldots,N}$. Each observer is a set of three equations of the following form:

$$\left\{\begin{array}{lll}\frac{dz}{dt}&=&A\left(u\right)z+b\left(u,z\right)-S\left(t\right)^{-1}C^TR_\theta^{-1}\left(Cz-y\left(t\right)\right)\\ \frac{dS}{dt}&=&-\left(A\left(u\right)+b^*\left(z,u\right)\right)'S-S\left(A\left(u\right)+b^*\left(z,u\right)\right)+C'R_\theta^{-1}C-SQ_\theta S\\ \frac{d\theta}{dt}&=&\lambda\left(1-\theta\right)\end{array}\right.$$

where $u=\left(L,V\right)$.

Due to the multi-output structure, with "Brunovsky-like" blocks of different dimensions (4 and 2), a way to make the proof of Theorem 1 work, is to take a matrix R depending also on θ, as shown below. This could be avoided by increasing the dimension of the state as explained in [10].

It is not hard to check that a good choice is to set:

$$\Delta = \operatorname{diag}\left(\frac{1}{\theta^2}, \frac{1}{\theta^3}, \frac{1}{\theta^2}, \frac{1}{\theta}, 1, \frac{1}{\theta^3}\right)$$

with $Q_\theta = \theta^2 \Delta^{-1} Q \Delta^{-1}$ and $R_\theta = \left(C\Delta^{-1}C'\right) R \left(C\Delta^{-1}C'\right)$.

In practice, we have chosen $N = 5$ observers, and we have taken a regular sampling $\frac{T}{N}$. That is to say, at each time step $k\frac{T}{N}$, the oldest observer is replaced by a new one (with $\theta = \theta_0$ and a new guess of state and covariance matrix). At the beginning of the simulation, we chose an initial value θ_0 of θ for each observer, such that the i^{th} observer has $\theta_i = 1 + e^{-\lambda\frac{(i-1)T}{N}}(\theta_0 - 1)$, see figure 3, where "crosses" represent reinitializations.

We have implemented our observer as described in the previous section. Since the state has dimension 6, each observer requires to solve 28 ordinary differential equations (for the state, the Riccati matrix, and the very simple equation for θ). Finally, our observer is a set of 140 ODE's. We have solved it in conjunction with the model (6 equations) using LSODAR from ODEPACK ([14]), without taking into account the possibility of decoupling these equations (which are indeed equivalent to five systems of 34 equations, including the model into each system). A simulation of 3 hours of real time takes about 40 seconds on a Pentium III machine.

4.3. **Simulation results.** We have chosen the following constant parameters:

- Hold-up $H_1 = 40$, $H_j = 10$ for $j = 2, 3, 4$ and $H_5 = 80$,
- Relative volatility $\alpha = 2$.

We have applied the following scenario:

- During the simulation, the state noise is simulated by the sum of several sine functions at some random frequencies representing a band limited noise with an amplitude of 10^{-8} before the time $t_2 = 116\,\mathrm{mn}\,40\,\mathrm{s}$ and 10^{-2} after this time,

- Moreover, at time $t_1 = 66\,\mathrm{mn}\,40\,\mathrm{s}$, we simulate a step in the feed quality Z_F from 0.45 to 0.60. Hence we can consider that there is no perturbation before time t_1, where a large "jump of the state" occurs,

- after that, nothing happens until time t_2 where a periodic perturbation on Z_F is applied.

We have also added a measurement noise at some random high frequencies and with amplitude of 10^{-2}. The effect of noise can be seen on Figure 1 (top and bottom lines).

To make the simulation more realistic, we have applied a very simple controller, which calculates the inputs L and V in order to regulate top and bottom qualities at a reasonable level (that is, 73% for the top quality and 23% for the bottom quality).

As we said already, the parameters of the observers where tuned in order to obtain good performances, and not caring about the theoretical bounds.

Practically, we have used $\theta_0 = 10$, $\frac{T}{N} = 10\,\mathrm{mn}$ and $\lambda = \frac{1}{600}\,\mathrm{s}^{-1}$, in such a way that the time of life of an observer is $T = 50\,\mathrm{mn}$, and then an old observer has $\theta \approx 1.16$. Also, there is always an observer with $\theta > 4.3$ which is running.

Finally, R is equal to 10^{-2} times the 2×2–identity matrix and Q is 10^{-9} times the 6×6–identity matrix.

First of all, the behavior of the observer is very good during the unmodelled transient as well as during smooth operation, see Figure 1: top and bottom quality measurements are plotted, as well as the unknown feed quality, each curve being represented by a continuous line. The overall estimation of the feed quality, corresponding to the estimation of the feed quality provided by the observer with the **smallest innovation,** is represented by a dashed line. It is very close to the actual feed quality.

A more accurate plot is presented on Figure 2 where we have only shown the relative estimation error of the feed quality. The estimation provided by the best observer (in our sense, that is to say, the observer with minimal innovation) is the continuous line. The crosses represent the estimation of Z_F provided by other observers every minute. One can see that our criteria on the innovation to select the right observer is a good choice, at least in this case.

Moreover, the behavior of the observer is very close to what we expected from the theoretical results:

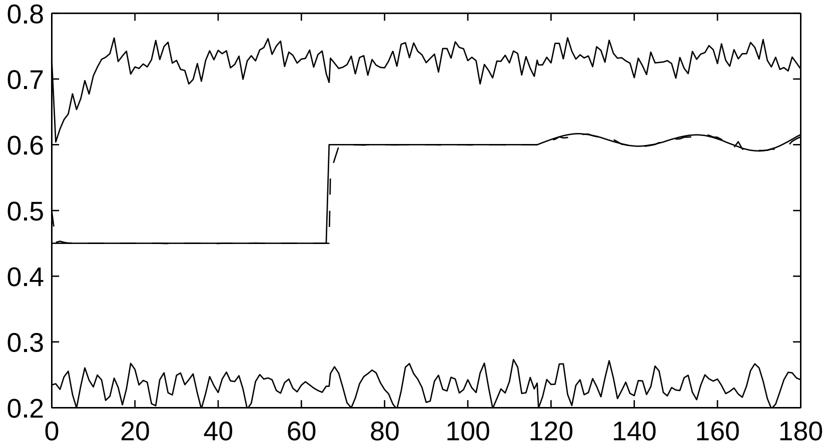


Figure 1. Measured output and estimation of the feed quality.

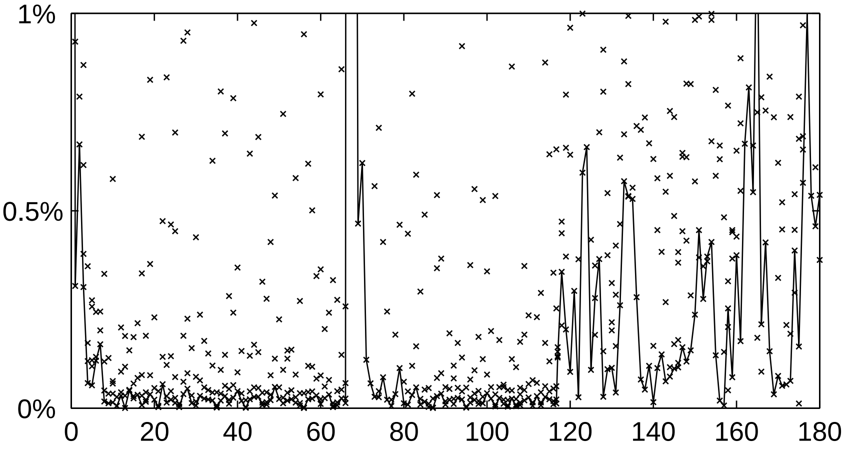


Figure 2. Relative error between the actual feed quality and its estimation by the selected observer (continuous line) and the others.

- When no perturbation arises, the best observer (that is to say the observer with the smallest innovation) is the one with the smallest value of θ *i.e.* the oldest observer which is also the observer which is the closest to the pure extended Kalman observer.

-If a large perturbation occurs (such as the feed change at time $t_1 = 66\,\text{mn}\,40\,\text{s}$), the best observer becomes the youngest one, *i.e.* the observer with the highest θ.

-Of course, small perturbations are well corrected by oldest or intermediate observers. This is very clear on the figure 4.

Our conclusion, from these simulations, is that even if the use of several observers in parallel requires the introduction of new tuning parameters (θ_0, λ, N and T), the choice of these new parameters is very easy, due to their very clear effect on the results.

From a practical point of view, θ_0, λ, N and T have to be chosen such that at any time, there is an HGEKF and an EKF-like observer running at the same time, that is to say such that $1 + e^{-\lambda\frac{T}{N}}(\theta_0 - 1)$ is large enough (to ensure that at least one observer is a HGEKF) and such that $1 + e^{-\lambda T}(\theta_0 - 1)$ is close to 1.

Also, an important point, for people that are used to tune Kalman's observers, is that the choice of the Q and R matrices is less crucial than with a single observer which has to be tuned in order to be efficient both with and without perturbations.

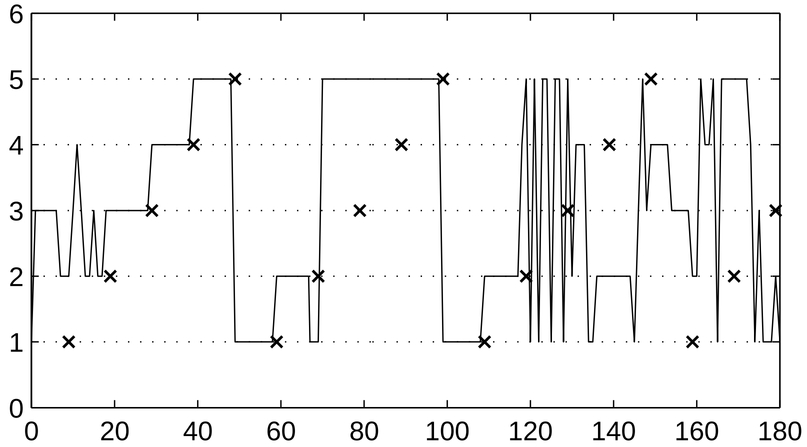


Figure 3. The 5 observers. Time of reinitialization of each observer (×), and the best one (continuous line).

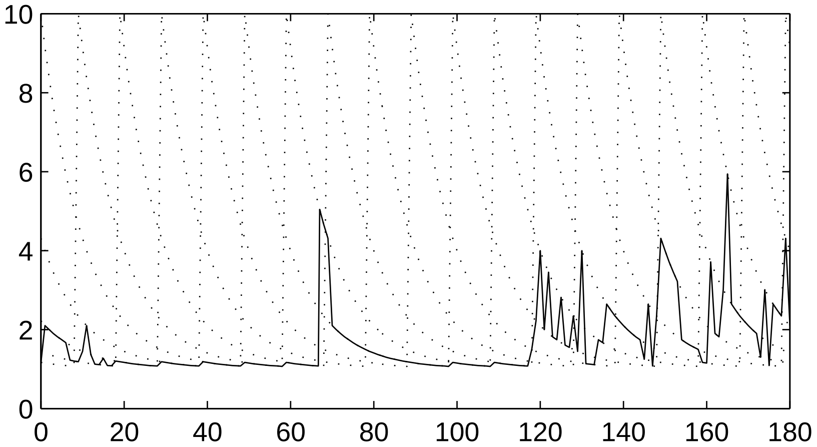


Figure 4. Various values of θ versus time (dotted lines), and best observer (continuous line).

Moreover, this approach allows us to obtain a diagnosis of abnormal behavior: if the smallest innovation is provided by the last reinitialized observer then one can conclude that the model has encountered a perturbation. If this happen for a long time then one can conclude that the model has some difficulties to deal with certain unmodelled perturbations. Indeed, the scenario that we have applied in our simulations can be easily deduced from the figure 4.

5. Appendix. Technical lemmas

Lemma 2. *Let $\{x(t) > 0,\, t \geq 0\} \subset \mathbb{R}^n$ be absolutely continuous, and satisfying:*

$$\frac{dx}{dt} \leq -\lambda x + kx\sqrt{x},$$

for almost all $t > 0$, for $\lambda, k > 0$. Then, as soon as $x(0) < \frac{\lambda^2}{4k^2}$, $x(t) \leq 4x(0)e^{-t\lambda}$.

Proof. We make the successive following changes of variables: $y = \sqrt{x}, z = 1/y$, $w(t) = e^{-\frac{\lambda}{2}t}z(t)$. Then, all the quantities $y(t)$, $z(t)$, $w(t)$ are positive and absolutely continuous, on any finite time interval $[0, T]$. We denote by $'$ the derivatives with respect to time.

Straightforward computations give, for almost all $t > 0$:

$$\begin{aligned} y' &\leq -\frac{\lambda}{2}y + \frac{k}{2}y^2, \\ z' &\geq \frac{\lambda}{2}z - \frac{k}{2}, \\ w' &\geq -e^{-\frac{\lambda}{2}t}\frac{k}{2}. \end{aligned} \tag{5.1}$$

Moreover, $w(0) = \frac{1}{\sqrt{x(0)}}$. Then, for all $t > 0$,

$$w(t) \geq \frac{1}{\sqrt{x(0)}} - \frac{k}{\lambda} + \frac{k}{\lambda}e^{-\frac{\lambda}{2}t}. \tag{5.2}$$

If $\frac{1}{\sqrt{x(0)}} - \frac{k}{\lambda} > 0$, then $w(t) > 0$,and we can go backwards in the previous inequalities:

$$\begin{aligned} w(t) &\geq \frac{1}{\sqrt{x(0)}} - \frac{k}{\lambda}(1 - e^{-\frac{\lambda}{2}t}), \\ z(t) &\geq e^{\frac{\lambda}{2}t}(\frac{1}{\sqrt{x(0)}} - \frac{k}{\lambda}) + \frac{k}{\lambda}, \\ y(t) &\leq \frac{1}{e^{\frac{\lambda}{2}t}(\frac{1}{\sqrt{x(0)}} - \frac{k}{\lambda}) + \frac{k}{\lambda}}, \\ x(t) &\leq \frac{x(0)e^{-\lambda t}}{(1 - \sqrt{x(0)}\frac{k}{\lambda})^2}. \end{aligned}$$

Hence, if $x(0) \leq \frac{\lambda^2}{4k^2}$, or $1 - \sqrt{x(0)}\frac{k}{\lambda} \geq \frac{1}{2}$, then:

$$x(t) \leq 4x(0)e^{-\lambda t}.$$

□

Lemma 3. *Let $B = \tilde{b}(z) - \tilde{b}(x) - \tilde{b}^*(z)\varepsilon$ be as in Section 3: $\varepsilon = z - x$, $\tilde{b}(x) = \Delta b(\Delta^{-1}x)$, $\tilde{b}^*(z) = \Delta b^*(\Delta^{-1}x)\Delta^{-1}$, where $b^*(x)$ is the Jacobian matrix of b at x, and where b is compactly supported. $\Delta = diag(1, \frac{1}{\theta}, ..., \frac{1}{\theta^{n-1}})$, $\theta \geq 1$. Then, $||B|| \leq K\ \theta^{n-1}||\varepsilon||^2$, for some $K > 0$.*

Proof. Let us consider a smooth expression $E(z, x)$ of the form:

$$E(z, x) = f(z) - f(x) - df(z)\varepsilon, \text{ with } \varepsilon = z - x,$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is compactly supported.

We have, for $t > 0$:

$$f(z - t\varepsilon) = f(z) - \sum_{i=1}^{p} \varepsilon_i \int_0^t \frac{\partial f}{\partial x_i}(z - \tau\varepsilon)d\tau,$$

and:

$$\frac{\partial f}{\partial x_i}(z - \tau\varepsilon) = \frac{\partial f}{\partial x_i}(z) - \sum_{j=1}^{p} \varepsilon_j \int_0^\tau \frac{\partial^2 f}{\partial x_i \partial x_j}(z - \theta\varepsilon)d\theta.$$

Hence,

$$f(z - \varepsilon) = f(z) - \sum_{i=1}^{p} \varepsilon_i \frac{\partial f}{\partial x_i}(z) + \sum_{i,j=1}^{p} \varepsilon_i\varepsilon_j \int_0^1 \int_0^\tau \frac{\partial^2 f}{\partial x_i \partial x_j}(z - \theta\varepsilon)d\theta d\tau.$$

Since f is compactly supported, we get:

$$|f(z) - f(z - \varepsilon) - df(z)\varepsilon| \leq \frac{M}{2}\sum_{i,j=1}^{p} |\varepsilon_i\varepsilon_j|,$$

where $M = \sup_x |\frac{\partial^2 f}{\partial x_i \partial x_j}(x)|$.

Now, we take $f = \tilde{b}_k$, and we use the facts that $\tilde{b}_k$ depends only on $x_1, ..., x_k$, and that $\theta \geq 1$:

$$|\frac{\partial^2 \tilde{b}_k}{\partial x_i \partial x_j}(x)| \leq \theta^{k-1}|\frac{\partial^2 b_k}{\partial x_i \partial x_j}(\Delta^{-1}x)|.$$

This gives the result. □

References

[1] M. BALDE, P. JOUAN, Observability of control affine systems, ESAIM/COCV, Vol. 3, pp. 345-359, 1998.

[2] J.S. BARAS, A. BENSOUSSAN, M.R. JAMES, "*Dynamic observers as asymptotic limits of recursive filters: special cases*", SIAM J. Appl. Math., **48**, (1988), 1147–1158.

[3] R. BUCY, P. JOSEPH, Filtering for stochastic processes with applications to guidance, Chelsea publishing company, 1968, second edition, 1987.

[4] F. DEZA, Contribution to the synthesis of exponential observers, Phd thesis, INSA de Rouen, France, June 1991.

[5] F.DEZA, E.BUSVELLE, J.P.GAUTHIER, High-gain estimation for nonlinear systems, Systems and Control Letters 18, pp. 295-299, 1992.

[6] J.P. GAUTHIER, H. HAMMOURI, I. KUPKA, Observers for nonlinear systems; IEEE CDC Conference, december, 1991, pp. 1483-1489; Brighton, England.

[7] J.P GAUTHIER, H. HAMMOURI, S. OTHMAN, A simple observer for nonlinear systems. IEEE Trans. Aut. Control, 37, pp. 875-880, 1992.

[8] J.P. GAUTHIER, I. KUPKA, Observability and observers for nonlinear systems. SIAM Journal on Control, vol. 32, N° 4, pp. 975-994, 1994.

[9] J.P. GAUTHIER, I. KUPKA, Observability for systems with more outputs than inputs. Mathematische Zeitschrift, 223, pp. 47-78, 1996.

[10] J.P. GAUTHIER, I. KUPKA, Deterministic observation theory and applications, book, to appear at Cambridge university press.

[11] A. JASWINSKY, Stochastic processes and filtering theory, Academic Press, New York, 1970.

[12] P. JOUAN, Singularités des systèmes non linéaires, observabilité et observateurs, PHD thesis, Université de Rouen, 1995.

[13] P. JOUAN, J.P. GAUTHIER, Finite singularities of nonlinear systems. Output stabilization, observability and observers. Journal of Dynamical and Control Systems, vol. 2, N° 2, 1996, pp. 255-288.

[14] A. C. HINDMARSCH, *Odepack, a systematized collection of ode solvers*, in scientific computing, r. s. Stepleman et al. (eds.), North-Holland, Amsterdam, 1983, pp. 55-64

[15] C.D. HOLLAND, Multicomponent Distillation, Englewood Cliffs, New-Jersey, USA: Prentice Hall, 1963.

[16] J. PICARD, "*Efficiency of the extended Kalman filter for nonlinear systems with small noise*", SIAM J. Appl. Math., **51**, No3, (1991), 843–885.

[17] H.H. ROSENBROCK, A Lyapunov function with applications to some nonlinear physical systems, Automatica, 1, pp. 31-53, 1962.

[18] P. ROUCHON, *Simulation dynamique et commande non linéaire des colonnes à distiller*, Thèse de l'école des mines de Paris, 1990

[19] F. VIEL, Stabilité des systèmes non linéaires controlés par retour d'état estimé. Application aux réacteurs de polymérisation et aux colonnes à distiller, Thèse de l'université de Rouen, 1994.

[20] F. VIEL, E. BUSVELLE, J.P. GAUTHIER, A stable control structure for binary distillation columns, International Journal on Control, Vol 67, N° 4, pp. 475-505, 1997.

ON DETERMINING UNKNOWN FUNCTIONS IN DIFFERENTIAL SYSTEMS, WITH AN APPLICATION TO BIOLOGICAL REACTORS.

ERIC BUSVELLE, JEAN-PAUL GAUTHIER

Abstract. In this paper, we consider general nonlinear systems with observations, containing a (single) unknown function φ. We study the possibility to learn about this unknown function via the observations: if it is possible to determine the [values of the] unknown function from any experiment [on the set of states visited during the experiment], and for any arbitrary input function, on any time interval, we say that the system is "identifiable".

For systems without controls, we give a more or less complete picture of what happens for this identifiability property. This picture is very similar to the picture of the "observation theory" in [7]:

1. If the number of observations is three or more, then, systems are generically identifiable.

2. If the number of observations is 1 or 2, then the situation is reversed. Identifiability is not at all generic. In that case, we add a more tractable infinitesimal condition, to define the "infinitesimal identifiability" property. This property is so restrictive, that we can almost characterize it (we can characterize it by geometric properties, on an open-dense subset of the product of the state space X by the set of values of φ). This, surprisingly, leads to a non trivial classification, and to certain corresponding "identifiability normal forms".

Contrarily to the case of the observability property, in order to identify in practice, there is in general no hope to do something better than using "approximate differentiators", as show very elementary examples. However, a practical methodology is proposed in some cases. It shows very reasonable performances.

As an illustration of what may happen in controlled cases, we consider the equations of a biological reactor, [2], [4], in which a population is fed by some substrate. The model heavily depends on a "growth function", expressing the way the population grows in presence of the substrate. The problem is to identify this "growth function". We give several identifiability results, and identification methods, adapted to this problem.

1. Introduction, main results, organization of the paper

1.1. **Systems under consideration.** In this paper, depending on the context, smooth will mean C^ω (real-analytic), or C^∞, or C^k, for an integer k large enough.

Very often, in nonlinear control systems, certain special variables with "physical meaning" appear, (called here the "**internal variables**") and the system depends on certain functions of these variables. These functions describe some physical characteristic inside the system, and it may happen that they are not well known, and have to be determined on the basis of experiments.

We consider general smooth nonlinear systems:

$$\Sigma : \begin{cases} \frac{dx}{dt} = f(x, u(t), \varphi \circ \pi(x(t))); \\ y = h(x, u(t), \varphi \circ \pi(x(t))), \end{cases} \tag{1.1}$$

or such systems that are "**uncontrolled**":

$$\begin{aligned} \frac{dx}{dt} &= f(x, \varphi \circ \pi(x(t))); \\ y &= h(x, \varphi \circ \pi(x(t))), \end{aligned} \tag{1.2}$$

where $x \in X$ denotes the state, y denotes the observation, $u(\cdot)$ is the control function, $z = \pi(x)$ is the "**internal variable**", π is called the **internal mapping**, $\pi : X \to Z$. We assume that X and Z are given analytic connected manifolds, both Hausdorf and paracompact. Here $\dim(X) = n$.

The (smooth) "**unknown function**" is denoted by $\varphi : Z \to I$. Here, I will denote a compact interval of $\mathbb{R}$. To finish, $y \in \mathbb{R}^{d_y}$, $u \in U \subset \mathbb{R}^{d_u}$, where U is a compact subanalytic subset of $\mathbb{R}^{d_u}$. Also, f is a $(u, \varphi)-$parametrized smooth vector field, The **observation mapping**, h, is a smooth mapping: $X \times U \times I \to \mathbb{R}^{d_y}$.

Date: April, 2003.

In the following, the systems $\Sigma = (f, h)$ will vary, but the manifolds X, Z, the (smooth) internal mapping π, the space $\mathbb{R}^{d_y}$, and the sets I and U are given and fixed.

1.2. **Rough definitions and statement of the main theoretical results.** Our main results in this paper are about uncontrolled systems. For these systems, we provide an almost complete theory of identifiability. Controlled systems may have very different identifiability properties. Just to show what may happen for controlled systems, we will treat a practical example: the model of a "biological reactor". In fact, the general study in this paper was originally motivated by this very interesting elementary example.

1.2.1. *Definitions.* Roughly speaking, we say that a system is **identifiable** if, for each experiment (input-output trajectory), one can reconstruct the piece (visited during the experiment) of the graph of the unknown function, on the basis of the input-output data.

Given any trajectory of the system:

$$\mathcal{T} = \{(u(t), \varphi \circ \pi(x(t)) = \hat{\varphi}(t), x(t), y(t))\},$$

defined on some time interval $[0, T]$, consider $\hat{\varphi}(t)$ as an (unknown) extra input of the system. The control function $u(\cdot)$ being fixed, we have a well defined mapping:

$$P_{\Sigma, u(\cdot)} : (\hat{\varphi}(\cdot), x_0) \to y(\cdot),$$

where x_0 is the initial condition for the system. Then, **identifiability** of Σ means that all these mappings (depending on $u(\cdot)$) are injective.

Our approach to the analysis of the identifiability property is very similar to the approach in [7], for the analysis of the observability property. The results we obtain are in the same spirit.

The injectivity of maps being not a stable property, it is not easy to handle. Hence it is reasonable to add to the injectivity requirement the additional requirement of "infinitesimal injectivity" (i.e. injectivity of all the tangent mappings $TP_{\Sigma, u(\cdot)}$), to get a more tractable notion of identifiability. This notion is called **uniform infinitesimal identifiability**.

We denote by $j^k y = (y(0), y'(0), \ldots, y^{(k-1)}(0))$, the k-jet at $t = 0$ of a smooth function $y(t)$. Another way to manage with the identifiability property is to consider the notion of **differential identifiability**, i.e. identifiability at the level of the k-jets. For this, consider the mappings:

$$j^k P_{\Sigma, u(\cdot)} : (j^k \hat{\varphi}, x_0) \to j^k y.$$

The system is said differentially identifiable (of order k), if:

$$j^k y^1 = j^k y^2 \Longrightarrow (\hat{\varphi}^1(0), x_0^1) = (\hat{\varphi}^2(0), x_0^2).$$

For uncontrolled systems, k-differential identifiability implies identifiability, **for all smooth internal mappings** π. (For controlled systems, the situation is much more complicated. We refer to [7] for similar questions in the context of the observability property).

Precise definitions of identifiability, uniform infinitesimal identifiability, and differential identifiability are given and discussed in section 2.1.

In the remaining of this section, we consider the "uncontrolled case" only, i.e. there is no control variable $u(t)$. In that case, we will get a lot of general results. To state them, let us endow the set S of uncontrolled systems, of the form (1.2) with the C^∞ Whitney topology[1].

[1] Consider subsets V_j of the vector space S^j of C^j functions (resp. vector fields, ...) over a C^∞ manifold X, of the following form: $V_j = \cap_{\alpha \in A} V_\alpha$, such that there is a locally finite family of compact sets $\{U_\alpha, \alpha \in A\}$, where V_α is a neighborhood of zero in the topology (non Hausdorf unless $U_\alpha = X$) of uniform C^j convergence over U_α. Sets of the form V_j are a basis of neighborhoods of zero of the C^j Whitney topology over S^j, $0 \leq j < \infty$. This topology is not metrizable (unless X is compact), nevertheless, it is Baire. The C^∞ Whitney topology is the union of the topologies induced on S^∞ by the C^j Whitney topologies over S^j via the inclusions $S^\infty \hookrightarrow S^j$. See [11] for details.

1.2.2. *Results for the generic case.* In section 3, we will show the following important result:

Theorem 1. *If the number of outputs is larger or equal to* 3, *then, differential identifiability of order* $2n+1$ *($n = \dim X$), is a generic property. This is true in the Baire sense only (i.e. the set of differentially identifiable systems is residual). This implies in particular that identifiability is a generic property.*

Remark 1. *This theorem has to be compared to the corresponding result for the observability theory (see* [7], [6]*). In the case of "single input observability", the number of outputs has to be larger or equal to* 2 *only. If one thinks about the function to be identified, as a control variable (which is what we do here in), it is surprising that the number of outputs for generic identifiability be 3. At a first glance, it should be 2.*

In section 3, we will show an open set of systems with 2 outputs, which is not identifiable. It is known ([7]) that single input, single output generic systems are already not observable. Hence, identifiability is not a generic property in the uncontrolled single output case.

1.2.3. *Results for the single output case.* This case will be the subject of section 4.

In that case, by the previous section, differential identifiability is not generic. Again similarly to the observability theory of [7], [5], we will consider the uniform infinitesimal identifiability property. It is a so restrictive property (infinite codimension, in fact) that it can be completely characterized.

We will show the following theorem, in the analytic case (i.e. the system is C^ω).

In the following, $\dim(X) = n$, and L_f is the Lie-derivative operator on X. Also, f_φ denotes the vector field $f(.,\varphi)$, for $\varphi \in I$, and $h_\varphi : X \to R^{d_y}$ is the map $h(.,\varphi)$. The symbol d_x means differential with respect to x only.

Theorem 2. *If* Σ *is uniformly infinitesimally identifiable, then, there is a subanalytic closed subset* Z *of* X, *of codimension 1 at least, such that on the open set* $X\backslash Z$, *the following two equivalent properties 1 and 2 below hold:*

1.a. $\frac{\partial}{\partial\varphi}\{(L_{f_\varphi})^k h_\varphi \equiv 0$, *for* $k = 0, \ldots, n-1$, *b.* $\frac{\partial}{\partial\varphi}\{(L_{f_\varphi})^n h_\varphi \neq 0$ *(in the sense that it* ***never*** *vanishes), c.* $d_x h_\varphi \wedge \ldots \wedge d_x L_{f_\varphi}^{n-1} h_\varphi \neq 0$,

2. any $x_0 \in X\backslash Z$ *has a coordinate neighborhood* $(x_1, \ldots, x_n, V_{x_0})$, $V_{x_0} \subset X\backslash Z$ *in which* Σ *(restricted to* V_{x_0}*) can be written:*

$$\left\{\begin{array}{c} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ . \\ . \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = \psi(x,\varphi); \\ y = x_1; \end{array}\right. \qquad (1.3)$$

where $\frac{\partial}{\partial\varphi}\psi(x,\varphi)$ *never vanishes.*

This theorem has the following pseudo-converse:

Theorem 3. *Assume that* Σ *meets the equivalent conditions of the previous theorem.*

Then, any x_0 *has a neighborhood* V_{x_0} *such that the restriction* $\Sigma_{|V_{x_0}}$ *of* Σ *to* V_{x_0} *is uniformly infinitesimally identifiable,* ***identifiable and differentially identifiable of order*** $n+1$.

Notice that Theorem 2 has a global character: it is almost everywhere on X, but it is global with respect to $\varphi \in I$.

1.2.4. *The 2-output case.* This case, for analytic systems, will be the purpose of Section 5.

Again, (differential) "identifiability" being not generic, we will consider the "uniform infinitesimal identifiability" property, that is very restrictive (infinite codimension), so that we will be able to characterize it completely in a geometric way.

Since this geometric description is non obvious, and since there is (surprisingly) a small but nontrivial zoology (3 distinct cases), we do not give the intrinsic geometric characterization here. This is done in Section 5 (Theorem 9). In this introduction, we state an equivalent theorem (Theorem 4) in terms of normal forms.

Theorem 4. *If Σ is uniformly infinitesimally identifiable, then, there is an open-dense semi-analytic subset $\tilde{U}$ of $X \times I$, such that each point (x_0, φ_0) of $\tilde{U}$, has a neighborhood $V_{x_0} \times I_{\varphi_0}$, and coordinates x on V_{x_0} such that the system Σ restricted to $V_{x_0} \times I_{\varphi_0}$, denoted by $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$, has one of the three following normal forms:*

-type 1 normal form: *(in that case, $n > 2k$)*

$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2, \\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4, \\ &\dots \\ \dot{x}_{2k-3} &= x_{2k-1},\ \dot{x}_{2k-2} = x_{2k}, \\ \dot{x}_{2k-1} &= f_{2k-1}(x_1, \dots, x_{2k+1}), \\ \dot{x}_{2k} &= x_{2k+1}, \\ &\dots \\ \dot{x}_{n-1} &= x_n, \\ \dot{x}_n &= f_n(x, \varphi),\ \text{with } \frac{\partial f_n}{\partial \varphi} \neq 0 \text{ (never vanishes)}, \end{aligned} \tag{1.4}$$

-type 2 normal form:

$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2, \\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4, \\ &\dots \\ \dot{x}_{2r-3} &= x_{2r-1},\ \dot{x}_{2r-2} = x_{2r}, \\ \dot{x}_{2r-1} &= \Phi(x, \varphi),\ \dot{x}_{2r} = F_{2r}(x_1, \dots, x_{2r+1}, \Phi(x, \varphi)), \\ \dot{x}_{2r+1} &= F_{2r+1}(x_1, \dots, x_{2r+2}, \Phi(x, \varphi)), \\ &\dots \\ \dot{x}_{n-1} &= F_{n-1}(x, \Phi(x, \varphi)), \\ \dot{x}_n &= F_n(x, \varphi), \end{aligned} \tag{1.5}$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0$ (hence $n > 2r$), $\dots$, $\frac{\partial F_{n-1}}{\partial x_n} \neq 0$,

-type 3 normal form:

$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2, \\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4, \\ &\dots \\ \dot{x}_{n-3} &= x_{n-1},\ \dot{x}_{n-2} = x_n, \\ \dot{x}_{n-1} &= f_{n-1}(x, \varphi),\ \dot{x}_n = f_n(x, \varphi), \end{aligned} \tag{1.6}$$

where $(\frac{\partial f_{n-1}}{\partial \varphi}, \frac{\partial f_n}{\partial \varphi})$ never vanishes.

Remark 2. *a) Among the normal forms in Theorems 2, 4, normal form "type 2" has a special feature: it is close to the uniform infinitesimal* ***observability*** *normal form obtained in* [5], [7]*: one of the outputs (y_1) is used to obtain $\Phi(x, \varphi)$ after successive differentiations. The remaining part of the normal form is exactly (after a number of direct differentiations) the uniform infinitesimal observability normal form, with $\Phi(x, \varphi)$ considered as an input.*

b) Type 1 and type 2 normal forms are normal up to a permutation of the two outputs y_1 and y_2.

In Section 5, we will give a lot of complementary results, examples, weak converses of Theorems 4, 9, and a few "global" results.

1.3. **Contents and Organization of the paper.** The purpose of the next section 2 is twofold: first we make clear in details the "rough definitions" of "Identifiability" just introduced. Second, we present our controlled example: the bioreactor. We state our main "identifiability result" for this example.

The sections 3, 4, 5 are devoted respectively to the proofs of Theorems 1, 2, 4, and to complementary results: weak converses of these theorems, examples, ...

In Section 6, we say a few words about the practical problem of identification, specially in the non generic cases: 1 and 2 outputs. In these cases, a reasonable practical methodology for identification is proposed.

Section 7 is devoted to the proof of our identifiability results for the "biological reactor" (we have several such results, corresponding to different choices for the observed variables). We also present a few numerical results for the biological reactor. Some of these results illustrate the general methodology we propose in Section 6, some others are specific to the example.

Acknowledgement 1. *We thank Jean-Luc Gouzé, from INRIA Sophia-Antipolis, for stating this problem, and for fruitful discussions.*

Acknowledgement 2. *We thank Witold Respondek for the great work he did when reading the paper, and for several improvements he suggested..*

2. Definitions, example

2.1. **Precise definitions of Identifiability.**

Associated to a system Σ of the form 1.1 or 1.2, we consider the "**input-output mapping**" P_Σ :

$$P_\Sigma : X \times L^\infty[U] \times L^\infty[I] \rightharpoonup L^\infty[\mathbb{R}^{d_y}];$$

$$(x_0, u(\cdot), \hat{\varphi}(\cdot)) \rightharpoonup y(\cdot),$$

where $L^\infty[U]$, $L^\infty[I]$, $L^\infty[\mathbb{R}^{d_y}]$ denote the set of $U-$valued (resp. $I-$valued, $\mathbb{R}^{d_y}-$ valued) measurable, bounded functions, defined on semi-open interval $[0, T_u[$, $[0, T_{\hat{\varphi}}[$, $[0, T_y[$. The mapping P_Σ is defined as follows:

For any input $u(\cdot) \in L^\infty[U]$, any $\hat{\varphi}(\cdot) \in L^\infty[I]$, any $x_0 \in X$, then, the solution $x(t)$ of the Cauchy problem:

$$\frac{dx(t)}{dt} = f(x(t), u(t), \hat{\varphi}(t)),\ x(0) = x_0; \tag{2.1}$$

is defined on a maximum semi-open interval $[0, e(x_0, u, \hat{\varphi})[$, where $0 < e(x_0, u, \hat{\varphi}) \leq \min(T_u, T_{\hat{\varphi}})$. If $e(x_0, u, \hat{\varphi}) < \min(T_u, T_{\hat{\varphi}})$, then $e(x_0, u, \hat{\varphi})$ is the positive escape time of x_0 for the time dependant vector field f in (2.1). For fixed $u, \hat{\varphi}$, the mapping $x_0 \to e(x_0, u, \hat{\varphi}) \in \bar{R}_*^+$ is lower semicontinuous ($\bar{R}_*^+ = \{t | 0 < t \leq \infty\}$).

$P_\Sigma(x_0, u, \hat{\varphi})$ is the function $\hat{y} : [0, e(x_0, u, \hat{\varphi})[\to \mathbb{R}^{d_y}$, defined by $\hat{y}(t) = h(x(t),\, u(t), \hat{\varphi}(t))$.

We say that $(u(\cdot), y(\cdot)) \in L^\infty[U] \times L^\infty[\mathbb{R}^{d_y}]$, with $T_u = T_y$, is an "**admissible input-output trajectory**", if there exits a couple $(x_0, \hat{\varphi}) \in X \times L^\infty[I]$, such that $y(t) = P_\Sigma(x_0, u, \hat{\varphi})(t)$ for almost all $t \in [0, T_u[$ (which means in particular $e(x_0, u, \hat{\varphi}) = T_u$), and $\hat{\varphi}(t) = \varphi \circ \pi(x(t))$, where φ is some smooth function, $\varphi : Z \to I$. (Of course, this φ depends on the input-output trajectory (or the "experiment") $(u(\cdot), y(\cdot))$).

Note that, as a consequence, if $(u(\cdot), y(\cdot))$ is an admissible i.o. trajectory, then, $\hat{\varphi}(\cdot)$ in the definition is in fact at least **absolutely continuous**. In the uncontrolled case, it is **smooth**.

We define now the notion of identifiability, already introduced in Section 1.2:

Definition 1. *The system Σ is said "identifiable at" $(u(\cdot), y(\cdot)) \in L^\infty[U] \times L^\infty[\mathbb{R}^{d_y}]$, with $T_u = T_y$, if there is* ***at most*** *a single couple $(x_0, \hat{\varphi}) \in X \times L^\infty([0, T_u[, I)$ such that, for almost all $t \in [0,\, T_u[$:*

$$P_\Sigma(x_0, u, \hat{\varphi})(t) = y(t),$$

and $\hat{\varphi}(t) = \varphi \circ \pi(x(t))$ for some smooth function $\varphi : Z \to I$.

Given a system Σ, the "identifiability set" of Σ is the subset of $L^\infty[U] \times L^\infty[\mathbb{R}^{d_y}]$ formed by the admissible i-o trajectories $(u(\cdot), y(\cdot))$ at which Σ is identifiable. If this set is exactly the set of admissible i-o trajectories, then Σ is said "identifiable".

Again, in this definition, the smooth function φ depends on the i.o. trajectory (or the "experiment") $(u(\cdot), y(\cdot))$. Also, again, $\hat{\varphi}$ in this definition is in fact at least **absolutely continuous**, and **smooth** in the uncontrolled case.

Of course, this definition is not very tractable in practice, and, as explained in Section 1.2, we will need a few other definitions.

For arbitrary $k-$jets of smooth functions $\hat{\varphi}, \hat{u}$ at $t=0$,

$$\hat{\varphi} : [0,\varepsilon[\rightarrow I, \hat{u} : [0,\varepsilon[\rightarrow U,$$
$$j^k(\hat{\varphi}) = (\hat{\varphi}(0), \hat{\varphi}'(0), \ldots, \hat{\varphi}^{(k-1)}(0)), j^k(\hat{u}) = (\hat{u}(0), \hat{u}'(0), \ldots, \hat{u}^{(k-1)}(0)),$$

and for any $x_0 \in X$, the corresponding $k-$jet $j^k\hat{y} = (\hat{y}, \hat{y}', \ldots, \hat{y}^{(k-1)})$ is well defined, in such a way that the mapping $\Phi_k^\Sigma : (x_0, j^k(\hat{u}), j^k(\hat{\varphi})) \rightarrow j^k\hat{y}$ be continuous. $\Phi_k^\Sigma : D_k\Phi = X \times (U \times \mathbb{R}^{(k-1)d_u}) \times (I \times R^{k-1}) \rightarrow \mathbb{R}^{kd_y}$.

Remark 3. *This mapping Φ_k^Σ is smooth with respect to x_0, $\hat{\varphi}(0)$, $\hat{u}(0)$, and algebraic with respect to $\hat{\varphi}'(0), \ldots, \hat{\varphi}^{(k-1)}(0)$ and $\hat{u}'(0), \ldots, \hat{u}^{(k-1)}(0)$. Also, Φ_k^Σ depends only on the $k-$jet $j^k\Sigma$ of Σ.*

We define $D_k\Phi_2^*$ as the set of couples $((x_1, j^k(\hat{u}), j^k(\hat{\varphi}_1)), ((x_2, j^k(\hat{u}), j^k(\hat{\varphi}_2)) = (z_1, z_2)$, with $z_1 = (x_1, j^k(\hat{u}), j^k(\hat{\varphi}_1)) \neq ((x_2, j^k(\hat{u}), j^k(\hat{\varphi}_2)) = z_2$.

Δ_k is the diagonal in $\mathbb{R}^{kd_y} \times \mathbb{R}^{kd_y}$. We denote by $\Phi_{k,2}^{\Sigma,*}$ the mapping: $D_k\Phi_2^* \rightarrow \mathbb{R}^{kd_y} \times \mathbb{R}^{kd_y}$, $(z_1, z_2) \rightarrow (\Phi_k^\Sigma(z_1), \Phi_k^\Sigma(z_2))$

Definition 2. *The system Σ is said "differentially identifiable" of order k if the mapping $\Phi_{k,2}^{\Sigma,*}$ has the following property: If*

$$\Phi_{k,2}^{\Sigma,*}(z_1, z_2) \in \Delta_k,$$

then, $(x_1, \hat{\varphi}_1(0)) = (x_2, \hat{\varphi}_2(0))$.

This means that, for all controls (sufficiently differentiable), all couples (initial state, value of φ) are distinguished between them, by the observations and their $k-1$ first derivatives, whatever the other time derivatives of φ are.

Equivalently, one can reconstruct the state and the values of φ, (at a point (x_0, u_0), visited by the system) in terms of the controls and their first derivatives, the outputs and their first derivatives.

Remark 4. *One could think that a good definition of differential identifiability of order k is just: the map Φ_k^Σ is injective. Unfortunately, this property is never generic (for uncontrolled systems), if there is less outputs than states ($d_y \leq n$).*

As stated in Section 1.2.2, Theorem 1, **differential identifiability** at certain orders k is generic, for $d_y \geq 3$ (uncontrolled case).

Following the ideas developed in the book [7], and in the papers [5], [6], (in the context of **observability**), we define now the **infinitesimal** notion of identifiability. For this purpose, we need an adequate concept of the "linearization" of a system.

The mapping $f : X \times U \times I \rightarrow TX$ induces a partial tangent mapping $T_{x,\varphi}f : TX \times TI \times U \rightarrow TTX$. (Here, $TI \sim I \times \mathbb{R}$). If ω denotes the canonical involution[2] of TTX, then, $\omega \circ T_{x,\varphi}f$ defines a parametrized vector field on TX, also denoted by $T_{x,\varphi}f$: it is parametrized by the elements $(u, (\varphi, \eta))$ of $U \times TI$.

Similarly, the function $h : X \times U \times I \rightarrow R^{d_y}$, has a partial differential $d_{x,\varphi}h$, and the **linearization of** Σ (or the **first variation** of Σ) is the following system on TX :

$$T\Sigma \left\{ \begin{array}{l} \frac{d\xi}{dt} = T_{x,\varphi}f(x, u, \varphi; \xi, \eta); \\ \hat{y} = d_{x,\varphi}h(x, u, \varphi; \xi, \eta), \end{array} \right. \tag{2.2}$$

where $(\xi, \eta) \in T_xX \times T_\varphi I$.

Denote by Π the canonical projection: $TX \rightarrow X$. If $\xi : [0, T_\xi[\rightarrow TX$ is a trajectory of $T\Sigma$ for the control $u(\cdot)$, function $\varphi(\cdot)$, and the "variation" $\eta(\cdot)$, (all belonging to $L^\infty_{[0,T_\xi[}$), then, $\Pi\xi$ is a trajectory of Σ corresponding to the same control $u(\cdot)$ and function $\varphi(\cdot)$. Conversely, if $\phi_\tau(x, u, \hat{\varphi}) : [0, e(x, u, \hat{\varphi})[\rightarrow X$ is a trajectory of Σ starting from x_0, and corresponding to the control u and the function $\hat{\varphi} : [0, T_{\hat{\varphi}}[\rightarrow I$, then, for all $0 < \tau < e(x, u, \hat{\varphi})$, the map:

$$(x, \varphi) \rightarrow \phi_\tau(x, u, \varphi);$$

is defined on a neighborhood of $(x_0, \hat{\varphi})$ in $X \times L^\infty([0, \tau[, I)$, and is differentiable at $(x_0, \hat{\varphi})$. This differential $T_{x,\varphi}\phi_\tau$ is in the usual sense with respect to x, and in the Frechet sense with respect to φ.

[2] If $f : X \rightarrow TX$ is a vector field (i.e. a section), then $Tf : TX \rightarrow TTX$ is not a section: for $\xi_x \in T_xX$, then $\Pi_{TTX}(Tf(\xi_x)) = f(x)$, although it should be equal to ξ_x. There is a well defined involution $\omega : TTX \rightarrow TTX$, called the "canonical involution" of TTX, such that $\omega \circ Tf$ is a section of TTX. See [1] for details.

For all $(\xi_0, \eta) \in T_{x_0}X \times L^\infty([0, \tau[, \mathbb{R})$, for almost all $t \in [0, \tau[$:

$$P_{T\Sigma}(\xi_0, \eta)(t) = d_{x,\varphi}h(u, \hat{\varphi}; T_{x,\varphi}\phi_t(u, \hat{\varphi}; \xi_0, \eta), \eta) = T_{x,\varphi}P_\Sigma(\xi_0, \eta)\,(t)\,. \tag{2.3}$$

The right-hand side of this equality is the differential of the mapping $P_\Sigma(x, u, \hat{\varphi})$ with respect to (x, φ), at the point $(x_0, u, \hat{\varphi})$, taken at (ξ_0, η), evaluated at time t.

Definition 3. *The system Σ is called "infinitesimally identifiable" at $(x_{0,}\ u\,(\cdot)\,,\ \hat{\varphi}\,(\cdot)) \in X \times L^\infty[U] \times L^\infty[I]$ if all the linear mappings $P_{T\Sigma} : T_{x_0}X \times L^\infty_{([0,\tau],\mathbb{R})} \to L^\infty_{([0,\tau],\mathbb{R}^{d_y})}$ in (2.3), (for all $0 < \tau < e(x_0, u, \hat{\varphi})$), are injective. It is called "uniformly infinitesimally identifiable", if it is infinitesimally identifiable at all $(x_{0,}\ u\,(\cdot)\,,\ \hat{\varphi}\,(\cdot)) \in X \times L^\infty[U] \times L^\infty[I]$.*

Remark 5. *The definitions 2, 3, do not depend on the internal variables and the internal mapping. On the contrary, Definition 1 does.*

A clear comparison between our 3 definitions of identifiability is not obvious at all at the level of this introduction. Definition 1 is a natural and general definition. Definitions 2, 3, are adapted respectively to the generic and non generic cases (uncontrolled), as we shall see.

It will be shown (Theorem 7, Section 3) that, **for uncontrolled systems**, differential identifiability at some order k implies identifiability, whatever the internal mapping π. Also, uniform infinitesimal identifiability implies identifiability of the restrictions of the system to certain open subsets. This last point will be proved in Sections 4, 5.

Very clearly, our identifiability properties are related to the notion of "left-invertibility", introduced by Hirschorn in the papers [12], [13]. One may also refer to the nice survey paper [15], specially Theorem 5.1. pages 164-165. In fact, Identifiability is a stronger property than left-invertibility, even for linear single input-output systems: left invertibility (in our uncontrolled case) is the injectivity of the maps $\hat{\varphi} \to P_\Sigma(\hat{\varphi})$, for x_0 chosen and fixed. This injectivity is required at the level of n-jets, and locally (which means locally around a fixed n-jet of $\hat{\varphi}$, and at all points of an open dense semi-analytic subset of the space of $(x_0, j^n\hat{\varphi})$. If a system is left invertible, roughly speaking, there is an "inverse system", that allows (except at some singularities, and locally), to recover the input of Σ, from the extension to n-jets of its output trajectory. Hence, note that, if people want to practically invert a system, they implicitly consider differentiation of the output. This is connected to the considerations in our section 6 below.

Obviously, in our (single input) case, invertibility as defined in these papers is generic, and moreover the set of noninvertible systems has infinite codimension.

2.2. **A comment on these definitions.** Let us limit ourselves to the uncontrolled case in this discussion.

We have 3 definitions (Identifiability, differential identifiability, Uniform Infinitesimal Identifiability). In all these definitions, the functions $\hat{\varphi}\,(\cdot)$ (or their jets) are considered as extra inputs for the system. These functions have to be smooth in two cases:

a. for Differential Identifiability, since the jets are jets of smooth functions,

b. in the definition of Identifiability, as a consequence of the definition.

But, in Definition 3, of Uniform Infinitesimal Identifiability, no smoothness for $\hat{\varphi}$ and its "variations" η is required (we assumed these functions to be in L^∞ only, which is natural, as control functions for a nonlinear system). One could think that enlarging in this way the class of "inputs" $\hat{\varphi}$ might restrict a lot the class of systems that are uniformly infinitesimally identifiable.

In fact, the class of $\hat{\varphi}$ and its "variations" η plays no role in this definition. This can be seen as a consequence of two distinct facts:

1. First, the proofs of our uniform infinitesimal identifiability results are all based upon a contradiction with the existence of smooth couples $(\hat{\varphi}, \eta)$, that make the system non infinitesimally identifiable along corresponding trajectories. Then, we could take $(\hat{\varphi}, \eta)$ smooth in the definition. But, a posteriori, we see on the resulting normal forms 1.3, 1.4, 1.5, 1.6, that such systems are also uniformly infinitesimally identifiable for the (bigger) class of $(\hat{\varphi}, \eta)$ that are L^∞ only.

2. Second, the following theorem can be proved:

Theorem 5. *If Σ (analytic) is not infinitesimally identifiable at $(x_0, \hat{\varphi})$, because of some variation $\eta\,(\cdot)$ such that $(\xi_0, \eta) \neq 0$ is in the Kernel of $T_{x_0,\hat{\varphi}}P_\Sigma$, with $(\hat{\varphi}, \eta)$ in L^∞, then Σ is also not infinitesimally identifiable at some other $(x_1, \hat{\varphi}_1)$, because of some other variation $\eta_1\,(\cdot)$ (such that $(\xi_0', \eta_1) \neq 0$ is in the Kernel of $T_{x_1,\hat{\varphi}_1}P_\Sigma$), and moreover $(\hat{\varphi}_1, \eta_1)$ are analytic.*

This theorem shows exactly that (fortunately), the class of $(\hat{\varphi}, \eta)$ in Definition 3 of Uniform Infinitesimal Identifiability, is in fact irrelevant. This result is stronger than the considerations in 1. above, that hold, as the normal forms, locally around points of an open dense subset of $X \times I$.

We will not give here the proof of this theorem: it is difficult, and far beyond the scope of this paper. It is related to ideas developed in [7] (where comparable results can be found), and also to ideas of "trajectory regularity" that can be found in [17].

2.3. **An example: biological reactors.**

2.3.1. *Presentation.* A simple biological reactor is a process where a population grows in presence of some substrate by eating the substrate. The concentration of the biomass (the population) in the mixture is denoted by x, and the concentration of the substrate is s. Here, x and s belong to $\mathbb{R}^+$. The reactor is fed continuously by some flow of substrate, with concentration S_{in}, and flowrate $D(t) > 0$. S_{in} is a constant (in first approximation), and $D(t)$ is usually a control variable.

The way the population grows in presence of substrate is described by a "growth function" $\mu(x, s)$. The total volume is maintained constant by perpetually throwing out the same volume of mixture as the volume of substrate entering the reactor.

Hence, the equations are:

$$\begin{aligned} \frac{ds}{dt} &= -\mu.x + D(S_{in} - s) \\ \frac{dx}{dt} &= (\mu - D)x. \end{aligned} \tag{2.4}$$

The questions of observation and control of such biological reactor has been treated by several authors. See for instance [4], and the book [2].

In the literature, there are many possible choices for the function μ. One can find a cornucopia of them in the book [2].

Here, we will consider the case where this function μ is an unknown function, of (either the variable s only, or the variable x only, or of both), and we will give several conclusion about its identifiability.

Remark 6. *The biological reactor will provide an example where the identifiability property depends on the internal mapping π, see section 7.2. Also, clearly, in that case, identifying the "open-loop" function $\varphi(t)$ is not equivalent to identifying the "closed-loop" function $\varphi \circ \pi(\chi(t))$, with $\chi = (x, s)$.*

In fact, the general study in this paper was initially motivated by this very interesting simple example.

2.3.2. *Main theoretical result for the biological reactor.* One of our results for the biological reactor will be like that: we will consider that there is a single observation, the concentration s of the substrate.

Then, the biological reactor is not identifiable in the sense of definition 1.

Also, if we consider only constant input functions, $(D(\cdot) = ct.)$, then, the (single output system is not uniformly infinitesimally identifiable in the sense of Definition 3, because it does not meet the necessary conditions of Theorem 2.

Assuming that the growth function depends on s only ($\pi : (x, s) \to s$ is the internal mapping), then, we prove (easily) the following theorem:

Theorem 6. *The bioreactor is identifiable at an admissible i-o trajectory $(y(\cdot), u(\cdot)) = (s(\cdot), D(\cdot)$ iff the output trajectory $y(\cdot) = s(\cdot)$ visits twice the same value.*

This result, with others, is proved in Section 7.

3. The Generic Case

The purpose of this section is to prove Theorem 1, and to give some complementary results and examples.

3.1. **Comparison between differential identifiability and identifiability.** We state a result in the **uncontrolled case** only. In the controlled case, the situation is much more complicated. It is possible to prove something very strong, using techniques developed in the book [7], for observability. But, this is a great complication, and it will be the purpose of another paper.

Theorem 7. ***(uncontrolled systems)*** *In the uncontrolled case, differential identifiability at some order implies identifiability, whatever the internal mapping π.*

Proof. Let us assume that an uncontrolled system Σ is not identifiable. It means that there is an admissible output trajectory $\hat{y} : [0, \tau[\to \mathbb{R}^{d_y}$, such that Σ is not identifiable at $\hat{y}$, i.e., there are two trajectories $x_1(t), x_2(t)$, and two corresponding functions $\hat{\varphi}_1(t), \hat{\varphi}_2(t)$, $\hat{\varphi}_1(t) = \varphi_1(\pi(x_1(t)))$, $\hat{\varphi}_2(t) = \varphi_2(\pi(x_2(t)))$, such that φ_1 and φ_2 are smooth, hence $\hat{\varphi}_1(t), \hat{\varphi}_2(t)$, and $\hat{y}(t)$ are smooth. Moreover, $\hat{\varphi}_1(t_0) \neq \hat{\varphi}_2(t_0)$ for some $t_0 \in [0, \tau[$, or $x_1(0) \neq x_2(0)$. Restricting the interval $[0, \tau[$, we may assume that $t_0 = 0$. Then, for an arbitrary positive integer k, consider the two k-jets at time zero, $j^k\hat{\varphi}_1$, $j^k\hat{\varphi}_2$. Consider the mapping Φ_k^Σ defined in the introduction. Set $z_1 = (x_1(0), j^k\hat{\varphi}_1)$, $z_2 = (x_2(0), j^k\hat{\varphi}_2)$. Then, $(\Phi_k^\Sigma(z_1), \Phi_k^\Sigma(z_2)) \in \Delta_k$, the diagonal of $\mathbb{R}^{kd_y} \times \mathbb{R}^{kd_y}$. Therefore, Σ is not differentially identifiable of order k, since $\hat{\varphi}_1(0) \neq \hat{\varphi}_2(0)$ or $x_1(0) \neq x_2(0)$. □

3.2. **Preliminaries for the proof of Theorem 1.** Let S denote the set of C^∞ systems $\Sigma = (f, h)$ of the form 1.2, i.e elements of S are couples (f, h) of φ-parametrized vector fields f and functions h. We endow S with the C^∞ Whitney topology. Let J^kS denote the bundle of $k-$jets of systems in S. It is the fiber product $J^kF \times_{X\times I} J^kH$ of the bundles J^kF, J^kH over $X \times I$, that are respectively the bundles of $k-$jets of smooth sections of $TX \times I \to X \times I$ and $X \times I \times \mathbb{R}^{d_y} \to X \times I$.

Let us also denote by $J^kS_2^*$ the restriction of $J^kS^2 = J^kS \times J^kS$ to $((X \times I) \times (X \times I))\backslash\Delta(X \times I)$, where $\Delta(X \times I)$ is the diagonal in $((X \times I) \times (X \times I))$.

Let us recall that the mappings Φ_k^Σ, $\Phi_{k,2}^{\Sigma,*}$, defined in the introduction, depend only on the $k-$jet $j^k\Sigma$ of Σ. When we want to stress this fact, we write $\Phi^{j^k\Sigma}$ in place of Φ_k^Σ.

3.2.1. *The bad sets.*

Definition 4. *$B_1(k)$ is the subset of J^kS^2 of all couples $(j^k\Sigma(p), j^k\Sigma(q))$, such that: (1) $p \neq q$, $p = (x_1, \varphi_1)$, $q = (x_2, \varphi_2)$, (2) $f(p) = f(q) = 0$, (3) $h(p) = h(q)$.*

Definition 5. *a. Let $\hat{B}_2(k)$ be the subset of $J^kS^2 \times \mathbb{R}^{(k-1)} \times \mathbb{R}^{(k-1)}$, of all tuples $(j^k\Sigma(p), j^k\Sigma(q), v_1, v_2)$ such that: (1) $p \neq q$, $p = (x_1, \varphi_1)$, $q = (x_2, \varphi_2)$, (2) $f(p) \neq 0$ or $f(q) \neq 0$, (3) $\Phi_k^\Sigma(x_1, (\varphi_1, v_1)) = \Phi_k^\Sigma(x_2, (\varphi_2, v_2))$,*

b. $B_2(k)$ denotes the canonical projection of $\hat{B}_2(k)$ in J^kS^2.

The two following lemmas are obvious.

Lemma 1. *$B_1(k)$, $\hat{B}_2(k)$, $B_2(k)$, are respectively partially semi-algebraic subbundles of $J^kS_2^*$, $J^kS_2^* \times \mathbb{R}^{(k-1)} \times \mathbb{R}^{(k-1)}$, $J^kS_2^*$ (this means that heir typical fiber is a semi-algebraic subset of the fibers of the ambient bundles).*

Lemma 2. *If the map*

$$\Theta : ((X \times I) \times (X \times I))\backslash\Delta(X \times I) \to j^kS_2^*$$
$$(p, q) \to (j^k\Sigma(p), j^k\Sigma(q)),\ p = (x_1, \varphi_1),\ q = (x_2, \varphi_2),$$

avoids $B_1(k) \cup B_2(k)$, then Σ is differentially identifiable of order k.

3.3. **Proof of Theorem 1.** a. **Estimation of the codimension of $B_1(k)$ in $J^kS_2^*$** : It is obvious that this codimension is $2\ n + d_y$.

b. **Estimation of the codimension of $B_2(k)$ in $J^kS_2^*$** : We treat only the case $f(p) \neq 0$. The other case $f(q) \neq 0$ is similar.

Let $(x, \varphi, y, \psi) \in ((X\times I)\times(X\times I))\backslash\Delta(X\times I)$. The typical fiber $\hat{B}_2(k, x, \varphi, y, \psi)$ of $\hat{B}_2(k)$ in $J^kS(x, \varphi)\times J^kS(y, \psi) \times \mathbb{R}^{k-1} \times \mathbb{R}^{k-1}$ is characterized by the following properties: (i) $f(p) \neq 0$ and (ii) $\Phi_k^\Sigma(x, (\varphi, v_1)) - \Phi_k^\Sigma(y, (\psi, v_2)) = 0$.

Let G be the subset of $J^kS(x,\varphi)\times J^kS(y,\psi)\times\mathbb{R}^{k-1}\times\mathbb{R}^{k-1}$ of all tuples $(j^k\Sigma(x,\varphi),v_1,\ j^k\Sigma(y,\psi),v_2)$ such that $f(x,\varphi)\neq 0$ and let $\chi:G\rightarrow\mathbb{R}^k$ be the mapping $\chi(j^k\Sigma(x,\varphi),v_1,\ j^k\Sigma(y,\psi),v_2)=\Phi_k^{\Sigma}(x,(\varphi,v_1))-\Phi_k^{\Sigma}(y,\ (\psi,v_2))$. Then, $\hat{B}_2(k,x,\varphi,y,\psi)=\chi^{-1}(0)$.

The map χ is an algebraic mapping, affine in $j^kh(x,\varphi)$.

By Lemma 3 in Appendix 3.5 below, for fixed $v_1\in\mathbb{R}^k$ and $j^kf(x,\varphi)$, the linear mapping $j^kh(x,\varphi)\in J^kH(x,\varphi)\rightarrow\Phi^{j^k\Sigma}(x,\varphi,v_1)$ is surjective. This shows that the map $\chi:G\rightarrow\mathbb{R}^k$ is a submersion. Since $\hat{B}_2(k,x,\varphi,y,\psi)=\chi^{-1}(0)$,

$$codim(\hat{B}_2(k,x,\varphi,y,\psi),J^kS(x,\varphi)\times J^kS(y,\psi)\times\mathbb{R}^{k-1}\times\mathbb{R}^{k-1})=kd_y.$$

Hence:

$$codim(\hat{B}_2(k),J^kS_2^*\times\mathbb{R}^{k-1}\times\mathbb{R}^{k-1})=kd_y.$$

It follows that $codim(B_2(k),J^kS_2^*)\geq k(d_y-2)+2$.

c. **Final estimation of codim**$(B_1(k)\cup B_2(k),J^kS_2^*)$.

It follows from a. and b. above that:

$$codim(B_1(k)\cup B_2(k),J^kS_2^*)\geq\min(2\ n+d_y,k(d_y-2)+2). \tag{3.1}$$

d. **Proof of Theorem** 1. Let $k\geq 2\ n+1$, and $d_y\geq 3$. Then, $codim(B_1(k)\cup B_2(k),J^kS_2^*)\geq 2n+3$. We apply the standard multijet transversality theorems ([1], [11]) to the map

$$\rho:S\times((X\times I)\times(X\times I))\backslash\Delta(X\times I)\rightarrow J^kS_2^*,$$

$$(\Sigma,x,\varphi,y,\psi)\rightarrow(j^k\Sigma(x,\varphi),\ j^k\Sigma(y,\psi)).$$

This allows to conclude that the set of $\Sigma\in S$ such that ρ_Σ is transverse to $B_1(k)\cup B_2(k)$ is residual. But, $\dim(((X\times I)\times(X\times I))\backslash\Delta(X\times I))=2n+2<codim(B_1(k)\cup B_2(k),J^kS_2^*)\geq 2n+3$.

Hence, the set of $\Sigma\in S$ such that ρ_Σ avoids $B_1(k)\cup B_2(k)$ is residual. By Lemma 2, this ends the proof of Theorem 1.

3.4. **A counterexample.** We consider, on the circle S^1, for values of φ in the interval $I=[-\pi,\pi]$, systems of the form: $\Sigma=\Sigma_0+\delta\Sigma$, where $\delta\Sigma$ is C^∞ but C^1 small, and Σ_0 is the system:

$$\Sigma_0\left\{\begin{array}{l}\dot{x}=\psi(x)-\varphi=f_0(x,\varphi)\\ y_1=\cos(x)=h_0^1(x,\varphi);\\ y_2=\sin(2x)=h_0^2(x,\varphi).\end{array}\right.$$

The cyclic coordinate on S^1 is x and $\psi:S^1\rightarrow\mathbb{R}$ is such that $\psi(x)=x$ on the interval $[-\frac{\pi}{2}-\varepsilon,\frac{\pi}{2}+\varepsilon]$, for $\varepsilon>0$ small. We want to solve the system of equations:

$$\begin{aligned}(f_0+\delta f)(x_1,\varphi_1)=0;\\ (f_0+\delta f)(x_2,\varphi_2)=0;\\ (h_1+\delta h_1)(x_1,\varphi_1)-(h_1+\delta h_1)(x_2,\varphi_2)=0;\\ (h_2+\delta h_2)(x_1,\varphi_1)-(h_2+\delta h_2)(x_2,\varphi_2)=0;\end{aligned} \tag{3.2}$$

around the trivial solution $\delta f=0,\delta h_1=0,\delta h_2=0,\ x_1=\frac{\pi}{2},\varphi_1=\frac{\pi}{2},\ x_2=-\frac{\pi}{2},\varphi_2=-\frac{\pi}{2}$.

Here, the set S of systems is identified with a subspace of the Banach space BS of C^1 triples $(\delta f,\delta h_1,\delta h_2)\in(C^1(S^1\times I))^3$ with the C^1 topology .

It is trivial to check that the Jacobian matrix of the system (3.2), with respect to the variables $(x_1,\varphi_1,x_2,\varphi_2)$ at the point $(\frac{\pi}{2},\frac{\pi}{2},-\frac{\pi}{2},-\frac{\pi}{2})$, is invertible. This means, applying the implicit function theorem in the Banach manifold $BS\times S^1\times I\times S^1\times I$, that for all C^∞ variations $\delta\Sigma$ small enough (C^1), we can find solutions to Equations (3.2), and these solutions are close to $x_1=\frac{\pi}{2},\varphi_1=\frac{\pi}{2},\ x_2=-\frac{\pi}{2},\varphi_2=-\frac{\pi}{2}$.

But, such a solution of (3.2) produces a system Σ and two couples $(x_1,\varphi_1)\neq(x_2,\varphi_2)$ such that, for the **constant** control functions φ_1, φ_2, and for the initial conditions x_1,x_2, the corresponding couples of outputs (y_1,y_2) are constant functions of the time, that are equal.

Hence, there is an open neighborhood (C^1 open in C^∞) of systems, that are not differentially identifiable at any order k (and also not identifiable).

Remark 7. *Notice that, in the formula 3.1, the 2 bounds on the codimension are $2n+2$ and 2, for $d_y = 2$. Hence, there should be **two typologies of counterexamples** for $d_y = 2$.*

In this example, we have exhibited the simplest one, corresponding to codimension $2n+2$. But, from the results of section 5, (the case $d_y = 2$), counterexamples corresponding to the other typology are easy to construct.

3.5. **Appendix.** For the sake of completeness, we give here a (very simple) lemma which is already contained in the paper [6], and in the book [7].

If V is a vector space, $Sym^a(V)$ denotes the space of symmetric tensors of degree a on V, that can be canonically identified with the space of homogeneous polynomials of degree a over V^*, dual space of V. The symbol $\odot$ means symmetric tensor product or power.

Let $(x, \varphi, v) \in X \times I \times \mathbb{R}^{k-1}$ be given, and set $p = (x, \varphi)$. Let $f \in F$ (the space of smooth $I-$parametrized vector fields on X), such that $f(p) \neq 0$.

Lemma 3. *The mapping $\Theta : J^k H \to \mathbb{R}^{kd_y}$, $j^k h \to \Phi^{j^k \Sigma}(x, \varphi, v)$ is linear and surjective.*

Proof. Let f be a representative of $j^k f(p)$. Take a coordinate system $(O, x_1, \ldots, x_n)$ for X at x. Then, if $y^{(r)}$ denotes the r^{th} derivative of the output at time 0, we have:

$$y^{(r)}(p, v) = d_x^r h(p; f(p)^{\odot r}) + \sum_{a=0}^{r-1} d_x^a h(p, R_{a,r}(j^k f(p), v) +$$
$$\sum_{\substack{l=1,\ldots,r,\\ s+l=r}} d_x^s d_\varphi^l h(p; T_{s,l}^r(j^k f(p), v)).$$

The expressions $R_{a,r}$, $T_{s,l}^r$, are universal polynomial mappings : $J^k F(p) \times \mathbb{R}^{k-1} \to Sym^a(T_x X)$ (resp. $J^k F(p) \times R^{k-1} \to Sym^s(T_x X) \otimes \mathbb{R}^l$). This implies the result because $f(p) \neq 0$. □

4. The single-output case.

The purpose of this section is to prove Theorem 2 and Theorem 3

4.1. **Proof of Theorem 2.** Let us assume (A) that $\frac{\partial}{\partial \varphi}(L_{f_\varphi})^i h_\varphi \equiv 0$, $i = 0, \ldots, k-1$, and $\frac{\partial}{\partial \varphi}(L_{f_\varphi})^k h_\varphi \neq 0$, for $k \leq n-1$.

Then, (B): the closed analytic subset of X : $Z_{k-1} = \{x | d_x h \wedge \ldots \wedge d_x L_f^{k-1} h(x) = 0\}$, has codimension 1 at least. Then, for all $x_0 \in X \backslash Z_{k-1}$, it can be completed by functions $h_k(x), \ldots, h_n(x)$, in order to form a local coordinate system on a neighborhood U_{x_0} of x_0.

Were it otherwise, then, by connectedness, $d_x h \wedge d_x L_f h \wedge \ldots \wedge d_x (L_f)^{k-1} h \equiv 0$. Take $j \leq k-1$, the first integer such that $d_x h \wedge d_x L_f h \wedge \ldots \wedge d_x (L_f)^j h \equiv 0$. Then, set, on some open subset of X : $x_1 = h, \ldots, x_j = (L_f)^{j-1} h$. Complete this set of functions by $x_{j+1}, \ldots, x_n$, in order to form a coordinate system.

In these coordinates, with a straightforward computation, the system Σ can be rewritten:

$$\Sigma \begin{cases} y = x_1, \\ \dot{x}_1 = x_2, \\ \quad . \\ \quad . \\ \dot{x}_{j-1} = x_j, \\ \dot{x}_j = \psi_j(x_1, \ldots, x_j), \\ \dot{x}_{j+1} = \psi_{j+1}(x, \varphi) \\ \quad . \\ \dot{x}_n = \psi_n(x, \varphi), \end{cases}$$

which is obviously not uniformly infinitesimally identifiable: take in the first variation, $\xi(0) \neq 0, \xi_1(0) = \cdots = \xi_j(0) = 0$, and the "variation" $\eta(\cdot) \equiv 0$. The output of the first variation system is identically zero, whatever $\varphi(\cdot)$. A contradiction with uniform infinitesimal identifiability.

This shows (B).

Now, let us show that (A) is impossible.

In the coordinate system defined as in (B), the system writes, on some open subset:

$$\sum\left\{\begin{array}{c} y = x_1, \\ \dot{x}_1 = x_2, \\ . \\ . \\ \dot{x}_{k-1} = x_k, \\ \dot{x}_k = \psi_k(x,\varphi), \\ . \\ . \\ \dot{x}_n = \psi_n(x,\varphi). \end{array}\right. \tag{4.1}$$

Now, the linearized system writes, in these coordinates (and the associated coordinates ξ_i on TX):

$$\left\{\begin{array}{c} \dot{x} = f(x,\varphi), \\ \dot{\xi}_1 = \xi_2, \\ \dot{\xi}_{k-1} = \xi_k, \\ \dot{\xi}_i = d_x\psi_i(x,\varphi)\xi + d_\varphi\psi_i(x,\varphi)\eta, \\ i = k,\dots,n, \end{array}\right. \tag{4.2}$$

where f is as in 4.1. Let us take $\xi(0) \neq 0$, $\xi_1(0) = 0, \dots, \xi_k(0) = 0$. Our assumption (A) implies that on an open dense semianalytic subset D of $X \times I$, $d_\varphi \psi_k(x,\varphi) \neq 0$. Around a point $p_0 = (x_0, \xi(0), \varphi_0) \in TX \times I$, $(x_0,\varphi_0) \in D$, with $\xi(0)$ just defined, we consider the feedback function for the system 4.2: $\eta(x,\varphi,\xi) = \frac{-d_x\psi_k(x,\varphi)\xi}{d_\varphi\psi_k(x,\varphi)}$. We consider also the function $\varphi(t) \equiv \varphi_0$. Then, the feedback system 4.2 has a unique trajectory $(\xi(t), x(t))$ around p_0, starting from (x_0, ξ_0) at time 0. By construction, this trajectory is such that $\xi_1(t), \dots, \xi_k(t) = 0$ for all t small enough. This contradicts the infinitesimal identifiability assumption. Hence, $\frac{\partial}{\partial\varphi}(L_{f_\varphi})^k h_\varphi = 0$, (A) is impossible.

Now, at this point, identically on $X \times I$, $\frac{\partial}{\partial\varphi}\{(L_{f_\varphi})^k h_\varphi = 0$, for $k = 0, \dots, n-1$. This is the property 1.a. of Theorem 2. Property 1.c. also holds, by the proof of (B) above, which works also in that case. Moreover, we already know that, on an open dense semianalytic subset D of X, our system can be rewritten (in local coordinates, around any point $x_0 \in D$):

$$\left\{\begin{array}{c} y = x_1, \\ \dot{x}_1 = x_2, \\ . \\ . \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = \psi(x,\varphi). \end{array}\right. \tag{4.3}$$

Consider the closed analytic subset $S \subset X \times I$, formed by the (x,φ)'s satisfying $\frac{\partial\psi}{\partial\varphi} = 0$ (or equivalently, $\frac{\partial}{\partial\varphi}\{(L_{f_\varphi})^n h_\varphi = 0)$. If this subset has nonempty interior, by analyticity and connectedness, it is the whole $X \times I$, and ψ does not depend on φ, which is easily seen as a contradiction with infinitesimal identifiability. Then, S has codimension 1 at least. Let ΠS be the projection of S on X, $\Pi : X \times I \to X$. Since I is compact, ΠS is subanalytic. By Hardt's Theorem ([9]), we can stratify the mapping $\Pi : S \to \Pi S$. Let S_1, ΠS_1 be two strata such that ΠS_1 has maximal dimension, and Π maps S_1 submersively onto ΠS_1. Let $\hat{\varphi} : \Pi S_1 \to S_1$ be a smooth section of Π. (See the footnote at the beginning of the appendix, Section 8).

Assume that $\dim(\Pi S_1) = \dim X = n$. Then, $\hat{\varphi}$ is a smooth mapping from an open subset ΠS_1 of X to $X \times I$.

Let us apply this (feedback) function $\hat{\varphi}$ to our linearized system restricted to $\Pi S_1 \times I$. In the coordinates defined above, it rewrites:

$$\begin{aligned} &y = x_1, \\ &\dot{x}_1 = x_2, \dots, \dot{x}_{n-1} = x_n, \dot{x}_n = \psi(x, \hat{\varphi}(x)), \\ &\dot{\xi}_1 = \xi_2, \dots, \dot{\xi}_{n-1} = \xi_n, \dot{\xi}_n = d_x\psi(x, \hat{\varphi}(x))\xi + 0. \\ &\hat{y} = \xi_1. \end{aligned} \tag{4.4}$$

This equation 4.4 is independent of η, the input of the linearized system. Then, let us take $\xi(0) = 0$, $\eta(t) \neq 0$. Then $\hat{y}(t)$ is identically zero, which contradicts the infinitesimal identifiability.

Therefore, ΠS_1 is not open in X, and therefore, it has codimension 1. ΠS has codimension 1, and this shows exactly the property 1.b. of Theorem 2, together with the property 2. (the normal form), in the statement of the theorem. This ends the proof.□

4.2. **Proof of Theorem 3.** Assume that Σ is a system in normal form 4.3, on some open neighborhood O of $x = 0$ in $\mathbb{R}^n$, with $\frac{\partial\psi}{\partial\varphi}(x)$ never vanishing.

It is clear that admissible output trajectories (there is no input in our case), are smooth. Given any smooth function $y(\cdot) : [0, T[\to \mathbb{R}$, an immediate computation with the normal form 4.3 shows that $x_1(t) = y(t)$, $x_2(t) = \dot{y}(t), \ldots, x_n(t) = \frac{d^{n-1}y}{dt^{n-1}}(t)$. Also,

$$\frac{d^n y}{dt^n}(t) = \psi(y, \dot{y}, \ldots, y^{(n-1)}(t), \varphi(t)). \tag{4.5}$$

Assume that $y(\cdot)$ is an admissible output trajectory. For $y, \dot{y}, \ldots, y^{(n-1)}$ fixed, set:

$$\psi_{y,\dot{y},\ldots,y^{(n-1)}}(\varphi) = \psi(y, \dot{y}, \ldots, y^{(n-1)}, \varphi).$$

The function $\psi_{y,\dot{y},\ldots,y^{(n-1)}}$ is monotonous: $\frac{\partial\psi}{\partial\varphi}(\cdot)$ never vanishes. Since $y(\cdot)$ is an admissible output trajectory, then, the equation 4.5 has a solution $\varphi(t)$ for all $t \in [0, T[$, (and this solution is smooth w.r.t. t). By the monotonicity, this solution is unique. This means that Σ is identifiable, for $Z = O$ and $\pi : O \to O$ being the trivial "internal mapping". For the same reason, it is also identifiable if $\pi : O \to Z$, is nontrivial.

To finish, by the normal form, it is just a matter of trivial computation to show that $\Sigma_{|O}$ is differentially identifiable of order $n + 1$, and the uniform infinitesimal identifiability of $\Sigma_{|O}$ is also obvious, from the normal form.

5. The two-output case

Here, we want to prove Theorem 4, give a series of intrinsic conditions corresponding to these normal forms, state and prove several weak converses of Theorem 4 for all these normal forms, and give a few examples.

5.1. **Preliminaries, definitions.** Here, as above, L is the Lie derivative operator on X. Hence, if $f(x, \varphi)$ is a φ-dependant vector field, L_f is the Lie derivative operator with respect to the vector fields $f_\varphi(x) = f(x, \varphi)$, for φ fixed in I.

Let $N(l)$ be the rank at generic points of $X \times I$ of the family E_l of one-forms on X:

$$E_l = \{d_x h_i, d_x L_f h_i, \ldots, d_x L_f^{l-1} h_i, i = 1, 2\}.$$

Set $N(0) = 0$.

This set of generic points U_l, is the intersection of the open sets $\tilde{U}_i$, $i \leq l$, where E_i has maximal rank. U_l is semianalytic, open and dense in $X \times I$. Moreover, $U_{l+1} \subset U_l$.

It is easy to check that $N(l)$ increases strictly by steps of 2, up to $l = k$, and after, (eventually), it increases by steps of 1 up to $l = l_M$, $N(l_M) \leq n$.

It may happen that $k = 0$, *i.e.* $N(1) = 1$.

Lemma 4. *If Σ is uniformly infinitesimally identifiable, then, $N(l_M) = n$.*

The idea in this lemma is that, if it is not true, then, for constant functions $\varphi(\cdot)$, infinitesimal identifiability will be contradicted.

Proof. If $N(l_M) < n$, let $(x_1, \ldots, x_n, \varphi)$ be a coordinate system in an open subset of U_{l_M} formed by φ, and by $N(l_M)$ functions of (x, φ) chosen, in the family $\{h_i, L_f h_i, \ldots, L_f^{l_M - 1} h_i, i = 1, 2\}$, (the $N(l_M)$ first coordinates), and other x-coordinates. In these coordinates, it is easy to see that, for the constant function $\varphi(\cdot) \equiv \varphi_0$, Σ can be rewritten as:

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, \ldots, x_{N(l_M)}, \varphi_0),\\
&.\\
\dot{x}_{N(l_M)} &= f_{N(l_M)}(x_1, \ldots, x_{N(l_M)}, \varphi_0),\\
\dot{x}_{N(l_M)+1} &= f_{N(l_M)+1}(x_1, \ldots, x_n, \varphi_0),\\
&.\\
\dot{x}_n &= f_n(x_1, \ldots, x_n, \varphi_0),\\
y_1 &= h_1(x_1, \ldots, x_{N(l_M)}, \varphi_0),\\
y_2 &= h_2(x_1, \ldots, x_{N(l_M)}, \varphi_0).
\end{aligned}$$

Then, taking $\eta \equiv 0$, and $\xi_1(0) = \cdots = \xi_{N(l_M)}(0) = 0$, $\xi_{N(l_M)+1}(0) \neq 0$, in the equation of the first variation, we see that the solution $\xi(t)$ verifies $\xi_1(t) = \cdots = \xi_{N(l_M)}(t) = 0$, on a small time interval $[0, T]$, $T > 0$. This implies that on this time interval, the output $\hat{y} = (\hat{y}_1, \hat{y}_2)$ of the first variation is identically zero. A contradiction with infinitesimal identifiability for $\varphi(\cdot) \equiv \varphi_0$. □

Definition 6. *We define r, the* **"order"** *of the system, as the first integer such that $d_\varphi L_f^r h$ does not vanish identically.*

Lemma 5. *If Σ is uniformly infinitesimally identifiable, then, $r \leq l_M$.*

Proof. Assume $r \geq l_M + 1$. Let us take again a coordinate system on an open subset of X, formed by functions of the family $\{h_i, L_f h_i, \ldots, L_f^{l_M - 1} h_i, i = 1, 2\}$. These functions are functions of x only, since $r \geq l_M + 1$. By the previous lemma, this is possible. It is obvious that, in these coordinates, the system can be rewritten:

$$\dot{x} = f(x), \quad y = h(x).$$

Let us take $\xi(0) = 0$, but a function $\eta(t)$ nonzero, in the first variation. Then, the output $\hat{y}$ of the first variation is identically zero, on some open time interval. This contradicts the infinitesimal identifiability. □

Definition 7. *A system Σ is* **regular** *if $N(l_M) = n$ and $r \leq l_M$.*

If a system is uniformly infinitesimally identifiable, then it is regular, by the 2 previous lemmas. From now on, in this section, we will assume that systems Σ under consideration are regular.

The integer k is the first with the following properties:

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^k h_1 \wedge d_x L_f^k h_2 \equiv 0, \quad \text{but}$$
$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_1 \wedge d_x L_f^{k-1} h_2 \neq 0 \text{ (not identically zero).}$$

If $r = k$, there are three possibilities:

A. $n = 2k$;

B.

B.1.

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_2 \neq 0$; or,

B.2.

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_1 \neq 0$;

C.

C.1

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_2 \equiv 0$, $d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \, \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \equiv 0$,

or

C.2

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_1 \equiv 0$, $d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \, \ldots \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \equiv 0$

Definition 8. *Let Σ be a* ***regular*** *system. We say that Σ has:*
*-**type 1** if $r > k$, or $r = k$ but **C**. is satisfied,*
*-**type 2** if $r < k$, or $r = k$ but **B**. is satisfied,*
*-**type 3** if $r = k$ and **A**. is satisfied.*

Lemma 6. *Types 1, 2 and 3 exhaust the class of regular systems, and form a partition of this class.*

Proof. For a system with $r \neq k$, it is clear that it is either of type 1 or type 2 and not both. There can be only problems for $r = k$. If we show that, for $r = k$:
i) cases B. and C. exhaust all regular systems with $n > 2k$,
ii) cases B. and C. do not intersect;
then, the theorem is proved since $n \geq 2k$.

Assume that Σ is simultaneously B. and C., then:

If Σ is B.1., it cannot be C.2. which contradicts $d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_1 \neq 0$, and it cannot be C.1., which contradicts $d_\varphi L_f^kh_2 \neq 0$.

On the same way, if Σ is B.2., it cannot be C.1. which contradicts $d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_2 \neq 0$, and it cannot be C.2., which contradicts $d_\varphi L_f^kh_1 \neq 0$.

This shows ii).

Proof of i): By definition of k :

$$d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^kh_1 \wedge d_xL_f^kh_2 \equiv 0,$$
$$d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_1 \wedge d_xL_f^{k-1}h_2 \neq 0.$$

Since $n > 2k$,either:

$$(a) \quad d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_1 \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_1 \neq 0,$$

or:

$$(b) \quad d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_1 \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_2 \neq 0,$$

or both: were it otherwise, $l_M = k, N(l_M) = 2k$, and $n = 2k$, by Lemma 4.

Assume that (a) and (b) hold simultaneously.

Then, since $d_\varphi L_f^kh = d_\varphi L_f^rh \neq 0$ by definition of r, either (α), $d_\varphi L_f^kh_1 \neq 0$, or (β), $d_\varphi L_f^kh_2 \neq 0$ Assume (α) (the case (β) is symmetric). Then we are in case B.2.

Assume that (a) holds, but not (b) (the case (b) holds but not (a) is symmetric). Then:

$$d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_1 \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_1 \neq 0,$$
$$d_xh_1 \wedge d_xh_2 \wedge d_xL_fh_1 \wedge \ldots \wedge d_xL_f^{k-1}h_1 \wedge d_xL_f^{k-1}h_2 \wedge d_xL_f^kh_2 \equiv 0.$$

If $d_\varphi L_f^kh_2 \neq 0$, we are in case B.1., if $d_\varphi L_f^kh_2 \equiv 0$, we are in case C.1. This ends the proof. □

Type 2 regular systems:

For a regular system of type 2, eventually interchanging the role of h_1, h_2, we can assume that $d_\varphi L_f^rh_2(x, \varphi) \neq 0$. In a neighborhood of a point $(x_0, \varphi_0) \in U_{l_M}$, such that $L_f^rh_2(x_0, \varphi_0) = u_0$ and $d\varphi L_f^rh_2(x_0, \varphi_0) \neq 0$, there is an analytic function $\Phi^*(x, u)$, such that $L_f^rh_2(x, \Phi^*(x, u)) = u$. Let us consider the "auxiliary system" Σ_A :

$$\Sigma_A \left\{ \begin{array}{l} \dot{x} = f(x, \Phi^*(x, \tilde{\varphi})) = F(x, \tilde{\varphi}) \\ y = h(x, \Phi^*(x, \tilde{\varphi})) = H(x, \tilde{\varphi}). \end{array} \right.$$

This system is well defined and intrinsic, over an open set $V_{x_0} \times V_{u_0} \subset X \times \mathbb{R}$.

By construction, the integer r (the order) associated with this auxiliary system is the same as the one of the given system Σ.

Moreover, the following flags D and D^A of integrable distributions over V_{x_0} :

$$D_0(x) = T_xX,\ D_1(x) = Ker(d_xh(x)), \ldots, D_r(x) = D_{r-1}(x) \cap Ker(d_xL_f^{r-1}h(x))$$
$$D = \{D_0 \supset D_1 \supset \ldots \supset D_r\};$$

and

$$D_0^A(x) = T_xX,\ D_1^A(x) = Ker(d_xH(x)), \ldots, D_r^A(x) = D_{r-1}^A(x) \cap Ker(d_xL_F^{r-1}H(x))$$
$$D^A = \{D_0^A \supset D_1^A \supset \ldots \supset D_r^A\},$$

are equal.

Let us "prolong" the auxiliary flag D^A, in the following way:

$$\begin{aligned}
D^A_{r+1}(x,\tilde{\varphi}) &= D^A_r(x) \cap Ker(d_x L^r_F H_1(x,\tilde{\varphi})),\\
D^A_{i+1}(x,\tilde{\varphi}) &= D^A_i(x) \cap Ker(d_x L^i_F H_1(x,\tilde{\varphi})),\\
D^A(\tilde{\varphi}) &= \{D^A_0 \supset D^A_1 \supset \ldots \supset D^A_r \supset D^A_{r+1}(\tilde{\varphi}) \supset \ldots \supset D^A_l(\tilde{\varphi}) = D^A_{l+1}(\tilde{\varphi})\},
\end{aligned}$$

where l is the first integer such that $D^A_l(x,\tilde{\varphi}) = D^A_{l+1}(x,\tilde{\varphi})$ at generic points.

Definition 9. *The auxiliary flag $D^A(\tilde{\varphi})$ is **regular** on an open subset $U \subset X \times I$, if $D^A_l(\tilde{\varphi}) = \{0\}$, and all the other $D^A_i(\tilde{\varphi})$ have constant rank first $n-2i$ $(i \leq r)$, second $n-r-i$ $(r < i < l)$, third, 0 $(i \geq l = n-r)$; on this open set.*

Definition 10. *The auxiliary flag $D^A(\tilde{\varphi})$ is **uniform** on an open subset $U \subset X \times I$, if it is regular, and independent of $\tilde{\varphi}$.*

5.2. **Normal form for a uniform auxiliary flag (systems of type 2).** Here, we consider a regular system Σ of type 2, with a uniform auxiliary flag over $X \times I$. The flag being integrable (in the sense that it is a flag of integrable distributions of constant rank), around each point of $X \times I$, we can find coordinates x such that:

$$\begin{aligned}
D^A_{l-1}(\tilde{\varphi}) &= Span\{\frac{\partial}{\partial x_n}\}, \ldots, D^A_{r+1}(\tilde{\varphi}) = Span\{\frac{\partial}{\partial x_n}, \ldots, \frac{\partial}{\partial x_{2r+2}}\},\\
D^A_r(\tilde{\varphi}) &= Span\{\frac{\partial}{\partial x_n}, \ldots, \frac{\partial}{\partial x_{2r+1}}\}, D^A_1(\tilde{\varphi}) = Span\{\frac{\partial}{\partial x_n}, \ldots, \frac{\partial}{\partial x_3}\}.
\end{aligned}$$

Moreover, we can take as x coordinates:

$$x_1 = h_1(x), x_2 = h_2(x), \ldots, x_{2r-1} = L^{r-1}_F h_1(x), x_{2r} = L^{r-1}_F h_2(x).$$

Then, the auxiliary system Σ_A can be written:

$$\begin{aligned}
&\dot{x}_1 = x_3,\ \dot{x}_2 = x_4,\\
&\quad ..\\
&\dot{x}_{2r-3} = x_{2r-1},\ \dot{x}_{2r-2} = x_{2r},\\
&\dot{x}_{2r-1} = F_{2r-1}(x,\tilde{\varphi}),\ \dot{x}_{2r} = F_{2r}(x,\tilde{\varphi}) = \tilde{\varphi},\\
&\quad ..\\
&\dot{x}_n = F_n(x,\tilde{\varphi}),\\
&y_1 = x_1,\ y_2 = x_2.
\end{aligned} \tag{5.1}$$

Since $D^A_{r+1}(\tilde{\varphi}) = Span\{\frac{\partial}{\partial x_n}, \ldots, \frac{\partial}{\partial x_{2r+2}}\}$, we must have $\frac{\partial F_{2r-1}}{\partial x_{2r+2}} = \cdots = \frac{\partial F_{2r-1}}{\partial x_n} = 0$, and $\frac{\partial F_{2r-1}}{\partial x_{2r+1}} \neq 0$, or, locally:

$$F_{2r-1} = F_{2r-1}(x_1, \ldots, x_{2r+1}, \tilde{\varphi}),\ \frac{\partial F_{2r-1}}{\partial x_{2r+1}} \neq 0.$$

Repeating this reasoning, we get that:

$$\begin{aligned}
F_{n-2} &= F_{n-2}(x_1, \ldots, x_{n-1}, \tilde{\varphi}),\ \frac{\partial F_{n-2}}{\partial x_{n-1}} \neq 0,\\
F_{n-1} &= F_{n-1}(x_1, \ldots, x_n, \tilde{\varphi}),\ \frac{\partial F_{n-1}}{\partial x_n} \neq 0,\\
F_n &= F_n(x,\tilde{\varphi}).
\end{aligned}$$

Conversely, if the system Σ^A is like that, then it is just a trivial computation to check that the auxiliary flag is uniform.

Hence, replacing $\tilde{\varphi}$ by $\Phi(x,\varphi) = L^r_f h_2(x,\varphi)$, and reversing the role of h_1, h_2, we get the following result:

Theorem 8. ***(Normal form for a uniform auxiliary flag)*** *A system Σ has a uniform auxiliary flag around (x_0, φ_0), iff there is a neighborhood $V_{x_0} \times I_{\varphi_0}$ of (x_0, φ_0), and coordinates on V_{x_0} such that Σ can be written:*

$$\begin{aligned}
y_1 &= x_1,\ y_2 = x_2,\\
\dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\
&..\\
\dot{x}_{2r-3} &= x_{2r-1},\ \dot{x}_{2r-2} = x_{2r},\\
\dot{x}_{2r-1} &= \Phi(x,\varphi),\ \dot{x}_{2r} = F_{2r}(x_1,\ldots,x_{2r+1},\Phi(x,\varphi)),\\
\dot{x}_{2r+1} &= F_{2r+1}(x_1,\ldots,x_{2r+2},\Phi(x,\varphi)),\\
&..\\
\dot{x}_{n-1} &= F_{n-1}(x,\Phi(x,\varphi)),\\
\dot{x}_n &= F_n(x,\varphi),
\end{aligned}$$

with $\frac{\partial\Phi}{\partial\varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0, \ldots, \frac{\partial F_{n-1}}{\partial x_n} \neq 0.$

5.3. Statement of the results for the two-output case.

Theorem 9. ***(main result in the 2-output case)*** *If Σ is uniformly infinitesimally identifiable, **(hence regular)**, then, there is an open-dense subanalytic subset $\tilde{U}$ of $X \times I$, such that at each point (x_0, φ_0) of $\tilde{U}$, Σ has the following properties, on a neighborhood of (x_0, φ_0):*
-If Σ has type 2, the auxiliary flag is uniform,
-If Σ has type 1, then, $N(r) = n$.

Remark 8. *a. In the case where Σ has type 3, then, there is no other requirement.*

*b. In the case of type 1 assume $r = k$. Then, we have **C.,** which implies $n > 2k$. But $N(r) = n = N(k) = 2k$. Then, **for type 1 systems,** $r = k$ **cannot happen for a uniformly infinitesimally identifiable system.***

This theorem is in fact equivalent to the theorem in the introduction, Theorem 4.

These two equivalent theorems (Theorems 9, 4) have a weak converse:

Theorem 10. *Assume that Σ satisfies the equivalent conditions of theorems 9, 4, on some subset $V_{x_0} \times I_{\varphi_0}$ of $X \times I$ (so that, taking V_{x_0}, I_{φ_0} small enough, the restriction $\Sigma_{|V_{x_0}\times I_{\varphi_0}}$ has one of the three normal forms above on $V_{x_0} \times I_{\varphi_0}$). Then, in case type 1, type 2, (normal forms 1.4, 1.5) $\Sigma_{|V_{x_0}\times I_{\varphi_0}}$ is uniformly infinitesimally identifiable **and identifiable**. In case type 3 (normal form 1.6), this is also true, eventually restricting the neighborhoods V_{x_0}, I_{φ_0}.*

Also, in the special case of type 1, there is a stronger result:

Theorem 11. *Assume Σ is uniformly infinitesimally identifiable, **(hence regular)**. Assume that Σ has type 1. Then, there is an open-dense subanalytic subset $\tilde{X}$ of X, such that each point x_0 of $\tilde{X}$, has a neighborhood V_{x_0}, and coordinates x on V_{x_0} such that the system Σ restricted to $V_{x_0} \times I$, denoted by $\Sigma_{|V_{x_0}}$, has the normal form 1.4 (globally over $V_{x_0} \times I$). Conversely, if it is the case, then, the restriction $\Sigma_{|V_{x_0}}$ is uniformly infinitesimally identifiable **and identifiable**.*

Several points in this last theorem are not true in the case of types 2 and 3.

Example 1. *(Type 2) consider the type 2 system:*

$$\begin{aligned}
y_1 &= x_1, y_2 = x_2,\\
\dot{x}_1 &= x_3\cos\varphi,\ \dot{x}_2 = x_3\sin\varphi,\\
\dot{x}_3 &= f(x).
\end{aligned}$$

where $x_3 > 0$ ($x \in X = \mathbb{R}^2 \times \mathbb{R}_+$), and $I = [-A, A]$, for $A > 0$, sufficiently large.

For this system, $k = 1$, $r = 1$, and **B**. is always satisfied.

(a) The normal form 1.5 is met only locally with respect to φ (changing the role of h_1 and h_2 depending on φ);

(b) For any open subset $\tilde{X}$ of X, $\Sigma_{|\tilde{X}}$ is never identifiable: for $\varphi(t)$ and $x(0)$ arbitrary, $(x(0), \varphi(t))$ and $(x(0), \varphi(t) + 2\pi)$ produce the same output.

Example 2. *(Type 3) Consider the type 3 system on* $\mathbb{R}^2$:

$$y_1 = x_1, y_2 = x_2,$$
$$\dot{x}_1 = \cos\varphi,\ \dot{x}_2 = \sin\varphi,$$

where, as above, $I = [-A, A]$, for $A > 0$, sufficiently large. Again, $(x(0), \varphi(t))$ and $(x(0), \varphi(t) + 2\pi)$ produce the same output.

Example 3. *(Type 2) Consider the system on a subset* $X \times I$ *of* $\mathbb{R}^2 \times \mathbb{R}$, *with* X *open*:

$$y_1 = \frac{1}{2}(\varphi - x_2)^2,\ y_2 = x_1$$
$$\dot{x}_1 = x_2,\ \dot{x}_2 = 0.$$

For this system, $r = 0$, $k = 1$.

If we take for X and I neighborhoods of zero in $\mathbb{R}^2$ and $\mathbb{R}$, then, the auxiliary flag is not uniform (on any open subset of X with zero in its closure), and the normal form 1.5 is not met. On the contrary, if X is a neighborhood of $(x_1(0), x_2(0))$, and I is a neighborhood of φ_0, provided that these neighborhoods are small and $\varphi_0 \neq x_2(0)$, the auxiliary flag is uniform, and the normal form is met.

5.4. **Proof of the results for the 2-output case.**

First, let us show that **theorems 9, 4 are equivalent.**

It is just a matter of simple computations to see that normal form 1.5 (resp. 1.4, 1.6) in Theorem 4 imply the conditions "type 2" (resp. "type 1", "type 3" of theorem 9.

Conversely it is easy to check that, in Theorem 9, conditions "type 1", "type 3" imply the corresponding normal forms in Theorem 4. For type 3, see the (trivial) details in the proof of Theorems 9, 4, below. Condition "type 2" is equivalent to the normal form 1.5 by Theorem 8.

Proof. (of Theorem 10).

1. Normal form 1.4. From the normal form, computing the first variation, it follows immediately that $\hat{y}(t) = 0$ on some time interval $[0, \varepsilon]$ ($\hat{y}$ the output of the first variation), implies $\xi(0) = 0$ on the same time interval, where $\xi(0) \in T_{x_0}X$ is the initial condition for the first variation. Deriving once more, we get that $\eta(\cdot)$ (the variation control) vanishes for almost all t. This shows that $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$ is uniformly infinitesimally identifiable.

Now, from the normal form, differentiating the output $y(t)$ a certain number of times, one reconstructs the full state $x(t) = (x_1(t), \ldots, x_n(t))$ of the system.

Knowing $x(t)$, the equation $\dot{x}_n(t) = f_n(x(t), \varphi(t))$ can be solved with respect to φ, at almost all $t \in [0, \varepsilon]$: if $\frac{\partial f_n}{\partial \varphi}$ never vanishes, then the function of φ: $\dot{x}_n(t) - f_n(x(t), \varphi)$ is a monotonous function of φ. Then, $\dot{x}_n(t)$ determines φ uniquely, for almost all t. This shows that $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$ is identifiable.

2. Normal form 1.6. Again, computing the first variation (with output $\hat{y}$), we see immediately that $\hat{y}(t) = 0$ on some time interval $[0, \varepsilon]$ implies that $\xi(t) = 0$ on this interval. The condition that $(\frac{\partial f_{n-1}}{\partial \varphi}$, $\frac{\partial f_n}{\partial \varphi})$ does not vanish implies that $\eta(t) = 0$ for almost all $t \in [0, \varepsilon]$ if $\hat{y}(\cdot) = 0$. This shows that $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$is uniformly infinitesimally identifiable.

Now, from the normal form, any output trajectory $y(t)$ determines $x(t)$ by differentiation. The condition that $(\frac{\partial f_{n-1}}{\partial \varphi}$, $\frac{\partial f_n}{\partial \varphi})$ does not vanish implies that (restricting the neighborhood $V_{x_0} \times I_{\varphi_0}$, and eventually exchanging the role of h_1, h_2) that $\frac{\partial f_{n-1}}{\partial \varphi}$ never vanishes. Then the same argument as for the normal form 1.4 shows that $\varphi(t)$ is determined almost everywhere by one more differentiation. $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$ is identifiable.

3. Normal form 1.5. Computing with the first variation, assuming that the output $\hat{y}(t)$ is identically zero on some interval $[0, \varepsilon]$ we get that $\xi_1(t), \ldots, \xi_{2r}(t)$ are identically zero on the same interval. The two equations:

$$\dot{x}_{2r-1} = \Phi(x, \varphi),\ \dot{x}_{2r} = F_{2r}(x_1, \ldots, x_{2r+1}, \Phi(x, \varphi)),$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0$, show that $d_x\Phi.\xi(t) + d_\varphi\Phi.\eta(t)$ and $\xi_{2r+1}(t)$ are identically zero for almost all t, and by continuity, $\xi_{2r+1}(t) = 0$ on $[0, \varepsilon]$. By induction, using the fact that $\frac{\partial F_{2r+i}}{\partial x_{2r+i+1}} \neq 0$, we get that $\xi_{2r+i+1}(t)$ is identically zero, and at the end, $\xi(t)$ is identically zero.

The equation of ξ_{2r-1} again, shows that η is zero almost everywhere: $0 = d_x\Phi.\xi(t) + d_\varphi\Phi.\eta(t)$ *a.e.* Hence, Σ is uniformly infinitesimally identifiable.

Now, the output $y(t) = (x_1, x_2)(t)$, by differentiation, determines $(x_3, x_4)(t)$ for all $t \in [0, \varepsilon]$. By differentiation of $(x_3, x_4)(t)$, we get $(x_5, x_6)(t)$, and so on. Once we know (x_{2r-1}, x_{2r}), with the same reasoning, we determine $\Phi(x, \varphi)$ and $F_{2r}(x_1, \ldots, x_{2r+1}, \Phi(x, \varphi))$ almost everywhere w.r.t. t. Now, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0$ (never vanishes), shows that F_{2r} is monotonous w.r.t. x_{2r+1}, the other variables being fixed. Hence, since we know Φ and $x_1, \ldots, x_{2r}$, we can determine uniquely $x_{2r+1}(t)$, $t \in [0, \varepsilon]$ (by continuity) from the knowledge of the values of F_{2r} a.e. By induction, we determine $x(t)$ for all $t \in [0, \varepsilon]$.

Solving the equation $\dot{x}_{2r-1} = \Phi(x(t), \varphi(t))$ (using again the monotonicity of Φ w.r.t φ, since $\frac{\partial\Phi}{\partial\varphi} \neq 0$), determines $\varphi(t)$ for almost all t, hence determines φ as an $L_\infty([0, \varepsilon], I_{\varphi_0})$ function. Hence, $\Sigma_{|V_{x_0} \times I_{\varphi_0}}$ is identifiable. □

Proof. (of Theorem 11.)

Assume that Σ is regular, type 1. Then, consider the subset $\tilde{X}$ of X where

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{k-1} h_1 \wedge d_x L_f^{k-1} h_2 \neq 0.$$

$\tilde{X}$ is semi-analytic, open dense. On a neighborhood of each point of $\tilde{X}$, we can chose as coordinates $(x_1, x_2) = h(x)$, ..., $(x_{2k-1}, x_{2k}) = L_f^{k-1} h(x)$. Let us assume that $r > k$ or **C.2** is satisfied (the case **C.1** is obtained by exchanging the role of h_1 and h_2). Then, Σ can be written locally in x :

$$\begin{aligned}
y_1 &= x_1,\ y_2 = x_2,\\
\dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\
&..\\
\dot{x}_{2k-3} &= x_{2k-1},\ \dot{x}_{2k-2} = x_{2k},\\
\dot{x}_{2k-1} &= f_{2k-1}(x_1, \ldots, x_{2k+1}),\\
\dot{x}_{2k} &= x_{2k+1},\\
&..\\
\dot{x}_{N(r)-1} &= x_{N(r)},\\
\dot{x}_{N(r)} &= f_{N(r)}(x, \varphi),\\
&..\\
\dot{x}_n &= f_n(x, \varphi).
\end{aligned}$$

with $\frac{\partial f_{N(r)}}{\partial\varphi}$ nonidentically zero.

Moreover, if $r = k$, then $f_{2k-1} = f_{2k-1}(x_1, \ldots, x_{2k})$, and $\dot{x}_{2k} = f_{2k}(x, \varphi) = f_{N(r)}(x, \varphi)$.

Assume that $N(r) < n$. Then, let us consider the initial condition $\xi(0)$ for the first variation: $\xi_1(0) = \cdots = \xi_{N(r)}(0) = 0$, $\xi_{N(r)+1}(0) \neq 0$. Chose the feedback function $\eta(x, \varphi, \xi) = -\frac{d_x f_{N(r)}(x,\varphi)\xi}{d_\varphi f_{N(r)}(x,\varphi)}$. For this, chose any function φ (φ constant for instance). We have, for the first variation:

$$\begin{aligned}
\dot{\xi}_1 &= \xi_3,\ \dot{\xi}_2 = \xi_4\\
&..\\
\dot{\xi}_{2k-1} &= \frac{\partial f_{2k-1}}{\partial x_1}\xi_1 + \cdots + \frac{\partial f_{2k-1}}{\partial x_{2k+1}}\xi_{2k+1},\\
\dot{\xi}_{2k} &= \xi_{2k+1},\\
&..\\
\dot{\xi}_{N(r)-1} &= \xi_{N(r)},\\
\dot{\xi}_{N(r)} &= 0 \ \text{ by construction.}
\end{aligned}$$

Moreover, if $r = k$, $\frac{\partial f_{2k-1}}{\partial x_{2k+1}} \equiv 0$, and $\dot{\xi}_{N(r)} = \dot{\xi}_{2k} = 0$.

Hence, for $t > 0$ small, $(\xi_1(t), \xi_2(t)) = \hat{y}(t) = 0$ (remind that $\hat{y}$ is the output of the first variation). This contradicts the uniform infinitesimal identifiability.

Hence, $N(r) = n$. Let $E = \{(x, \varphi) | d_\varphi L_f^r h(x, \varphi) = 0\}$. Let πE be the projection of E on X. $X\backslash\pi E$ is subanalytic. Assume that πE contains an open set $\tilde{X}$. By Hardt's Theorem on the stratification of mappings, there is a smooth (analytic) mapping $\hat{\varphi} : \tilde{X} \to E$ (restricting $\tilde{X}$ eventually). Then, choosing for the first variation the initial condition $\xi(0) = 0$, and any nonzero variation $\eta(t)$, but the feedback "control" $\varphi(t) = \hat{\varphi}(x(t))$, (for small times), we get by construction that the trajectory $\xi(t)$ of the first variation is in the zero section of TX. A contradiction with the uniform infinitesimal identifiability.

We conclude that πE has codimension 1, and $\tilde{X} = X \backslash \pi E$ contains an open dense set. This shows the first part of the theorem.

Since on $\tilde{X} \times I$, $d_\varphi L_f^r h(x, \varphi)$ never vanishes by construction, the proof of the last part of the theorem is exactly the same as the proof of Theorem 10, part "type 1". □

Proof. (of Theorems 9, 4.)

We already know, by the beginning of this section, that these theorems are equivalent.

The proof of Theorem 4, **"type 1"**, is already contained in the proof of Theorem 11 (which is stronger, for type 1 systems).

Type 3: (the most simple case). It is clear that if $r = k, n = 2k$, the system can be locally written under normal form 1.6 around any point of an open dense semi-analytic subset of $X \times I : x_1 = h_1(x)$, $x_2 = h_2(x), \ldots, x_{2k-1} = L_f^{k-1} h_1(x)$, $x_{2k} = L_f^{k-1} h_2(x)$ are adequate coordinates around any point where they are independent. The set of these points is semianalytic open, dense in X. The fact that $r = k$ implies that $d_\varphi L_f^k h(x, \varphi)$ is not identically zero. Hence, it is never zero on a semi-analytic open dense subset of $X \times I$. On the intersection of these two semianalytic subsets of $X \times I$, (the first one can be considered as such), in the coordinates just defined, the system is under the normal form 1.6.

Type 2: This is the most difficult case. In that case, by the definition of type 2, eventually interchanging h_1, h_2, we may assume that, around any point (x_0, φ_0) of the complement $(X \times I) \backslash \tilde{U}$ of the codimension 1 analytic set $\tilde{U} \subset X \times I$:

$$\tilde{U} =$$

$$\{(x, \varphi) | d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge \ldots \wedge d_x L_f^{r-1} h_2 \wedge d_x L_f^r h_1 = 0\}$$
$$\cup \{(x, \varphi) | d_\varphi L_f^r h_2 = 0\},$$

we can find coordinates x such that the system can be written:

$$\begin{aligned}
&y_1 = x_1, \ y_2 = x_2, \\
&\dot{x}_1 = x_3, \ \dot{x}_2 = x_4, \\
&\quad .. \\
&\dot{x}_{2r-3} = x_{2r-1}, \ \dot{x}_{2r-2} = x_{2r}, \\
&\dot{x}_{2r-1} = f_{2r-1}(x, \varphi), \dot{x}_{2r} = f_{2r}(x, \varphi), \\
&\quad .. \\
&\dot{x}_n = f_n(x, \varphi),
\end{aligned} \tag{5.2}$$

with $\frac{\partial f_{2r}}{\partial \varphi} \neq 0$, and $dx_1 \wedge \ldots \wedge dx_{2r} \wedge d_x f_{2r-1} \neq 0$.

Here, in case $r = k$, we treat **B.1**. only. In case $r < k$, we chose h_2 for $d_\varphi L_f^r h_2 \neq 0$, and h_1 is the other. This ensures that $\tilde{U}$ has codimension 1.

Let $\hat{y}(t)$ be an output function of the first variation of the system, which is identically zero on some time interval $[0, \varepsilon]$. Let $x(t)$, $\xi(t)$, $\varphi(t)$, $\eta(t)$ be the corresponding trajectories. We know that the couple $(\xi(t), \eta(t))$ has to be identically zero, by the uniform infinitesimal identifiability. Differentiating $\hat{y}(t) = 0$ r times, we get:

$\xi_1(t) = \cdots = \xi_{2r}(t) = 0$ for all $t \in [0, \varepsilon]$, and:

$$\begin{aligned}
d_x f_{2r-1}(x(t), \varphi(t))\xi(t) + d_\varphi f_{2r-1}(x(t), \varphi(t))\eta(t) = 0, \\
d_x f_{2r}(x(t), \varphi(t))\xi(t) + d_\varphi f_{2r}(x(t), \varphi(t))\eta(t) = 0,
\end{aligned} \tag{5.3}$$

for almost all $t \in [0, \varepsilon]$.

Hence, the system of equations (5.3), must have no smooth solution $(\eta, \varphi)(x, \xi)$, in a neighborhood of $(x_0, \xi(0))$, $\xi(0) \neq 0$, in $X \times \mathbb{R}^{n-2r} = \{(x, \xi(0)) | \xi_1(0) = \cdots = \xi_{2r}(0) = 0\}$, with φ close to φ_0 and η arbitrary:

Indeed, assume that there is a solution $(\eta, \varphi)(x, \xi)$. Then, consider the feedback system:

$$\begin{aligned}
\dot{x} &= f(x, \varphi(x, \xi)),\\
\dot{\xi}_{2r+1} &= d_x f_{2r+1}(x, \varphi(x, \xi))\xi + d_\varphi f_{2r+1}(x, \varphi(x, \xi))\eta(x, \xi),\\
&\dots\\
\dot{\xi}_n &= d_x f_n(x, \varphi(x, \xi))\xi + d_\varphi f_n(x, \varphi(x, \xi))\eta(x, \xi),
\end{aligned}$$

in which $\xi_1 = \dots = \xi_{2r} = 0$.

This is a smooth differential equation on an open subset of $X \times \mathbb{R}^{n-2r}$. It has a smooth solution $x(t), \xi(t)$. We set $\hat{\varphi}(t) = \varphi(x(t), \xi(t))$, $\hat{\eta}(t) = \eta(x(t), \xi(t))$, with $\xi(t) \neq 0$.

By construction, we have:

$$\begin{aligned}
\dot{x} &= f(x, \hat{\varphi}(t)),\\
\dot{\xi} &= T_x f(x, \hat{\varphi}(t))\xi + d_\varphi f(x, \hat{\varphi}(t))\hat{\eta}(t),
\end{aligned}$$

because, for $\xi_1, \dots, \xi_{2r}$, these equations read $\xi_1(t) = \dots = \xi_{2r}(t) = 0$ for all t (small).

Hence, in particular $\hat{y}_1(t) = 0$, $\hat{y}_2(t) = 0$. This is impossible, by uniform infinitesimal identifiability.

Therefore, by the crucial Lemma 7, Section 8, we get that:

there are neighborhoods V_{x_0}, V_{φ_0}, and coordinates $\tilde{x}$ on V_{x_0}, with $\tilde{x}_1 = x_1, \dots, \tilde{x}_{2r} = x_{2r}$, such that:

$$\tilde{x}_{2r+1} = \Phi_{\varphi_0}(\tilde{x}_1, \dots, \tilde{x}_{2r}, f_{2r-1}(\tilde{x}, \varphi), f_{2r}(\tilde{x}, \varphi)), \tag{5.4}$$

for all $(\tilde{x}, \varphi) \in V_{x_0} \times V_{\varphi_0}$.

Moreover,

$$\left(\frac{\partial \Phi_{\varphi_0}}{\partial f_{2r-1}}, \frac{\partial \Phi_{\varphi_0}}{\partial f_{2r}}\right) \text{ never vanishes,} \tag{5.5}$$

$$\frac{\partial f}{\partial \varphi} \bar{\wedge} \frac{\partial f}{\partial \tilde{x}_{2r+1}} \text{ never vanishes,} \tag{5.6}$$

on $V_{x_0} \times V_{\varphi_0}$, where $\frac{\partial f}{\partial \varphi} \bar{\wedge} \frac{\partial f}{\partial \tilde{x}_{p+1}}$ denotes the determinant of the 2×2 matrix formed by $\frac{\partial (f_{2r-1}, f_{2r})}{\partial \varphi}$ and $\frac{\partial (f_{2r-1}, f_{2r})}{\partial \tilde{x}_{2r+1}}$.

In fact, $\frac{\partial \Phi_{\varphi_0}}{\partial f_{2r-1}} \neq 0$, because, if it is zero, then $\frac{\partial \Phi_{\varphi_0}}{\partial f_{2r}} \neq 0$ by (5.5), and, differentiating (5.4) with respect to φ, we get a contradiction (since $\frac{\partial f_{2r}}{\partial \varphi} \neq 0$).

Therefore, we can apply the implicit function theorem to (5.4): restricting may be our neighborhood $V_{x_0} \times V_{\varphi_0}$, we find a smooth function $\bar{\Phi}$ such that

$$f_{2r-1}(\tilde{x}, \varphi) = \bar{\Phi}(\tilde{x}_1, \dots, \tilde{x}_{2r}, \tilde{x}_{2r+1}, f_{2r}(\tilde{x}, \varphi)), \tag{5.7}$$

with $\frac{\partial \bar{\Phi}}{\partial \tilde{x}_{2r+1}} \neq 0$.

Now, set $\Delta = \frac{\partial (f_{2r-1}, f_{2r})}{\partial \varphi} \bar{\wedge} \frac{\partial (f_{2r-1}, f_{2r})}{\partial \tilde{x}_{2r+1}}$. An easy computation shows that $\Delta = -\frac{\partial f_{2r}}{\partial \varphi} \frac{\partial \bar{\Phi}}{\partial \tilde{x}_{2r+1}}$. This says no more than $\frac{\partial f_{2r}}{\partial \varphi} \neq 0$, which we already know. Equation (5.3) becomes:

$$\begin{aligned}
&1. \quad \frac{\partial \bar{\Phi}}{\partial \tilde{x}_{2r+1}} \xi_{2r+1} + \frac{\partial \bar{\Phi}}{\partial f_{2r}} d_x f_{2r}(\tilde{x}, \varphi)\xi + \frac{\partial \bar{\Phi}}{\partial f_{2r}} d_\varphi f_{2r} \eta = 0,\\
&2. \quad d_x f_{2r}(\tilde{x}, \varphi)\xi + d_\varphi f_{2r} \eta = 0.
\end{aligned} \tag{5.8}$$

Equation (5.8, 2) implies $\eta = -\frac{d_x f_{2r}(\tilde{x}, \varphi)\xi}{d_\varphi f_{2r}}$, which replaced in (5.8,1) gives $\frac{\partial \bar{\Phi}}{\partial \tilde{x}_{2r+1}} \xi_{2r+1} = 0$, $\xi_{2r+1} \equiv 0$. We have shown that $\hat{y} \equiv 0$ implies $(\xi_1, \dots, \xi_{2r+1}) \equiv 0$, and, making the change of notations $x := \tilde{x}$, (5.2) rewrites:

$$
\begin{aligned}
y_1 &= x_1,\ y_2 = x_2,\\
\dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\
&\dots\\
\dot{x}_{2r-3} &= x_{2r-1},\ \dot{x}_{2r-2} = x_{2r},\\
\dot{x}_{2r-1} &= f_{2r-1}(x_1,\dots,x_{2r+1}, f_{2r}(x,\varphi)), \dot{x}_{2r} = f_{2r}(x,\varphi),\\
\dot{x}_{2r+1} &= f_{2r+1}(x,\varphi),\\
&\dots\\
\dot{x}_n &= f_n(x,\varphi),
\end{aligned} \tag{5.9}
$$

$\frac{\partial f_{2r}}{\partial \varphi} \neq 0$, and $\frac{\partial f_{2r-1}}{\partial x_{2r+1}} \neq 0$. With $\xi_1(0) = \dots = \xi_{2r+1}(0) = 0$, we obtain, for the first variation:

$$
\begin{aligned}
\hat{y}_1 &= \xi_1,\ \hat{y}_2 = \xi_2,\\
\dot{\xi}_1 &= \xi_3,\ \dot{\xi}_2 = \xi_4\\
&\dots\\
\dot{\xi}_{2r-3} &= \xi_{2r-1},\ \dot{\xi}_{2r-2} = \xi_{2r},\\
\dot{\xi}_{2r-1} &= d_x f_{2r-1}(\xi_1,..,\xi_{2r+1}) + \frac{\partial f_{2r-1}}{\partial f_{2r}}(d_x f_{2r}\xi + d_\varphi f_{2r}\eta),\\
\dot{\xi}_{2r} &= d_x f_{2r}\xi + d_\varphi f_{2r}\eta,\\
\dot{\xi}_{2r+1} &= d_x f_{2r+1}\xi + d_\varphi f_{2r+1}\eta,\\
&\dots\\
\dot{\xi}_n &= d_x f_n\xi + d_\varphi f_n\eta.
\end{aligned} \tag{5.10}
$$

If $n = 2r+1$, our theorem is already proved (exchanging h_1, h_2). Assume $n > 2r+1$.

If we can find a feedback solution $(\hat{\eta},\hat{\varphi})(x,\xi)$ of :

$$
\begin{aligned}
d_x f_{2r}(x,\varphi)\xi + d_\varphi f_{2r}(x,\varphi)\eta = 0,\\
d_x f_{2r+1}(x,\varphi)\xi + d_\varphi f_{2r+1}(x,\varphi)\eta = 0,
\end{aligned}
$$

for $(\xi_1,\dots,\xi_{2r+1}) = 0$, $(\xi_{2r+2},\dots,\xi_n) \neq 0$, on some open set in the space of the other variables $\xi_{2r+2},\dots,\xi_n, x$, then, this will contradict the uniform infinitesimal identifiability, with the same reasoning as above. Therefore, another application of Lemma 7 shows that, we can change the coordinates x, keeping $x_1,\dots,x_{2r+1}$ unchanged, for:

$$f_{2r+1} = f_{2r+1}(x_1,\dots,x_{2r+2}, f_{2r}(x,\varphi)),$$

where $\frac{\partial f_{2r+1}}{\partial x_{2r+2}} \neq 0$ (never vanishes on some neighborhood $V_{x_0} \times V_{\varphi_0}$). Iterating the process, and at the end, exchanging the roles of $y_1 = x_1$ and $y_2 = x_2$, ends the proof of the theorem. □

6. Identification

We will not consider the case where identifiability is a generic property ($d_y \geq 3$). In this case, there is a new difficulty, and it will be treated in another paper. For the cases $d_y = 1$, $d_y = 2$, we will be very short, and give only the ideas, leaving all details to the reader. Also, we will focus on the problem of ”on line” identification (that is, the process of learning about the unknown function runs simultaneously to the process of observing the data $y(t)$).

In the case $d_y = 1$, (normal form 1.3), the most simple example is:

$$y^{(n)} = \varphi.$$

This trivial example shows that there is in general no hope to do something better than approximate differentiation: the problem of learning about the graph of φ is just the problem of estimating the n first derivatives of the output.

6.1. Identification using approximate derivators.

Theorem 12. *We consider, on $X \times I$, (X an open subset of $\mathbb{R}^n$), a system which is globally in normal form 1.3, ($d_y = 1$) or in normal forms 1.4, 1.5, ($d_y = 2$), or 1.6 with the additional requirement that $\frac{\partial f_n}{\partial \varphi}$ never vanishes ($d_y = 2$). The points of the graph of the functions $\varphi(x)$, or $\varphi \circ \pi(x)$ visited during some experiment can be reconstructed by (a finite number of) differentiations of the output(s).*

Proof. Several facts in this theorem have been proved already. Nevertheless, we shortly prove everything.

1. **Normal forms 1.3, 1.4, 1.6.**

Differentiating the outputs, we can reconstruct the state $x(t)$ of the system. Differentiating once more, we reconstruct respectively $\psi(x, \varphi)(t)$ in (1.3), and $f_n(x, \varphi)(t)$ in (1.4), (1.6). These functions, as functions of φ, are monotonous. Hence, $x(t)$ being known, we can reconstruct $\varphi(t)$ from the knowledge of their values.

2. **Normal form** 1.5.

Differentiating a certain number of times, we reconstruct $x_1, \ldots, x_{2r}$, and $\Phi(x, \varphi)$. Differentiating once more, we reconstruct $F_{2r}(x_1, \ldots, x_{2r+1}, \Phi)$. It is a monotonous function of x_{2r+1}. Hence, x_{2r+1} can be obtained from these values. Iterating the result, we reconstruct $x(t)$. Once $x(t)$ is known, the function Φ being monotonous with respect to φ, we get $\varphi(t)$. □

In fact, this procedure is not so far from what is done for the identification of linear systems.

6.2. Identification using nonlinear observers.

We may assume, along the trajectories visited, a "local model" for φ as a function of time. For instance, the most simple model is $\varphi^{(k)} \equiv 0$, i.e. φ is a polynomial function of the time. Of course, the coefficients of this polynomial will be perpetually reestimated. And hence, the question is not that they model the function φ globally as a function of time, but only locally, on reasonable time intervals (reasonable with respect to the performances of the observer).

Let us consider again the 4 normal forms 1.3, 1.4, 1.5, 1.6, (with the same additional requirement, in case 1.6, that $\frac{\partial f_n}{\partial \varphi}$ never vanishes). Adding φ and its k first derivatives as extra state variables, the problem is now reduced to the problem of estimation of the state.

It turns out that, in all these cases, the extended systems we get have **very strong observability properties**, and that the "High Gain Construction", presented in the book [7], generalizes to these cases, allowing to reconstruct (approximately) the state of the extended system, and hence to estimate the corresponding points of the graph of φ.

6.3. A more robust solution.

High gain observers may be rather sensitive to noise. A more robust solution is proposed in [3]. This construction also works for all the cases under consideration here in, if $d_y = 1$ or 2.

As an example, let us consider just the case of the normal form (1.3), which gives, adding derivatives of φ as state variables:

$$\begin{aligned} \frac{d^n y}{dt^n} &= \psi(y, \ldots, y^{(n-1)}, \varphi), \\ \dot{\varphi} &= \varphi_1, \\ &.. \\ \dot{\varphi}_{k-2} &= \varphi_{k-1}, \\ \dot{\varphi}_{k-1} &= 0, \end{aligned} \tag{6.1}$$

where, as usual, $\frac{\partial \psi}{\partial \varphi}$ never vanishes.

The map $\Xi : \mathbb{R}^{n+k} \to \mathbb{R}^{n+k}$, $(y, \dot{y}, \ldots, y^{n-1}, \varphi, \ldots, \varphi^{k-1}) \to (y, \dot{y}, \ldots, y^{(n+k-1)})$, is a diffeomorphism, as it is easily checked. Hence, (see again [7] for details), this system is diffeomorphic to a system on $\mathbb{R}^{n+k}$, of the form:

$$\dot{\xi}_1 = \xi_2, \ldots, \dot{\xi}_{n+k-1} = \xi_{n+k}, \dot{\xi}_{n+k} = \hat{\psi}(\xi), \quad y = \xi_1,$$

for some smooth function $\hat{\psi}$. This is exactly what is needed for applying the technique developed in [3].

We will exploit a variation of this idea for the biological reactor in the next section.

7. The biological reactor

7.1. **The model and its basic properties.** Let us recall the equations of the model of bioreactor presented in the introduction:

$$\frac{ds}{dt} = -\mu.x + D(S_{in} - s) \qquad (7.1)$$
$$\frac{dx}{dt} = (\mu - D)x.$$

The growth function, $\mu(s,x)$, is smooth, positive or zero, and, for obvious physical reasons, $\mu(0,x) = 0$: if there is no substrate to eat, the population cannot grow. On the contrary, if there is something to eat, the population grows: $\mu(s,x) > 0$ for $s > 0$. The control function $D(t)$ verifies $D(t) > \varepsilon > 0$: this means that the reactor is always fed. The constant S_{in} is assumed to be strictly positive.

Now, the subset $X = \mathbb{R}^+ \times \mathbb{R}^+$ is invariant by the dynamics (7.1) of the bioreactor: if $x = 0$, then $\dot{x} = 0$, and if $s = 0$, then $\dot{s} = D.\ S_{in} > 0$. Therefore, we may consider that $x(t)$, $s(t)$ are always strictly positive.

The function $\mu(s,x)$ is often considered in the literature as a function of s only, $\mu(0) = 0$, $\mu \geq 0$. This means that the internal variable is $s \in \mathbb{R}^+$, and the internal mapping $\pi : X \to \mathbb{R}^+$ is the mapping $(x,s) \to s$.

Typical expressions of the growth function in that case are the Monod model:

$$\mu(s) = \frac{\mu_0 s}{k_m + s}, \qquad (7.2)$$

or the Haldane model:

$$\mu(s) = \frac{\mu_0 s}{k_m + s + \frac{s^2}{k_i}}. \qquad (7.3)$$

7.2. **Observation of s only.** Setting $X = x + s$, we get:

$$\frac{ds}{dt} = -\mu.x + D(S_{in} - s), \qquad (7.4)$$
$$\frac{dX}{dt} = D(S_{in} - X),$$

hence, setting $\tilde{D}(t) = \int_0^t D(\tau)d\tau$, we get:

$$X = e^{-\tilde{D}(t)}X_0 + (1 - e^{-\tilde{D}(t)})S_{in}. \qquad (7.5)$$

Let us set:

$$(7.6) \qquad (1) \quad \Lambda(t) = e^{\tilde{D}(t)}(s - S_{in}) + S_{in}, \quad \text{or:}$$
$$(2) \qquad s = e^{-\tilde{D}(t)}\Lambda(t) + S_{in}(1 - e^{-\tilde{D}(t)}), \quad s(0) = \Lambda(0).$$

By (7.4), we get:

$$\dot{\Lambda} = -e^{\tilde{D}(t)}(X - s)\mu, \qquad (7.7)$$

and with (7.5):

$$\dot{\Lambda} = (\Lambda - X_0)\mu. \qquad (7.8)$$

Let us assume, as we said in the previous section, that μ is a function of s only. Let us also assume that $s(\cdot)$ visits twice the same value, i.e, $T_0 < T_1$, $s(T_0) = s(T_1)$. Then, with $X_0 = X(0) = x(0) + s(0)$:

$$\mu(s(T_0)) = \frac{\dot{\Lambda}(T_0)}{\Lambda(T_0) - X_0} = \frac{\dot{\Lambda}(T_1)}{\Lambda(T_1) - X_0} = \mu(s(T_1)).$$

Observe, with Equation (7.8), that Λ is **everywhere continuously differentiable**, even if $D(\cdot)$ is only measurable, bounded. >From (7.7), and the fact that $\mu(s) > 0$ for $s > 0$, we get that $\Lambda(t)$ is a strictly decreasing function, and $\Lambda(0) = s(0)$, $X_0 = x(0) + s(0)$, $x(0) > 0$, implies that $\Lambda(t) - X_0$ is never zero for $t \geq 0$.

Also, $s(t)$ being observed, and $D(t)$, the control, being known, we may consider, by the definition (7.6, 1) of Λ, that Λ is an observed function. Then, X_0 can be computed:

$$X_0 = \frac{\dot{\Lambda}(T_0)\Lambda(T_1) - \dot{\Lambda}(T_1)\Lambda(T_0)}{\dot{\Lambda}(T_0) - \dot{\Lambda}(T_1)}, \tag{7.9}$$

indeed, $\dot{\Lambda}(T_0) - \dot{\Lambda}(T_1) = (\Lambda(T_0) - \Lambda(T_1))\mu(s(T_0)) \neq 0$.

Now, X_0 being known,

$$\mu(s(t)) = \frac{\dot{\Lambda}(t)}{\Lambda(t) - X_0} \tag{7.10}$$

since $\Lambda(t) - X_0$ never vanishes.

Conversely, let us consider a trajectory, defined on $[0, T]$, such that:

(H.a.) s never visits twice the same value on $[0, T]$, (i.e. s is strictly monotonous),

or, stronger,

(H.b.) $D(\cdot)$ is smooth, (from what it follows that $s(\cdot), x(\cdot), \Lambda(\cdot), X(\cdot)$, are all smooth functions), and $\frac{ds}{dt}(t) \neq 0$ for all $t \in [0, T]$.

Initial conditions corresponding to this trajectory are s_0, x_0, X_0 (all strictly > 0).

Let us consider another arbitrary value, $\tilde{X}_0 = \tilde{x}_0 + s_0 > 0$, close to X_0. Then, since $\Lambda(t) = e^{\tilde{D}(t)}(s - S_{in}) + S_{in}$ is C^1 (even in case, (H.a.)), as we know by (7.8), we may compute $\tilde{\mu}(t) = \frac{\dot{\Lambda}(t)}{\Lambda(t) - \tilde{X}_0}$, on the interval $[0, T]$: indeed, since $\tilde{X}_0$ is close to X_0, then $\Lambda(t) - \tilde{X}_0$ is close to $\Lambda(t) - X_0$, and then never vanishes. Then, under Assumption (H.a.), there is a continuous function $t(s)$, which is the inverse of $s(t)$, and which is smooth under assumption (H.b.). Set $\bar{\mu}(s) = \tilde{\mu}(t(s))$. In case (H.b.), $\bar{\mu}$ is smooth, in case (H.a.), it is continuous only. Set also $\tilde{x}(t) = \tilde{X}(t) - s(t) = e^{-\tilde{D}(t)}\tilde{X}_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t)$. In these circumstances, we claim that:

Claim: (a) $s(t)$, $\tilde{x}(t)$ are solutions of the system:

$$\begin{aligned} (1) \quad \frac{ds}{dt} &= -\bar{\mu}(s(t))\tilde{x} + D(t)(S_{in} - s(t)), \\ (2) \quad \frac{d\tilde{x}}{dt} &= (\bar{\mu}(s(t)) - D(t))\tilde{x}(t), \end{aligned} \tag{7.11}$$

(b) if $\tilde{X}_0$ is sufficiently close to X_0, $\tilde{x}(t) > 0$ for all $t \in [0, T]$.

Proof. (of the claim) (a): Let us show first that $\mu(t)x(t) = \tilde{\mu}(t)\tilde{x}(t)$ for all $t \in [0, T]$. By construction:

$$\tilde{\mu}(t) = \frac{\Lambda(t) - X_0}{\Lambda(t) - \tilde{X}_0}\mu(t),$$

then it is sufficient to check that $\frac{\Lambda(t) - X_0}{\Lambda(t) - \tilde{X}_0}\mu(t)\tilde{x}(t) = \mu(t)x(t)$, or, since $\mu(t) > 0$:

$$(\Lambda(t) - X_0)\tilde{x}(t) = (\Lambda(t) - \tilde{X}_0)x(t). \tag{7.12}$$

But $x(t) = X(t) - s(t) = e^{-\tilde{D}(t)}X_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t)$, and $\tilde{x}(t) = \tilde{X}(t) - s(t) = e^{-\tilde{D}(t)}\tilde{X}_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t)$. Replacing by these expressions and by the expression (7.6, 1) of $\Lambda(t)$, just shows that (7.12) is true. Therefore, since $\frac{ds}{dt} = -\mu.x + D(S_{in} - s)$, we get that also, $\frac{ds}{dt} = -\tilde{\mu}.\tilde{x} + D(S_{in} - s)$. This is (a, 1).

Now,

$$\begin{aligned} \tilde{x}(t) &= \tilde{X}(t) - s(t) = e^{-\tilde{D}(t)}\tilde{X}_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t), \\ \frac{d\tilde{x}}{dt}(t) &= -D(t)(\tilde{x}(t) - S_{in} + s(t)) + \mu(t)x(t) - D(t)(S_{in} - s(t)), \\ &= -D(t)\tilde{x}(t) + \mu(t)x(t) = -D(t)\tilde{x}(t) + \tilde{\mu}(t)\tilde{x}(t), \end{aligned}$$

by the proof of (a, 1). Hence, $\frac{d\tilde{x}}{dt}(t) = (\tilde{\mu}(t) - D(t))\tilde{x}(t)$. This is (a, 2).

To prove (b), let us just observe that: $\tilde{x}(t) = e^{-\tilde{D}(t)}\tilde{X}_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t)$, and $x(t) = e^{-\tilde{D}(t)}X_0 + (1 - e^{-\tilde{D}(t)})S_{in} - s(t)$, then, $\tilde{x}(t) - x(t) = e^{-\tilde{D}(t)}(\tilde{X}_0 - X_0)$. Hence, for $\tilde{X}_0$ sufficiently close to X_0, or equivalently $\tilde{x}_0$ sufficiently close to x_0, $\tilde{x}(t)$ is strictly positive. □

We have shown the following theorem (more precise version of Theorem 6 in Section 2.3):

Theorem 13. *a. The bioreactor is identifiable at any admissible i.o. trajectory* $(s(t), D(t))$, $t \in [0, T]$ *such that* $s(t)$ *visits twice the same value,*

b. Conversely, if $(s(t), D(t))$, $t \in [0, T]$ *is an admissible i.o. trajectory along which* $s(t)$ *is strictly monotonous, then, there is an infinity of corresponding couples* $(x(\cdot), \mu)$, *with* μ *continuous. If moreover* $D(\cdot)$ *is smooth, and* $\dot{s}(t) \neq 0$ *for all* $t \in [0, T]$, *then, there is an infinity of corresponding couples* $(x(\cdot), \mu)$, *with* μ *smooth.*

In Section 7.5, some numerical investigation will be made on the basis of this theorem.

7.3. **Observation of both s and x.** In that case, assuming $D(\cdot)$ constant, we are in the situation $d_y = 2$, of Section 5, and our system is regular, and Type 3 (normal form 1.6), with the additional requirement of Theorem 12.

Therefore, we may use the ideas explained briefly in Section 6, Subsections 6.2, 6.3. In fact, in that case, we can do better:

1. we can adapt these ideas to the case where $D(\cdot)$ is not constant,
2. making a small change of variables, we can use a standard linear Kalman observer, in place of a high-gain one:

Let us set $z(t) = \mu(s(t)).x(t)$, and let us, (as explained in Section 6), assume a local model for $z(\cdot)$, of the form $\frac{d^k z}{dt^k} = 0$. Then, our system becomes:

$$\begin{aligned} \dot{s} &= -z + D(t)(S_{in} - s); \\ \dot{x} &= z - D(t).x, \\ \dot{z} &= z_1, \ldots, \dot{z}_{k-2} = z_{k-1}, \dot{z}_{k-1} = 0; \\ y_1 &= s, \quad y_2 = x. \end{aligned} \tag{7.13}$$

Assuming only that $0 \leq D(t) \leq D_{\max}$, this is a linear time-dependant system, which is uniformly observable in the sense of the linear theory. Hence, classical versions of the (time dependant) Linear Kalman Filter work, even in a stochastic context.

This method will be also investigated numerically in Section 7.5.

7.4. **Observation of x only.** This case is often considered in practice.

If μ would be a function $\mu(x)$ of x, then, we could consider the system:

$$\begin{aligned} \dot{x} &= (\mu(x) - D)x, \\ y &= x, \end{aligned}$$

for $x \in \mathbb{R}^+$. Then, assuming $D(\cdot)$ constant, we are exactly in the case of uniformly infinitesimally identifiable systems, in the case $d_y = 1$, normal form 1.3. Our ideas of Section 6 can then be used, and work as perfectly as in the case of the previous section 7.3, (even for $D(\cdot)$ non constant).

But, in the case where μ depends only on s, as we assume here, (and as is often assumed in the literature), or if $\mu = \mu(s, x)$ depends on both s and x, then, the system:

$$\begin{aligned} \dot{x} &= (\mu - D)x, \\ \dot{s} &= -\mu.x + D(S_{in} - s), \\ y &= x; \end{aligned}$$

is not uniformly infinitesimally identifiable (it does not verify the necessary conditions of Theorem 2).

Moreover, clearly, it is not identifiable: assume D constant (for simplicity only), and $\mu(x, s)$, or $\mu(s)$ given, smooth. Given a smooth trajectory $(x(t), s(t)), t \in [0, T]$, $x(0) > 0, s(0) > 0$, of this system. Then,

$$s(t) = e^{-Dt} s(0) + \int_0^t e^{-D(t-\tau)} (-(\mu x)(\tau) + DS_{in}) d\tau.$$

Let us chose another $\tilde{s}_0 \neq s(0)$, ($\tilde{s}(0)$ close to $s(0)$), and consider:

$$\tilde{s}(t) = e^{-Dt} \tilde{s}_0 + \int_0^t e^{-D(t-\tau)} (-(\mu x)(\tau) + DS_{in}) d\tau.$$

Consider now the trajectory $(x(t), \tilde{s}(t)), t \in [0, T]$, and a smooth function $\tilde{\mu}(x, s)$, or $\tilde{\mu}(s)$, such that $\tilde{\mu}(x(t), \tilde{s}(t)) = \mu(x(t), s(t))$ for $t \in [0, T]$. This is possible, eventually restricting the interval $[0, T]$ to some $[T_0, T_1] \subset [0, T]$. In fact, $(x(t), \tilde{s}(t))$ is a solution of:

$$\frac{dx}{dt} = (\tilde{\mu} - D)x,$$
$$\frac{d\tilde{s}}{dt} = -\tilde{\mu}.x + D(S_{in} - \tilde{s}),$$

$x(0) = x_0$, $\tilde{s}(0) = \tilde{s}_0$.

Hence, some people identify systems that are not identifiable: It is possible, by differentiation of the output x, to obtain complete information on the function $\mu(x(t), s(t))$, as a function of time. But it is not possible to deduce from this, any information about the function $\mu(s, x)$, or $\mu(s)$.

In fact, the reason why they obtain "some results" in practice (even for $D(\cdot)$ nonconstant) is the following: because of Equation 7.5, we see that, whatever the control $D(\cdot) > \varepsilon > 0$, $X(t)$ tends to S_{in} when $t \to +\infty$. Hence, after some time, $X = x + s$ is close to S_{in}, and s is close to $S_{in} - x$. Then, the estimate $\hat{\mu}(t)$, obtained after differentiation of the output $x(t)$, is such that $(x(t), S_{in} - x(t), \hat{\mu}(t))$, is close to a point of the graph of μ, for t large enough.

7.5. **Numerical simulations.**

7.5.1. *Simulation with complete measurements.* In this section, we simulate the bioreactor model using the Monod growth function (7.2)

$$\mu(s) = \frac{0.15\, s}{2 + s}$$

We assume S_{in} constant, equal to 5. $D(t)$ is a periodic function with period shown on Figure 1. Initial conditions are $x(0) = 3$ and $s(0) = 2$.

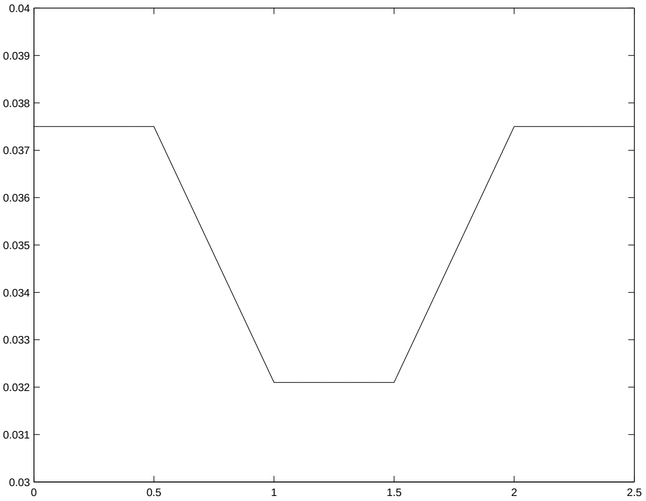


FIGURE 1. $D(t)$

In this first case, we assume that both the substrate and the biomass concentration are measured. Hence the output is $y(t) = (s(t), x(t))$. A colored noise is added to both variables to simulate noisy measurement. More precisely, this colored noise has been simulated using

$$dU_t = -a\, U_t\, dt + \sigma\sqrt{2a}\, dW_t$$

where W_t is a normalized Brownian motion. So U_t is an Ornstein–Uhlenbeck process, that is, a stationary process with mean 0 and with covariance function

$$\Gamma_U(t, s) = E[U_t U_s] = \sigma^2 \mathrm{e}^{-a|t-s|}$$

and hence U_t is a reasonable approximation of a realistic noise.

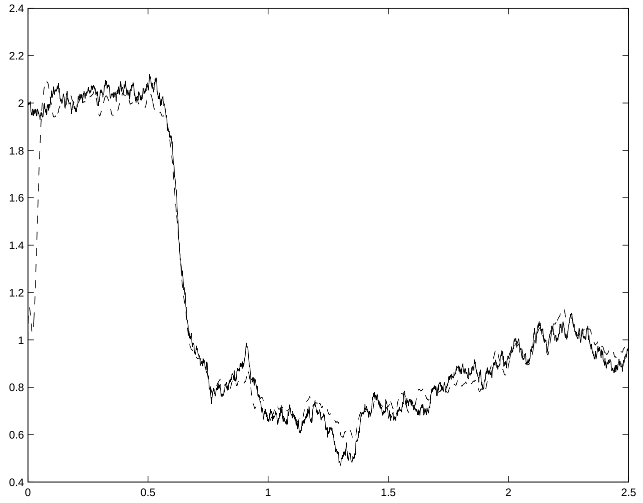


Figure 2. $s\left(t\right)$

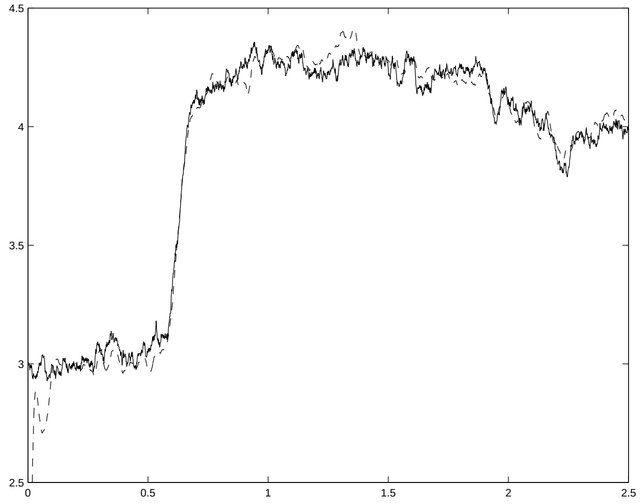


Figure 3. $x\left(t\right)$

As explained in section 7.3, we set $z\left(t\right)=\mu\left(s\left(t\right)\right)\,x\left(t\right)$, and we assume a local model for $z\left(t\right)$ of the form

$$\left\{\begin{array}{rcl}\dfrac{dz\left(t\right)}{dt} & = & z_1\left(t\right)\\ \dfrac{dz_1\left(t\right)}{dt} & = & z_2\left(t\right)\\ \dfrac{dz_2\left(t\right)}{dt} & = & 0\end{array}\right.$$

We add to these equations the two equations of the bioreactor

$$\left\{\begin{array}{rcl}\dfrac{ds\left(t\right)}{dt} & = & -z\left(t\right)+D\left(t\right)\,\left(S_{\text{in}}-s\left(t\right)\right)\\ \dfrac{dx\left(t\right)}{dt} & = & z\left(t\right)-D\left(t\right)\,x\left(t\right)\end{array}\right.$$

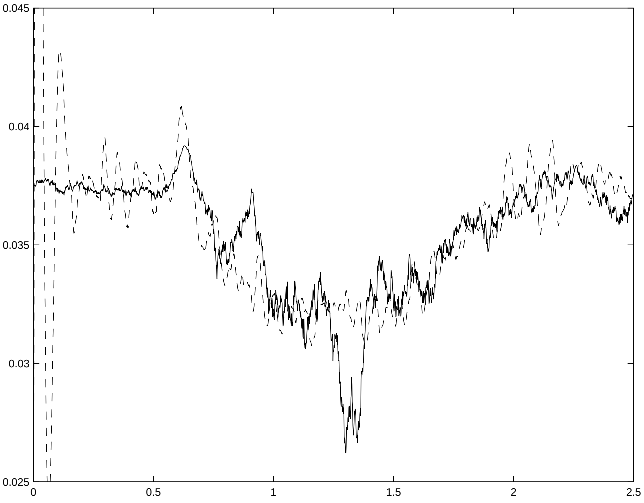


FIGURE 4. $\mu(t)$

and we apply the classical Kalman filter to these five equations, *i.e.* setting $X(t) = (s(t), x(t), z(t), z_1(t), z_2(t))$

$$\begin{aligned}\frac{d\widehat{X}(t)}{dt} &= A(t)\widehat{X}(t) + b(t) + PC^TR^{-1}\left(y(t) - C\widehat{X}(t)\right)\\ \frac{dP(t)}{dt} &= A(t)P(t) + P(t)A(t)^T + Q - P(t)C^TR^{-1}CP(t)\end{aligned}$$

with

$$A(t) = \begin{pmatrix} -D(t) & 0 & -1 & 0 & 0\\ 0 & -D(t) & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0\end{pmatrix},\ b(t) = \begin{pmatrix} D(t)s_{\text{in}}\\ 0\\ 0\\ 0\\ 0\end{pmatrix}$$

$$\text{and } C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\end{pmatrix}$$

Measured outputs are shown on figure 2 and 3 where the continuous line represents noisy outputs and the dashed line represents observer estimation.

Figure 4 represents $\mu(t)$ and $\widehat{\mu}(t) = \dfrac{\widehat{z}(t)}{\widehat{x}(t)}$ with the same convention as in the previous figures. Clearly, after the initial transient, the behavior of the Kalman filter looks good.

7.5.2. *Simulation with a single measurement $s(t)$.* If s is the only measurement, we can still expect to reconstruct $\mu(s)$ for all visited value of s provided that at least one value of s has been visited twice, as already explained in section 7.2.

Since this identification process does not depend on the estimation of x, we use the estimation of μ to estimate x. Moreover, once μ is known, the system is a linear system with a matrix depending on the output s, therefore we can apply linear Kalman filter to estimate x.

Our algorithm is divided in two steps:

- estimation of μ using the redundant visited values of the measurement s. The estimation of the function $s \to \mu(s)$ at time t is denoted by $s \to \widehat{\mu_t}(s)$;
- estimation of $x(t)$ on the basis of $s(t)$ and $\widehat{\mu_t}$.

The second part of our algorithm is simply standard Kalman filtering. Let us explain the first part.

We first choose a sample time Δt, arbitrarily fixed to 0.2. At each sample time $t = k\Delta t$, we consider the $N+1$ values of s at time $k\Delta t$, $(k-1)\Delta t, \ldots,$ $(k-N)\Delta t$, *i.e.* we consider only the history up to time $T = N\Delta t$ in the past.

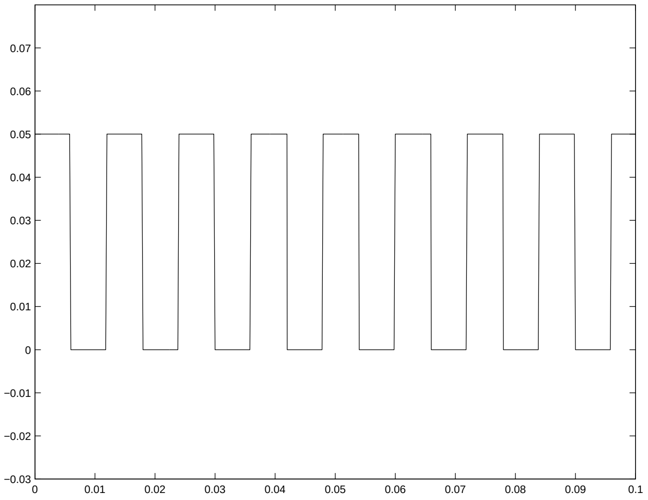


Figure 5. $D(t)$

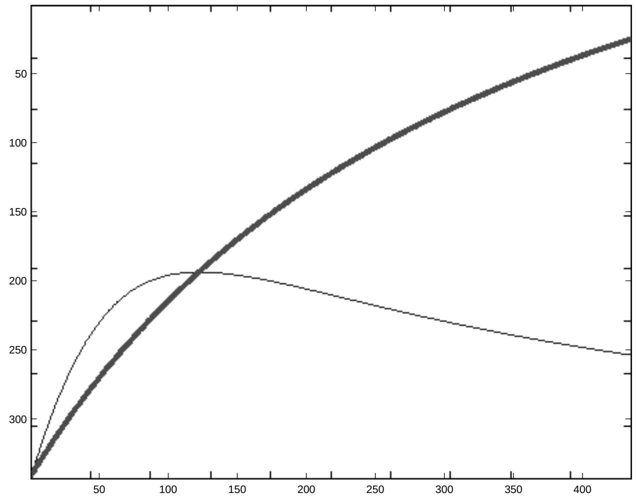


Figure 6. $\mu(s)$ (thin line) and its initial guess $\widehat{\mu_0}(s)$ (thick line)

Since we want to estimate a function, we have to discretize the coordinate space of s in a range corresponding to physical values, with a small enough discretization step, to ensure good accuracy. In this way, we chose $s_{\min}$ and $s_{\max}$ and a step size Δs.

At each sample time $k\Delta t$, we replace the measured trajectory

$$s((k-N)\Delta t), \ldots, s(k\Delta t) \tag{7.14}$$

by a linearly interpolated trajectory

$$s(t_1), s(t_2), \ldots, s(t_n) \tag{7.15}$$

such that $(k-N)\Delta t \leq t_1 < t_2 < \cdots < t_n \leq k\Delta t$ and each $s(t_j)$ is of the form $s_{\min} + p\Delta s$. Let us explain more precisely how we build the list 7.15. Let us assume that we have already build the list from t_1 to t_j considering measurements from time $(k-N)\Delta t$ to time $(k-i)\Delta t$. Then between time $(k-i)\Delta t$ and

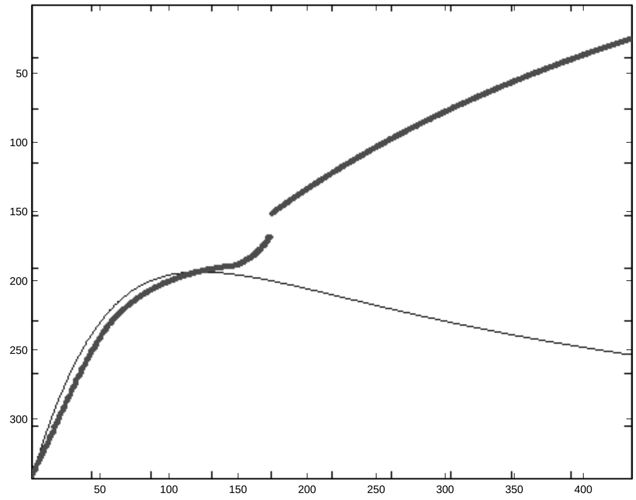


FIGURE 7. $\mu(s)$ and $\widehat{\mu_t}(s)$ at time $t = 0.019$

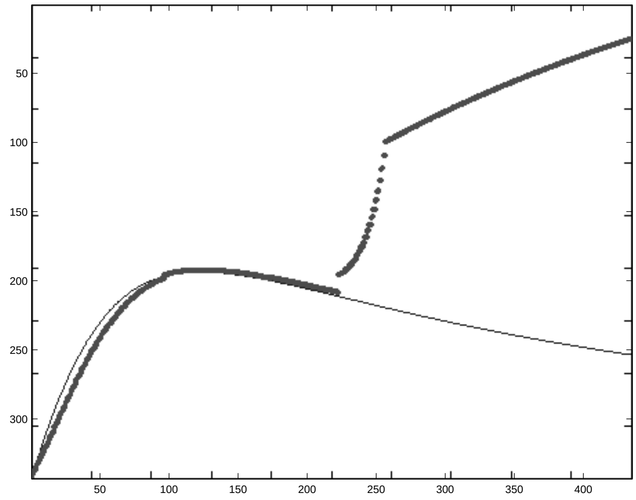


FIGURE 8. $\mu(s)$ and $\widehat{\mu_t}(s)$ at time $t = 0.042$

$(k-i+1)\,\Delta t$, l and r are such that

$$\begin{aligned} s_{\min} + l\Delta s < s\left((k-i)\,\Delta t\right) \le s_{\min} + (l+1)\,\Delta s \le s_{\min} + (l+2)\,\Delta s \le \cdots \\ \le s_{\min} + (l+r)\,\Delta s \le s\left((k-i+1)\,\Delta t\right) \le s_{\min} + (l+r+1)\,\Delta s \end{aligned}$$

Now we interpolate $t_{j+1}, \dots, t_{j+r}$ so that $s(t_{j+p}) \approx s_{\min} + (l+p)\,\Delta s,\ p = 1, \dots, r$ and we add $s(t_{j+1}), \dots, s(t_{j+r})$ to the list.

The list 7.15 is then used to look for values appearing at least twice and to estimate X_0 and μ at sample values $s_{\min} + (l+p)\,\Delta s$.

Nevertheless, in order to give some weight to *a priori* knowledge of μ and to increase robustness with respect to measurement noise, we do not modify completely $\mu(s)$ when a new value is provided by equations (7.9) and (7.10) and the previous algorithm. In fact, we use a first order filter to actualize $\widehat{\mu_t}$,

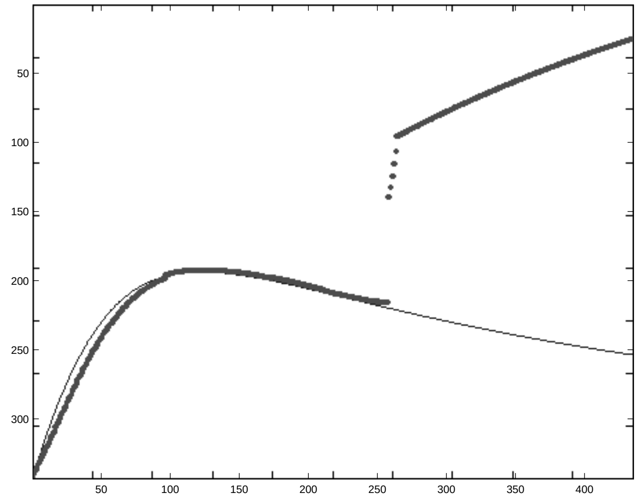


FIGURE 9. $\mu(s)$ and $\widehat{\mu_t}(s)$ at time $t = 0.051$

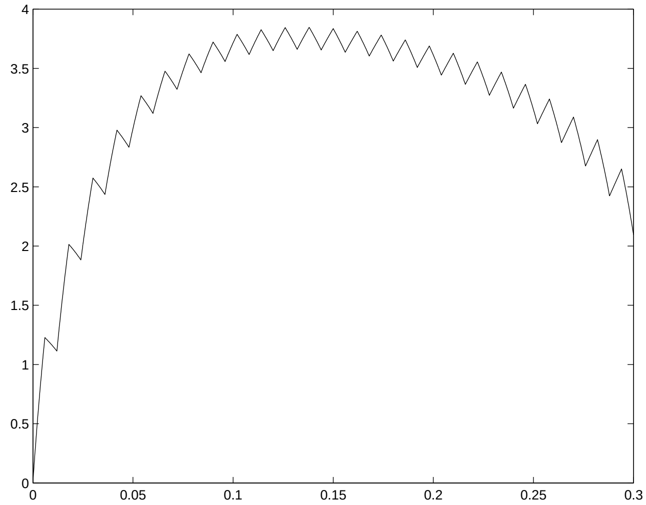


FIGURE 10. $s(t)$

that is, if a new $\mu_t(s)$ is obtained from (7.10) at time t for a discretized s, we modify $\widehat{\mu_t}(s)$ using

$$\widehat{\mu_t}(s) = (1-\beta)\ \widehat{\mu_t}(s) + \beta\, \mu_t(s)$$

Despite this filtering, both on measurement and actualization of μ, this algorithm looks rather sensitive to noise. Here, we illustrate our approach without adding any noise on the output s.

Figure 5 shows $D(t)$ which is the control. Figures 6 to 9 show the estimation $\widehat{\mu_t}(s)$ of $\mu(s)$ at different times. At the beginning of the simulation, we set $\widehat{\mu_0}$ to be the Monod law, although the actual unknown law $\mu(s)$ is the Haldane law. Thin lines represent the actual Haldane growth function. Thick lines and dots represent the estimation of the function μ.

At each sample time, as already mentioned, we use the bioreactor model with $\widehat{\mu_t}(s)$ as growth function to estimate $x(t)$ using a linear Kalman filter. At the beginning of our simulation, since $\widehat{\mu_0}$ is wrong (Monod law instead Haldane law), our observer does not estimate x accurately and there is a bias between the actual $x(t)$ and its estimate $\widehat{x}(t)$. It is expected that when μ is correctly identified, the observer gives

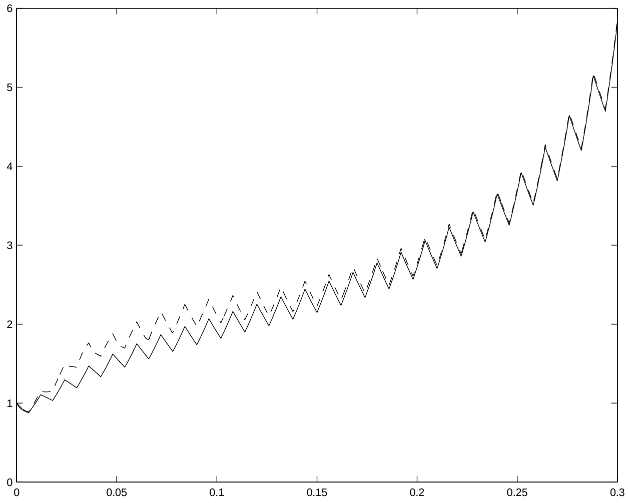


FIGURE 11. $x\,(t)$

an unbiased estimation of x. Indeed, at time 0.15 approximately, $s\,(t)$ begins to decrease (Figure 10) and then visits again a domain where μ has already been identified. Therefore $\mu\,(s) \approx \widehat{\mu_t}\,(s)$ for $t > 0.15$. It is then expected that after time 0.15, estimation of $x\,(t)$ will be unbiased. This actually happens, see Figure 11.

8. Appendix

8.1. **A crucial lemma .**

Notations:

1. In this section, we keep the notation $\wedge$ for the exterior product of differential forms on X or on $X \times I$, and, for $V_1, V_2 \in \mathbb{R}^2$, we denote by $V_1 \,\overline{\wedge} V_2$ the determinant of the 2×2 matrix formed by the vectors V_1, V_2.

2. Again, in this section, for a smooth function f of two variables (x, φ), $x \in \mathbb{R}^n$, $\varphi \in \mathbb{R}^p$, we denote by $d_x f$ (resp $d_\varphi f$), the differentials with respect to the x variable (resp. φ variable) only.

3. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, we denote by $\underline{x_p} \in \mathbb{R}^p$ the vector $\underline{x_p} = (x_1, \ldots, x_p)$. For a function $h : \mathbb{R}^n \to \mathbb{R}$, of the variables $x_1, \ldots, x_n$, the notation $h^{\widehat{i_1, \ldots, i_p}}$ means that h does not depend on $x_{i_1}, \ldots, x_{i_p}$.

Lemma 7. *Consider $h : U \times I \to \mathbb{R}^2$, U an open connected subset of $\mathbb{R}^n$, with a given C^ω coordinate system $x = (x_1, \ldots, x_n)$, and $h(x, \varphi) = (h_1(x, \varphi), h_2(x, \varphi))$, such that $d_\varphi h$ never vanishes on $U \times I$, and the equation*

$$(8.1) \qquad d_x h(x, \varphi)\xi + d_\varphi h(x, \varphi)\eta = 0, \quad \textit{with} \ \ (\xi_1, \ldots, \xi_p) = 0, \ \ p < n,$$

has no smooth solution $(\eta, \varphi)(x, \xi)$, on any open subset of $U \times \mathbb{R}^{n-p} \subset \{(x, \xi) | (\xi_{p+1}, \ldots, \xi_n) \neq 0\}$.

Then, there is $Z \subset U$, a closed subanalytic subset of codimension 1, such that, for all $x_0 \in U \backslash Z$, it does exist a neighborhood V_{x_0} of x_0 and coordinates $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ on V_{x_0}, with $\tilde{x}_1 = x_1, \ldots, \tilde{x}_p = x_p$, and:

for all $\varphi_0 \in I$, there is a neighborhood V_{φ_0} of φ_0, a C^ω real function Φ_{φ_0}, with open domain D in $V_{x_0} \times \Theta_1 \times \Theta_2$, Θ_1, Θ_2 open subsets of $\mathbb{R}$,

with:

$$(F)\ \tilde{x}_{p+1} = \Phi_{\varphi_0}(\underline{\tilde{x}_p}, h_1(\tilde{x}, \varphi), h_2(\tilde{x}, \varphi)),$$

for all $(\tilde{x}, \varphi) \in V_{x_0} \times V_{\varphi_0}$.

Moreover,

$$(G)\ \ (\frac{\partial\Phi_{\varphi_0}}{\partial h_1}, \frac{\partial\Phi_{\varphi_0}}{\partial h_2})\ \textit{never vanishes on } D,$$

$$(H)\ \ \frac{\partial h}{\partial\varphi} \barwedge \frac{\partial h}{\partial\tilde{x}_{p+1}}\ \textit{ never vanishes on } (U\backslash Z)\times I. \tag{8.2}$$

Remark 9. *(G) is a consequence of (F), since $\tilde{x}$ is a C^ω coordinate system on X.*

Proof. (of Lemma 7).

Proof for $p+1=n$ **:** Let $E=\{(x,\varphi)|\frac{\partial h}{\partial\varphi}\barwedge\frac{\partial h}{\partial x_n}(x,\varphi)=0\}$, and let $\pi_E : E\to U$. By Hardt's theorem on the stratification of proper subanalytic mappings between subanalytic sets, if π_E contains an open set, then, there is , on a (may be smaller) open set Θ, a smooth (C^ω) function $\hat{\varphi}:\Theta\to E$.[3]

Hence, $\frac{\partial h}{\partial\varphi}\barwedge\frac{\partial h}{\partial x_n}(x,\hat{\varphi}(x))=0$, for $x\in\Theta$. We chose $x_0\in\Theta$, $\varphi_0=\hat{\varphi}(x_0)$.

If $d_\varphi h_1(x_0,\varphi_0)\neq 0$ (we can assume this by the statement of the lemma, eventually changing h_1 for h_2), then, set:

$$\hat{\eta}(x,\xi)=-\frac{(d_x h_1)(x,\hat{\varphi}(x))\xi}{(d_\varphi h_1)(x,\hat{\varphi}(x))},\ \text{for } \xi=(0,\ldots,0,\xi_n)\neq 0.$$

Then the couple $(\hat{\eta},\hat{\varphi})(x,\xi_n)$ solves Equation (8.1). This is a contradiction. Therefore, $Z=\pi_E$ has codimension 1, and on $(U\backslash Z\times I)$, $\frac{\partial h}{\partial\varphi}\barwedge\frac{\partial h}{\partial x_n}$ never vanishes by construction. This proves (H).

Now, $(x_1,\ldots,x_p,h_1,h_2)$ is a coordinate system over small open subsets of $(U\backslash Z)\times I$:

$$dx_1\wedge\ldots.\wedge dx_p\wedge dh_1\wedge dh_2=$$
$$dx_1\wedge\ldots.\wedge dx_p\wedge d\varphi\wedge dx_n(\frac{\partial h}{\partial\varphi}\barwedge\frac{\partial h}{\partial x_n})\neq 0.$$

Hence, on such a small open set, $x_{p+1}=\Phi(x_1,\ldots,x_p,h_1,h_2)$.

Proof for $p+1<n$ **:** There exists $i>p$ such that $\frac{\partial h}{\partial\varphi}\barwedge\frac{\partial h}{\partial x_i}$ does not vanish identically: were it otherwise, for $\xi\neq 0$, $\xi_1=0,\ldots,\xi_p=0$, and in a neighborhood of (x,φ) such that $d_\varphi h_1\neq 0$, $\hat{\eta}=-\frac{d_x h_1.\xi}{d_\varphi h_1}$ solves Equation (8.1), which is impossible.

Now, for $\xi\neq 0$, for $\xi_1=0,\ldots,\xi_p=0$,

$$\begin{aligned}&(1)\quad d_x h.\xi\barwedge d_\varphi h(x,\varphi)=0\ \text{ implies:}\\&(2)\ d\varphi(d_x h.\xi\barwedge d_\varphi h(x,\varphi))=0.\end{aligned}\tag{8.3}$$

Indeed, if it is not true, by the implicit function Theorem, one can solve (8.3,1) with respect to φ, and obtain a smooth solution $\hat{\varphi}(x,\xi)$, on some open set. Setting

$$\hat{\eta}(x,\xi)=-\frac{(d_x h_1)(x,\hat{\varphi}(x,\xi))\xi}{(d_\varphi h_1)(x,\hat{\varphi}(x,\xi))},$$

we solve again Equation (8.1) on an open set, a contradiction.

Statement (8.3) can be rewritten:

$$\sum_{i=p+1}^{n}(d_{x_i}h\barwedge d_\varphi h(x,\varphi))\xi_i=0\Rightarrow\sum_{i=p+1}^{n}d_\varphi(d_{x_i}h\barwedge d_\varphi h(x,\varphi))\xi_i=0.$$

But, one of the $d_{x_i}h\barwedge d_\varphi h(x,\varphi)$does not vanish identically. Hence, in a neighborhood V of some (x_0,φ_0), we have, for all $i=p+1,\ldots,n$, and for a certain analytic function $\lambda(x,\varphi)$:

$$d_\varphi(d_{x_i}h\barwedge d_\varphi h(x,\varphi))=\lambda(x,\varphi)(d_{x_i}h\barwedge d_\varphi h(x,\varphi)),$$

and then, integrating this (linear) differential equation in φ :

$$d_{x_i}h\barwedge d_\varphi h(x,\varphi)=\Lambda(x,\varphi)\omega_i(x),\ \Lambda(x,\varphi)\neq 0,\ i=p+1,\ldots,n.$$

This implies, on V :

$$\omega_j(x)\ d_{x_i}h\barwedge d_\varphi h(x,\varphi)-\omega_i(x)d_{x_j}h\barwedge d_\varphi h(x,\varphi)=0,$$

for all $i,j>p$. Hence, since $d_\varphi h$ never vanishes,

$$\omega_j(x)\ d_{x_i}h-\omega_i(x)d_{x_j}h=\lambda_{i,j}(x,\varphi)d_\varphi h,\ i,j>p.\tag{8.4}$$

[3] Here, in fact, Sard's Theorem plus the implicit function Theorem is enough to obtain this function $\hat{\varphi}$. But Hardt's Theorem is more explicit, and we crucially use subanalyticity elsewhere.

This is true again on V, and for analytic functions $\omega_i(x)$, $\lambda_{i,j}(x,\varphi)$. Moreover, one of the functions $\omega_i(x)$ does not vanish ($\omega_{p+1} \neq 0$, say): remember that $\omega_i(x_0) = d_{x_i}h \,\bar{\wedge}\, d_\varphi h(x_0,\varphi_0)$, which is nonzero for some i.

Taking $i = p+1, j = p+2$, this implies in particular that h is a ($\mathbb{R}^2-$valued) first integral of the vector field:

$$(\overrightarrow{C}) \quad \varpi_{p+2}(x)\frac{\partial}{\partial x_{p+1}} + \frac{\partial}{\partial x_{p+2}} + \bar{\lambda}(x,\varphi)\frac{\partial}{\partial \varphi}.$$

This is true on V, and for certain analytic functions $\varpi_{p+2}(x)$, $\bar{\lambda}(x,\varphi)$.

The flow of the "characteristic vector field" $\overrightarrow{C}$ is:

$$\exp(t\ \overrightarrow{C})(x) = (x_1,\dots,x_p,R_{p+2}(t,x),t+x_{p+2},x_{p+3},\dots,x_n,S_{p+2}(t,x,\varphi)),$$

with $\frac{\partial R_{p+2}}{\partial x_{p+1}} \neq 0$, $\frac{\partial S_{p+2}}{\partial \varphi} \neq 0$. Therefore, $h(\exp(t\ \overrightarrow{C})(x) = h(x)$, and, setting ($x_{p+2} := 0$; $t := x_{p+2}$), we get:

$$h(x_1,\dots,x_p,\bar{R}_{p+2}(x),x_{p+2},\dots,x_n,\bar{S}_{p+2}(x,\varphi)) = h(x^{\widehat{p+2}},\varphi),$$

for some analytic functions $\bar{R}_{p+2}(x)$, $\bar{S}_{p+2}(x,\varphi)$, with $\frac{\partial \bar{R}_{p+2}}{\partial x_{p+1}} \neq 0$, $\frac{\partial \bar{S}_{p+2}}{\partial \varphi} \neq 0$. Hence, using the implicit function Theorem:

$$h(x,\varphi) = h(x_1,\dots,x_p,R^*_{p+2}(x),0,x_{p+3},\dots,x_n,S^*_{p+2}(x,\varphi)),$$

or, setting $\tilde{x}_{p+1} = R^*_{p+2}(x)$, $\tilde{x}_i = x_i$ for $i \neq p+1$,

$$h(\tilde{x},\varphi) = H(\tilde{x}^{\widehat{p+2}},S^*_{p+2}(\tilde{x},\varphi)),\ \frac{\partial S^*_{p+2}}{\partial \varphi} \neq 0.$$

Let us denote these new coordinates $\tilde{x}$ by x. We get:

$$h(x,\varphi) = H(x^{\widehat{p+2}},S_{p+2}(x,\varphi)),\ \frac{\partial S_{p+2}}{\partial \varphi} \neq 0.$$

Again (for the same reason as above), there is $i > p$ such that $d_{x_i}h \,\bar{\wedge}\, d_\varphi h(x,\varphi)$ does not vanish identically. Let us assume $d_{x_{p+1}}h \,\bar{\wedge}\, d_\varphi h(x,\varphi) \neq 0$. Hence, $(d_{x_{p+1}}H + d_S H\frac{\partial S_{p+2}}{\partial x_{p+1}}) \,\bar{\wedge}\, d_S H\frac{\partial S_{p+2}}{\partial \varphi} \neq 0$, which implies $d_{x_{p+1}}H \,\bar{\wedge}\, d_S H \neq 0$.

Also for the same reason as above, we obtain (8.4), for $i = p+1$, $j > p+2$, and dividing by $\omega_i(x) \neq 0:$

$$\omega_j(x)\ d_{x_{p+1}}h - d_{x_j}h = \lambda_{p+1,j}(x,\varphi)d_\varphi h,$$

which gives:

$$\omega_j(x)\ d_{x_{p+1}}(H(x,S(x,\varphi))) - d_{x_j}(H(x,S(x,\varphi))) = \lambda_{p+1,j}(x,\varphi)d_\varphi(H(x,S(x,\varphi))),$$
$$d_\varphi S \neq 0.$$

or:

$$\omega_j(x)\ (d_{x_{p+1}}H)(x,\theta) - (d_{x_j}H)(x,\theta) = \bar{\lambda}_{p+1,j}(x,\theta)(d_\theta H)(x,\theta).$$

For $j = p+3$, with the same reasoning as above, we get:

$$H(x_1,\dots x_p,\bar{R}_{p+3}(x),x_{p+3},\dots,x_n,\bar{S}_{p+3}(x,\varphi)) = H(x^{\widehat{p+2,p+3}},\varphi),$$

with $\frac{\partial \bar{R}_{p+3}(x)}{\partial x_{p+1}} \neq 0$, $\frac{\partial S_{p+3}(x)}{\partial \varphi} \neq 0$, or:

$$H(x_1,\dots x_p,x_{p+1},x_{p+3},\dots,x_n,\varphi)$$
$$= H(x_1,\dots x_p,R^*_{p+3}(x),x_{p+4},\dots,x_n,S^*_{p+3}(x,\varphi)).$$

Making the change of coordinates $x_{p+1} := R^*_{p+3}(x)$, we get:

$$H(x,\varphi) = H(x_1,\dots x_p,x_{p+1},x_{p+4},\dots,x_n,S^*_{p+3}(x,\varphi)).$$

At the end, iterating the process, we get that:

$$h(x,\varphi) = H(x_1,\dots,x_{p+1},S(x,\varphi)),\ \frac{\partial S}{\partial \varphi} \neq 0, \tag{8.5}$$

on some open subset of $U \times I$. The coordinates x_i, $i = 1,\dots,p$, are unchanged.

As a consequence, we finally get that there is an open dense subanalytic subset $U\backslash Z$ of U, and for each $x_0 \in U\backslash Z$, a coordinate neighborhood of x_0, (U_{x_0},x), with coordinates x_i, $i = 1,\dots,p$, unchanged, and with:

$$dh_1 \wedge dh_2 \wedge dx_1 \wedge \dots \wedge dx_{p+1} \equiv 0,$$

identically on $U_{x_0} \times I$, by analyticity.

Now, let $E_{x_0} = \{(x, \varphi) | dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_p = 0\}$, and let πE_{x_0} be the canonical projection of E_{x_0} on U_{x_0}. If πE_{x_0} contains an open set, again by Hardt's Theorem on stratification of proper subanalytic maps, we can find another open subset Θ of U, and a smooth mapping $\hat{\varphi} : \Theta \to E_{x_0}$. Then:

$$(dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_p)_{|(x,\hat{\varphi}(x))} = 0,$$

and, in particular, for $j > p$, $(\frac{\partial h_1}{\partial x_j}\frac{\partial h_2}{\partial \varphi} - \frac{\partial h_1}{\partial \varphi}\frac{\partial h_2}{\partial x_j})_{|(x,\hat{\varphi}(x))} = 0$. Therefore, since $d_\varphi h$ is nonzero, Equation (8.1) can still be solved, in the following way:

$$\eta = -\frac{d_x h_1(x, \hat{\varphi}(x))\xi}{d_\varphi h_1(x, \hat{\varphi}(x))},$$

for $\xi \neq 0$ (if $d_\varphi h_1(x_0, \hat{\varphi}(x_0)) \neq 0$, and using h_2 if not).

This contradicts the assumptions of the lemma. Hence, there is a codimension 1 subanalytic closed subset of U, called again Z, such that, over $U \backslash Z$, each x_0 has a coordinate neighborhood (U_{x_0}, x), x_i, $i = 1, \ldots, p$, unchanged, where $dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_p$ never vanishes, and $dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_{p+1}$ is everywhere zero, and this is true over $U_{x_0} \times I$.

Since (if $d_\varphi h_1 \neq 0$), $dx_1 \wedge \ldots \wedge dx_n \wedge dh_1 \neq 0$, $dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_{p+1} \equiv 0$ implies:

$$h_2 = H_2(h_1, x_1, \ldots, x_{p+1}). \tag{8.6}$$

The condition $dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_p \neq 0$ can be rewritten $\frac{\partial H_2}{\partial x_{p+1}} \neq 0$. Hence:

$$d_\varphi h \,\bar{\wedge}\, dx_{p+1} h = d_\varphi h_1 . d_{x_{p+1}} H_2 \neq 0.$$

This is (H), in (8.2).

Now, since $\frac{\partial H_2}{\partial x_{p+1}} \neq 0$,

$$\begin{aligned} & dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_p \wedge dx_{p+2} \wedge \ldots \wedge dx_n \\ & = (d_\varphi h \,\bar{\wedge}\, dx_{p+1} h) d\varphi \wedge dx_{p+1} \wedge dx_1 \wedge \ldots \wedge dx_p \wedge dx_{p+2} \wedge \ldots \wedge dx_n \neq 0. \end{aligned}$$

This shows that $h_1, h_2, x_1, \ldots, x_p, x_{p+2}, \ldots, x_n$ is a coordinate system on some neighborhood $U_{x_0,\varphi_0} \times V_{\varphi_0}$, for all $\varphi_0 \in I$, and then, since $dh_1 \wedge dh_2 \wedge dx_1 \wedge \ldots \wedge dx_{p+1}$ is identically zero,

$$x_{p+1} = \Phi_{\varphi_0}(h_1, h_2, \underline{x_p}).$$

Now, I being compact, we may cover $\{x_0\} \times I \subset X \times I$ by a finite number of such open neighborhoods $U_{x_0,\varphi_0} \times V_{\varphi_0}$, on which (F), (G), (H) are satisfied. Hence the neighborhood U_{x_0} can be taken fixed, independantly of φ_0. This ends the proof. □

References

[1] R. ABRAHAM, J. ROBBIN, Transversal mappings and flows ; W.A. Benjamin, Inc., 1967.

[2] G. BASTIN, D. DOCHAIN, Adaptive control of bioreactors, Elsevier, 1990.

[3] E. BUSVELLE, J.P. GAUTHIER, High-Gain and Non High-Gain Observers for nonlinear systems, Contemporary Trends in Nonlinear Geometric Control Theory, pp. 257-286, 2002, World Scientific, Anzaldo-Meneses, Bonnard, Gauthier, Monroy-Perez, editors.

[4] J.P GAUTHIER, H. HAMMOURI, S. OTHMAN, A simple observer for nonlinear systems. Applications to Bioreactors. IEEE Trans. Aut. Control, 37, pp. 875-880, 1992.

[5] J.P. GAUTHIER, I. KUPKA, Observability and observers for nonlinear systems. SIAM Journal on Control, vol. 32, N° 4, pp. 975-994, 1994.

[6] J.P. GAUTHIER, I. KUPKA, Observability for systems with more outputs than inputs. Mathematische Zeitschrift, 223, pp. 47-78, 1996.

[7] J.P.GAUTHIER, I. KUPKA, Deterministic Observation Theory and Applications, Cambridge University Press, 2001.

[8] M. GORESKY, R. Mc PHERSON, Stratified Morse Theory, Springer Verlag, 1988.

[9] R. HARDT, Stratification of real analytic mappings and images, Invent. Math. 28 (1975), pp. 193-208.

[10] H. HIRONAKA, Subanalytic Sets, Number Theory, Algebraic Geometry and Commutative Algebra, in honor of Y. Akizuki ; Kinokuniya, Tokyo, 1973, pp. 453-493.

[11] M.W. HIRSCH, Differential Topology, Springer-Verlag, Graduate texts in maths, 1976.

[12] R.M. HIRSCHORN, Invertibility of control systems on Lie Groups, SIAM Journal on Control and Opt., Vol 15, N°6, 1977, pp 1034-1049.

[13] R.M. HIRSCHORN, Invertibility of Nonlinear Control Systems, SIAM J. Control and Opt., Vol 17, N°2, 1979, pp. 289-297.

[14] P. JOUAN, J.P. GAUTHIER, Finite singularities of nonlinear systems. Output stabilization, observability and observers. Journal of Dynamical and Control Systems, vol. 2, N° 2, 1996, pp. 255-288.

[15] W. RESPONDEK, Right and Left Invertibility of Nonlinear Control Systems, in Nonlinear Controllability and Optimal Control, H.J. Sussmann ed., Marcel Dekker, New-York, 1990, pp. 133-176.

[16] M. SHIOTA, Geometry of Subanalytic and Semi-Algebraic Sets, Birkhauser, P.M. 150, 1997.

[17] H.J. SUSSMANN, Some Optimal Control Applications of Real-Analytic Stratifications and Desingularization, Singularities Symposium Lojiasiewicz 70, Banach Center Publications, Vol. 43, 1998.

OBSERVATION AND IDENTIFICATION TOOLS FOR NONLINEAR SYSTEMS. APPLICATION TO A FLUID CATALYTIC CRACKER.

ERIC BUSVELLE, JEAN-PAUL GAUTHIER

Abstract. In this paper, we recall general methodologies we developed for observation and identification in nonlinear systems theory, and we show how they can be applied to real practical problems.

In a previous paper, we introduced a filter which is intermediate between the extended Kalman filter in its standard version and its high-gain version, and we applied it to certain observation problems. But we were missing some important cases. Here, we show how to treat these cases.

We also apply the same technique in the context of our identifiability theory.

As non academic illustrations, we treat a problem of observation and a problem of identification, for a fluid catalytic cracker (FCC). This FCC unit is one of the most crucial from the economic point of view, in petroleum industry.

Submitted for publication to **The International Journal of Control**

Authors' address: LE2I, UMR CNRS 5158, Université de Bourgogne, Aile des Sciences de l'Ingénieur
BP 47870 - 21078 Dijon Cedex, France
Phone + 33 (0)3-80-39-58-38
Fax + 33 (0)3-80-39-58-90
e-mail: busvelle@u-bourgogne.fr, gauthier@u-bourgogne.fr

1. Introduction

1.1. **The observation and identification problems:** In this paper, we address the problems of observation and identification of general nonlinear control systems.

By observation, we mean reconstruction of the state trajectory of the system, on the basis of some "observed data", produced by output functions. Roughly speaking, we say that a system is observable if this reconstruction is possible. A device realizing the observation task is called an "observer". Usually, these devices are realized under the guise of a differential system, fed by the observed data.

The problem of identification is a bit different: very often, practical control systems depend on some functions, (with physical meaning), that are not well known, and that have to be determined on the basis of experiments.

If x denotes the state of the system, if $\varphi(x)$ is the unknown function, and $y(t)$ is the observed data, the identification problem is the problem of reconstructing the piece of the graph of $\varphi(.)$,visited during the experiment. That is, for an experiment of duration T, we want to determine the couples $(x(t),\ \varphi(x(t)))$, for all $t \in [0,T]$, using only the observed data $\{y(t),\ t \in [0,T]\}$. We say that a system is identifiable if this is possible, whatever the experiment.

An identifier is a device performing this task. We will be interested with "on-line identifiers" only, i.e. identifiers that estimate the graph of φ simultaneously to the experiment.

The two problems, of observation and identification, are of course strongly connected.

1.2. **Our previous results.**

1.2.1. *Observability and observers.* In the book [6], a general observability theory has been exposed, together with a methodology for constructing observer systems. (These observers we construct are called "high gain" observers).

Otherwise, there is a practical tool, used for long by engineers, to construct observers for nonlinear systems (in a stochastic context): the extended Kalman filter (EKF). The idea is just to use the classical equations of the linear Kalman filter, and to apply them to the linearization of the system along the **estimate** trajectory. This is not a well defined procedure (since we linearize along the trajectory we

Date: December 2003.

are just estimating, and not along the real trajectory). Nevertheless, and despite a lack of theoretical justification, the EKF gives often very good results.

One version of our high-gain construction in [6] is connected with the EKF.

In the paper [3], we propose an observer system that has the advantages of both approaches: in presence of big disturbances, or "state jumps", it has the good properties of the high gain observers, to recover the state of the system arbitrarily fast. On the contrary, when the estimation error is small, it behaves exactly as the EKF (with good performances in front of noise).

1.2.2. *Identifiability and identifiers.* In the paper [2], we establish the main results of an identification theory and we propose an observer-based strategy for identification.

A remarkable (but not surprising) fact is that the identification theory is perfectly parallel to the observation theory. To compare, let us consider, **for the observation problem, single input systems** only, and **for the identification problem, the case of a single unknown function** to be identified, and **no control**. Then, for observability:

-If the number of outputs is two or more, systems are generically observable.

-If there is only one output, then, observability is a nongeneric property, so strong that it can be characterized by a very rigid normal form.

For identifiability:

-If the number of outputs is three ore more, then, identifiability is a generic property,

-If there is only one or two outputs, then, identifiability is a nongeneric property, so strong that it can be characterized by 4 very rigid normal forms.

1.3. **Purpose of the paper.** Our purpose here is twofold:

First, we want to recall, compare and summarize the main results of both theories of observability and identifiability. We want also to present our observers, and specially the final uppermentionned observer, mixing the high-gain construction with the EKF.

Therefore, we will summarize the results of [6], [2], [3].

At the same time, we will give some improvement of our method, allowing to apply it to certain important cases that we missed in our previous works.

Second, we want to show, through a practical nonacademic application, that these theories are really applicable in practice, and give very good results, both for the observation and the identification problem.

1.4. **The application.** The process we consider is a FCC unit, i.e. a Fluid Catalytic Cracker process, used in petroleum industry to convert heavy petroleum residues into gasoline.

There is at least one such FCC in any refinery, and it is a very strategic process, from the economic point of view (may be the most one).

It is a highly nonlinear process, rather hard to control.

As usual in petroleum industry, a few measurements are available (mostly temperature, pressure and flowrate measurements).

We will apply our observation methodology to recover the state of the system on the basis of these measurements, and specially, our purpose is to reconstruct a crucial variable inside the system: the Carbon Conradson factor.

Also, the FCC model depends on a certain "oxygen reaction rate" function, which is very important, and is in practice not well known. Also, due to the degree of simplification of the model under consideration, even if this "oxygen reaction rate" function is known in theory, it has to be adapted.

We will apply our identification procedure to estimate this function.

1.5. **Organization of the paper.** First, in the remaining of this introduction, we will fix precisely the systems under consideration, the classes of controls we consider, and we will define a few notions that are crucial for our work.

Section 2 will be devoted to the observability theory: we state the definitions of the various notions of observability we use, and we give a summary of the main observability results.

In Section 3, we do the same for our identification theory.

In Section 4, we present our results on the construction of observers, and on practical observer-based identification. We also present certain important simple improvement of our methodology.

Section 5 is devoted to the application to the Fluid Catalytic Cracker unit.

1.6. **Notations, systems under consideration.**

1.6.1. *Conventions.* All along the paper, for a smooth real-valued function $h(x, y)$, or for a smooth mapping $f(x, y)$, the notation $d_x h$ (resp. $D_x f$) means the differential of h (resp. the tangent mapping to f) w.r.t. the variable x only. In coordinates, they are represented by the Jacobian matrices of h, f w.r.t. x only.

Also, all along the paper, the notation A' means the transpose of the matrix A.

1.6.2. *Systems.* We will consider general finite dimensional nonlinear controlled systems, of the form

$$(\Sigma^1)\begin{cases} \dot{x} = f(x, u), \\ y = h(x, u), \end{cases} \tag{1.1}$$

or,

$$(\Sigma^2)\begin{cases} \dot{x} = f(x, u, \varphi(x)), \\ y = h(x, u, \varphi(x)), \end{cases} \tag{1.2}$$

where the state $x \in \mathbb{R}^n$, or more generally to a n-dimensional analytic differentiable manifold X, where u denotes the control variable , $u \in U$, some "regular" compact subset of $\mathbb{R}^p$ of dimension p (a polyhedron for instance) with nonempty interior, and $y \in \mathbb{R}^m$ is the output variable.

The function φ in Σ^2 is an unknown function to be identified. In this paper, we will restrict to the case where it is $\mathbb{I}$-valued, where $\mathbb{I}$ is a compact subinterval of $\mathbb{R}$ (i.e., we consider only the case of a single function to be identified) but the theory we developed in [2], clearly has extensions to higher dimension.

In practice, very often, this function φ represents some "physical characteristic" inside the system, that has to be determined on the basis of experiments. It may happen that φ does not depend on the whole state x of the system, but only on some projection $\pi(x)$ ($\pi : X \to Z$ a known fixed smooth function).

Remark 1. *Usually, output functions do not depend on the controls. It might seem only a (practically void) mathematical assumption to consider this dependence. Unfortunately, assuming the non-dependence would lead to more complicated unnatural statements, for many of our results below.*

1.6.3. *Topologies.* The sets of systems $S^1 = \{\Sigma^1 = (f, h)\}$, $S^2 = \{\Sigma^2 = (f, h)\}$, are given the C^∞ Whitney topology, (which is the relevant topology for our considerations). Basic neighborhoods of a system $\Sigma = (f, h)$ in this topology are determined by the data of functions $\varepsilon^j(z) > 0$, and are formed by the systems $\Sigma' = (f', h')$ such that all the partial derivatives, w.r.t. all variables, up to order j of $(f' - f, h' - h)$, have norm at $z = (x, u)$ (resp. $z = (x, u, \varphi)$) smaller than $\varepsilon(z)$. Because of the fact that ε depends on z, this topology "controls" the behavior at infinity of the systems. The counterpart of this fact is that (unless $X \times U \times \mathbb{I}$ is compact), it is not metrizable. Nevertheless, it has the nice "Baire property" that a countable intersection of open dense subsets is still dense. A subset of S^1, S^2 is called **residual** if it is such a countable intersection of open dense subsets. A subset is said generic if it contains a residual subset. (Contrarily to the way it sounds, "residual" means very big).

1.6.4. *Controls and outputs.* Control and output functions $u(t), y(t)$ of systems Σ^1 will be defined on semi-open intervals $[0, T_{u(.)}[$ depending on the control. May be $T_{u(.)} = +\infty$. Controls $u(.)$ and outputs $y(.)$ are measurable functions, bounded on any compact subinterval $[0, T] \subset [0, T_{u(.)}[$. The space of such functions is denoted by $L^\infty(U)$, (resp. $L^\infty(\mathbb{R}^m)$).

For systems $\Sigma^2 = (f(x, u, \varphi), h(x, u, \varphi)) \in S^2$, it will be convenient to consider φ not as a function of x, but as an extra input function $\varphi(t)$ of the time. Then, we will often consider in (1.2) that $\varphi = \varphi(t)$, $\varphi(.) \in L^\infty(\mathbb{I})$.

1.6.5. *State-output, input-state-output mappings and their "first variations".* For a system $\Sigma \in S^1$, the control $u(.)$ being fixed, we may consider the **state-output mapping** $PX_{\Sigma,u}$, which to the initial condition x_0 associates the output trajectory $y(.) \in L^\infty(\mathbb{R}^m)$.

$$PX_{\Sigma,u} : x_0 \to y(.).$$

If $u(.) \in L^\infty[0, T_u[$, it may happen that the output is only defined on $[0, T_y[$, $T_y < T_u$, i.e. up to the "**explosion time**" T_y. In that case, $\lim_{t\to T_y} x(t) = \infty$, (obviously, here, $x(t)$ is the state trajectory corresponding to initial condition x_0 and control $u(.)$). For all $T_0 < T_y$, there is a neighborhood $V_{x_0} \times V_u \subset$

$X \times L^{\infty}[0, T_0]$, such that for all $(x_0, u(.)) \in V_{x_0} \times V_u$, the corresponding trajectories $x(t), y(t)$ are well defined on $[0, T_0]$.

Moreover, the mapping $PX_{\Sigma,u}$ (defined on a neighborhood of x_0 as we said) is differentiable with respect to x_0. Let $TPX_{\Sigma,u}|x_0$ be this differential (it is well defined as a linear mapping from $T_{x_0}X$ to $L^{\infty}([0, T_0], \mathbb{R}^m)$, $T_0 < T_y$).

In fact, this differential $TPX_{\Sigma,u}|x_0$ is also the state-output mapping of another system TX_{Σ} , called the **first (state) variation** of Σ, with state space TX, the tangent bundle of X (or on $\mathbb{R}^n \times \mathbb{R}^n$)):

$$(1.3) \qquad (TX_{\Sigma}) \begin{cases} \dot{x} = f(x, u), \\ \dot{\xi} = D_x f(x, u)\xi, \\ \hat{y} = d_x h(x, u)\xi. \end{cases}$$

Here, $(x, \xi) \in TX$ (or $\mathbb{R}^n \times \mathbb{R}^n$) is the state of TX_{Σ}.

Let $PTX_{\Sigma,u}$ denote the state-output map of TX_{Σ}. Then, the linear mapping $\xi \to PTX_{\Sigma,u}(x_0, \xi)$ is the differential at x_0, $TPX_{\Sigma,u}|x_0$, of the state-output mapping .

Now, for the purpose of explaining the main features of our identification theory, we will define the **input-state-output mapping**, and the first (input-state-output) variation of Σ^2 in a similar way.

In fact, in our case, for systems $\Sigma^2 \in S^2$, we will be led to consider φ as an extra control function, as we said, the (eventual) usual control $u(.)$ remaining fixed.

Then, the input-state-output mapping is the mapping $PX\mathbb{I}_{\Sigma,u}$ (or $PX\mathbb{I}_{\Sigma}$ if there is no control u)

$$PX\mathbb{I}_{\Sigma,u} : (x_0, \varphi(.)) \to y(.),$$

which to the initial state x_0 and the extra control function $\varphi(.)$ associates the output function.

Assume that $PX\mathbb{I}_{\Sigma,u} : D \subset X \times L^{\infty}(\mathbb{I}) \to L^{\infty}(\mathbb{R}^m)$, is defined at a point (x_0, φ_0) on the time interval $[0, T_y[$. Then, for all $T < T_y$, it is defined on an open neighborhood of (x_0, φ_0) in $X \times L^{\infty}([0, T], \mathbb{I})$, and it is differentiable in the Frechet sense at (x_0, φ_0) on $X \times L^{\infty}([0, T], \mathbb{I})$. The differential is denoted by $T_{(x_0,\varphi_0)}PX\mathbb{I}_{\Sigma,u}$.

The **first (input-state-output) variation** of Σ^2 is the system $TX\mathbb{I}_{\Sigma,u}$,

$$(1.4) \qquad (TX\mathbb{I}_{\Sigma,u}) : \begin{cases} \dot{x} = f(x, u, \varphi_0), \\ \dot{\xi} = D_x f(x, u, \varphi_0)\xi + D_{\varphi} f(x, u, \varphi_0)\eta \\ \hat{y} = d_x h(x, u, \varphi_0)\xi + d_{\varphi} h(x, u, \varphi_0)\eta, \end{cases}$$

with (variational) control $\eta \in L^{\infty}(\mathbb{R})$.

If we take initial conditions $(x_0, \xi_0) \in TX$, and control functions $\varphi_0(.)$, $u(.)$ as above, then, the input-state-output mapping $PTX\mathbb{I}_{\Sigma,u}$ of $TX\mathbb{I}_{\Sigma,u}$ is the mapping:

$$(\xi_0, \eta(.)) \to \hat{y}(.),$$
$$T_{x_0}X \times L^{\infty}[\mathbb{R}] \to L^{\infty}[\mathbb{R}^m].$$

This mapping also coincides (on the small enough finite time intervals $[0, T]$ considered above) with the tangent mapping $T_{(x_0,\varphi_0)}PX\mathbb{I}_{\Sigma,u}$.

1.6.6. *k-jet extensions of state-output and input-state-output mappings.* Let us consider k-jets $j^k\hat{\varphi}$, $j^k\hat{u}$, of smooth functions $\hat{\varphi}$, $\hat{u}$ at $t = 0$,

$$\hat{\varphi} : [0, \varepsilon[\to \mathbb{I}, \ \ \hat{u} : [0, \varepsilon[\to U,$$
$$j^k\hat{\varphi} = (\hat{\varphi}(0), \hat{\varphi}'(0), ..., \hat{\varphi}^{(k-1)}(0)), \quad j^k\hat{u} = (\hat{u}(0), \hat{u}'(0), ..., \hat{u}^{(k-1)}(0)).$$

Then, for any $x_0 \in X$, the corresponding k-jet $j^k\hat{y} = (\hat{y}(0), \hat{y}'(0), ..., \hat{y}^{(k-1)}(0))$ is well defined, in such a way that the (extension to k-jets) mappings

$$\Phi_k^{\Sigma^1} : (x_0, j^k\hat{u}) \to j^k\hat{y}; \quad X \times U \times \mathbb{R}^{(k-1)p} \to \mathbb{R}^{km},$$
$$\Phi_k^{\Sigma^2} : (x_0, j^k\hat{u}, j^k\hat{\varphi}) \to j^k\hat{y}; \quad X \times (U \times \mathbb{R}^{(k-1)p}) \times (\mathbb{I} \times \mathbb{R}^{(k-1)}) \to \mathbb{R}^{km},$$

are continuous. We call these mappings $\Phi_k^{\Sigma^1}$, $\Phi_k^{\Sigma^2}$ the **k-jets state-output mappings, (resp. k-jets input-state output mappings)** associated to $\Sigma^1 \in S^1$ (resp. $\Sigma^2 \in S^2$).

2. Observation theory

2.1. **Definitions .** In this section, the relevant set of systems is S^1, i.e. there is no unknown function φ. We summarize the main observability results of the observation theory developed in [6].

We are not very precise in definitions and results with the explosion times, intervals of definitions of input and output functions, but everything is natural, and details can be found in [6].

Definition 1. *The system $\Sigma^1 = (f_1, h_1)$ is said uniformly* ***observable****, or just* ***observable****, w.r.t. a certain class $\mathcal{C}$ of inputs ($L^\infty(U)$ in most cases) if, for each $u(.) \in \mathcal{C}$, the state output mapping $PX_{\Sigma,u}$ is injective.*

Observability means that we can reconstruct the complete information about the system (i.e. the full state trajectory $x(.)$), from the knowledge of the input-output data $(u(.), y(.))$.

Injectivity is not a very tractable property, since it is not stable (even for standard mappings between finite dimensional spaces -example: $x \to x^3, \mathbb{R} \to \mathbb{R}$). Therefore, in order to state results, we need a few other definitions.

Notice also that, the bigger the class $\mathcal{C}$, the more restrictive observability property. For example, $L^\infty(U)$-observability is very strong, and implies observability in the C^k class, $k = 0, .., \infty, \omega$. A major property of C^ω (analytic) systems is the following, that expresses that in fact, no matter the class:

Theorem 1. [6] *For C^ω systems, C^ω observability implies L^∞ observability.*

This theorem is in fact very hard to prove.

For usual smooth mappings between finite dimensional spaces, a way to make the injectivity property stable is to add the requirement of infinitesimal injectivity (i.e. injectivity of all the tangent mappings). This is done for the study of differential mappings in differential topology. In the same spirit, let us define uniform infinitesimal observability.

Definition 2. *System Σ^1 is said uniformly infinitesimally observable if, for each $u(.) \in L^\infty(U)$, each $x_0 \in X$, all the tangent mappings $TPX_{\Sigma^1,u}|x_0$ are injective. By Section 1.6.5, it is equivalent to require that the state-output mappings $PTX_{\Sigma^1,u}$ of the first variation of Σ^1 are injective.*

Another way to be more effective is to "pass to k-jets":

Definition 3. *System Σ^1 is said differentially observable (of order k) if for all $j^k\hat{u}$, the extension to k-jets mapping $\Phi_k^{\Sigma^1} : x_0 \to j^k\hat{y};\ \ X \to \mathbb{R}^{km}$ is injective.*

Again, this definition will become more effective if one adds an "infinitesimal injectivity" requirement:

Definition 4. *System Σ^1 is said strongly differentially observable (of order k) if for all $j^k\hat{u}$, the extension to k-jets mapping $\Phi_{k,j^k\hat{u}}^{\Sigma^1} : x_0 \to j^k\hat{y};\ \ X \to \mathbb{R}^{km}$ is an injective immersion (immersion means that all the tangent mappings $T_{x_0}\Phi_{k,j^k\hat{u}}^{\Sigma^1}$ to this map, have full rank n at each point).*

Clearly, strong differential observability implies differential observability, which implies observability for the C^∞ class, which -for analytic systems- implies L^∞-observability by Theorem 1.

It is also a consequence of the theory that -for analytic systems- uniform infinitesimal observability implies observability of the restrictions of Σ^1 to small open subsets of X, the union of which is dense in X (but this is a priori non-obvious).

2.2. **The generic case.** We consider here the case where $m > p$ (i.e. the number of outputs is strictly larger than the number of inputs). Then in that case, strong differential observability is generic:

Theorem 2. . *a).The set of systems that are strongly differentially observable of order $2n+1$ is residual in S^1;*

b) The set of ***analytic*** *strongly differentially observable systems (of order $2n+1$) that are moreover L^∞-observable is dense in S^1.*

For people who know about these topics, it could seem that b) is a consequence of a) and of Theorem 1, using some general result of "approximation of smooth by analytic". It is not at all the case, and this part b) is difficult in itself.

This theorem has a nice consequence:

Theorem 3. *The following is a generic (residual) property on S^1 : Set $k = 2n+1$.*

For all sufficiently smooth $u(.)$, set $j^k u(t) = (u(t), \dot{u}(t), ..., u^{(k-1)}(t))$. Chose an arbitrarily large relatively compact open subset Γ of X, and an arbitrary bound on $u, \dot{u}, ..., u^{(k)}$, the control and its first k derivatives. Then the mappings $\Phi^{\Sigma^1}_{k,j^k u} : x(t) \mapsto (y, \dot{y}(t), ..., y^{(k-1)}(t))$ are smooth injective immersions that map the trajectories of the system Σ^1 (restricted to Γ) to the trajectories of the following system:

$$
\begin{aligned}
y &= z_1, \\
\dot{z}_1 &= z_2, \\
&\cdot \\
&\cdot \\
\dot{z}_{k-1} &= z_k, \\
\dot{z}_k &= \varphi_K(z_1, ..., z_k, u, \dot{u}, ..., u^{(k)}).
\end{aligned}
\tag{2.1}
$$

A system under the form (2.1) is called a "phase-variable representation" (of order k). It means that, in restriction to a compact subset K of X,

$$y^{(k)} = \varphi_K(y, ..., y^{(k-1)}, u, \dot{u}, ..., u^{(k)}).$$

Theorem 3 claims that, generically, in restriction to (arbitrarily large) compact subset, a system $\Sigma^1 \in S^1$ can be embedded into a phase-variable one, and the state $x(t)$ of Σ^1 can be recovered from the state $z(t)$ of (2.1), by the inverse mapping of $\Phi^{\Sigma^1}_{k,j^k u}$, which is also smooth.

Of course, if we consider $(u, \dot{u}, ..., u^{(k)}) = v$ as the control, systems of the form (2.1), $y^{(k)} = \varphi_K(y, ..., y^{(k-1)}, v)$, are observable, strongly differentially observable, uniformly infinitesimally observable. Hence, if $m > p$, (and for sufficiently smooth inputs), generic systems are subsystems of other systems that are very good from the point of view of observability.

2.3. The nongeneric case $m \leq p$.

2.3.1. *The canonical flag.* In this discussion, we will restrict to the case where $m = 1, \ \ p \geq 1$. Results for $m > 1$ are less clear. We will restrict also to analytic systems in S^1, but this is a purely technical assumption that can be avoided.

Associated to $\Sigma = (f, h) \in S^1$, we may define the **canonical flag** $D(u)$ of distributions as follows:

$$
\begin{aligned}
D(u) &= \{D^0(u) \supset D^1(u) \supset ... \supset D^{n-1}(u)\}, \\
D^0(u) &= Ker(d_x h), \quad D^{k+1}(u) = D^k(u) \cap Ker(d_x L_f^{k+1} h),
\end{aligned}
\tag{2.2}
$$

where $L_f h$ is the Lie derivative of the function h w.r.t. the vector field f, the control u being considered as fixed.

The flag $D(u)$ is a flag of possibly singular distributions, depending on the value of the control u.

If the distributions $D^i(u)$ have constant rank $n - i - 1$ and are independent of u, then, the canonical flag $D(u)$ is said to be **uniform**.

Theorem 4. *The system Σ has a uniform canonical flag if and only if, for all $x^0 \in X$, there is a coordinate neighborhood of x^0, (V_{x^0}, x), such that, in these coordinates, the system $\Sigma_{|V_{x^0}}$ (Σ restricted to V_{x^0}) can be written as:*

$$
\begin{aligned}
y &= h(x_1, u); \\
\dot{x}_1 &= f_1(x_1, x_2, u), \\
\dot{x}_2 &= f_2(x_1, x_2, x_3, u), \\
&\vdots \\
\dot{x}_{n-1} &= f_{n-1}(x_1, x_2, .., x_n, u), \\
\dot{x}_n &= f_n(x_1, x_2, ..., x_n, u),
\end{aligned}
\tag{2.3}
$$

where moreover

$$\frac{\partial h}{\partial x_1} \ \text{and} \ \frac{\partial f_i}{\partial x_{i+1}}, i = 1, .., n-1, \tag{2.4}$$

are never zero on $V_{x_0} \times U$.

The property to have a uniform canonical flag is highly non-generic (it has codimension ∞).

It is easily seen from this normal form that, if a system Σ has a uniform canonical flag, then, when restricted to neighborhoods $V_{x_0} \times U$ where it is under the normal form (2.3, 2.4), it is **infinitesimally observable, observable, and differentially observable of order** n.

2.3.2. *Characterization of uniform infinitesimal observability.* The main result is the following: a necessary condition for a system Σ to be uniformly infinitesimally observable, is that, on an open-dense subset of X, it has a uniform canonical flag.

Theorem 5. *If Σ is uniformly infinitesimally observable, then, on the complement of a subanalytic subset of X of codimension 1, Σ has a uniform canonical flag.*

In other term, the canonical form (2.3, 2.4) characterizes uniform infinitesimal observability.

2.3.3. *Control affine case.* In the control affine case, where Σ can be written:

$$\begin{aligned} y &= h(x); \\ \dot{x} &= f(x) + \sum_{i=1}^{p} g_i(x) u_i, \end{aligned} \tag{2.5}$$

there is a stronger result. Set $\Phi = (h, L_f h, ..., L_f^{n-1} h)$, $\Phi : X \rightarrow \mathbb{R}^n$. First, it is an elementary exercise to show that, if Σ is observable, then Φ has to have maximum rank n on an open dense subset V of X. Then, consider any subset $W \subset X$ in restriction to which Φ is a diffeomorphism.

Theorem 6. *Assume that Σ is observable. Then, the restriction $\Phi_{|W}$ maps Σ into a system of the form:*

$$\begin{aligned} y &= x_1; \\ \dot{x}_1 &= x_2 + \sum_{i=1}^{p} g_{1,i}(x_1) u_i, \\ \dot{x}_2 &= x_3 + \sum_{i=1}^{p} g_{2,i}(x_1, x_2) u_i, \\ &. \\ &. \\ \dot{x}_{n-1} &= x_n + \sum_{i=1}^{p} g_{n-1,i}(x_1, x_2, .., x_{n-1}) u_i, \\ \dot{x}_n &= \psi(x) + \sum_{i=1}^{p} g_{n,i}(x_1, x_2, .., x_{n-1}, x_n) u_i. \end{aligned} \tag{2.6}$$

Conversely, if a system is under the form (2.6) on an open subset $\Omega \subset \mathbb{R}^n$, then it is observable.

This normal form (2.6) is of course a special case of the uniform infinitesimal observability canonical form (2.3, 2.4).

Notice that both theorems 5, 6 have a global character: they are local almost everywhere w.r.t. x, but global w.r.t. u.

3. Identification theory

We will restrict to the uncontrolled case, i.e. our systems $\Sigma^3 = (f, h) \in S^3$ are of the form:

$$\Sigma^3 : \left\{ \begin{array}{l} y = h(x, \varphi(x)); \\ \dot{x} = f(x, \varphi(x)), \end{array} \right. \tag{3.1}$$

with $\varphi : X \rightarrow \mathbb{I}$, i.e. there is no control, and a single function to identify. These are the results presented in [2]. But now, we already have results for more general systems, with controls, and with several φ's.

The results of our identifiability theory are very comparable to the results of Section 2 above.

3.1. **Definitions.** We will give several definitions of identifiability, starting form a general natural one, but not very tractable. In our definitions, as we said, $\varphi(.)$ will not be considered as a function of x, but as an extra input, function of the time t. Some of these definitions are with respect to the class of functions $\varphi(.)$ that are measurable bounded only (although, $\varphi(x(t))$ is smooth, and even analytic in t if Σ^3 is analytic). This choice (of a largest class of φ's) is in fact justified by the following property: if a system -analytic- is identifiable (or uniformly infinitesimally identifiable) in the sense defined below, for C^ω inputs $\varphi(t)$, then it is also identifiable in the same sense, for general L^∞ inputs. This is discussed in details in our paper [2].

In presence of controls u (for systems in S^2, which we do not address here), the natural class for the φ's as functions of t is the class of absolutely continuous functions: $\varphi(x)$ is smooth and $x(t)$ is absolutely continuous.

Definition 5. *The system $\Sigma \in S^3$ is said identifiable at $y(.) \in C^\infty[0, T_y[$, if there is at most a single couple $(x_0, \varphi(.))$, with $\varphi(.) \in C^\infty[0, T_y[$, such that, for all $t \in [0, T_y[$,*

$$PX\mathbb{I}_\Sigma(x_0, \varphi)(t) = y(t).$$

Σ is said identifiable if it is identifiable at all $y(.) \in C^\infty[0, T_y[$.

In other terms, Σ is identifiable if its input-state-output mapping is injective.

Now, let us consider $\Phi_k^\Sigma : X \times \mathbb{I} \times \mathbb{R}^{(k-1)} \to \mathbb{R}^{km}$, the k-jet input-state output mapping of Σ, $(x_0, j^k\hat{\varphi}) \to j^k\hat{y}$.

Definition 6. *The system Σ is said differentially identifiable of order k, if,*

$$\Phi_k^\Sigma(x_0^1, j^k\hat{\varphi}^1) = \Phi_k^\Sigma(x_0^2, j^k\hat{\varphi}^2)$$

implies that $(x_0^1, \hat{\varphi}^1(0)) = (x_0^2, \hat{\varphi}^2(0))$.

This property is weaker than the injectivity of Φ_k^Σ. It means that all couples (initial state, value of φ) are distinguished between them by the observations and their $k-1$ first derivatives. But, it may happen that certain couples $(x_0, j^k\hat{\varphi})$ are not distinguished between them.

For the purpose of getting a genericity result similar to Theorem 2 for observability, this is the adequate notion. (one could think that the injectivity of Φ_k^Σ is the right notion for this purpose, but it is never generic).

Also, the following is more or less obvious:

Theorem 7. *Differential identifiability at some order implies identifiability.*

Now, we will define the infinitesimal notion of identifiability.

We consider $TX\mathbb{I}_\Sigma$, the first input-state-output variation of $\Sigma \in S^3$, and its input-state-output map $PTX\mathbb{I}_\Sigma$,

$$\begin{gathered}(\xi_0, \eta(.)) \to \hat{y}(.),\\ T_{x_0}X \times L^\infty[\mathbb{R}] \to L^\infty[\mathbb{R}^m].\end{gathered}$$

It is equivalent to consider the tangent mapping $T_{(x_0,\varphi_0)}PX\mathbb{I}_\Sigma$ of the input-state-output mapping $PX\mathbb{I}_\Sigma$ of Σ.

Definition 7. *Σ is said uniformly infinitesimally identifiable if, for all $(x_0, \varphi_0(.)) \in X \times L^\infty[\mathbb{I}]$, the tangent mapping $PTX\mathbb{I}_\Sigma$ is injective (as a mapping $T_{x_0}X \times L^\infty([0, t], \mathbb{R}) \to L^\infty([0, t], \mathbb{R}^m)$, for all $t < T_{y_0}$, where $y_0(.) \in L^\infty(\mathbb{R}^m)$ is defined on $[0, T_{y_0}[$).*

That is, uniform infinitesimal identifiability means that all the tangent mappings to the input-state-output mapping are injective.

It will be a consequence of the theory that (in all the cases under consideration) uniform infinitesimal identifiability implies identifiability of the restrictions of the system to certain small open subsets of $X \times \mathbb{I}$, the union of which is dense in $X \times \mathbb{I}$.

3.2. **The generic case.** We have the fundamental following result, comparable to Theorem 2 for observability.

Theorem 8. *If the number of outputs m is larger or equal to* 3*, then, differential identifiability of order* $2n+1$ *is a generic property. In particular, identifiability is a generic property.*

Of course, this theorem is false if $m = 1, 2$. On the contrary, identifiability becomes a property of infinite codimension.

3.3. **The nongeneric cases** $m = 1, 2$**.** Again here, and also for purely technical reasons, we consider systems that are analytic only.

3.3.1. *The single output case.* We denote by L_f (or L_{f_φ}, when φ is fixed) the Lie-derivative operator on X. Also, f_φ denotes the vector field $f(.,\varphi)$, for $\varphi \in \mathbb{I}$, and $h_\varphi : X \to \mathbb{R}$ is the map $h(.,\varphi)$.

Theorem 9. *If* Σ *is uniformly infinitesimally identifiable, then, there is a subanalytic closed subset* Z *of* X*, of codimension 1 at least, such that on the open dense set* $X\backslash Z$*, the following two equivalent properties 1 and 2 below hold:*

1.a. $\frac{\partial}{\partial\varphi}\left\{(L_{f_\varphi})^k h_\varphi\right\} \equiv 0$*, for* $k = 0,...,n-1$*, b.* $\frac{\partial}{\partial\varphi}\left\{(L_{f_\varphi})^n h_\varphi\right\} \neq 0$ *(in the sense that it* ***never*** *vanishes), c.* $d_x h_\varphi \wedge ... \wedge d_x L_{f_\varphi}^{n-1} h_\varphi \neq 0$*,*

2. any $x_0 \in X\backslash Z$ *has a coordinate neighborhood* $(x_1, ..., x_n,\ V_{x_0})$*,* $V_{x_0} \subset X\backslash Z$ *in which* Σ *(restricted to* V_{x_0}*) can be written:*

$$\left\{\begin{array}{c} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ . \\ . \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = \psi(x,\varphi); \\ y = x_1; \end{array}\right. \tag{3.2}$$

where $\frac{\partial}{\partial\varphi}\psi(x,\varphi)$ *never vanishes.*

This theorem has the following pseudo-converse:

Theorem 10. *Assume that* Σ *meets the equivalent conditions of the previous theorem.*

Then, any x_0 *has a neighborhood* V_{x_0} *such that the restriction* $\Sigma_{|V_{x_0}}$ *of* Σ *to* V_{x_0} *is uniformly infinitesimally identifiable,* ***identifiable and differentially identifiable of order*** $n+1$*.*

Notice that, again, Theorem 9 has a global character: it is almost everywhere on X, but it is global with respect to $\varphi \in \mathbb{I}$.

3.3.2. *The two-output case.* Let us first state the results in a non invariant way.

Theorem 11. *If* Σ *is uniformly infinitesimally identifiable, then, there is an open-dense subanalytic subset* $\tilde{U}$ *of* $X \times \mathbb{I}$*, such that each point* (x_0, φ_0) *of* $X \times \mathbb{I}$*, has a neighborhood* $V_{x_0} \times \mathbb{I}_{\varphi_0}$*, and coordinates* x *on* V_{x_0} *such that the system* Σ *restricted to* $V_{x_0} \times \mathbb{I}_{\varphi_0}$*, denoted by* $\Sigma_{|V_{x_0}\times\mathbb{I}_{\varphi_0}}$*, has one of the three following normal forms:*

*-**type 1 normal form:** (in that case,* $n > 2k$*)*

$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2, \\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4, \\ &.. \\ \dot{x}_{2k-3} &= x_{2k-1},\ \dot{x}_{2k-2} = x_{2k}, \\ \dot{x}_{2k-1} &= f_{2k-1}(x_1, ..., x_{2k+1}), \\ \dot{x}_{2k} &= x_{2k+1}, \\ &.. \\ \dot{x}_{n-1} &= x_n, \\ \dot{x}_n &= f_n(x,\varphi),\ \textit{with}\ \frac{\partial f_n}{\partial\varphi} \neq 0\ \textit{(never vanishes)}, \end{aligned} \tag{3.3}$$

-type 2 normal form:

(3.4)
$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2,\\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\ &..\\ \dot{x}_{2r-3} &= x_{2r-1},\ \dot{x}_{2r-2} = x_{2r},\\ \dot{x}_{2r-1} &= \Phi(x,\varphi),\ \dot{x}_{2r} = F_{2r}(x_1...,x_{2r+1},\Phi(x,\varphi)),\\ \dot{x}_{2r+1} &= F_{2r+1}(x_1...,x_{2r+2},\Phi(x,\varphi)),\\ &..\\ \dot{x}_{n-1} &= F_{n-1}(x,\Phi(x,\varphi)),\\ \dot{x}_n &= F_n(x,\varphi), \end{aligned}$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0, ..., \frac{\partial F_{n-1}}{\partial x_n} \neq 0$,

-type 3 normal form:

(3.5)
$$\begin{aligned} y_1 &= x_1,\ y_2 = x_2,\\ \dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\ &..\\ \dot{x}_{n-3} &= x_{n-1},\ \dot{x}_{n-2} = x_n,\\ \dot{x}_{n-1} &= f_{n-1}(x,\varphi),\ \dot{x}_n = f_n(x,\varphi), \end{aligned}$$

where $(\frac{\partial f_{n-1}}{\partial \varphi}, \frac{\partial f_n}{\partial \varphi})$ *never vanishes.*

Notice that the type 2 normal form 3.4 is very comparable to the observability normal form (2.3, 2.4).

Now, we will give the intrinsic characterization of the conditions "type 1, type 2, type 3".

We define two integers r and k, attached to a two-output system $\Sigma \in S^3$. The first one r is called the order of the system. It is the first integer such that $d_\varphi L_f^r h$ does not vanish identically on $X \times \mathbb{I}$.

Set $h = (h_1, h_2)$.

Let $N(l)$ be the rank at generic points of $X \times \mathbb{I}$ of the family E_l of one-forms on X:

$$E_l = \{d_x h_i, d_x L_f h_i, ..., d_x L_f^{l-1} h_i, i = 1, 2\}.$$

Set $N(0) = 0$.

This set of generic points U_l, is the intersection of the open sets $\tilde{U}_i$, $i \leq l$, where E_i has maximal rank. U_l is semianalytic, open and dense in $X \times \mathbb{I}$. Moreover, $U_{l+1} \subset U_l$.

It is easy to check that $N(l)$ increases strictly by steps of 2, up to $l \overset{\text{def.}}{=} k$, and after, (eventually), it increases by steps of 1 up to $l \overset{\text{def.}}{=} l_M$, $N(l_M) \leq n$.

It may happen that $k = 0$, i.e. $N(1) = 1$.

Lemma 1. *If* Σ *is uniformly infinitesimally identifiable, then,* $N(l_M) = n$ *and* $r \leq l_M$.

Definition 8. *A system* Σ *is* ***regular*** *if* $N(l_M) = n$ *and* $r \leq l_M$.

Lemma 1 says that, if a system is uniformly infinitesimally identifiable, then it is regular. From now on, in this section, we will assume that systems Σ under consideration are regular.

The integer k is the first with the following properties:

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^k h_1 \wedge d_x L_f^k h_2 \equiv 0, \quad \text{but}$$
$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_1 \wedge d_x L_f^{k-1} h_2 \neq 0 \text{ (not identically zero).}$$

If $r = k$, there are three possibilities:

A. $n = 2k$;

B.

B.1.

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_2 \neq 0$; or,

B.2.

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_1 \neq 0$;

C.

C.1

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_2 \equiv 0$, $d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \equiv 0$,

or

C.2

$$d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_2 \neq 0$$

(hence $n > 2k$) and $d_\varphi L_f^k h_1 \equiv 0$, $d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \wedge ... \wedge d_x L_f^{k-1} h_2 \wedge d_x L_f^k h_1 \equiv 0$

Definition 9. *Let Σ be a **regular** system. We say that Σ has:*
*-**type 1** if $r > k$, or $r = k$ but **C**. is satisfied,*
*-**type 2** if $r < k$, or $r = k$ but **B**. is satisfied,*
*-**type 3** if $r = k$ and **A**. is satisfied.*

Lemma 2. *Types 1, 2 and 3 exhaust the class of regular systems, and form a partition of this class.*

Type 2 regular systems:

For a regular system of type 2, eventually interchanging the role of h_1, h_2, we can assume that $d_\varphi L_f^r h_2(x,\varphi) \neq 0$. In a neighborhood of a point $(x_0, \varphi_0) \in U_{l_M}$, such that $L_f^r h_2(x_0,\varphi_0) = u_0$ and $d_\varphi L_f^r h_2(x_0,\varphi_0) \neq 0$, there is an analytic function $\Phi^*(x,u)$, such that $L_f^r h_2(x, \Phi^*(x,u)) = u$. Let us consider the "auxiliary system" Σ_A :

$$\Sigma_A \left\{ \begin{array}{l} \dot{x} = f(x, \Phi^*(x,\tilde{\varphi})) = F(x,\tilde{\varphi}) \\ y = h(x, \Phi^*(x,\tilde{\varphi})) = H(x,\tilde{\varphi}). \end{array} \right.$$

This system is well defined and intrinsic, over an open set $V_{x_0} \times V_{u_0} \subset X \times \mathbb{R}$.

By construction, the integer r (the order) associated with this auxiliary system is the same as the one of the given system Σ.

Moreover, the following flags D and D^A of integrable distributions over V_{x_0} :

$$D_0(x) = T_x X,\ D_1(x) = Ker(d_x h(x)), ...D_r(x) = D_{r-1}(x) \cap Ker(d_x L_f^{r-1} h(x)),$$
$$D = \{D_0 \supset D_1 \supset ... \supset D_r\};$$

and

$$D_0^A(x) = T_x X,\ D_1^A(x) = Ker(d_x H(x)), ...D_r^A(x) = D_{r-1}^A(x) \cap Ker(d_x L_F^{r-1} H(x)),$$
$$D^A = \{D_0^A \supset D_1^A \supset ... \supset D_r^A\},$$

are equal.

Let us "prolong" the auxiliary flag D^A, in the following way:

$$D_{r+1}^A(x,\tilde{\varphi}) = D_r^A(x) \cap Ker(d_x L_F^r H_1(x,\tilde{\varphi})),$$
$$D_{i+1}^A(x,\tilde{\varphi}) = D_i^A(x) \cap Ker(d_x L_F^i H_1(x,\tilde{\varphi})),$$
$$D^A(\tilde{\varphi}) = \{D_0^A \supset D_1^A \supset .. \supset D_r^A \supset D_{r+1}^A(\tilde{\varphi}) \supset .. \supset D_l^A(\tilde{\varphi}) = D_{l+1}^A(\tilde{\varphi})\},$$

where l is the first integer such that $D_l^A(x,\tilde{\varphi}) = D_{l+1}^A(x,\tilde{\varphi})$ at generic points.

Definition 10. *The auxiliary flag $D^A(\tilde{\varphi})$ is **regular** on an open subset $U \subset X \times \mathbb{I}$, if $D_l^A(\tilde{\varphi}) = \{0\}$, and all the other $D_i^A(\tilde{\varphi})$ have constant rank first $n - 2i$ $(i \leq r)$, second $n - r - i$ $(r < i < l)$, third, 0 $(i \geq l = n - r)$; on this open set.*

Definition 11. *The auxiliary flag $D^A(\tilde{\varphi})$ is **uniform** on an open subset $U \subset X \times \mathbb{I}$, if it is regular, and independent of $\tilde{\varphi}$.*

This property of having a uniform auxiliary flag (for identifiability) is the equivalent of having a uniform canonical flag for observability: it is the necessary (and almost sufficient condition) for uniform infinitesimal identifiability (type 2):

Theorem 12. ***(Normal form for a uniform auxiliary flag)*** *A system Σ has a uniform auxiliary flag around (x_0, φ_0), iff there is a neighborhood $V_{x_0} \times \mathbb{I}_{\varphi_0}$ of (x_0, φ_0), and coordinates on V_{x_0} such that Σ can be written:*

$$\begin{aligned}
y_1 &= x_1,\ y_2 = x_2,\\
\dot{x}_1 &= x_3,\ \dot{x}_2 = x_4,\\
&..\\
\dot{x}_{2r-3} &= x_{2r-1},\ \dot{x}_{2r-2} = x_{2r},\\
\dot{x}_{2r-1} &= \Phi(x,\varphi),\ \dot{x}_{2r} = F_{2r}(x_1 ..., x_{2r+1}, \Phi(x,\varphi)),\\
\dot{x}_{2r+1} &= F_{2r+1}(x_1 ..., x_{2r+2}, \Phi(x,\varphi)),\\
&..\\
\dot{x}_{n-1} &= F_{n-1}(x, \Phi(x,\varphi)),\\
\dot{x}_n &= F_n(x,\varphi),
\end{aligned}$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_{2r}}{\partial x_{2r+1}} \neq 0, ..., \frac{\partial F_{n-1}}{\partial x_n} \neq 0.$

Theorem 13. ***(intrinsic result in the 2-output case)*** *If Σ is uniformly infinitesimally identifiable,* ***(hence regular)****, then, there is an open-dense subanalytic subset $\tilde{U}$ of $X \times \mathbb{I}$, such that at each point (x_0, φ_0) of $\tilde{U}$, Σ has the following properties, on a neighborhood of (x_0, φ_0):*
-If Σ has type 2, the auxiliary flag is uniform,
-If Σ has type 1, then, $N(r) = n$.

This theorem is in fact equivalent to Theorem 11.

These two equivalent theorems (Theorems 13, 11) have a weak converse:

Theorem 14. *Assume that Σ satisfies the equivalent conditions of theorems 13, 11, on some subset $V_{x_0} \times \mathbb{I}_{\varphi_0}$ of $X \times \mathbb{I}$ (so that, taking V_{x_0}, $\mathbb{I}_{\varphi_0}$ small enough, the restriction $\Sigma_{|V_{x_0} \times \mathbb{I}_{\varphi_0}}$ has one of the three normal forms above on $V_{x_0} \times \mathbb{I}_{\varphi_0}$). Then, in case type 1, type 2, (normal forms 3.3, 3.4) $\Sigma_{|V_{x_0} \times \mathbb{I}_{\varphi_0}}$ is uniformly infinitesimally identifiable* ***and identifiable****. In case type 3 (normal form 3.5), this is also true, eventually restricting the neighborhoods V_{x_0}, $\mathbb{I}_{\varphi_0}$.*

Also, in the special case of type 1, there is a stronger result:

Theorem 15. *Assume Σ is uniformly infinitesimally identifiable,* ***(hence regular)****. Assume that Σ has type 1. Then, there is an open-dense subanalytic subset $\tilde{X}$ of X, such that each point x_0 of $\tilde{X}$, has a neighborhood V_{x_0}, and coordinates x on V_{x_0} such that the system Σ restricted to $V_{x_0} \times \mathbb{I}$, denoted by $\Sigma_{|V_{x_0}}$, has the normal form 3.3 (globally over $V_{x_0} \times \mathbb{I}$). Conversely, if it is the case, then, the restriction $\Sigma_{|V_{x_0}}$ is uniformly infinitesimally identifiable* ***and identifiable****.*

4. Observer and identifier design

4.1. **Observer design.** Here, mainly, we recall the results of [6], and of the paper [3]. We add an improvement that makes the strategy proposed in [3] effective for uniformly infinitesimally observable systems, i.e. systems in normal form (2.3, 2.4). This improvement uses a crucial observation of Hammouri and all., in [7].

4.1.1. *Luenberger-type high-gain observers.* We present the basic construction of high gain observers. It works for general uniformly infinitesimally observable systems in normal form (2.3, 2.4). As a consequence, it works also for control affine systems in the normal form (2.6), or for systems in phase-variable representation (2.1), (therefore, it works in all cases -generic or not- of "observable" systems exhibited by the theory). In the 2 last cases, the construction is more explicit than in the case of Normal-form (2.3, 2.4).

Let us consider a system on $\mathbb{R}^n$, which is globally under the form (2.3, 2.4), or under the affine normal form (2.6), or under the phase-variable form:

$$\begin{aligned}
y &= Cx = x_1,\\
\dot{x}_1 &= x_2, ..., \dot{x}_{N-1} = x_N,\\
\dot{x}_N &= \psi(x,u);
\end{aligned} \tag{4.1}$$

In that case, we may have several outputs (i.e. $m > 1$, each $x_i \in \mathbb{R}^m$, $n = Nm$), and the control in the normal form (2.1), that was $(u, \dot{u}, ..., u^{(N)})$ is now denoted by u. This practically means that the observer system will be fed not only by the control, but also by certain of its derivatives.

We will make the following technical assumptions, that will be discussed below:
Assumptions (A):
Case of Normal-form (2.3, 2.4):
A1. $0 < \alpha \leq \frac{\partial h}{\partial x_1} \leq \beta$, $0 < \alpha \leq \frac{\partial f_i}{\partial x_{i+1}} \leq \beta$, $i = 1, ..., n-1$
A2. Each of the maps f_i, $i = 1, ..., n-1$, is globally Lipschitz w.r.t. $(x_1, ..., x_i))$, uniformly w.r.t. u and x_{i+1}.
A3. The control u is bounded,
Case of Normal-forms (2.6) **and** (4.1):
A1. The control u (and a certain number of first derivatives of u, if any) is bounded
A2. All the functions ψ, $g_{i,j}$ appearing in the normal forms are compactly supported.

Comment 1 about these assumptions:
The assumption that u is bounded, seems to be a non avoidable requirement. In fact, since we assumed U compact, it is automatically satisfied. We recall it here for clarity. Also, in the case of a phase variable representation 2.1, it means that certain derivatives of u have to be bounded.

Comment 2 about these assumptions:
The other assumptions can be always realized, provided that we restrict the observation problem to a compact subset $K \subset X$ (which means that we want to estimate the state $x(t)$ of Σ as long as the trajectory $x(t)$ remains in K only).
Indeed:
1. in the case of Normal forms (2.6) and (4.1), all the functions under consideration can be multiplied by a smooth cut-off function, which is equal to one on K;
2. in the case of Normal-form (2.3, 2.4), h and f can be smoothly prolonged outside K, for they satisfy A1, A2. (This last point is not immediate, and it is shown in [6]).

Under these assumptions, let us consider another system of the form:

$$\frac{d\hat{x}}{dt} = f(\hat{x}, u) - \Delta_\theta \Omega(h(\hat{x}, u) - y(t)); \tag{4.2}$$

where Ω is a certain constant $n \times m$ matrix, θ is a real parameter, and Δ_θ is the block-diagonal matrix:

$$\Delta_\theta = Block - diag(\theta Id_m, ..., \theta^N Id_m).$$

Here, Id_m denotes the $m \times m$ identity matrix, and $m = 1$ in other cases than (4.1).

Theorem 16. *There is an Ω with the following properties: for all $\hat{\beta} > 0$, there exists a θ (large enough), such that, whatever $\hat{x}(0)$, the solution $\hat{x}(t)$ of (4.2) satisfies:*

$$||\hat{x}(t) - x(t)|| \leq k(\hat{\beta}) e^{-\hat{\beta} t} ||\hat{x}(0) - x(0)||,$$

as long as $x(t)$ remains in K. The function $k(.)$ has polynomial growth.

Remark 2. *The function k having polynomial growth, it implies that the "estimation error" $||\hat{x}(t) - x(t)||$ can be made arbitrarily small in arbitrary short time (a polynomial against an exponential).*

The construction of the matrix Ω is not so hard in the case (2.3, 2.4), See [6]. It is specially simple in the cases (2.6), (4.1): any Ω such that $(A - \Omega C')$ is Hurwitz does the job (where (A, C) is the canonical linear system in Brunowsky form).

This very simple "luenberger-type" observer shows already good performances in many cases. Since it is high-gain (θ is large, then the "correction gain" $\Delta_\theta \Omega$ might be big), it may be sensitive to noise.

4.1.2. *Kalman-filter type high-gain observers.* The high-gain extended Kalman filter is another solution of high-gain type. Since it is high gain, it might be also sensitive to noise (although, very often, it works well in practice). It is related to the classical extended Kalman filter, and it can be shown that (in a stochastic context), it is a nonlinear filter with bounded variance (See [4]).

In [6] and in [3], it is applied to a less general class of nonlinear systems than the Luenberger high-gain observer. In Section 4.1.4 below, we will show how to apply it to the same class of systems. But in this section, let us stay at the level of the results of [6] and [3].

We assume that the system $\Sigma \in S^1$ is (on $\mathbb{R}^n$) in normal form (2.6), or more generally in the following (multi-output) normal form:

$$y = Cx = x_1, \qquad (4.3)$$
$$\dot{x}_1 = Ax + b(x,u),$$

where $C : \mathbb{R}^n \to \mathbb{R}^m$, $C = (Id_m, 0, ..., 0)$, where A is the $Nm \times Nm = n \times n$ block-antishift matrix:

$$A = \begin{pmatrix} 0 & Id_m & 0 & . & . & 0 \\ 0 & 0 & Id_m & 0 & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & . & . & . & 0 & Id_m \\ 0 & 0 & . & . & . & 0 \end{pmatrix},$$

and where $x = (x_1, ..., x_N)$, $x_i \in \mathbb{R}^m$, and $b(x,u)$ is lower block-triangular: the i^{th} component b_i depends only on $(x_1, ..., x_i)$.

Notice that this normal form includes the case (2.1), (4.1) of a phase-variable representation.

The only case of observable systems which is not covered by this normal form is the "uniform infinitesimal observability normal form" (2.3, 2.4). But, as we said, we will remedy to this in Section 4.1.4.

Again, we need additional technical assumptions:

Assumption B:

B1. the components $b_i(x,u)$ are compactly supported with respect to all their respective arguments;

B2. u is bounded (void assumption if $u \in U$ which is assumed to be compact, but non void assumption if $u = (u, \dot{u}, ..., u^{(k)})$ for a phase-variable representation).

Remark 3. *As in the previous section 4.1.1, the assumption B1 can be realized by smooth prolongation of b out of any compact subset $K \subset X \times U$. In that case, the observer we construct will work as long as the trajectories $x(t)$ of the system remain inside K.*

Consider the Kalman-type equations:

$$\frac{dS}{dt} = -(A + D_{\hat{x}}b(\hat{x},u))'S - S(A + D_{\hat{x}}b(\hat{x},u)) + C'r^{-1}C - SQ_\theta S, \qquad (4.4)$$
$$\frac{d\hat{x}}{dt} = A\hat{x} + b(\hat{x},u) - S^{-1}C'r^{-1}(C\hat{x} - y).$$

Here, $Q_\theta = \theta^2\Delta^{-1}Q\Delta^{-1}$, $\Delta = Block - diag(Id_m, \frac{1}{\theta}Id_m, ..., (\frac{1}{\theta})^{N-1}Id_m)$. Here, θ is a real parameter. The matrix S, as usual, lies in the cone of symmetric positive definite matrices. This equation is called the "high-gain extended Kalman filter". If $\theta = 1$, it is just the standard Extended Kalman filter equation, and in a stochastic context, Q and r (both symmetric positive definite) are the covariance matrices of the state and output noise respectively.

Theorem 17. *Let $x(t)$, $t \geq 0$ be a semi-trajectory of (4.3). Let θ be large enough. Let $\hat{x}(0)$, $S(0)$ (positive definite) be initial conditions for (4.4). Let $\hat{x}(t)$, $S(t)$ be the corresponding semi-trajectories ($t \geq 0$). Then:*

a. $S(t)$ remains in a compact subset of the cone of positive definite symmetric matrices,

b. for all $\beta > 0$, there is a θ (large), such that:

$$||\hat{x}(t) - x(t)|| \leq H(S(0))k(\beta)e^{-\beta t}||\hat{x}(0) - x(0)||,$$

where H is a smooth mapping, and k is a function with polynomial growth.

Again, the exponential against the polynomial ensure that the estimation error, $||\hat{x}(t) - x(t)||$ can be made arbitrarily small in arbitrary short time.

4.1.3. *A mixed solution.* We present here the version of the extended Kalman filter proposed in [3], that mixes the standard extended Kalman filter and the high-gain extended Kalman filter of the previous section.

We obtain an observer with the following properties:

a) in presence of big disturbances (state jumps), it has the high-gain behavior: the estimation error can be made arbitrarily small very quickly;

b) when the error is small, it behaves as the regular extended Kalman filter (good performances w.r.t. noise).

This observer behaves extremely well in practice, as we will show in Section 5 on a non-academic example, and it can be also used for identification, as explained in section 4.2.

We consider exactly the same class of systems on $\mathbb{R}^n$, globally in normal form (4.3), as in the previous section 4.1.2, and meeting assumption **B** of this section. Again, the compact support assumption for b, if not realized, can be obtained by smoothly modifying b out of a large compact subset of X.

Then, the equations of the observer are:

$$\begin{aligned} \frac{dS}{dt} &= -(A + D_{\hat{x}}b(\hat{x}, u))'S - S(A + D_{\hat{x}}b(\hat{x}, u)) + C'r^{-1}C - SQ_{\theta}S, \\ \frac{d\hat{x}}{dt} &= A\hat{x} + b(\hat{x}, u) - S^{-1}C'r^{-1}(C\hat{x} - y), \\ \frac{d\theta}{dt} &= \lambda(1 - \theta). \end{aligned} \tag{4.5}$$

Now, θ is not constant anymore, hence we have an initial condition θ_0, and λ is a positive real, to be chosen not too large as we shall see.

Intuitively, this observer should behave like that:

a. When $t \to +\infty$, $\theta \to 1$, and then, the behavior of the observer for large t is the same as the one of the usual extended Kalman filter (known as good w.r.t. noise in practice [1]).

b. On the contrary, if θ_0 is large, when t is small, the behavior is the one of the high-gain EKF: the error can be made arbitrarily small in arbitrary short time.

This is true, as stated in the following theorem:

Theorem 18. *1. For all $0 \leq \lambda \leq \lambda_0$, ($\lambda_0$ a certain positive real), for all θ_0 large enough, (depending on λ), for all $S_0 \geq c\ Id$, for all $\bar{K} \subset \mathbb{R}^n$, $\bar{K}$ a compact subset, for all $\varepsilon_0 = \hat{x}_0 - x_0$, $\varepsilon_0 \in \bar{K}$, the following estimation holds, for all $\tau \geq 0$.*

Long term estimation:

$$\begin{aligned} ||\varepsilon(\tau)||^2 &\leq R(\lambda, c)e^{-a\ \tau}||\varepsilon_0||^2\Lambda(\theta_0, \tau, \lambda), \\ \Lambda(\theta_0, \tau, \lambda), &= {\theta_0}^{2(n-1)+\frac{a}{\lambda}}e^{-\frac{a}{\lambda}\theta_0(1-e^{-\lambda\tau})}, \end{aligned} \tag{4.6}$$

where $a > 0$. $R(\lambda, c)$ is a decreasing function of c.

2. Set $\theta(T) = 1+(\theta_0-1)e^{-\lambda T}$. For all $T > 0$, $\tau \leq T$, for all $\theta_0 \geq \bar{\theta}_0$, $\bar{\theta}_0$ a certain positive real depending on λT, for a_1, a_2 certain positive constants, and with $H(c)$, a certain decreasing positive function of c, where $S(0) \geq c\ Id$, the following estimation holds.

Short term estimation:

$$||\varepsilon(\tau)||^2 \leq \theta(\tau)^{2(n-1)}H(c)e^{-(a_1\theta(T)-a_2)\tau}||\varepsilon(0)||^2. \tag{4.7}$$

This theorem has been proved in [3].

Comments.

a. Note that the function $\Lambda(\theta_0, \tau, \lambda)$ is a decreasing function of τ, and that, for all $\tau > 0$, $\lambda > 0$, $\Lambda(\theta_0, \tau, \lambda)$ can be made arbitrarily small, increasing θ_0.

b. This means that, provided that λ is smaller than a certain constant λ_0, and θ_0 is large in front of λ, the estimation error goes exponentially to zero, and can be made arbitrarily small in arbitrary short time.

[1] No more than in practice: there is no theoretical result (but negative) on the stochastic behavior of the EKF, even for small noise. See [13].

c. The asymptotic behavior of the observer is the one of the extended Kalman filter,
d. The "short term behavior" is the one of the "high gain extended Kalman filter".

Practical implementation:
The problem with this observer is that its behavior is "time dependant": for small time, it is the high-gain EKF, and for large time it is the ordinary EKF.

A way to remedy this evil is to proceed as follows.

We consider a one parameter family $\{O_\tau, \tau \geq 0\}$ of observers of type (4.5), indexed by the time, each of them starting from S_0, θ_0, at the current time τ. In fact, in practice, it will be sufficient to consider, at current time t, a slipping window of time, $[t-T, t[$, and a finite set of observers $\{O_{t_i},\, t-T \leq t_i \leq t\}$, with $t_i =$ largest multiple of $\frac{T}{N}$ smaller than $t-(i-1)\frac{T}{N}$, $i = 1, ..., N$.

As usual, we call the term $I(\tau) = \hat{y}(\tau) - y(\tau)$, (the difference at time τ between the estimate output and the real output), the **"innovation".** Here, for each observer O_{t_i},we have an innovation $I_{t_i}(\tau)$.

Our suggestion (very natural and very simple), is to take as the estimate of the state, the estimation given by the observer O_{t_i} that minimizes a norm of the innovation.

Let us analyze what will be the effect of this procedure in a deterministic setting:

1. Let us assume that there is no "jump" of the state. Then, clearly, the best estimation will be given by the "oldest" observer in the window, O_{t_N}. Then, the error will be given by the "long term" and "short term" estimates at time T :

$$||\varepsilon(\tau+T)||^2 \leq R(\lambda, c)e^{-a\ T}||\varepsilon(\tau)||^2\Lambda(\theta_0, T, \lambda),$$
$$||\varepsilon(\tau+T)||^2 \leq \theta(T)^{2(n-1)}H(c)e^{-(a_1\theta(T)-a_2)T}||\varepsilon(\tau)||^2.$$

a. If T is large enough, the asymptotic behavior will be the one of the "extended Kalman filter".
b. At the beginning, the transient is the one of the HGEKF.
c. the error can be made arbitrarily small in arbitrary short time, provided that θ_0 is large enough.

2. If at a certain time we have a "jump" of the state, then, the innovation of the "old observers" will become large. The "youngest" one will be chosen, and the transient will be the same as the one of the HGEKF, first, and of the EKF, after T.

4.1.4. *The case of the uniform infinitesimal observability normal form (2.3).* In fact, the observer presented in the previous section can be also applied to uniformly infinitesimally observable systems, under the general normal form (2.3, 2.4), provided that the control functions $u(.)$ are (globally on $[0, +\infty[$) Lipschitz functions of the time. We will prove this now.

A first remark is that systems under this normal form (2.3, 2.4) are strongly differentially observable of order n (this is easily checked). A consequence is that they have a phase variable representation of order n (of type 2.1). This is a way to apply the results of the previous section, but this method involves complicated changes of coordinates, and above all, it uses the derivatives of the control $\dot{u}, ..., u^{(n)\cdot}$.

In fact, there is a way to transform systems in normal form (2.3, 2.4) into systems in normal form (4.3), by using only $\dot{u}$, the first derivative of the control.

Consider such a system Σ on $\mathbb{R}^n$, and set:

$$z = \Phi_u(x) = (h(x,u), L_f h(x,u), ..., L_f^{n-1}h(x,u)). \tag{4.8}$$

Let $K \subset \mathbb{R}^n$ be any fixed open relatively compact subset. As previously in this paper, we deal with semi-trajectories of Σ that remain in K, only. It follows from (2.4) that, for all $u \in U$, Φ_u is an injective immersion (this is easily checked by induction on the components of Φ_u). Therefore, Φ_u is a u-dependent diffeomorphism from K onto its image. Consider the image Σ' of the system $\Sigma_{|K}$ (Σ restricted to K) by the time dependant diffeomorphism Φ_u. It is of the form:

$$\begin{aligned} y &= z_1, \\ \dot{z} &= F(z,u) + G(z,u)\dot{u}, \end{aligned} \tag{4.9}$$

and moreover, it is in fact in the form (4.3):

$$(4.10)\qquad \begin{aligned} y &= z_1, \\ \dot{z} &= Az + \bar{G}(z, u, \dot{u}), \end{aligned}$$

where A is the antishift matrix, and where $\bar{G}$ is smooth and depends in a triangular way of z : $\bar{G}_1(x_1, u, \dot{u}), ...$

There is the very small difficulty that $\bar{G}$ is not defined on the whole of $\mathbb{R}^n \times U \times \mathbb{R}^p$, and then, has to be smoothly prolonged to $\mathbb{R}^n \times U \times \Omega$, where Ω is a compact subset of $\mathbb{R}^p$ (Ω : set of values of $\dot{u}$, which we may take compact as will be justified below). Moreover, this prolongation can be taken with compact support, in order to meet assumption **B1** of the two previous sections.

Consider any semi-trajectory $x(.)$: $[0, T[$ of Σ (possibly $T = +\infty$), corresponding to a Lipschitz control u, with Lipschitz constant K_u. Then u is absolutely continuous, and its derivative is bounded by K_u (which explains that we may take Ω compact: $\dot{u}$ is bounded by the Lipschitz constant of u). Consider $Z(t) = \Phi_{u(t)}(x(t))$.

Then, since x, u are absolutely continuous, $Z(.)$ is also absolutely continuous. But, by construction, $Z(t)$ satisfies almost everywhere (4.10) and (4.9), provided that $x(t)$ remains in K, which we assume. Then, $Z(.)$ is the unique absolutely continuous solution of (4.10), corresponding to the controls $u, \dot{u}$, and the initial condition $z_0 = \Phi_{u(0)}(x_0)$.

We have shown the following:

Lemma 3. *For Lipschitz controls, (with given Lipschitz constant), semi-trajectories of* Σ *that remain in* K *are mapped by* Φ_u *in the semi-trajectories of the systems (4.9, 4.10).*

Then, at the price of using the first derivative of u, the EKF and the "mixed" observer of the previous section 4.1.3, can be used for general uniformly infinitesimally observable systems.

This observation, in fact, comes from the paper [7] (where it is not stated precisely, but the idea is present).

4.2. **Identifier design.** In this section, we will remain in the context of Section 3, i.e. systems without controls. Nevertheless, the application considered in Section 5 will concern systems with controls, but the effect of the controls will be transparent, since the systems are in the normal forms studied previously (but control-dependent).

The following simple single-output example shows that there is no chance to avoid approximate differentiation for effective identification:

$$(4.11)\qquad \begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2, .., \dot{x}_{n-1} = x_n, \\ \dot{x}_n &= \varphi(x). \end{aligned}$$

In fact, in this example, identifying (i.e. **reconstructing the piece of the graph of** φ **visited during the experiment**) is just **equivalent** to differentiate the output n times.

Also, we will not consider the generic case $m \geq 3$, but only the cases corresponding to $m = 1$ (normal form 3.2) and $m = 2$, systems of type 1,2,3.

Our basic idea is the same in all cases, and leads to **the use of the nonlinear observers** developed in the previous section 4.1: we assume, along the trajectories visited, **a local model for** φ. For instance, a simple local model is: $\varphi^{(k)} = 0$.

This does not mean, at the end, that we will identify φ as a polynomial in t: the question is not that this polynomial models the function φ globally as a function of t, but only locally, on reasonable time intervals (reasonable w.r.t. the performances of the observer that we will use).

4.2.1. *The single output case.* Let us consider a system Σ in the identifiability normal form 3.2. Adding the local model for φ, we get the system:

$$
\begin{aligned}
&y = x_1, \\
&\dot{x}_1 = x_2, ..., \dot{x}_{n-1} = x_n, \\
&\dot{x}_n = \Psi(x, \varphi_1), \dot{\varphi}_1 = \varphi_2, ..., \dot{\varphi}_{k-1} = \varphi_k, \dot{\varphi}_k = 0,
\end{aligned} \tag{4.12}
$$

$$
\frac{\partial \Psi}{\partial \varphi_1} \neq 0 \ \ \text{(never vanishes)}. \tag{4.13}
$$

This is a system on $\mathbb{R}^{n+k}$, which is not controlled (however, for the considerations that follow, Ψ could depend on a control u), and this system is under the normal form (2.3, 2.4).

Therefore, we may apply the high gain Luenberger observer, or we may apply the trick of the previous section 4.1.4. Then, for instance, the observer of Section 4.1.3 may be applied to this system. It will provide estimations of $x(t), \varphi(t)$, that is, just an estimation of the piece of the graph of φ visited during the experiment. (It provides also estimations of $\dot{\varphi}, ..., \varphi^{(k)}$, which we don't care about).

4.2.2. *The two-output case.* The cases of normal forms (3.3), (3.4), (3.5), corresponding to Type 1 to 3 systems can be treated in a similar way to the single-output case, with some more or less easy adaptations of the methods of the previous sections.

Let us just consider one example: the case of a Type 2 system, with $r = 0$, which is very illustrative, and goes directly back to the observation problem for uniformly infinitesimally observable systems:

$$
\begin{aligned}
y_1 &= \Phi(x, \varphi), \quad y_2 = h(x_1, \Phi(x, \varphi)), \\
\dot{x}_1 &= F_1(x_1, x_2, \Phi(x, \varphi)), \\
&\ .. \\
\dot{x}_{n-1} &= F_{n-1}(x, \Phi(x, \varphi)), \\
\dot{x}_n &= F_n(x, \varphi),
\end{aligned} \tag{4.14}
$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_1}{\partial x_2} \neq 0, ...,\frac{\partial F_{n-1}}{\partial x_n} \neq 0, \frac{\partial h}{\partial x_1} \neq 0$.

In that case, Let us set $\Phi(x, \varphi) = \Phi(t)$ $(= y_1(t))$. Then, forgetting about y_1, and since $\Phi_x(\varphi)$ is an invertible function of φ for x fixed, we have the system:

$$
\begin{aligned}
y_2 &= h(x_1, \Phi(t)) \\
\dot{x}_1 &= F_1(x_1, x_2, \Phi(t)), \\
&\ .. \\
\dot{x}_{n-1} &= F_{n-1}(x, \Phi(t)), \\
\dot{x}_n &= F_n(x, \Phi_x^{-1}(\Phi(t))).
\end{aligned} \tag{4.15}
$$

The function $\Phi(t)$ being known, the identification problem is just the observation problem for this new system, which is in uniform infinitesimal observability normal form: having an estimate $\hat{x}(t)$ of the state, we get an estimate of $\varphi(t)$ by $\hat{\varphi}(t) = \Phi_{\hat{x}(t)}^{-1}(y_1(t))$.

Then, we may apply the high-gain Luenberger observer (4.2), or at the price of a single (approximate) differentiation of y_1, we may apply again the observer of section 4.1.3.

5. Application to a fluid catalytic cracker

A fluid catalytic cracker (FCC) is a process used in refineries to produce gasoline from heavy petroleum residues.

The well known model of FCC used in this paper is adapted from [9]. It has been used by several authors as an example of a highly nonlinear system with a lot of strong interactions, see for instance [10, 11]. This model has already been used by the first author, to study the performances of high gain observers, in [1].

A FCC unit, as depicted on Figure 1, is composed of a reactor and a regenerator. The heavy petroleum residues feed the FCC by the bottom of the reactor (called the riser). Long molecules are broken thanks to a catalyst, that circulates between the reactor and the regenerator. When the long molecules are broken, carbon is produced (coke), and is fixed on the catalyst, that becomes dirty. In the regenerator, the coke on the catalyst is burned, regenerating the catalyst. The cracking reactions inside the riser and the reactor are endothermic, and use the heat produced when regenerating the catalyst. This "heat flowrate" is driven from the regenerator to the riser of the reactor, by the flow of regenerated catalyst.

Therefore it is clear that the thermal balance between the reactor and the regenerator results in a strong coupling between these two parts, that affects crucially the behavior of the unit.

5.1. **Description of the model, and purpose of the study.** Our model is closely related to the Kurihara model [9], as described in [11] for dynamic optimization. The Kurihara FCC model has also been used in [1] in order to estimate the "carbon Conradson factor" F_{cf}, an important parameter from the point of view of operation: it characterizes in some sense the propensity of the catalyst to become inefficient. In particular, it becomes worse when the catalyst is old, and it is not very well known. **Our first objective here, will be also the estimation of this parameter F_{cf}.**

But here, we will use less measurements than in [1]: indeed, we will only use temperature measurements, from both the reactor and the regenerator. We will not use the measurement of the carbon concentration in the regenerator, which was supposed to be reconstructed in [1] thanks to a measurement of the concentration of oxygen in flue gas.

Another problem is that the local model of oxygen combustion in the regenerator is not very accurate (may be not very well known, in fact). Therefore, the purpose of the second part of this study will be to **identify this model.**

The control of the FCC unit is performed via **two control variables**, namely the "air flowrate" (R_{ai}) (air of combustion of the coke on the catalyst, inside the regenerator) and the "catalyst circulation rate" (R_c).

The Kurihara model is really a two time-scales model: the evolutions of the rate of carbon in the reactor and in the regenerator are modelled by three differential equations. However, one of these equations (modelling the catalytic carbon balance) has a very short time-constant with respect to other dynamics inside the system. Here, we have replaced this differential equation by an algebraic one. Comparisons of solutions between both models lead to very similar results.

Another simplification that we make is the following: We assume steady-state for the catalyst flowrates between the reactor and the regenerator. These two simplifications are analyzed from a chemical-engineering point of view in [10]. In this paper ([10]), authors use the model to illustrate the fact that the FCC may admit multiple equilibria.

In fact, these two modifications of the Kurihara FCC model and the local model of oxygen burning in the regenerator are the main differences between [9] and [10] FCC models.

As we shall see, in the case of temperature measurements only, this simplified model is observable, even if F_{cf} is considered as an unknown parameter, and added to the system as a constant state variable (indeed, this parameter varies very slowly: it represents the long-term degradation of the catalyst, as we said).

Finally, the model we consider consists of a set of four differential equations representing the evolution of the reactor temperature (5.1), regenerator temperature (5.8), carbon concentration on spent catalyst (5.4) and carbon concentration on regenerated catalyst (5.13).

5.1.1. *Reactor model.*
Temperature in the reactor:

$$\begin{aligned} S_c H_{ra} \dot{T}_{ra} = {} & S_c R_c \left(T_{rg} - T_{ra}\right) + S_{tf} R_{tf} \left(T_{tf} - T_{ra}\right) \\ & - \Delta H_{fv} R_{tf} - \Delta H_{cr} R_{tf} C_{tf} \end{aligned} \tag{5.1}$$

$$C_{tf} = \frac{R_{cr}}{R_{cr} + R_{tf}} \tag{5.2}$$

$$R_{cr} = \frac{\sqrt{k_{cr} R_c P_{ra} H_{ra}}}{10\, C_{rc}^{0.12}} \exp\left(-\frac{1}{2}\frac{A_{cr}}{R T_{ra}}\right), \tag{5.3}$$

with:

- Reactor operating conditions $T_{ra}|_{t=0} = 775\,\mathrm{K}$, $H_{ra} = 1.85\,10^{-4}\,\mathrm{kg}$, $P_{ra} = 211.7\,\mathrm{kPa}$,
- Feed properties $R_{tf} = 41\,\mathrm{kg}\,/\,\mathrm{s}$, $T_{tf} = 492.8\,\mathrm{K}$, $S_{tf} = 3140\,\mathrm{J}\,/\,(\mathrm{kg}\,.\,\mathrm{K})$,
- Catalyst recirculation $R_c|_{t=0} = 290\,\mathrm{kg}\,/\,\mathrm{s}$, $0 < R_c^{\min} \leq R_c \leq R_c^{\max}$, $S_c = 1047\,\mathrm{J}\,/\,(\mathrm{kg}\,.\,\mathrm{K})$,
- Heat constants $\Delta H_{cr} = 4.65\,10^5\,\mathrm{J}\,/\,\mathrm{kg}$, $\Delta H_{fv} = 1.74\,10^5\,\mathrm{J}\,/\,\mathrm{kg}$, $\Delta H_{rg} = 3.02\,10^7\,\mathrm{J}\,/\,\mathrm{kg}$,

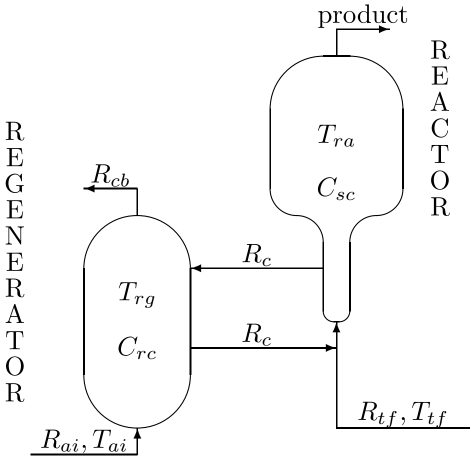


Figure 1. FCC Unit

- $k_{cr} = 25.96\,\mathrm{kPa}^{-1}\,\mathrm{s}^{-1}$, $A_{cr} = 83.8\,10^3\,\mathrm{J}\,/\,\mathrm{mol}$
- $R = 8.314\,\mathrm{J}\,/\,(\mathrm{mol}\,.\,\mathrm{K})$

Carbon concentration on spent catalyst in the reactor:

$$H_{ra}\dot{C}_{sc} = R_c\,(C_{rc} - C_{sc}) + 100\,R_{cf} \tag{5.4}$$

$$R_{cf} = R_{cc} + R_{ad} \tag{5.5}$$

$$R_{cc} = \frac{\sqrt{k_{cc}R_cP_{ra}H_{ra}}}{10\,C_{rc}^{0.03}}\exp\left(-\frac{1}{2}\frac{A_{cc}}{RT_{ra}}\right) \tag{5.6}$$

$$R_{ad} = F_{cf}R_{tf} \tag{5.7}$$

with:

- $C_{sc}|_{t=0} = 1.2$
- $k_{cc} = 2.66\,10^{-4}\,\mathrm{kPa}^{-1}\,\mathrm{s}^{-1}$, $A_{cc} = 4.18\,10^4\,\mathrm{J}\,/\,\mathrm{mol}$

5.1.2. *Regenerator model.*

Temperature in the regenerator:

$$S_cH_{rg}\dot{T}_{rg} = S_cR_c\,(T_{ra} - T_{rg}) + S_{ai}R_{ai}\,(T_{ai} - T_{rg}) + \Delta H_{rg}R_{cb} \tag{5.8}$$

$$R_{cb} = \frac{R_{ai}}{242}\,(21 - O_{fg}) \tag{5.9}$$

$$O_{fg} = 21\exp\left(\frac{-\frac{P_{rg}H_{rg}}{R_{ai}}}{\frac{1}{K_{od}} + \frac{1}{K_{or}C_{rc}}}\right) \tag{5.10}$$

$$K_{od} = 6.34\,10^{-9}R_{ai}^2 \tag{5.11}$$

$$K_{or} = 1.16\,10^{-5}\exp\left(\frac{A_{or}}{R\left(\frac{1}{866.7} - \frac{1}{T_{rg}}\right)}\right) \tag{5.12}$$

with:

- Regenerator operating conditions $T_{rg}|_{t=0} = 943\,\mathrm{K}$, $H_{rg} = 1.53\,10^5\,\mathrm{kg}$, $P_{rg} = 254.4\,\mathrm{kPa}$,
- Air properties $R_{ai}|_{t=0} = 26\,\mathrm{kg}\,/\,\mathrm{s}$, $0 < R_{ai}^{\min} \leq R_{ai} \leq R_{ai}^{\max}$, $T_{ai} = 394\,\mathrm{K}$, $S_{ai} = 1130\,\mathrm{J}\,/\,(\mathrm{kg}\,.\,\mathrm{K})$
- $A_{or} = 1.47\,10^5\,\mathrm{J}\,/\,\mathrm{mol}$

Carbon concentration on regenerated catalyst in the regenerator:

$$H_{rg}\dot{C}_{rc} = R_c\,(C_{sc} - C_{rc}) - 100\,R_{cb} \tag{5.13}$$

with $C_{rc}|_{t=0} = 0.3$

5.2. **Estimation of the Carbon Conradson factor.** Here, we will use this model to estimate unmeasured state variables. We will first study the observability, and, after positive answer to this question, we will apply the observer construction described above. Moreover, in the next section, we will assume that a part of the model is unknown and we will show that the unknown function is identifiable. Therefore the system can be transformed to a certain observability canonical form, similar to the type 2 normal form 3.4.

In fact, it can also be put under a form such that our "mixed" observer construction of Section 4.1.3, can be applied in order to estimate simultaneously both the state variables and unknown function.

However, this section is devoted to observation only: we will assume that our knowledge-based model is perfectly known and we will use it to **estimate unknown state variables** (C_{rc}, C_{sc} and F_{cf}) thanks to **temperature measurements** (T_{ra} and T_{rg}).

5.2.1. *Observer construction.* First of all, an elementary analysis shows that the system is observable and infinitesimally observable: Indeed, since T_{ra} is measured, the derivative $\dot{T}_{ra}$ allows to compute C_{rc} from C_{tf}. Then $\dot{C}_{rc}$ gives C_{sc}. Finally, $\dot{C}_{sc}$ gives R_{cf} and hence F_{cf}. Moreover, as soon as the function $C_{tf} \mapsto C_{rc}$ is one to one (the other variables being fixed), these considerations are global, and the system is globally observable (using only the fact that each physical parameter is positive). If this function $C_{tf} \mapsto C_{rc}$ has nonvanishing derivative, we get global observability (on the physical domain where state variables are positive), plus **uniform infinitesimal observability**.

Notice that the **measurement** T_{rg} **is not used** here. We will use it in the next section in order to **identify the unknown function.**

This analysis of infinitesimal observability shows us that the system may already be written under the (single output, as we said) observability canonical form 2.3, 2.4: by the theory, this can be done at least locally (see [6] and Section 2.3).

Nevertheless, in order to apply our mixed high-gain extended Kalman filter construction 4.1.3, we have to apply the trick presented in Section 4.1.4. To do this, we have to find a coordinate-change, such that the system in new coordinates is under the canonical form 4.10, as explained in Section 4.1.4. To do this, we will consider successive time derivatives of the output, and at each step, we will obtain a new coordinate, corresponding to a new variable obtained by derivation.

Let us consider first the reactor temperature. Its time derivative is given by the right hand side of 5.1. Since both T_{ra} and T_{rg} are outputs, the new information about the state, provided by $\dot{T}_{ra}$, is the value of C_{tf} which is a function of T_{ra} and C_{rc}. Let us observe that:

$$C_{tf} = \frac{R_{cr}}{R_{cr} + R_{tf}} = 1 - R_{tf} \frac{1}{R_{cr} + R_{tf}},$$

so that,

$$\dot{T}_{ra} = \frac{\Delta H_{cr} R_{tf}^2}{S_c H_{ra}} \frac{1}{R_{cr} + R_{tf}} + \frac{1}{S_c H_{ra}} \left(S_c R_c \left(T_{rg} - T_{ra}\right)\right.$$
$$\left. + S_{tf} R_{tf} \left(T_{tf} - T_{ra}\right) - \left(\Delta H_{fv} + \Delta H_{cr}\right) R_{tf}\right)$$

Denoting the two measured state variables $x_1 = T_{ra}$ and $x_5 = T_{rg}$, the first control variable $u_1 = R_c$ and defining the new state variable $x_2 = \frac{1}{R_{cr}+R_{tf}}$ then

$$\dot{x}_1 = a_2 x_2 + g_1 \left(x_1, T_{rg}, u_1\right) \tag{5.14}$$

with $a_2 = \frac{1}{S_c H_{ra}} \Delta H_{cr} R_{tf}^2$.

It is clear that (other variables being fixed), the function $C_{rc} \mapsto R_{cr} \mapsto \frac{1}{R_{tf}+R_{cr}}$ is a diffeomorphism from any open interval $]\varepsilon, +\infty[$, $\varepsilon > 0$, to its image. Practically, C_{rc} is the concentration of carbon on spent catalyst hence it is a positive variable, and it can be assumed to have a strictly positive lower bound.

In fact, there is more than that. We have **mathematical coherence** of the model with this property, in the sense that the domains $\{C_{rc} > \varepsilon\}$, are **positively invariant under the dynamics**. Let us check this property now.

Proposition 1. *Certain domains* $\{C_{rc} > \varepsilon\}$, *for* ε *small, are positively invariant*

Proof. Let us consider both temperatures. Let us assume that $T_{ra} = 273\,\mathrm{K}$ and $T_{rg} \geq 273\,\mathrm{K}$ then using (5.1),

$$\begin{aligned}\frac{1}{R_{tf}} S_c H_{ra} \dot{T}_{ra} &\geq S_{tf}\left(T_{tf} - 273\right) - \left(\Delta H_{fv} + \Delta H_{cr}\right)\\ &= 3140\,(492.8 - 273) - 1.74\,10^5 - 4.65\,10^5 > 0\end{aligned}$$

and also if $T_{ra} \geq 273\,\mathrm{K}$ and $T_{rg} = 273\,\mathrm{K}$ then using (5.8),

$$S_c H_{rg} \dot{T}_{rg} \geq S_a R_{ai}\left(T_{ai} - 273\right) > 0.$$

Hence T_{ra} and T_{rg} are bounded from below by $273\,\mathrm{K}$. Then we will prove that there exist ε_1 such that $C_{cs} \geq C_{rc} + \varepsilon_1$. Using (5.4,5.13),

$$\frac{d}{dt}\left(C_{sc} - C_{rc}\right) = -R_c\left(\frac{1}{H_{ra}} + \frac{1}{H_{rg}}\right)\left(C_{sc} - C_{rc}\right) + \frac{100}{H_{ra}} R_{cf} + \frac{100}{H_{rg}} R_{cb}$$

but $R_{cf} = R_{cc} + R_{ad}$ and since $0 < R_c^{\min} \leq R_c \leq R_c^{\max}$, R_{cc} can be bounded from below by a positive decreasing function of C_{rc} (recall that $T_{ra} \geq 273\,\mathrm{K}$). Moreover, R_{cb} can be bounded from below by a positive increasing function of C_{rc} (using $T_{rg} \geq 273\,\mathrm{K}$ and $0 < R_{ai}^{\min} \leq R_{ai} \leq R_{ai}^{\max}$ in (5.9) to (5.12)). Therefore, there exist ε_1 such that $\frac{100}{H_{ra}} R_{cf} + \frac{100}{H_{rg}} R_{cb} \geq \varepsilon_1 R_c^{\max}\left(\frac{1}{H_{ra}} + \frac{1}{H_{rg}}\right)$ hence

$$\frac{d}{dt}\left(C_{sc} - C_{rc} - \varepsilon_1\right) \geq -R_c\left(\frac{1}{H_{ra}} + \frac{1}{H_{rg}}\right)\left(C_{sc} - C_{rc} - \varepsilon_1\right)$$

We can chose ε_1 such that $C_{sc} - C_{rc}|_{t=0} \geq \varepsilon_1$ and so $C_{sc} - C_{rc} \geq \varepsilon_1$ along the trajectory.

Finally

$$\begin{aligned}H_{rg}\dot{C}_{rc} &= R_c\left(C_{cs} - C_{rc}\right) - 100\,R_{cb}\\ &\geq R_c^{\min}\varepsilon_1 - 100\,R_{cb}\end{aligned}$$

and since $R_{cb} \overset{C_{rc}\to 0}{\longrightarrow} 0$ there exist ε_2 such that $R_{cb} \leq \frac{1}{100} R_c^{\min}\varepsilon_1$ if $C_{rc} \leq \varepsilon_2$ hence $H_{rg}\dot{C}_{rc} \geq 0$ for $C_{rc} = \varepsilon_2$ and therefore, C_{rc} is bounded from below by ε_2. □

Due to this analysis, the diffeomorphism $C_{rc} \to x_2$ can be smoothly prolonged to all of $\mathbb{R}$ without changing the trajectories on the physical domain. This property could be important, in order to reach the assumption B that is needed for the construction of our observer (Section 4.1.2). In fact, in practice, we observe in simulations that our estimations of C_{rc} never vanish. Hence, we just don't make any prolongation, and in fact, we don't use this property, which nevertheless is crucial from the theoretical point of view.

Let us remark that our change of variable already depends on the control variable R_c since $x_2 = \frac{1}{R_{cr}+R_{tf}}$ and R_{cr} depends explicitly on R_c. Then, as expected, the first derivative of the control will appear after the coordinate-change.

We have to calculate the derivative of x_2 with respect to time:

$$\begin{aligned}\dot{x}_2 &= \frac{1}{\left(R_{cr} + R_{tf}\right)^2}\left(-\dot{R}_{cr}\right)\\ &= -x_2^2\left(\frac{dR_{cr}}{dC_{rc}}\dot{C}_{rc} + \frac{dR_{cr}}{dT_{ra}}\dot{T}_{ra} + \frac{dR_{cr}}{dR_c}\dot{R}_c\right)\\ &= -x_2^2\,(-0.12)\,\frac{R_{cr}}{C_{rc}}\frac{R_c}{H_{rg}} C_{sc} + g_2\left(x_1, x_2, x_5, u_1, \dot{u}_1, u_2\right)\\ &= a_3 x_3 + g_2\left(x_1, x_2, x_5, u_1, \dot{u}_1, u_2\right), \qquad (5.15)\end{aligned}$$

with $x_3 = x_2^2 \dfrac{R_{cr}}{C_{rc}} R_c C_{sc}$ and $a_3 = \dfrac{0.12}{H_{rg}}$.

To finish, differentiating once more, we obtain:

$$\begin{aligned}\dot{x}_3 &= \frac{x_3}{C_{sc}}\frac{100\,R_{tf} F_{cf}}{H_{ra}} + g_3\left(x_1, x_2, x_3, x_5, u, \dot{u}\right)\\ &= a_4 x_4 + g_3\left(x_1, x_2, x_3, x_5, u, \dot{u}\right), \qquad (5.16)\end{aligned}$$

where $x_4 = x_3 \dfrac{F_{cf}}{C_{sc}} = x_2^2 \dfrac{R_{cr}}{C_{rc}} R_c F_{cf}$ and $a_4 = \dfrac{100\,R_{tf}}{H_{ra}}$.

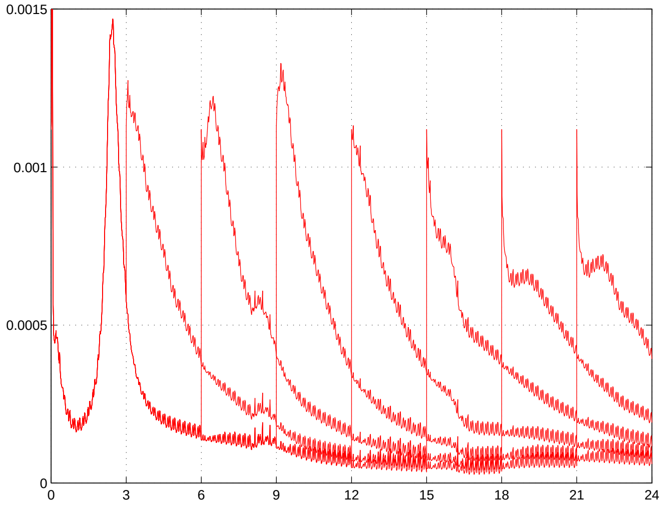


FIGURE 2. Euclidian norm of the gain

Finally, our (control depending) change of coordinates is

$$\psi_u\left(T_{ra}, C_{rc}, C_{sc}, F_{cf}, T_{rg}\right) = \left(x_1, x_2, x_3, x_4, x_5\right)$$

Here u denotes the control variables $u = (R_c, R_{ai})$ but ψ_u does not depend explicitly on R_{ai}.

Again, a certain number of "theoretical precautionary measures" have to be taken: it is easy to check that ψ_u is a smooth function in the interior of the physical domain (positive variables). Therefore, it would be possible to prolong ψ outside this domain in order that it becomes (smoothly) everywhere defined. In practice, we don't do this: we do not prolong ψ_u but we keep in mind that our simulation results are justified only if the state variables remain in the physical domain.

The only problem that may occur -and that occurs in practice- concerns the fact that, **temporarily**, **the estimations of C_{rc}** may vanish or become negative, which may have unpleasant consequences in the other equations. To palliate this difficulty, we just introduce a smooth cut-off function χ_ε such that χ_ε is one to one from $\mathbb{R}$ into $\left]\frac{\varepsilon}{2}, +\infty\right[$ and $\chi_\varepsilon(z) = z$ if $z > \varepsilon$. Then we replace C_{rc} in (5.3) by $\chi_\varepsilon(C_{rc})$ with ε small enough. This is just an artificial way to correct irrelevant estimations of C_{rc}.

Finally, our system is equivalent up to a diffeomorphism to the following system

$$\begin{array}{rcl} y_1 & = & x_1, \quad y_2 = x_5 \\ \dot{x}_1 & = & a_2 x_2 + g_1(x_1, x_5, u) \\ \dot{x}_2 & = & a_3 x_3 + g_2(x_1, x_2, x_5, u, \dot{u}) \\ \dot{x}_3 & = & a_4 x_4 + g_3(x_1, x_2, x_3, x_5, u, \dot{u}) \\ \dot{x}_4 & = & g_4(x) \end{array}$$

Then we may apply our "mixed" extended Kalman filter to this system. To do this, we don't care about the theoretical bounds of the tuning parameters θ, λ that come from the theory. We just tune these parameters in order to obtain reasonable practical performances: increasing θ results in better performances in high-gain mode, decreasing λ makes the filtering mode (performances w.r.t. noise) be good a long time after the occurrence of large perturbations.

5.2.2. *Tuning of parameters.* As we explained, we want to build an observer mixing the good properties of the high-gain EKF with respect to large unmodelled disturbances and the good properties of the classical EKF with respect to noise. In order to achieve this goal, we have to tune each parameter very carefully. let us now explain how to achieve this goal.

(1) As a first step, we just use a single classical EKF (that is $\theta = 1$) and we tune Q and r – in (4.5) – in order to obtain best possible performances with respect to measurement noise. During this first step, we do not simulate disturbances and, of course, we initialize our EKF at the right value of the state. Nevertheless, we choose Q and r such that the EKF reaches also good performances when θ is slightly larger than 1 (for instance $\theta \approx 2$). As a consequence, when several observers

will be working together, a number of them will reach good performances with respect to noise (those among them for which θ will be close to 1), similar to performances of a classical EKF.

(2) The second step is to tune the high-gain EKF. We use the same matrices Q and r as in the first step and we use θ_0 and λ to achieve our purposes, that is fast convergence. We have simulated several "physical" disturbances. Since the rate of convergence might be theoretically arbitrary, we have simulated a very small noise, and asked for a – fast but – reasonable convergence. During this step, we keep in mind that the high-gain EKF should become a classical EKF as fast as possible. Therefore, despite the fact that λ should be small enough to ensure exponential convergence, it should not be too small. The price to pay for a too small λ will be **a large number of observers**. As in the first step, we check that the performances are reasonable even for slightly lower values than θ_0. We will denote by θ_1 the minimal value of θ which ensures good performance in presence of disturbances.

(3) The last step consists of using several observers as explained in Section 4.1.3. Since at each current time, we want to have at least an observer working in "high-gain mode", and another one working in "filtering mode", the number of observers will depend on the values obtained for θ_0, θ_1 and λ. The time between two consecutive initializations of an observer will be the time necessary for $\theta\left(t\right)$ to reach θ_1 starting from θ_0 and satisfying $\dot{\theta} = \lambda\left(1-\theta\right)$. The number of observers will be high enough in order to be sure that at any time, there exist observers with a current value of θ almost equal to 1 (at least less than 2, according to first step). It can be useful to plot informations concerning the actual gain value versus the time in order to check that the gain is actually high enough when θ is high.

Indeed, since the gain is obtained as the solution of a Ricatti equation, and since this equation itself depends on an exponentially converging parameter, it is not at all obvious to get intuition of how the gain is varying. We decided to plot the Euclidian norm of the correction applied to the state, (which is equal to the correction gain times a normalized innovation (equal to $1\,\mathrm{K}$)), see figure 2. Here, the ratio between high gain and non-high-gain is approximately equal to 15.

Practically, we used 5 observers running in parallel, each of them with a life-time equal to $15\,\mathrm{h}$ (hence we initialize a new observer each $3\,\mathrm{h}$). This "3 hours" is comparable to the average response time of the FCC to perturbations. The initial value of θ has been set to $\theta_0 = 10$ and we have set $\lambda = 0.27\,\mathrm{h}^{-1}$, such that at any time, there is an observer with a corresponding value of θ greater than 2. The value $\theta_1 = 5$ looks sufficient to ensure convergence of a high-gain observer for any initial condition (from simulation results).

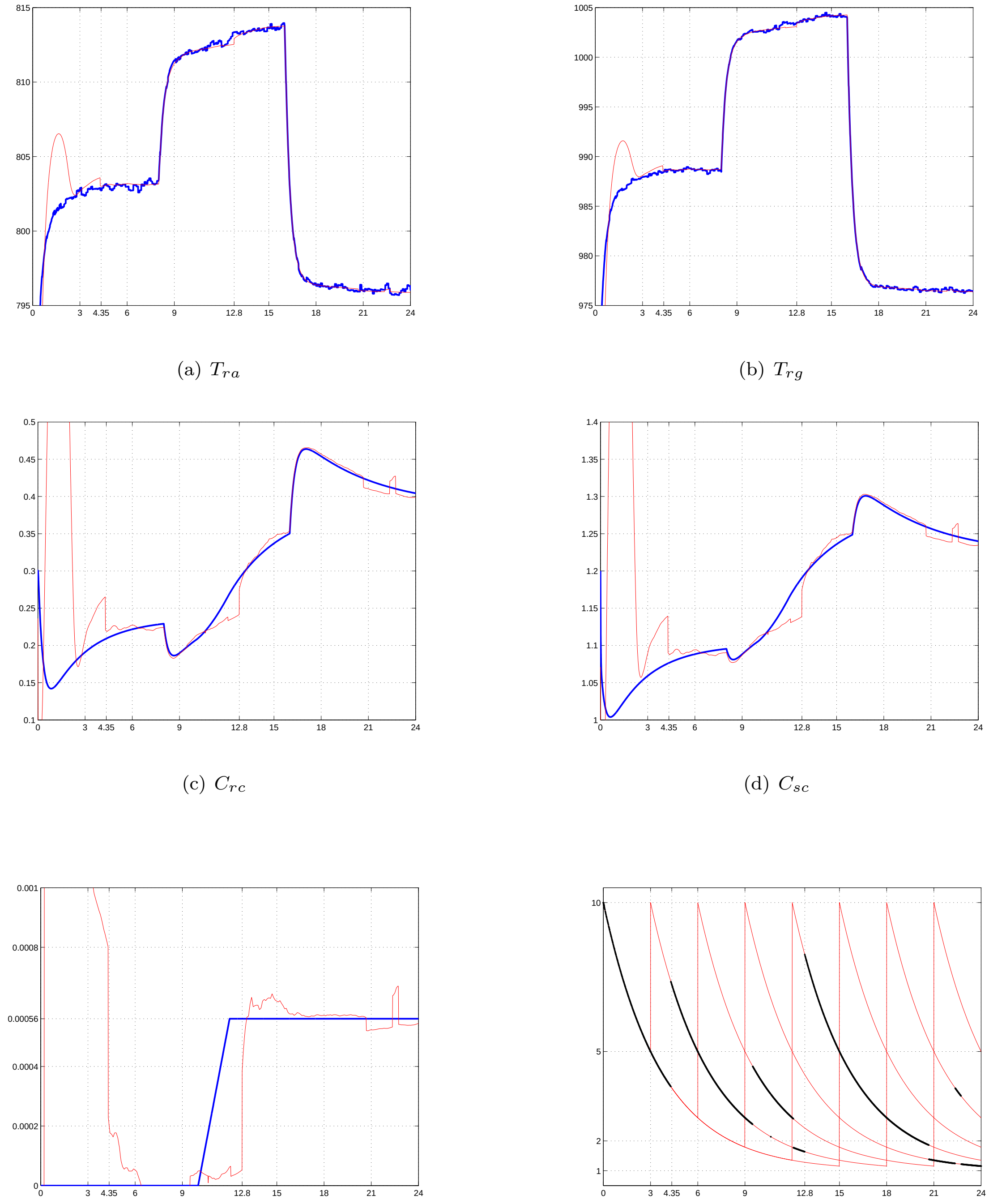
All simulations below were performed using Simulink® [12].

5.2.3. *Discussion of numerical results.* Figures 3(a) and 3(b) represent respectively T_{ra} and T_{rg} measurements. Figures 3(c), 3(d) and 3(e) represent the (unknown) state variables C_{rc}, C_{sc} and F_{cf} respectively. We also plotted on figures 3(a) to 3(e) the estimation provided by the best observer. On figure 3(f), we have plotted five curves corresponding to the value of θ for each observer. Hence, it is simply five exponentially decreasing functions, each of them obtained from the others by a time shift. On the same figure, we have also plotted the value of θ corresponding to the best observer at current time (see below).

We simulate some measurement-noise on each temperature. We simulate also a disturbance consisting of a ramp on F_{cf}, starting at $t = 10\,\mathrm{h}$ from $F_{cf} = 0$ to $F_{cf} = 5.6\,10^{-4}$ at $t = 12\,\mathrm{h}$. This unmodelled and unmeasured disturbance is larger than realistic actual disturbances: the coke formation factor usually varies very slowly. The tracking of the parameter is fast and accurate (figure 3(e)).

Moreover, figure 3(f) shows that our multiple high-gain extended Kalman filter does exactly what we expected from the theory. Indeed, we have represented on figure 3(f) the behavior of $\theta\left(t\right)$ for each of our five observers (thin lines). At each time, the value of $\theta\left(t\right)$ corresponding to the observer with smallest innovation is plotted with a thick line. So it is clear which kind of observer (high-gain or classical extended Kalman filter) has the best performance at each time. At the beginning, each observer is a high-gain observer. Then, since the model is accurate (no disturbances) the observer with the smallest value of $\theta\left(t\right)$ becomes more robust to measurement noise and so has the smallest innovation. When suddenly $F_{cf}\left(t\right)$ begins to vary according to a ramp, the observer with highest gain becomes more accurate and this behavior illustrates the well-known ability of high-gain observers to track the state in presence of unmodelled disturbances. When $F_{cf}\left(t\right)$ stabilizes and is correctly estimated, the classical extended Kalman filter becomes better again, thanks to its good (optimal) local properties.

5.3. Identification of reaction rate model of oxygen.



(a) T_{ra}

(b) T_{rg}

(c) C_{rc}

(d) C_{sc}

(e) F_{cf}

(f) $\theta\left(t\right)$ and the best observer versus time

FIGURE 3. FCC simulation 1

5.3.1. *Identifiability analysis.* Now, we will assume that the function describing the reaction rate of oxygen is unknown. As a matter of fact, this part of the model is usually very dependant from the FCC unit under consideration, and/or more simply, from the author of the paper.

Indeed, the main difference between the model described in [11] and the simplest one in [10] concerns the oxygen reaction rate model O_{fg} which gives the rate R_{cb} of burning carbon. Therefore, **we will assume that** K_{or} (in Formula 5.12) **is now an unknown function of** T_{rg}. We use the same approach as in the previous section to check the (infinitesimal) identifiability of the system. Since K_{or} is now considered as an unknown function, we will use T_{rg} to estimate it:

Mainly, by differentiation, $\dot{T}_{rg}$ allows to compute R_{cb} and R_{cb} gives O_{fg}. Here, O_{fg} is a nonlinear function (5.10) of our unknown function K_{or}. This function is not defined at $K_{or} = 0$. But, in restriction to the set of values that are physically relevant, it is bijective. Hence, it is possible, and useful to modify it outside a "relevant set", as we shall see below.

Therefore, let us set $K_{or} = \varphi\left(T_{rg}\right)$, and $R_{cb} = \Phi\left(C_{rc}, R_{ai}, \varphi\left(T_{rg}\right)\right)$ so that (5.15) becomes:

$$\dot{x}_2 = a_3 x_3 + g_2\left(x_1, x_2, x_5, \Phi\left(x_2, u, \varphi\left(x_5\right)\right), u, \dot{u}\right).$$

Starting from this point, it would be easy to transform directly the FCC system into Identifiability canonical form. But, in order to illustrate the intrinsic characterizations given in Section 3 (details in the paper [2]), we will apply the theory to the FCC model.

At first, we have to calculate the two indices, k and r, as defined in section 3.3.2. Briefly, let us consider a system of the form:

$$\left\{\begin{array}{lcl} \dot{x} & = & f\left(x, \varphi\right) \\ y_1 & = & h_1\left(x, \varphi\right) \\ y_2 & = & h_2\left(x, \varphi\right), \end{array}\right.$$

and recall that:

- k is the first index such that the respective ranks $\left(N_l\right)_{l=0,1,\ldots}$ of the family $\left(E_l\right)_{l=0,1,\ldots}$ of one-forms
$$E_l = \operatorname{span}\left(\left\{d_x h_i, d_x L_f h_i, \ldots, d_x L_f^{l-1} h_i, i = 1, 2\right\}\right)$$
are such that $N_k = 2k$ and $N_{k+1} < 2k+2$;
- r is the order of the system with respect to φ that is the first index such that $d_\varphi L_f^r h_1$ or $d_\varphi L_f^r h_2$ does not vanish identically.

In the FCC case, $h_1\left(x, \varphi\right) = T_{ra} = x_1$ and $h_2\left(x, \varphi\right) = T_{rg} = x_5$, hence $E_1 = \operatorname{span}\left(\left\{dT_{ra}, dT_{rg}\right\}\right)$. But since the only new state variable appearing in T_{ra} and T_{rg} is C_{rc}, then $E_2 = \operatorname{span}\left(\left\{dT_{ra}, dT_{rg}, dC_{rc}\right\}\right)$. As a consequence, $N_1 = 2$ and $N_2 = 3$ therefore $k = 1$. Moreover, thanks to our observability analysis, let us observe that $N_3 = 4$ and $N_4 = 5 = n$.

Remark 4. *The order of the system is $r = 1$ since $L_f h_2$ is a function of O_{fg} and therefore of the unknown function K_{or} ($d_\varphi L_f h_2 \neq 0$). Nevertheless, $d_\varphi L_f h_1 \equiv 0$.*

This remark will be used for the construction of the observer

The system is regular (see Section 3.3.2, or [2]), but we have to look further in order do decide if the system is of type 1, 2 or 3. Here, since $d_x h_1 \wedge d_x h_2 \wedge d_x L_f h_1 \neq 0$ then **hypothesis B.1.** is satisfied hence the system has **type 2**.

More precisely, it may be locally written under the following form:

$$\begin{array}{lcl lcl} y_1 & = & x_1 & y_2 & = & x_5 \\ \dot{x}_1 & = & F_1\left(x_1, x_2, x_5, u\right) & \dot{x}_5 & = & \Phi\left(x, \varphi\right) \\ \dot{x}_2 & = & F_2\left(x_1, x_2, x_3, x_5, \Phi\left(x, \varphi\right), u\right) & & & \\ \dot{x}_3 & = & F_3\left(x_1, x_2, x_3, x_4, x_5, \Phi\left(x, \varphi\right), u\right) & & & \\ \dot{x}_4 & = & F_4\left(x, \varphi, u\right) & & & \end{array} \tag{5.17}$$

with $\frac{\partial \Phi}{\partial \varphi} \neq 0$, $\frac{\partial F_1}{\partial x_2} \neq 0$, $\frac{\partial F_2}{\partial x_3} \neq 0$, $\frac{\partial F_3}{\partial x_4} \neq 0$.

Notice that this canonical form has a particularity: $\dot{x}_1$ does not depend of Φ since $d_\varphi L_f h_1 \equiv 0$.

5.3.2. *Model for identification.* We will estimate the function Φ. In order to construct our exponentially converging observer of "mixed" high-gain extended Kalman filter type, we will use a "local" second order model for Φ that is to say we will assume that $\frac{d^3\Phi}{dt^3} = 0$ (locally: on reasonable time intervals, Φ is accurately approximated by a 2rd order polynomial, which doesn't mean that it is globally a third order polynomial).

Hence we add three state variables to the original state variables, $x_6 = \Phi$ $(= R_{cb})$, $x_7 = \dot{\Phi}$ and $x_8 = \ddot{\Phi}$ such that $\dot{x}_6 = x_7$, $\dot{x}_7 = x_8$ and $\dot{x}_8 = 0$.

It has to be noticed that this local model becomes not valid when the control variable R_{ai} has jumps, since R_{ai} appears both in R_{cb} as a function of O_{fg} and in O_{fg} as a nonlinear function of K_{or}. We will tune the model in order to obtain very fast convergence so that this kind of disturbance has no large effect on estimation.

In order to retrieve φ from Φ, the function $\varphi \mapsto \Phi(x,\varphi)$ should be one to one. More precisely, the following function

$$z \mapsto \exp\left(\frac{-a}{b+\frac{c}{z}}\right),$$

has to be modified out of the set of "physical values", in such a way that it becomes a smooth diffeomorphism for any positive value of the parameters a, b and c, bounded from below. We leave the reader to check that this function can be smoothly modified outside any domain $\{A > z > 0, A \text{ large}\}$ in order to obtain a global diffeomorphism from $\mathbb{R}$ to $\mathbb{R}$.

Moreover, we observed that, during simulations, the observer internal variables remain in the physical state domain except at the very beginning of the simulation, so that, again, (and this seems to be a general fact in that type of applications), our carefulness is not justified here. Brute force computation appears to be enough, in most cases.

Finally, in agreement with the theory developed in the first part of this paper, our system can be written globally,

$$(5.18)\qquad \begin{array}{lcllcl} y_1 &=& x_1 & y_2 &=& x_5 \\ \dot{x}_1 &=& a_2x_2 + g_1(x_1,x_5,u) & \dot{x}_5 &=& x_6 + g_5(x_1,x_5,u) \\ \dot{x}_2 &=& a_3x_3 + g_2(x_1,x_2,x_5,x_6,u,\dot{u}) & \dot{x}_6 &=& x_7 \\ \dot{x}_3 &=& a_4x_4 + g_3(x_1,x_2,x_3,x_5,u,\dot{u}) & \dot{x}_7 &=& x_8 \\ \dot{x}_4 &=& g_4(x). & \dot{x}_8 &=& 0 \end{array}$$

5.3.3. *Identification algorithm.* As we said, the system (5.17) is clearly in the type 2 identifiability normal form. However, we cannot apply our (mixed Kalman) observer directly on the system (5.18) because x_6 appears in $\dot{x}_2$ which is not allowed in our method. To overcome this (small) difficulty, we decided to proceed as follows.

We consider the two following systems (Σ_1) and (Σ_2) independently

$$(\Sigma_2)\left\{\begin{array}{lcl} y_2 &=& \xi_1 \\ \dot{\xi}_1 &=& \alpha_1(u)\xi_1 + \alpha_2\xi_2 + \alpha_3(u)y_1 + \beta(u) \\ \dot{\xi}_2 &=& \xi_3 \\ \dot{\xi}_3 &=& \xi_4 \\ \dot{\xi}_4 &=& 0, \end{array}\right.$$

where the equation of $\dot{\xi}_1$ comes immediately from (5.8), and

$$(\Sigma_1)\left\{\begin{array}{lcl} y_1 &=& x_1 \\ \dot{x}_1 &=& a_2x_2 + g_1(x_1,y_2,u) \\ \dot{x}_2 &=& a_3x_3 + g_2(x_1,x_2,\xi_1,\xi_2,u,\dot{u}) \\ \dot{x}_3 &=& a_4x_4 + g_3(x_1,x_2,x_3,\xi_1,u,\dot{u}) \\ \dot{x}_4 &=& g_4(x) \end{array}\right.$$

The observer of Subsystem (Σ_2) will be mainly an approximate derivator of $\xi_2 = \Phi\ (= R_{cb})$.

But (Σ_2) is very simple: it is linear, time dependant. This will allow us to "filter" the output T_{rg} and its derivative, using a **standard linear Kalman filter**. This filter will provide an **accurate estimation** of $\xi_1 = T_{rg}$ and $\xi_2 = \Phi$.

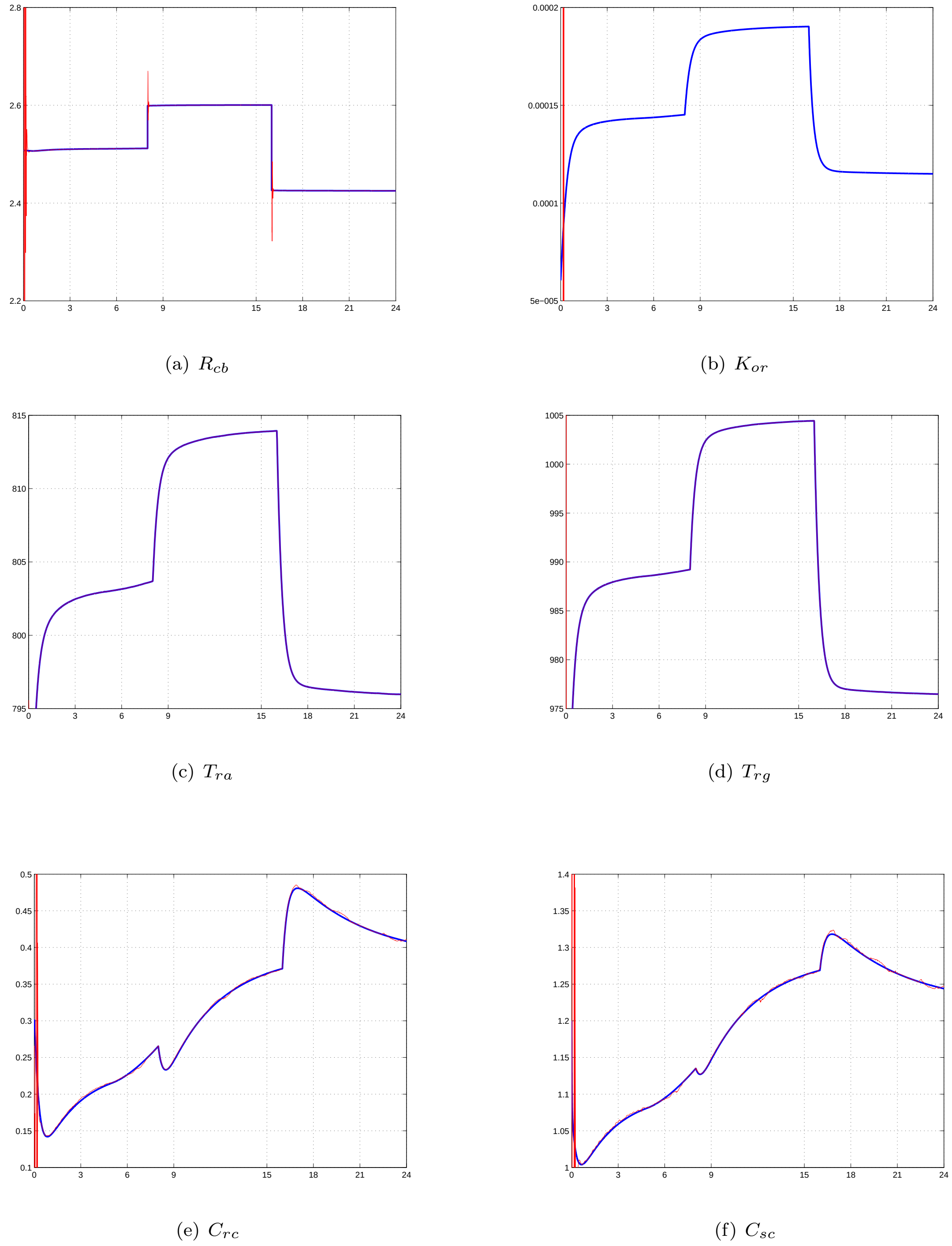
We will use directly this estimation of (T_{rg},Φ) inside (Σ_1), to which we will apply the (mixed) observer of Section 4.1.3, just considering ξ_1 and ξ_2 as new inputs.

Remark 5. *For general results concerning that type of cascade systems, and on the way to apply exponentially converging high-gain observers (mostly Luenberger-type) to cascade systems, see* [8].

5.3.4. *Simulation.* We used exactly the same scenario as in Section 5.2.

Identification of R_{cb} is not a hard task, even if it requires one derivative of $y_2 = T_{rg}$. But the estimation of K_{or} requires a very good estimation of R_{cb}: this is due to the high "sensitivity" of the nonlinear function $R_{cb}(K_{or})$ (i.e. large values of the derivative of R_{cb} w.r.t. K_{or}).

The model – supposed to be unknown – used for simulation is the equation (5.12), coming from [11]. Figures 4(c) to 5 represent the state estimation.



(a) R_{cb}

(b) K_{or}

(c) T_{ra}

(d) T_{rg}

(e) C_{rc}

(f) C_{sc}

FIGURE 4. FCC simulation 2

The noise being not too large, estimation of F_{cf} is not hard. Therefore, we were able to tune our observers for they estimate faster than in Section 5.2. The result of this new choice of parameters appears clearly on figure 5, where one can see the very quick estimation of F_{cf} (comparing to figure 3(e)).

We also plotted the estimation (coming from a standard linear Kalman filter, as we said) of R_{cb} on figure 4(a) and the estimation of K_{or} resulting from the estimation of R_{cb}, on figure 4(b).

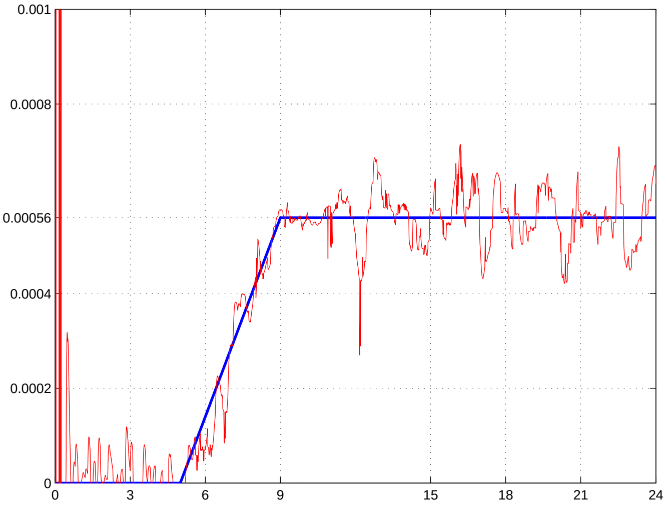


FIGURE 5. F_{cf}

We have plotted the results of the identification at time 0 h, 1 h, 9 h and 17 h on figures 6(a) to 6(d). The continuous line is the actual function K_{or} versus T_{rg}. The dotted line represents the estimation of K_{or} as a function of T_{rg}. The figure 6(a) represents the unknown function to be identified.

On figure 6(b), we have no information about this function for temperatures larger than 985 K, because $T_{rg}\,(t)$ does not pass beyond 985 K. However, the function has been identified between 970 K and 985 K. This interval represents the range of regenerator temperatures during the first hour of operation, as it can be seen on figure 4(d). After 9 h, the function has been identified between 970 K and 1005 K (figure 6(c)). At the end of the simulation (figure 6(d)), the function has been also identified between 970 K and 1005 K, that is for each values reached by T_{rg} during the simulation. It can be seen on figure 6(d) that there are two values for which $K_{or}\,(T_{rg})$ is very badly estimated (around $T_{rg} = 990\,\mathrm{K}$ and $T_{rg} = 1005\,\mathrm{K}$). These values correspond to discontinuities of R_{ai} : the estimation of R_{cb} becomes bad during a short transient.

An accurate estimation of the unknown function can be obtained by some regularization of the informations collected during simulation. We just applied some outliers removal procedure followed by some smoothing procedure, in order to obtain the estimation shown on figure 7.

REFERENCES

[1] BUSVELLE, E., DEZA, F., MOKRANI, F., and HAPIAK, S. 1992, Application of a nonlinear observer to a fluid catalytic cracker, NOLCOS'92, Juin 1992 Bordeaux, France

[2] BUSVELLE, E., and GAUTHIER, J.-P., 2003, On determining unknown functions in differential systems, with an application to biological reactors., *COCV*, **9**, 509-552.

[3] BUSVELLE, E., and GAUTHIER, J.-P., 2002, High-Gain and Non High-Gain Observers for nonlinear systems, In Contemporary Trends in Nonlinear Geometric Control Theory, (World Scientific, Anzaldo-Meneses, Bonnard, Gauthier, Monroy-Perez, editors), pp. 257-286.

[4] DEZA, F., 1991, Contribution to the synthesis of nonlinear observers, PHD Thesis, INSA de Rouen, France.

[5] GAUTHIER, J.-P., HAMMOURI, H., and OTHMAN, S., 1992, A simple observer for nonlinear systems. Applications to Bioreactors. *IEEE Trans. Aut. Control*, **37**, 875-880.

[6] GAUTHIER, J.-P., and KUPKA, I., 2001, *Deterministic Observation Theory and Applications* (Cambridge University Press).

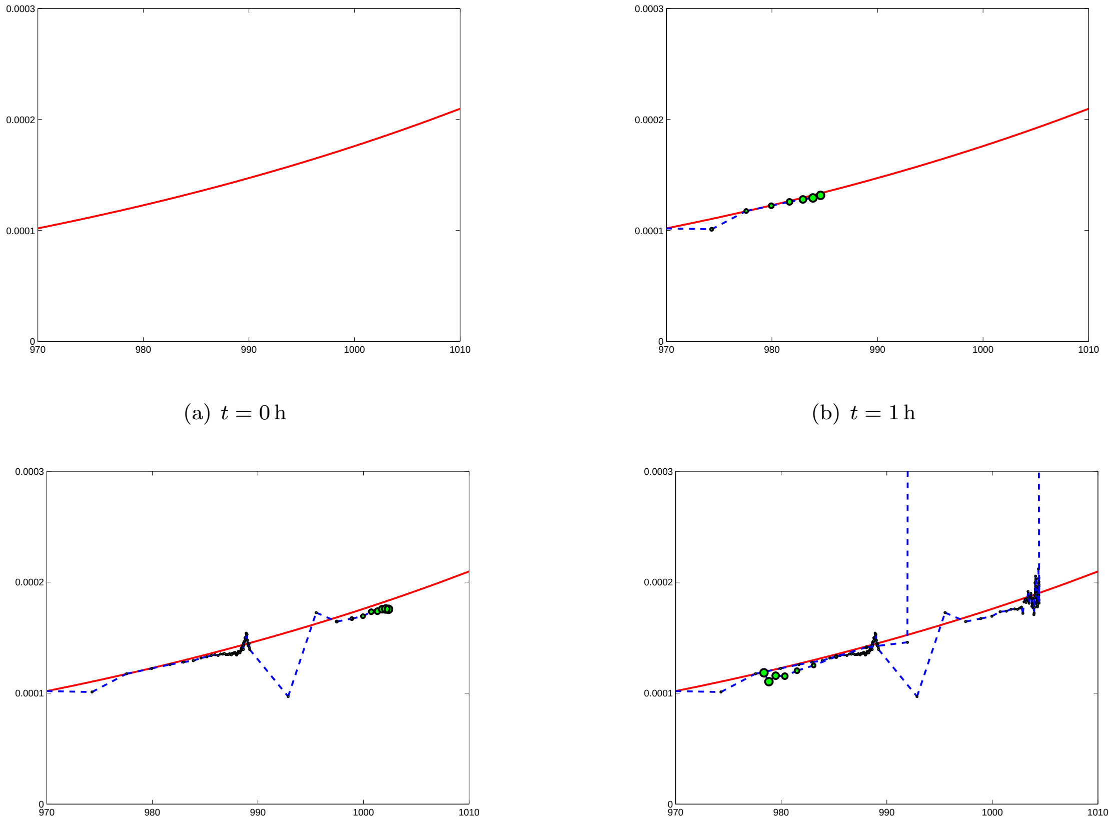
[7] HAMMOURI, H., and FARZA, M., 2003, Nonlinear observers for locally uniformly observable systems, *COCV*, **9**, 343-352.

[8] SHIM, H., SON, Y. I., and SEO, J. H., 2001, Semi-global observer for multi-output nonlinear systems, *Systems Control Lett.* **42** 233–244.

[9] KURIHARA, H., 1967, Optimal control of fluid catalytic cracking processes, Sc. D. Dissertation, M.I.T. Cambridge, MA.

[10] LEE, W., and KUGELMAN, A. M., 1973, Number of steady–state operating points and local stability of open–loop fluid catalytic cracker, *Ind. Eng. Chem. Process Des. Develop.*, **12**, N° 2, 197–204.

[11] MCFARLANE, R. C., and BACON, D. W., 1989, Adaptive optimizing control of multivariable constrained chemical processes. 2. Application studies, *Ind. Eng. Chem. Res.*, **28**, N° 12, 1834–1845.



(a) $t = 0\,\text{h}$

(b) $t = 1\,\text{h}$

(c) $t = 9\,\text{h}$

(d) $t = 17\,\text{h}$

FIGURE 6. K_{or} as a function of T_{rg}

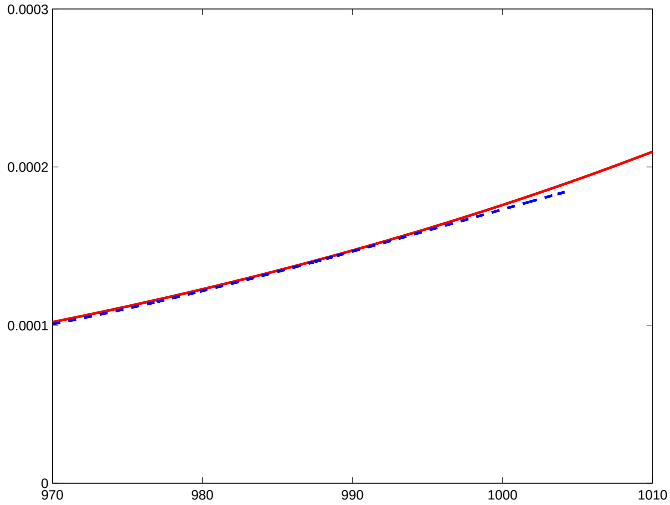


FIGURE 7. K_{or} as a function of T_{rg} after smoothing

[12] Matlab® and Simulink®, The MathWorks, Inc., http://www.mathtools.net/MATLAB/index.html

[13] Picard, J., 1991, Efficiency of the extended Kalman filter for nonlinear systems with small noise, *SIAM Journal on Appl. Math.*, **51**, 3, 843-885.

GEOMETRIC OPTIMAL CONTROL OF THE ATMOSPHERIC ARC FOR A SPACE SHUTTLE

B. BONNARD, E. BUSVELLE, G. LAUNAY

Laboratoire d'Analyse Appliquée et d'Optimisation, Université de Bourgogne, UFR des Sciences, 21000 Dijon, France.

Abstract. We give preliminary remarks concerning the optimal control of the atmospheric arc for a space shuttle (earth re-entry or Mars sample return project). The system governing the trajectories is 6–dimensional, the control is the bank angle, the cost–integrand is the thermal flux and we have state constraints on the thermal flux and the normal acceleration. Our study is geometric and founded on the analysis of the solutions of a minimum principle and direct evaluation of the small–time reachable set for the problem taking into account the state constraints.

1. Introduction

The objective of this article is to make a preliminary analysis of the optimal control of the atmospheric arc for a space shuttle where the cost is the total thermal flux. The control is the bank angle (the angle of attack being hold fixed) and we have state constraints on the thermal flux and the normal acceleration. A pure numerical approach to the problem is presented in [2] where the analysis is also simplified because the terminal condition is relaxed to a condition on the modulus of the speed. Our aim is to analyze the problem with fixed end-point conditions which leads to a complex control law due to the number of switchings (or the number of rotations) we need to match the boundary conditions.

This article is only a first step in the analysis in order to introduce the geometric tools to handle the problem and the necessary optimality conditions. In particular we shall restrict our computations to a 3 dimensional subsystem where the state variables are the modulus of the velocity, the altitude and the flight path angle. Also we shall localize the analysis to a small neighborhood of any point in the flight domain. This will allows to give local bounds to the number of switchings. It must be completed by numerical simulations to get a global bound.

Our approach is geometric and use necessary optimality conditions and direct evaluation of the small time reachable set in the spirit of [11] but using normal forms as in [2] where the constraints are taken into account. It is well illustrated by the following planar example. Consider the time optimal control problem for the system $\dot{q} = X\left(q\right) + uY\left(q\right)$, $q = (x, y)$, $|u| \leq 1$. Let γ_+ (resp. γ_-) be an arc corresponding to $u = +1$ (resp. $u = -1$) and denote $\gamma_1\gamma_2$ an arc γ_1 followed by γ_2. Take a generic point q_0, then the small time reachable set starting from q_0 is a cone bounded by arcs γ_+ and γ_- and each optimal trajectory is of the form $\gamma_+\gamma_-$ or $\gamma_-\gamma_+$, see figure 1; moreover along a trajectory the time can be measured using Miele's form: $\omega = p\,dq$ where p is given by $\langle p, X\rangle = 1$, $\langle p, Y\rangle = 0$.

Assume $q_0 = 0$ and the trajectories constrained to the domain $C : y \geq 0$. Let $\gamma_b\left(t\right)$, $t \in [0, T]$ be a boundary arc starting from $q_0 = 0$ and contained in the frontier $y = 0$; assume that the corresponding control u_b is admissible and not saturating. Let $B = \gamma_b\left(t\right)$, $T > 0$ small enough. Consider the arcs $\gamma_+\gamma_-$ and $\gamma_-\gamma_+$ joining 0 to B, one is time minimal (and the other is time maximal) for the problem without state constraint, and we have two possibilities for the constrained problem, see figure 1,(b). Assume it is $\gamma_+\gamma_-$, then if it is contained in $y \geq 0$, it is admissible and the boundary arc is not optimal, the optimal synthesis near q_0 for the constrained system being $\gamma_+\gamma_-$. If $\gamma_+\gamma_-$ is not contained in $y \geq 0$ the boundary arc is time optimal and the optimal synthesis is $\gamma_+\gamma_b\gamma_-$. The analysis can be carried out in full details using the model $\dot{x} = 1 + ay$, $\dot{y} = c + u$ and not the Miele's form ω defined only for planar systems.

A major problem when analyzing optimal control problems with state constraints is to derive necessary optimality conditions. Indeed the constraints can be penalized in the cost in several manners and this

Date: November 6, 2000.

1991 *Mathematics Subject Classification.* AMS Classif. 49K15,70H20.

Key words and phrases. Optimal control, Nonlinear systems with state constraints, Minimum principle, Reachable set, Space shuttle re–entry problem and Mars sample return project.

E-mail address: bbonnard@u-bourgogne.fr, busvelle@u-bourgogne.fr.

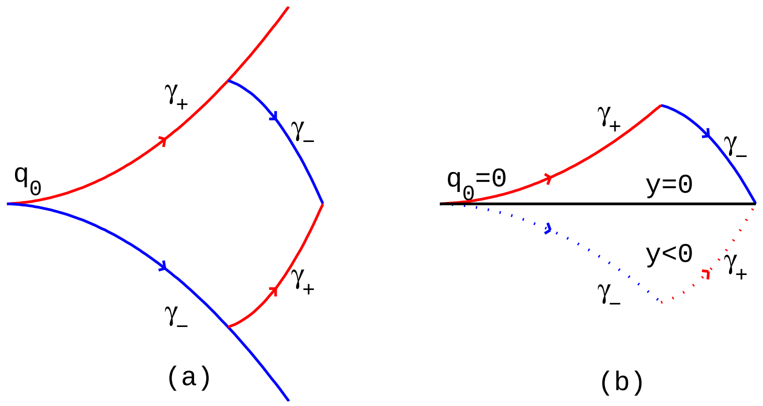


Figure 1. Reachable set with and without constraints

leads to introduce the concept of order of the constraints. Also it is the basic concept to construct normal forms and evaluate the reachable sets for the system with the constraints. We shall formulate a minimum principle due to [8, 12], adapted to analyze the optimal trajectories for the space shuttle. It concerns single input control systems and we need regularity assumptions. It is much more precise than the general minimum principle of [12], where an optimal trajectory is the projection of a trajectory in cotangent bundle depending of a measure supported by the constraints.

2. The model

The problem is to control the atmospheric arc nearby a planet which can be the Earth (re–entry problem) or Mars (sample return project). In both cases the equations are the same, except for constants related to the planet (radius, mass, angular velocity, atmosphere). In our computations we shall assume that the planet is the Earth. In order to modelize the problem, we use the laws of classical mechanics, a model of the gravitational force, a model of atmosphere and a model of the aerodynamic force which decomposes into a drag force and a lift force.

The equations are simplified by choices of orthonormal moving frames that we explain below.

2.1. Moving frames.

We denote by $E = (e_1, e_2, e_3)$ a standard Galilean frame whose origin O is the center of the Earth and let $R_1 = (I, J, K)$ be a rotating frame centered at 0 where K is the axis N–S of rotation of the Earth, the angular velocity being Ω and I is chosen to intersect Greenwich meridian.

Let R be the Earth radius and let G be the center of mass of the shuttle. We denote by $R_1' = (e_r, e_l, e_L)$ the frame associated to spherical coordinates of $G = (r, l, L)$, $r \geq R$ being the distance OG and l, L being respectively the longitude and latitude.

We introduce the following moving frame $R_2 = (i, j, k)$ whose center is G. Let $\zeta : t \longrightarrow (x(t), y(t), z(t))$ be the trajectory of G measured in the frame R_1 and let $\overrightarrow{v}$ be the relative speed $v = \dot{x}I + \dot{y}J + \dot{z}K$. To define $\overrightarrow{i}$, we set $\overrightarrow{v} = |v|\,\overrightarrow{i}$. The vector j is a vector in the plane (i, e_r), j is perpendicular to i and oriented by $j.e_r > 0$. We take $k = i \wedge j$. The vector i is parametrized in the frame $R_1' = (e_r, e_l, e_L)$ by two angles:

- γ: flight path angle
- Ξ: azimuth

defined on figure 2.

2.2. Model of the forces.

For the atmospheric arc we assume the following

Assumption 1. *There is no thrust: the shuttle is a glider.*

Assumption 2. *The speed of the atmosphere is the speed of the Earth, i.e. the relative speed of the shuttle with respect to the atmosphere is the speed $\overrightarrow{v}$.*

We must consider two types of forces acting on the shuttle.

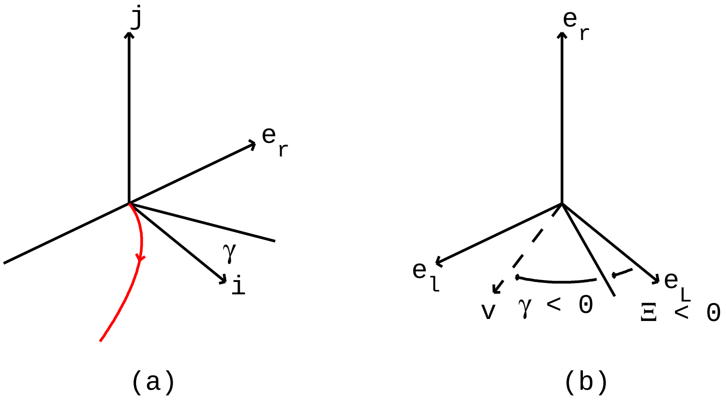


FIGURE 2. Moving frames: flight path angle and azimuth

- Gravitational force. We assume that the Earth is spherical so that the gravitational force is oriented along e_r.It is written in the moving frame R_2

$$\overrightarrow{P} = -mg\,(i \sin\gamma + j\cos\gamma)$$

where $g = \frac{\mu_e}{r^2}$.
- Aerodynamic force. The effect of the atmosphere on the shuttle is on aerodynamic force which decomposes into
  - A drag force colinear to the speed $\overrightarrow{v}$ and of the form

$$\overrightarrow{T} = -\left(\frac{1}{2}\rho S C_D v^2\right) i$$

  - A lift force perpendicular to $\overrightarrow{v}$ and given by

$$\overrightarrow{P}_T = \frac{1}{2}\rho S C_L v^2\,(j\cos\mu + k\sin\mu)$$

  and μ is called the bank angle, where $\rho = \rho\,(r)$ is the atmospheric density, S is a constant and C_D, C_L are respectively the drag and lift coefficient.

Assumption 3. *Both coefficients C_D and C_L are depending upon the angle of attack α which parametrized the orientation of the speed v with respect to the normal of an element of area of the shuttle. We assume that for the atmospheric arc the angle of attack is kept constant. This is very restrictive but it is worth to point out that in the numerical simulations of* [2] *where α is a control, in the optimal solution it is a constant.*

Hence the only control is the angle of bank μ.

2.3. **System equations.** The atmospheric arc is governed by the following system

$$\frac{dr}{dt} = v\sin(\gamma) \tag{2.1a}$$

$$\frac{dv}{dt} = -g\sin(\gamma) - \frac{1}{2}\rho\frac{S\,C_D}{m}v^2 + \Omega^2 r\cos L\left(\sin\gamma\cos L - \cos\gamma\sin L\cos\Xi\right) \tag{2.1b}$$

$$\frac{d\gamma}{dt} = \cos(\gamma)\left(-\frac{g}{v} + \frac{v}{r}\right) + \frac{1}{2}\rho\frac{S\,C_L}{m}v\cos(\mu) + 2\Omega\cos L\sin\Xi \tag{2.1c}$$

$$+\,\Omega^2\frac{r\cos L}{v}\left(\cos\gamma\cos L + \sin\gamma\sin L\cos\Xi\right) \tag{2.1d}$$

$$\frac{dL}{dt} = \frac{v}{r}\cos\gamma\cos\Xi \tag{2.1e}$$

$$\frac{dl}{dt} = -\frac{v}{r}\frac{\cos\gamma\sin\Xi}{\cos L} \tag{2.1f}$$

$$\frac{d\Xi}{dt} = \frac{1}{2}\rho\frac{SC_L}{m}\sin\mu\frac{v}{\cos\gamma} + \frac{v}{r}\cos\gamma\tan L\sin\Xi \tag{2.1g}$$

$$+\,2\Omega\left(\sin L - \tan\gamma\cos L\cos\Xi\right) + \Omega^2\frac{r}{v}\frac{\sin L\cos L\sin\Xi}{\cos\gamma} \tag{2.1h}$$

where the control is the bank angle μ and the state space is $q = (r, v, \gamma, L, l, \Xi)$

2.4. **Atmospheric model.** Atmospheric density is tabulated for Earth, Mars and Venus and we take an exponential model

$$\rho = \rho_0 \mathrm{e}^{-\beta r}$$

3. The control problem

3.1. **Control and control bounds.** The control can be either μ or $\dot{\mu}$. In the first case we can have the following bounds: $\mu \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ or $\mu \in [-\pi, \pi]$. We set $u_1 = \cos\mu$ and u_1 is a direct control on the flight path angle γ. We let $u_2 = \sin\mu$ and u_2 control the azimuth, the sign of u_2 allows the glider to turn left or right.

3.2. **State constraints.** There are several state constraints but in the first step of our analysis we shall consider two constraints:

- Constraint on the thermal flux

$$\varphi = C_q\sqrt{\rho}v^3 \leq \varphi^{\max} \tag{3.1}$$

  where C_q is a given constant
- Constraint on the normal acceleration

$$\gamma_n = \gamma_{n_0}(\alpha)\,\rho v^2 \leq \gamma_n^{\max} \tag{3.2}$$

3.3. **Optimal cost.** Several choices are allowed and we make the analysis for

$$J(\mu) = \int_0^T C_q\sqrt{\rho}v^3 dt \tag{3.3}$$

which represents the total thermal flux, the duration T of the atmospheric arc being not fixed. We introduce the differential equation

$$\frac{d\widetilde{q}_0}{dt} = C_q\sqrt{\rho}v^3 \tag{3.4}$$

with $\widetilde{q}_0(0) = 0$.

3.4. **Boundary conditions.** The transfer time T is free and we have two choices for the boundary conditions:

- Fixed boundary conditions at $t = 0$ and $t = T$ for $q = (r, v, \gamma, L, l, \Xi)$.
- $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ at $t = 0$ with the constraint that a preliminary maneuver on the Keplerian arc allows this possibility.

3.5. **State domain for the atmospheric arc.** The flight domain D for the Earth re–entry of the shuttle is the following:

- Altitude: $h = r - R \in [40\,\mathrm{km}, 120\,\mathrm{km}]$
- Velocity amplitude $v \in [2000\,\mathrm{m}\,/\,\mathrm{s}, 8000\,\mathrm{m}\,/\,\mathrm{s}]$
- The flight path angle domain is $0 > \gamma > -15\,°$.

Assumption 4. *(controllability assumption) The Earth angular velocity Ω is small and hence*

$$\frac{d\gamma}{dt} \cong \cos(\gamma)\left(-\frac{g}{v} + \frac{v}{r}\right) + \frac{1}{2}\rho\frac{S\,C_L}{m}v\cos(\mu) \tag{3.5}$$

We shall denote by D_c the subset of D where the lift force can at each point compensated the gravitational force that is

$$\frac{1}{2}\rho\frac{S\,C_L}{m}v > \frac{g}{v}$$

and (3.5) is feedback linearizable in the domain.

4. The minimal principle without state constraints – Extremal curves

4.1. **Problem statement and notations.** Let the single–input control system

$$\dot{q} = F(q,u) \tag{4.1}$$

and a cost to be minimized of the form

$$J(u) = \int_0^T \varphi(q)\,dt \tag{4.2}$$

where the transfer time T is free and φ is not depending upon u. The set of admissible controls is the set $\mathcal{U}$ of measurable mappings $u : [0,T] \to U$. The state domain is a subset of $\mathbb{R}^n$ with the state constraints:

- Constraint on the thermal flux

$$c_1(q) = \varphi(q) \leq \alpha_1 \tag{4.3}$$

- Constraint on the normal acceleration

$$c_2(q) = \gamma_n(q) \leq \alpha_2 \tag{4.4}$$

The boundary conditions are of the form:

- $q(0) = q_0$ and $q(T) = q_1$ fixed

or

- if $q = (r, v, \gamma, L, l, \Xi)$ then $\gamma(0) \in [\gamma_1, \gamma_2]$, $\gamma_1 < \gamma_2 < 0$.

We denote by $R(q_0, t)$ the reachable set at time $t > 0$ fixed and $R(q_0) = \bigcup_{t \text{ small enough}} R(q_0, t)$ the small time reachable set.

4.2. **Minimum principle.** We recall the minimum principle [13] which allows to parametrize the boundaries of the reachable sets [11].

We introduce the Hamiltonian

$$H(q,p,u) = \langle p, F(q,u)\rangle + \widetilde{p}_0\varphi(q)$$

where $q = (r, v, \gamma, L, l, \Xi)$ and $p = (p_r, p_v, p_\gamma, p_L, p_l, p_\Xi)$ is the adjoint vector and $\widetilde{p}_0$ is a constant such that $(p, \widetilde{p}_0) \neq 0$.

Definition 1. *If $\widetilde{p}_0 \neq 0$ we are in the normal case and if $\widetilde{p}_0 = 0$ we are in the abnormal case.*

Definition 2. *We call extremal a triplet (q,p,u) solution of the minimum principle*

$$\dot{q} = F(q,u) = \frac{\partial H}{\partial p} \tag{4.5}$$

$$\dot{p} = -p\frac{\partial F}{\partial q} - \widetilde{p}_0\frac{\partial \varphi}{\partial q} = -\frac{\partial H}{\partial q} \tag{4.6}$$

$$H(q,p,u) = \min_{w \in \mathcal{U}} H(q,p,w) \tag{4.7}$$

Proposition 1. *An optimal solution for the problem without state constraint is a projection on the state space of an extremal solution. Moreover $\widetilde{p}_0 \geq 0$. Since the transfer time T is free it is exceptional, that is $H = 0$. If moreover γ is free at $t = 0$, the adjoint vector p satisfy the transversality condition*

$$p_\gamma(0) = 0 \textit{ if } \gamma(0) \in \,]\gamma_1, \gamma_2[\tag{4.8}$$

4.3. **Definition of subsystem (I).** Observe that Ω is small with respect to the velocity of the shuttle. Hence if we neglect the transport terms $O\left(\Omega^2\right)$ and the Coriolis terms $O(\Omega)$ our system can be decomposed with $q_1 = (r, v, \gamma)$ and $q_2 = (L, l, \Xi)$ into

$$\begin{aligned} \dot{q}_1 &= F_1(q_1, u_1) \\ \dot{q}_2 &= F_2(q, u_2) \end{aligned}$$

where $u_1 = \cos\mu$, $u_2 = \sin\mu$, $u = (u_1, u_2)$ and

$$U = \left\{u_1^2 + u_2^2 = 1 \text{ and } u_1 \geq 0 \text{ if } \mu \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]\right\}$$

The adjoint system (4.6) is the decomposed into

$$\begin{pmatrix} \dot{p}_1 & \dot{p}_2 & 0 \end{pmatrix} = -\begin{pmatrix} p_1 & p_2 & \widetilde{p}_0 \end{pmatrix} \begin{pmatrix} \frac{\partial F_1}{\partial q_1} & 0 & 0 \\ \frac{\partial F_2}{\partial q_1} & \frac{\partial F_2}{\partial q_2} & 0 \\ \frac{\partial \varphi}{\partial q_1} & 0 & 0 \end{pmatrix}$$

If we relax the end–point condition on $q_2 = (L, l, \Xi)$ we obtain using the transversality condition $p_2(T) = 0$ and hence $p_2(t) \equiv 0$. The analysis of extremals reduces to the analysis of the solutions of

$$\begin{aligned} \dot{q}_1 &= F_1(q_1, u_1) \\ \dot{p}_1 &= -p_1 \frac{\partial F_1}{\partial q_1} - \widetilde{p}_0 \frac{\partial \varphi}{\partial q_1} \end{aligned}$$

It is associated to the optimal control of system (I) given below:

$$\text{(I)} \quad \left\{ \begin{array}{lcl} \dfrac{dr}{dt} & = & v\sin(\gamma) \\ \dfrac{dv}{dt} & = & -g\sin(\gamma) - \dfrac{1}{2}\rho\dfrac{SC_D}{m}v^2 \\ \dfrac{d\gamma}{dt} & = & \cos(\gamma)\left(-\dfrac{g}{v} + \dfrac{v}{r}\right) + \dfrac{1}{2}\rho\dfrac{SC_L}{m}v\cos(\mu) \end{array} \right.$$

Note that $q_1 = (r, v, \gamma)$ appears only in the state–constraints. We shall concentrate in a first step our analysis on the subsystem (I). It is related to the numerical simulation of [2].

4.4. **Analysis of extremals of system (I).**

4.4.1. *Problem reduction and definitions.* Consider a single input affine system

$$\dot{q} = X + uY, \ |u| \leq 1 \tag{4.9}$$

and a cost to be minimized of the form

$$J(u) = \int_0^T \varphi(t)\, dt \tag{4.10}$$

Assume moreover that $\varphi(q) > 0$ in the state domain. Introduce the equation

$$\left\{ \begin{array}{lcl} \dot{\widetilde{q}}_0 & = & \varphi(q) \\ \widetilde{q}_0(0) & = & 0 \end{array} \right.$$

and $\widetilde{q} = (q, \widetilde{q}_0)$ is the enlarged state. Hence (4.9), (4.10) can be written

$$\dot{\widetilde{q}} = \widetilde{X}(\widetilde{q}) + u\widetilde{Y}(\widetilde{q}), \ |u| \leq 1 \tag{4.11}$$

and let s be the new time parameter defined by

$$ds = \varphi(q(t))\, dt \tag{4.12}$$

and if q' denote the derivative of q with respect to s, (4.9) can be written

$$q' = \overline{X}(q) + u\overline{Y}(q), \ |u| \leq 1 \tag{4.13}$$

where $\overline{X} = \psi X$, $\overline{Y} = \psi Y$ and $\psi = \frac{1}{\varphi}$. The optimal control problem becomes a time minimum control problem.

Definition 3. *Consider $\dot{q} = X + uY$. A singular trajectory of the system (X, Y) is a projection of the following equations*

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} \\ \dot{p} &= -\frac{\partial H}{\partial q} \\ \langle p, Y \rangle &= 0 \end{aligned} \tag{4.14}$$

where $H = \langle p, X + uY \rangle$, $p \neq 0$. It is called exceptional if $H = 0$ and, admissible if $|u| \leq 1$ and strictly admissible if $u \in \left]-1, +1\right[$.

Notation 1. *If X_1 and X_2 are two smooth vector fields, we denote by $[X_1, X_2]$ the Lie bracket computed with the convention*

$$[X_1, X_2](q) = \frac{\partial X_2}{\partial q}(q)\, X_1(q) - \frac{\partial X_1}{\partial q}(q)\, X_2(q)$$

Assumption 5. *Throughout this article we shall assume that $g = \frac{\mu_e}{r^2}$ is constant.*

Proposition 2. *In the domain $\cos\gamma \neq 0$ there is no exceptional singular arc for the system (X, Y).*

Proof. The singular extremals are located on $\langle p, Y(q) \rangle = 0$. Differentiating twice with respect to t one gets

$$\begin{aligned} \langle p, [X, Y](q) \rangle &= 0 \\ \langle p, [X, [X, Y]](q) \rangle + u \langle p, [Y, [X, Y]](q) \rangle &= 0 \end{aligned}$$

We must compute the Lie brackets Y, $[X, Y]$, $[Y, [X, Y]]$ and $[X, [X, Y]]$ where

$$\begin{aligned} X &= v \sin\gamma \frac{\partial}{\partial r} - \left(g \sin\gamma + k\rho v^2\right) \frac{\partial}{\partial v} + \cos\gamma \left(-\frac{g}{v} + \frac{v}{r}\right) \frac{\partial}{\partial \gamma} \\ Y &= \overline{k}\rho v \frac{\partial}{\partial \gamma} \end{aligned}$$

and k , $\overline{k}$ are defined by the equations (2.1b) and (2.1c). Since the concept of singular arc is feedback invariant we can replace in our computations X and Y by

$$\begin{aligned} X &= v \sin\gamma \frac{\partial}{\partial r} - \left(g \sin\gamma + k\rho v^2\right) \frac{\partial}{\partial v} \\ Y &= \frac{\partial}{\partial \gamma} \end{aligned}$$

We have then

$$\begin{aligned} [X, Y] &= -v \cos\gamma \frac{\partial}{\partial r} + g \cos\gamma \frac{\partial}{\partial v} \\ [Y, [X, Y]] &= v \sin\gamma \frac{\partial}{\partial r} - g \sin\gamma \frac{\partial}{\partial v} \end{aligned}$$

hence $[X, Y]$ and $[Y, [X, Y]]$ are colinear. Moreover

$$[X, [X, Y]] = k\rho v^2 \cos\gamma \frac{\partial}{\partial r} + \left(-kv^3 \rho' \cos\gamma + 2k\rho g v \cos\gamma\right) \frac{\partial}{\partial v}$$

The singular extremals are located on Σ': $\langle p, Y \rangle = \langle p, [X, Y] \rangle = 0$ that is $p_\gamma = p_v g - p_r v = 0$. We introduce

$$\begin{aligned} D &= \det\left(Y, [X, Y], [Y, [X, Y]]\right) \\ D' &= \det\left(Y, [X, Y], [X, [X, Y]]\right) \\ D'' &= \det\left(Y, [X, Y], X\right) \end{aligned}$$

From our previous computations singular arcs are located on $D = D' = 0$ and moreover if they are exceptional they satisfy $D'' = 0$. We have

$$\begin{aligned} D &\equiv 0 \\ D' &= kv^2 \cos^2\gamma \left(\rho' v^2 - 3\rho g\right) \\ D'' &= k\rho v^3 \cos\gamma \end{aligned}$$

Since cos$\gamma \neq 0$ the proposition is proved. □

Moreover we have for system (4.9)

Lemma 1. *If cos$\gamma \neq 0$ then*

(1) Y *and* $[X, Y]$ *are independent;*
(2) $[Y, [X, Y]] \in \operatorname{Span}\{Y, [X, Y]\}$

4.4.2. *Analysis of extremals.* Consider the time minimum control problem for system (4.13):

$$q' = \overline{X}(q) + u\overline{Y}(q)\,, \ |u| \leq 1$$

We introduce the following definitions

Definition 4. *The set* Σ: $\left\langle p, \overline{Y}(q)\right\rangle = 0$ *is called the switching surface. Let* (q, p, u) *be an extremal defined on* $[0, T]$*; it is called singular if it is contained in* Σ*, bang if* $u = +1$ *or* $u = -1$ *and bang–bang if* $u(t)$ *is piecewise constant and given a.e. by* $u(t) = -\operatorname{sign}\langle p(t), Y(q(t))\rangle$*. We denote respectively by* γ_+ *(resp.* γ_-*,* γ_s*) a smooth arc associated to* $u = +1$ *(resp.* $u = -1$*,* u *singular control) and* $\gamma_+\gamma_-$ *represents an arc* γ_+ *followed by an arc* γ_-*.*

Let us calculate Lie brackets. We have

$$\begin{aligned} \overline{X} &= \psi\left(v \sin\gamma \frac{\partial}{\partial r} - \left(g \sin\gamma + k\rho v^2\right)\frac{\partial}{\partial v} + \cos\gamma\left(-\frac{g}{v} + \frac{v}{r}\right)\frac{\partial}{\partial \gamma}\right) \\ \overline{Y} &= \psi\overline{k}\rho v \frac{\partial}{\partial\gamma} \end{aligned}$$

where $\psi = \varphi^{-1}$. Since $\overline{X} = \psi X$, $\overline{Y} = \psi Y$ using for f_1, f_2 smooth functions the formula

$$[f_1 X, f_2 Y] = f_1 f_2 [X, Y] + f_1 (X f_2) Y - f_2 (Y f_1) X$$

where $Zf = \frac{\partial f}{\partial q} Z(q)$ is the Lie derivative we get

$$\left[\overline{X}, \overline{Y}\right] = \psi^2 [X, Y] + \psi (X\psi) Y - \psi (Y\psi) X$$

Since $Y = \overline{k}\rho v \frac{\partial}{\partial\gamma}$ and $\psi = f(\rho, v)$ we have $Y\psi = 0$. Hence $\left[\overline{X}, \overline{Y}\right] = \psi^2 [X, Y] + \psi (X\psi) Y$. Computing $\left[\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right]$ as before we obtain

Lemma 2. (1) *The set* Σ': $\left\langle p, \overline{Y}\right\rangle = \left\langle p, \left[\overline{X}, \overline{Y}\right]\right\rangle = 0$ *is given by* $\langle p, Y\rangle = \langle p, [X, Y]\rangle = 0$*.*
(2) $\left[\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right] = \psi^3 [Y, [X, Y]] \mod \operatorname{Span}\{Y, [X, Y]\}$ *and hence* $\left[\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right] \in \operatorname{Span}\{Y, [X, Y]\}$*.*

Moreover,

$$\overline{D} = \det\left(\overline{Y}, \left[\overline{X}, \overline{Y}\right], \left[\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right]\right) \equiv 0$$

and

$$\overline{D}'' = \det\left(\overline{Y}, \left[\overline{X}, \overline{Y}\right], \overline{X}\right) = \frac{k\overline{k}^2 \rho \cos\gamma}{C_q^4 v^7}$$

and hence $\overline{Y}$, $\left[\overline{X}, \overline{Y}\right]$, $\overline{X}$ are a frame in the flight domain where $\cos\gamma \neq 0$. So there exists a, b, c such that

$$\left[\overline{X}, \left[\overline{X}, \overline{Y}\right]\right] = a\,\overline{X} + b\,\overline{Y} + c\,\left[\overline{X}, \overline{Y}\right] \tag{4.15}$$

Long computations give us the crucial result

Lemma 3. *If* $\cos\gamma \neq 0$ *we have*

(1) $\overline{D}' = \det\left(\overline{Y}, \left[\overline{X}, \overline{Y}\right], \left[\overline{X}, \left[\overline{X}, \overline{Y}\right]\right]\right) = -\dfrac{\beta}{2}\dfrac{k\overline{k}^3\rho}{C_q^6 v^{11}} \cos^2\gamma \neq 0$

(2) $a = -\dfrac{\beta}{2}\dfrac{\overline{k}\sqrt{\rho}}{C_q^2 v^4} \cos\gamma < 0$

Corollary 1. *If* $\cos\gamma \neq 0$ *there exists no singular trajectory.*

4.4.3. *Application to the classification of extremals near* Σ. Let (q,p,u) be a smooth extremal on $[0,T]$. Differentiating the switching function $\Phi : t \to \left\langle p\left(t\right), \overline{Y}\left(q\left(t\right)\right)\right\rangle$ we have

$$\begin{aligned}\dot{\Phi}\left(t\right) &= \left\langle p\left(t\right), \left[\overline{X},\overline{Y}\right]\left(q\left(t\right)\right)\right\rangle \\ \ddot{\Phi}\left(t\right) &= \left\langle p\left(t\right), \left[\overline{X},\left[\overline{X},\overline{Y}\right]\right]\left(q\left(t\right)\right) + u\left(t\right)\left[\overline{Y},\left[\overline{X},\overline{Y}\right]\right]\left(q\left(t\right)\right)\right\rangle\end{aligned}$$

We use the results of [10] to classify the extremals near a point $z_0 = (q_0, p_0)$.

(1) Ordinary points. If z_0 belong to $\left\langle p, \overline{Y}\right\rangle = 0$, $\left\langle p, \left[\overline{X},\overline{Y}\right]\right\rangle \neq 0$, the point z_0 is called of order 1 or ordinary and each extremal curve is locally of the form $\gamma_+\gamma_-$ or $\gamma_-\gamma_+$.

(2) Points of order 2. Let $z_0 \in \Sigma'$: $\left\langle p, \overline{Y}\right\rangle = \left\langle p, \left[\overline{X},\overline{Y}\right]\right\rangle = 0$. Then if (q,p,u) is a smooth extremal through z_0 the switching function satisfies at z_0:

$$\Phi\left(t\right) = \dot{\Phi}\left(t\right) = 0$$

and

$$\begin{aligned}\ddot{\Phi}\left(t\right) &= \left\langle p\left(t\right), \left[\overline{X},\left[\overline{X},\overline{Y}\right]\right]\left(q\left(t\right)\right) + u\left(t\right)\left[\overline{Y},\left[\overline{X},\overline{Y}\right]\right]\left(q\left(t\right)\right)\right\rangle \\ &= \left\langle p\left(t\right), \left[\overline{X},\left[\overline{X},\overline{Y}\right]\right]\left(q\left(t\right)\right)\right\rangle\end{aligned}$$

from lemma 2 which is non zero from lemma 3. Hence both curves corresponding to $u = +1$ and $u = -1$ have a contact of order 2 with respect to Σ and the extremal solutions are represented on figure 3. According to the classification of [10] the point z_0 is a parabolic point and each extremal is locally bang–bang and of the form $\gamma_+\gamma_-\gamma_+$ or $\gamma_-\gamma_+\gamma_-$.

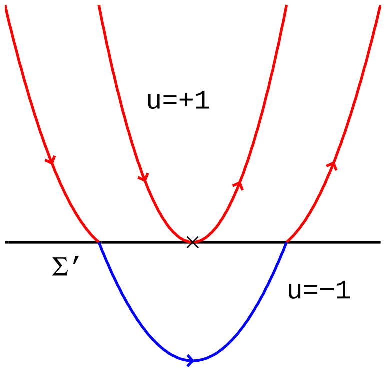


FIGURE 3. extremal solutions $(a > 0)$

From this analysis and from the minimum principle we can conclude about small time optimal policy.

Theorem 1. *If* $\cos\gamma \neq 0$ *each small time optimal policy is of the form* $\gamma_-\gamma_+\gamma_-$ *where* γ_+ *is an arc corresponding to* $u = \cos\mu = +1$ *and* γ_- *an arc corresponding to* $u = -1$ *(or* $u = 0$ *if* $\mu \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$*).*

Proof. According to the time minimum principle, an optimal arc has to satisfy $\widetilde{H} = 0$, $p_0 \geq 0$ where $\widetilde{H} = \left\langle p, \overline{X}\left(q\right) + u\overline{Y}\left(q\right)\right\rangle + p_0 = 0$. Hence p can be oriented at $z_0 \in \Sigma'$ according to $\left\langle p, \overline{X}\left(q\right)\right\rangle \leq 0$. Write

$$\left[\overline{X},\left[\overline{X},\overline{Y}\right]\right] = a\overline{X} + b\overline{Y} + c\left[\overline{X},\overline{Y}\right]$$

and $a < 0$ from lemma 3. Hence from figure 3, only an extremal $\gamma_-\gamma_+\gamma_-$ can be optimal. The assertion is proved. □

4.5. **Geometry of the small time reachable set.** Consider again system in 3–dimension

$$\frac{dq}{ds} = \overline{X}(q) + u\overline{Y}(q)$$

and its time extension in 4–dimension by adding the cost $\frac{ds'}{ds} = 1$. We denote respectively by $R(q_0)$ the small time reachable set and by $\widetilde{R}(q_0, 0)$ the small time reachable set for the extended system. One major research program undertake in [14, 11] using original ideas from Lobry is to evaluate in small dimensions the small time reachable set and its boundary. In particular the following result is basic:

Lemma 4. *Consider system $(\overline{X}, \overline{Y})$ in dimension 3 and let $g_1 = \overline{X} + \overline{Y}$ and $g_2 = \overline{X} - \overline{Y}$. Assume g_1, g_2 and $[g_1, g_2]$ linearly independent at q_0 then $R(q_0)$ is bounded by the two surfaces $\gamma_+\gamma_-(q_0)$ and $\gamma_-\gamma_+(q_0)$ and moreover $R(q_0) = \bigcup \gamma_+\gamma_-\gamma_+(q_0)$ (or $\bigcup \gamma_-\gamma_+\gamma_-(q_0)$)*

Actually in theorem 1 we proved more (see also [14]):

Lemma 5. *If $\cos\gamma \neq 0$ the boundary of the small time reachable set for the extended system $\widetilde{R}(q_0, 0)$ is an union of $\widetilde{\gamma}_-\widetilde{\gamma}_+\widetilde{\gamma}_-(q_0, 0)$ and $\widetilde{\gamma}_+\widetilde{\gamma}_-\widetilde{\gamma}_+(q_0, 0)$ where $\widetilde{\gamma}$ denotes the time extended trajectory.*

4.6. **Optimal control of the atmospheric arc.** If we consider the complete set of equations it can be written as a time optimal control problem for a 6–dimension system of the form

$$q' = \overline{X}(q) + u_1\overline{Y}_1(q) + u_2\overline{Y}_2(q)$$

where $u_1 = \cos\mu$ and $u_2 = \sin\mu$. We can use two points of view related to the control device:

- We set $\dot{\mu} = w$ where w is taken as a control bounded by M. The system is then a single input affine control system on the 7–dimensional state space (q, μ).
- We can consider the original system on the 6–dimensional state space. The control $u = (u_1, u_2)$ satisfies $u_1^2 + u_2^2 = 1$. If $\mu \in [0, 2\pi]$, the optimal control problem is equivalent to a sub-Riemannian problem with drift. Indeed if we set $\psi_i = \langle p, \overline{Y}_i(q)\rangle$, $i = 1, 2$ an extremal normal control is given by $u = \frac{1}{\|\psi\|}(\psi_1, \psi_2)$.

We must analyze the existence of abnormal extremals and the number of oscillations and switchings of optimal trajectories.

4.7. **Conclusion about this section.** Using minimum principle, Lie brackets and geometric methods we have evaluated the small time reachable set and solved the small time optimal control problem for system I. In particular we have obtained bounds on the number of switchings. The main property of system I is that $[\overline{Y}, [\overline{X}, \overline{Y}]]$ belong to Span $\{\overline{Y}, [\overline{X}, \overline{Y}]\}$. This is connected to the feedback linearizability if system I. For the global aspect one needs to analyze the global proportion of the switchings function Φ using convexity analysis and Rolle theorem. Our study is a preliminary step in order to evaluate the reachable set for the full system of equation without the state constraints and nearby the state constraints. We shall analyze the structure of the reachable set for system nearby the constraints in the next section.

5. Optimal control with state constraints

In this section we analyze the optimal control problem for system I, taking into account the constraints. We recall a minimum principle from [12] adapted to our situation. Our contribution is to make a direct evaluation of the small time reachable set for the constrained system using the previous computations of section 4 and a normal form.

When dealing with constrained systems the main concept is the concept of order of the constraint that we define next before to state the minimum principle adapted to our analysis.

5.1. **A minimum principle.** We consider the single input affine control system

$$\dot{q} = f(q) + ug(q) \quad |u| \leq 1$$

and a cost to be minimized of the form

$$J(u) = G(x(T))$$

where the transfer time T is fixed and q is constrained to

$$c(q) \leq 0$$

The boundary conditions are

$$q\left(0\right)=q_0$$
$$\Phi\left(x\left(T\right)\right)=0$$

The problem is denoted by $(\mathcal{P}_0)$ and can be imbedded into the one parameter family of problems $(\mathcal{P}_\alpha)$ where the constraints set is taken as

$$c\left(q\right)\leq\alpha,\ \alpha\text{ small}$$

The important concept is the concept of order of the constraint.

Definition 5. *The absolute (or generic) order of the constraint is the first integer $n+1$ such that*

$$g\left(f^0c\right)\equiv g\left(f^1c\right)\equiv\cdots\equiv g\left(f^{n-1}c\right)\equiv 0$$
$$g\left(f^nc\right)\neq 0$$

where the vector fields f, g acts on c by Lie derivative.

Definition 6. *A boundary arc $t\mapsto\gamma_b\left(t\right)$ is a solution of the system contained in $c=0$. If the constraint is of order n it can be generically computed by differentiating n times the constraint and solving the linear equation*

$$c^{(n)}=f^nc+ug\left(f^{n-1}c\right)=0 \tag{5.1}$$

A boundary arc is contained in

$$c=\dot{c}=\cdots=c^{(n-1)}=0 \tag{5.2}$$

and the constraints $\dot{c}=\cdots=c^{(n-1)}=0$ are called the secondary constraints.

We denote by u_b the feedback control $\frac{-f^nc}{g(f^{n-1}c)}$ which allows to remain in the constrains set.

Let's now formulate the Maurer minimum principle [12]:

Assumption 6 (General assumption)**.** *We assume that the following conditions hold on a boundary arc $s\mapsto\gamma_b\left(s\right)$, $s\in[0,t]$:*

(H$_6$): $g\left(f^{n-1}c\right)|_{\gamma_b}\neq 0$ *(n being the order)*

(H$_7$): $|u_b\left(t\right)|<1$ *i.e. the boundary feedback control is admissible and not saturating.*

Necessary conditions. Define the Hamiltonian by

$$H\left(q,u,p,\eta\right)=\langle p,f+ug\rangle+\eta c \tag{5.3}$$

where η is a Lagrange multiplier of the constraint set. The necessary conditions of the minimum principle are the following:

Condition 1.

(1) *There exists $\eta\left(t\right)\geq 0$, a real number $\eta_0\geq 0$ and δ such that the adjoint (row) vector satisfies*

$$\dot{p}=-p\left(\frac{\partial f}{\partial q}+u\frac{\partial g}{\partial q}\right)-\eta\frac{\partial c}{\partial q} \tag{5.4}$$

$$p\left(T\right)=\eta_0\frac{\partial\Phi}{\partial q}\left(q\left(T\right)\right)+\delta\frac{\partial G}{\partial q}\left(x\left(T\right)\right) \tag{5.5}$$

(2) *The function $\eta\left(t\right)$ satisfies $\eta\left(t\right)c\left(q\left(t\right)\right)=0$ $\forall t\in[0,T]$ and is continuous on the interior of the boundary arc.*

(3) *The jump condition at a contact point or a junction time t_1 is*

$$p\left(t_1^+\right)=p\left(t_1^-\right)-\nu_1\frac{\partial c}{\partial q}\left(q\left(t_1\right)\right),\ \nu_1\geq 0 \tag{5.6}$$

(4) *The optimal control $u\left(t\right)$ minimizes the Hamiltonian, i.e.*

$$H\left(q\left(t\right),u\left(t\right),p\left(t\right),\eta\left(t\right)\right)=\min_{|u|\leq 1}H\left(q\left(t\right),u,p\left(t\right),\eta\left(t\right)\right) \tag{5.7}$$

Remark 1. *In this minimum principle, only the constraint c is penalized in H; others choices are possible using the secondary constraints, see* [8, 13].

Remark 2. *There exist a general minimum principle without assumption (*H_6*), see for instance* [9] *where the adjoint equation (5.4) takes the form*

$$p(t) = -\int p(s)\left(\frac{\partial f}{\partial q} + u\frac{\partial g}{\partial q}\right) ds - \int \frac{\partial c}{\partial q} d\mu_i$$

where μ_i *is a measure supported on the set* $c = 0$*. Our principle is much more precise because from (5.4) the measure is of the form*

$$d\mu_i = \eta(t)\, dt$$

where η *is* $\mathcal{C}^0$*. This additional regularity comes from assumption (*H_6*) and at non generic point where* $g\left(f^{n-1}c\right)$ *vanishes* η *can blow up.*

The case where T is not fixed can be deduced from the case where T is fixed. We introduce a new variable $z = T$ and the system

$$\frac{dt}{ds} = z$$
$$\frac{dq}{ds} = (f(q) + ug(q))\, z$$
$$\frac{dz}{ds} = 0$$

We have $s = \frac{t}{T}$ and the trajectories are parametrized by $s \in [0,1]$. The new transfer time is 1.

An important research program is to analyze the solutions of the minimum principle with constraints. This analysis is outlined in [12]. An interesting point of view is to analyze the open loop solution deduced from the problem without constraints by analyzing the bifurcation of an unconstrained optimal solution when the constraint $c(q) \leq \alpha$ becomes active.

Next we adopt a different approach based on the evaluation of the small time reachable set near the constraints. It will provide necessary and sufficient optimality conditions.

5.2. **A direct approach.**

5.2.1. *Order of the constraint.* Consider in the shuttle problem the constraint on thermal flux

$$c_1 = C_q\sqrt{\rho}v^3 \leq \alpha,\ \rho = \rho_0 \mathrm{e}^{-\beta r}$$

and $\dot{c}_1 = \varphi_1(r,v) + \varphi_2(r,v)\sin\gamma = 0$ is a secondary constraint. Moreover $\ddot{c}_1 = \varphi_3(r,v,\gamma) + u\varphi_4(r,v)\cos\gamma$ where $\varphi_4(r,v) = -\overline{k}C_q\rho^{\frac{3}{2}}\left(3gv^3 + \frac{\beta}{2}v^5\right) \neq 0$.

Similarly for the normal acceleration $c_2 = \gamma_{n_0}\rho v^2$ we get

$$\dot{c}_2 = -\gamma_{n_0}\left(2k\rho^2v^3 + \left(\beta\rho v^3 + 2g\rho v\right)\sin\gamma\right)$$

i.e. $\dot{c}_2 = \varphi_5(r,v) + \varphi_6(r,v)\sin\gamma = 0$, $\varphi_6(r,v) = -\gamma_{n_0}\rho\left(\beta v^3 + 2gv\right)$ is a secondary constraint and $\ddot{c}_2 = \varphi_7(r,v,\gamma) + u\varphi_8(r,v)\cos\gamma$ with $\varphi_8(r,v) = -\overline{k}\gamma_{n_0}\rho^2\left(\beta v^4 + 2gv^2\right) \neq 0$. Hence we prove:

Lemma 6. *For the space shuttle if* $\cos\gamma \neq 0$ *the constraints on the thermal flux and on the normal acceleration are of order* 2 *and (*H_6*) is satisfied along a boundary arc.*

5.2.2. *Evaluation of the small time reachable set for the constrained system.* It is based on the following normal form. Consider system $\dot{q} = X + uY$, $|u| \leq 1$, $q = (x,y,z)$ and the constraint $c(q) \leq \alpha$, $\alpha \simeq 0$. We compute a normal form in the geometric configuration of the shuttle system near $q_0 \in c(q) = 0$.

- **Normalization 1.** We let $q_0 = 0$. Assume $Y(0) \neq 0$. Hence Y can be identified locally to $\frac{\partial}{\partial z}$. Diffeomorphisms preserving Y are $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ where $\frac{\partial \Phi_1}{\partial z} = \frac{\partial \Phi_2}{\partial z} = 0$ and $\frac{\partial \Phi_3}{\partial z} = 1$. In our problem the constraint is of order 2, hence $Yc = 0$ near 0 and Y is tangent to all the surfaces $c = \alpha$. Hence $\frac{\partial c}{\partial z} = 0$.
- **Normalization 2.** Since c is not depending upon z using a diffeomorphism preserving $Y \sim \frac{\partial}{\partial z}$ we can normalize c to $c(x,y) = x$. The system can be written

$$\dot{x} = X_1(q)$$
$$\dot{y} = X_2(q)$$
$$\dot{z} = X_3(q) + u$$

and $c = x$. The secondary constraint is $\dot{x} = 0$ and we assume that $x = \dot{x} = 0$ is an arc σ passing through $q_0 = 0$. If we keep the affine approximation sufficient for our analysis we obtain a system which can be written

$$\begin{aligned}\dot{x} &= a_1 x + a_2 y + a_3 z \\ \dot{y} &= b_0 + b_1 x + b_2 y + b_3 z \\ \dot{z} &= c_0 + c_1 x + c_2 y + c_3 z + u\end{aligned}$$

where σ is approximated by the straight line $x = a_2 y + a_3 z$. If $b_0 \neq 0$ (generic case) we can assume $b_0 = 1$.

- **Normalization 3.** Changing z into $-z$ and u into $-u$ if necessary and using a transformation of the form $Z = \alpha y + z$ one can identify σ to $x = z = 0$ and the system can be written

$$\begin{aligned}\dot{x} &= a_1 x + a_3 z \\ \dot{y} &= 1 + b_1 x + b_2 y + b_3 z \\ \dot{z} &= c_0 + c_1 x + c_2 y + c_3 z + u\end{aligned} \tag{5.8}$$

where $a_3 > 0$. If moreover the boundary arc is admissible and not saturating (assumption (H_7)) we have the condition $|c_0| < 1$.

Theorem 2. *Consider the problem of time minimization in* $\dot{q} = \overline{X}(q) + u\overline{Y}(q)$, $q \in \mathbb{R}^3$ *subject to* $c(q) \leq 0$. *Let* $q_0 \in \{c = 0\}$ *and assume the following:*

1. *Near* q_0, $\left[\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right] \in \operatorname{Span}\left\{\overline{Y}, \left[\overline{X}, \overline{Y}\right]\right\}$
2. $\overline{X}, \overline{Y}, \left[\overline{X}, \overline{Y}\right]$ *are linearly independent at* q_0 *and*
$$\left[\overline{X}, \left[\overline{X}, \overline{Y}\right]\right](q_0) = a\overline{X}(q_0) + b\overline{Y}(q_0) + c\left[\overline{X}, \overline{Y}\right](q_0)$$
with $a < 0$
3. *The constraint* $c = 0$ *is of order* 2 *and assumption* (H_6) *and* (H_7) *are satisfied at* q_0

then the boundary arc through q_0 *is small–time optimal if and only if* $\gamma_-(q_0)$ *is contained in the domain* $c \geq 0$.

Proof. From lemma 4, we know that each small time reachable point from q_0 can be reached by an arc $\gamma_+\gamma_-\gamma_+$ and $\gamma_-\gamma_+\gamma_-$ and from theorem 1 we know that the small time optimal arc is of the form $\gamma_-\gamma_+\gamma_-$ for the unconstrained system.

Let the constrained system written as (5.8) in the normal coordinates where $q_0 = 0$ and the boundary arc $\gamma_b(t)$ is identified to $(0, t, 0)$. Let $B = \gamma_b(t)$, $t > 0$ small enough. Let $u = +1$ or $u = -1$. For a trajectory with $q(0) = 0$ we have the following approximations:

$$\begin{aligned}z(t) &= (c_0 + u)\,t + o(t) \\ x(t) &= a_3 (c_0 + u)\frac{t^2}{2} + o(t)\end{aligned}$$

Hence the projections of the arcs $\gamma_+\gamma_-\gamma_+$ and $\gamma_-\gamma_+\gamma_-$ joining 0 to B in the plane (x, z) are loops denoted $\widetilde{\gamma}_+\widetilde{\gamma}_-\widetilde{\gamma}_+$ and $\widetilde{\gamma}_-\widetilde{\gamma}_+\widetilde{\gamma}_-$ and are represented on figure 4.

In particular, we proved the following.

Lemma 7. *The loops* $\widetilde{\gamma}_-\widetilde{\gamma}_+\widetilde{\gamma}_-$ *(resp.* $\widetilde{\gamma}_+\widetilde{\gamma}_-\widetilde{\gamma}_+$*) are contained in the domain* $x < 0$ *(resp.* $x > 0$*).*

We can now end the proof of the theorem (the assertions concern system $\left(\overline{X}, \overline{Y}\right)$).

If the arc $\gamma_-\gamma_+\gamma_-$ to join 0 to B is contained in the domain $c \leq 0$ it is time minimal and the boundary arc is not optimal. If the arc $\gamma_-\gamma_+\gamma_-$ is contained in $c \geq 0$ then we can join 0 to B by an arc $\gamma_+\gamma_-\gamma_+$ in $c \leq 0$. But the analysis of section 4 replacing $\min t$ by $\max t$ shows that such an arc is time maximal. Hence a bang–bang arc $\gamma_+\gamma_-\gamma_+$ in the domain $c \leq 0$ joining 0 to B cannot be optimal. Then the boundary arc γ_b is optimal. □

Moreover

Corollary 2. *If a boundary arc* γ_b *is small time optimal then there exist optimal trajectories of the form* $\gamma_-\gamma_+\gamma_b\gamma_+\gamma_-$.

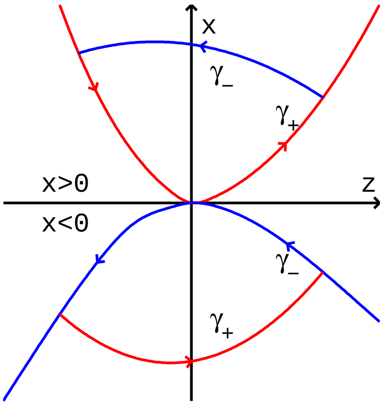


FIGURE 4. Projection of the arcs $\gamma_+\gamma_-\gamma_+$ and $\gamma_-\gamma_+\gamma_-$

5.2.3. *Application to the shuttle.* For the shuttle we have $a < 0$, so we have to consider loops $\gamma_-\gamma_+\gamma_-$ where γ_- corresponds to $\cos\mu = 0$ or $\cos\mu = -1$. From the computations of section 5.2.1 we have for both constraints c_1, c_2:

$$\ddot{c}_i = \Phi\left(r, v, \gamma\right) + u\cos\gamma\overline{\Phi}\left(r, v\right)$$

where $\overline{\Phi} < 0$. Hence $\ddot{c}_i$ is minimal when $\mu = 0\,°$ i.e. $u = +1$. Assume that the parameters of the problem are such that assumption (H$_7$) is satisfied. Then the arcs γ_- through the boundary points are contained in the non admissible domain and the boundary arc is optimal.

6. Conclusion

We have outlined the geometric research program to analyze the optimal control of the atmospheric arc for the space shuttle. Our tools are necessary optimality conditions and evaluation of the small time reachable set. Near the constraints the evaluation is related to the classification of pairs of vector fields near a surface. This problem is common to several problems met in optimal control: classification of extremals near the switching surface, optimal control with targets and so on...

References

[1] O. Bolza, *Calculus of variations*, Chelsea, 1973
[2] F. Bonnans and G. Launay, *Large scale direct optimal control applied to the re–entry problem*, Journal of guidance, control and dynamics, Vol. 21, N. 6, 1998, pp.996–1000
[3] B. Bonnard, G. Launay, *Time minimal control of batch reactors*, COCV–ESAIM, Vol. 3, 1998, pp.407–467
[4] B. Bonnard and I. Kupka, *Théorie des singularités de l'application entrée/sortie et optimalité des singulières*, Forum Math., 5, 1993, pp.11–159
[5] A. E. Bryson and Y. C. Ho, *Applied optimal control*, Hemisphere Pub. Corporation, 1975
[6] CNES, Mécanique spatiale, tome 1/2, Cépaduès Eds, 1995
[7] J.-M. Coron and L. Praly, *Guidage en rentrée atmosphérique*, Rapports 4/5, CNES, Octobre 2000
[8] D. H. Jacobson and al., *New necessary conditions of optimality for control problems with state–variable inequality constraints*, J. Math and Appl., 35, 1971, pp.255–284
[9] A. D. Ioffe and V. M. Tikhomirov, *Theory of extremal problems*, North Holland, 1979
[10] I. Kupka, *Geometric theory of extremals in optimal control problems*, TAMS 299, 1973, pp. 225–243
[11] A. J. Krener and H. Schättler, *The structure of small time reachable sets in small dimensions*, SIAM J. on Control and Op. 27, 1989, pp. 120–147
[12] H. Maurer, *On optimal control problems with bounded state variables and control appearing linearly* SIAM J. Control Optimization 15, 1977, pp. 345–362
[13] V. Pontriaguine and al., *Méthodes mathématiques des processus optimaux*, Ed. MIR, 1974
[14] J. H. Schättler, *The local structure of time optimal trajectories in dim. 3 under generic conditions*, SIAM J. Control Optimization, 26, No. 4, 1988, pp. 899–918
[15] H. Sussmann, *The structure of time–optimal trajectories for single input systems in the plane: the C^∞ non singular case*, SIAM J. Control Opt. 25, 1987, pp. 856–905

NUMERICAL INTEGRATION OF DIFFERENTIAL EQUATIONS IN THE PRESENCE OF FIRST INTEGRALS: OBSERVER METHOD

E. BUSVELLE, R. KHARAB, A. J. MACIEJEWSKI, J.-M. STRELCYN

Abstract. We introduce a simple and powerful procedure – the observer method – in order to obtain a reliable method of numerical integration over an arbitrary long interval of time for systems of ordinary differential equations having first integrals. This aim is achieved by a modification of the original system such that the level manifold of the first integrals becomes a local attractor. We provide the theoretical justification of this procedure. We report many tests and examples dealing with a large spectrum of systems with different dynamical behavior. The comparison with standard and symplectic methods of integration is also provided.

1. Introduction

Frequently we need to integrate numerically a system of ordinary differential equations (ODE's) having some first integrals or an evolutionary partial differential equation admitting some conservation laws. In what follows we will consider exclusively the ODE's case, but undoubtedly, our method can also be applied to the case of partial differential equations.

Many examples of systems of ODE's having first integrals can be found in classical mechanics, where Hamiltonian equations are of this type, always having the Hamiltonian as a first integral. The Newtonian many bodies problem with at least three bodies is an example of a non-integrable system having many first integrals.

Another class of systems of ODE's that always have first integrals is the class of the so called Euler equations on Lie algebra ([32, 33, 63, 65–70]). The prototype of equations of this class is the system of the standard Euler equations of motion of rigid body with fixed point in absence of gravity. When the Lie algebra is $\mathbb{R}^{2n}$ endowed with trivial Lie bracket, we recover the class of usual Hamiltonian systems.

The Euler equations on Lie algebras are particularly interesting, because, as it was discovered recently, many systems of ODE's governing the motion of rigid bodies in different circumstances, described already by L. Euler, D. Poisson, G. R. Kirchoff, V. A. Steklov and some others, belong to this class ([48, 65–70]). Many physically interesting systems of Euler equations on Lie algebra were recently described and studied by O. I. Bogoyavlensky ([11–20]).

When the system of ODE's is not integrable, but admits some first integrals, its numerical study is particularly delicate. Indeed, when studying a non integrable system of ODE's without any known first integrals, except for standard precaution measures, we do not have any control on the numerical errors, especially in the chaotic regions. When first integrals exist, one can demand from the numerical integration procedure to preserve, at least, the first integrals.

It can be easily understood that this is really a very difficult problem if one sees the huge amount of literature devoted to this problem (see for example [6, 7, 10, 28, 29, 36, 38, 51, 57–61]) and references cited in the above works. Large parts of these papers were written by astronomers, because this problem was recognised for the first time when studying the Newtonian many bodies problem and in particular the Kepler problem.

The recent paper ([45]) proves that even for an integrable systems this problem can be extremely delicate.

The main aim of this paper is to introduce a simple numerical procedure, or rather a class of them, called by us *the observer method*, derived from the control theory (see for example [39]), which preserves very well the first integrals over arbitrary long integration time. This last property will be rigorously proven, at least for some range of parameters. The idea of this method consists of the perturbation of the original system in such a way that the surface of constant integrals value, where the interesting solution lie, is an attractor for the perturbed system.

After some preliminary numerical tests on the planar Kepler problem and the uncoupled harmonic oscillators the observer method is applied to the numerical study of the following four examples. The elliptic orbits with big eccentricities of the spatial Kepler problem, the Gavrilov-Shil'nikov system ([34]), the so-called non-Manakov case of Euler equations on Lie algebras *so*(4). ([1,11,12,32,40,41,49,69,80,82]), and finally the geodesic flow on compact orientable surface with constant curvature -1 ([22]). These examples represent four main different types of dynamical behaviour: stable completely integrable systems, completely integrable unstable systems, systems with large regions of chaotic behaviour presenting the so called 'coexistence' (see for example [78]) and the Anosov systems ([73]). It is important to note that the results of numerical computations reported in Sections 4–8 clearly indicate that the use of the observer method increases the reliability of the numerical integration of the considered system on the level manifold of the first integrals. As was noted by M. Balabane, one can drastically simplify the observer method reducing it to the method called by us *the penalty method*. Although the penalty method is, from a computational point of view, substantially simpler than the observer method, its field of applications is limited and more careful numerical investigations indicate that contrarily to the observer method, it does not preserve the first integrals as well as the observer method. Seeing the simplicity of the penalty method, we are very astonished by its absence in the literature. We know only the paper of [61], where some vaguely related numerical procedure is described. See also paper of [36]. Let us also note the evident relation between the problem studied in this paper and the problem of physical realisation of constraints in the classical mechanics (see for example [2]. The paper is organised as follows: in Section 2 we describe the observer and penalty methods. In this Section we also briefly discuss the case of so called *partial first integrals* known also under the name of *invariant relations* (see for example [44,55,79]), which generalise the globally defined first integrals. In Section 3 we relate rigorously the observer method to the numerical integration of ODE's. In Section 4 the preliminary tests on observer method is provided. In Section 5 we study numerically the spatial Kepler problem, while in Section 6 the Gavrilov-Shil'nikov system is considered. In Section 7, we study the Euler equations on the Lie algebra *so*(4), and finally in Section 8 the geodesic flow on the surface of constant negative curvature. In sections 4–8 we report many comparisons of the observer method with different standard Runge-Kutta methods, extrapolation method, penalty method and different symplectic methods of integration.

2. The Observer Method

Let us consider a system of ODE's written in vector form

$$\frac{dx}{dt} = F(x), \tag{2.1}$$

where $F \in C^1(U, \mathbb{R}^p)$ and U is an open subset of $\mathbb{R}^p$, together with the initial condition

$$x(0) = x_0, \tag{2.2}$$

for some $x_0 \in U$. We will call the problem (2.1)–(2.2)—the problem (P).

Let us suppose that $H_1, \ldots, H_q \in C^1(U)$, $1 \leq q < p$, are the first integrals of the system (2.1). Let us denote

$$H = \begin{pmatrix} H_1 \\ \vdots \\ H_q \end{pmatrix}$$

$H \in C^1(U, \mathbb{R}^q)$, and by D_H the derivative of H i.e., its Jacobian matrix.

Let $\|\cdot\|$ denotes the Euclidean norm. Let us consider $x_0 \in U$. Let us denote $H_0 = H(x_0)$. Let us define the level set

$$\Gamma = \{x \in U; H(x) = H_0\}. \tag{2.3}$$

We will suppose that the integrals $H_1, \ldots, H_q$ are functionally independent on Γ, i.e. for $x \in \Gamma$, $\mathrm{rank}(D_H(x)) = q$. Under this assumption, Γ is a smooth submanifold of U. In general Γ admits many connected components.

Let us underline that in concrete examples it can be difficult to verify if the integrals $H_1, \ldots, H_q$ are functionally independent on a given level set Γ (see for example [4, 5, 74, 75]). In what concerns the practical side of the observer method, this verification can be neglected.

To carry forth successfully the theoretical considerations of this paper we will suppose Γ compact.

Now, from the implicit function theorem one easily deduce that for sufficiently small ϵ, $0 < \epsilon \leq \epsilon_0$

$$\tilde{\Gamma}_\varepsilon = \left\{ x \in U : r\left(x, \Gamma\right) \overset{def}{=} \|H_0 - H\left(x\right)\| \leq \varepsilon \right\} \tag{2.4}$$

is also a compact subset of U on which the integrals $H_1, \ldots, H_q$ are functionally independent.

The above compactness condition occurs quite frequently and, in particular, this is the case of the Kepler problem for strictly negative energies and of the Euler equations on $so(n)$, $n \geq 3$.

It can be easily seen that if A is a $q \times p$ matrix, $q \leq p$, then the $q \times q$ symmetric matrix AA^T is invertible if and only if A is of maximal rank, i.e. of rank q, where A^T denotes the matrix transposed of A. Thus for every $\xi \in \Gamma$, $D_H(\xi) D_H^T(\xi)$ is invertible. Let us also note that if the matrix AA^T is invertible, then the matrix $A^T(AA^T)^{-1}$ is nothing else but the Moore-Penrose generalised inverse of A.

As the submanifold $\Gamma = \{x \in U; H(x) = H_0\}$ is compact, then it is well known (see for example [43]) that the solution $x = x(t)$ of the problem (P) is defined for all $t \in \mathbb{R}$, when $x_0 \in \Gamma$. Let us consider such a solution. For every $t \in \mathbb{R}$ one has $H(x(t)) = H(x_0) = H_0$.

Together with the problem (P), let us consider the following perturbed initial value problem (called problem (PP)):

$$\begin{array}{rcl} \frac{d\xi}{dt} & = & F(\xi) + D_H^T(\xi) \left[D_H(\xi) D_H^T(\xi) \right]^{-1} K(\xi) \left(H_0 - H(\xi) \right) \\ \xi(t_0) & = & \xi_0 \end{array} \tag{2.5}$$

where $\xi_0 \in U$ and $K(\xi)$ is a diagonal matrix given by

$$K(\xi) = \begin{array}{cccc} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & \theta_q \end{array} \tag{2.6}$$

where $\theta_1, \ldots, \theta_q$ are continuous functions on Γ_ϵ such that

$$\eta = \min_{1 \leq i \leq q} \min_{\xi \in \Gamma_\epsilon} \theta_i(\xi) > 0 \tag{2.7}$$

On the submanifold Γ, where $H(\xi) = H_0$, the system of ODE's (2.5) reduces to the initial system (2.1). Consequently, the orbits of both systems coincide on Γ.

We will show now that Γ is the global attractor of the system (2.5) restricted to the Γ_ϵ with an exponential rate of convergence of orbits towards Γ which is exactly controlled. Our main point is contained in the following very simple lemma.

Main Lemma

Let $\xi_0 \in \Gamma_\epsilon$. Then the unique solution of problem (PP) is defined for every $t \geq 0$ and for such t

$$r(\xi(t), \Gamma) \leq r(\xi_0, \Gamma) e^{-\eta t} \tag{2.8}$$

where $r(\xi, \Gamma)$ is defined in (2.4). Consequently $\lim_{t \to \infty} H(\xi_t) = H_0$.

Proof

If $\xi_0 \in \Gamma$, then $r(\xi(t), \Gamma) = 0$ and (2.8) is true. Now, let $\xi_0 \in \Gamma_\epsilon$ and $\xi_0 \notin \Gamma$. We will prove both parts of lemma simultaneously. To prove that every solution $\xi = \{\xi(t)\}$ of the problem (PP) is defined for every $t > 0$, because Γ_ϵ is compact, it is sufficient to prove that for every $T > 0$ such that the solution $\{\xi(t)\}_{0 \leq t \leq T}$ is defined, $\{\xi(t)\}_{0 \leq t \leq T} \subset \Gamma_\epsilon$ (see Chapter II of [43]). Thus, consider the solution $\{\xi(t)\}_{0 \leq t \leq T}$ and let $\epsilon_t = H_0 - H(\xi(t))$. Then, taking into account (2.5), as $D_H F$ is equal to zero because H is a first integral of the initial non-perturbed equation (2.1), one can write:

$$\begin{array}{rcl} \frac{d}{dt} \|\epsilon_t\|^2 & = & -2\epsilon_t^T D_H(\xi(t)) \frac{d\xi(t)}{dt} \\ & = & -2\epsilon_t^T \left[D_H(\xi(t)) F(\xi(t)) + K(\xi(t)) \left(H_0 - H(\xi(t)) \right) \right] \\ & = & -2\epsilon_t^T K(\xi(t)) \epsilon_t \leq -2\eta \|\epsilon_t\|^2 \end{array}$$

Therefore

$$\|\epsilon_t\|^2 \leq \|\epsilon_0\|^2 e^{-2\eta t}$$

and then the inequality (2.8) holds for $0 \leq t \leq T$. Consequently for such t, it was shown that $\xi(t) \in \Gamma_\epsilon$ and this finishes the proof.

Let us suppose that instead of integrating numerically on Γ the problem (P), we integrate numerically the perturbed problem (PP). Taking into account the Main Lemma we can predict intuitively the behaviour of first integrals thanks to the correction term, and hope that if the computed orbits of the perturbed system leave Γ, then the correction term pushes them toward Γ. Hence we can expect to bound the error of the computed first integrals. In the next section we will establish rigorously the above qualitative feature. The method of control of integration errors through the use of problem (PP), with perhaps other matrices $K(\xi)$ than (2.6), is *the observer method*.

In what follows we will use two kinds of matrix $K(\xi)$.

– The first one when $\theta_1, \ldots, \theta_q$ are constants. In this case we will say that one applies *the simple observer.*

– The second one when

$$\theta_i = \alpha_i \| D_H^T(\xi)[D_H(\xi)D_H^T(\xi)]^{-1} L(\beta)\|^{-1},$$

$1 \leq i \leq q$, where α_i, β_i are some strictly positive numbers, and $L(\beta)$ is the $q \times q$ diagonal matrix with $\beta_1, \ldots, \beta_q$ on the diagonal. In this case the obtained method will be called *the normalised observer.*

As it will be seen in Section 5 when studying the spatial Kepler problem, the simple observer fails for the elliptic orbits with big eccentricities, while the normalised observer works very well. What concerns practical computations with the simple observer method let us notice that for its application it is enough to solve a system of q linear equations for one calculation of right hand side of the equation (2.5). In fact, let us denote

$$P(\xi) = D_H^T(\xi)[D_H(\xi)D_H^T(\xi)]^{-1}K(\xi)(H_0 - H(\xi))$$

the observer perturbation term (see (2.5)). It can be written equivalently as

$$P(\xi) = D_H^T(\xi)y(\xi),$$

where $y(\xi)$ is the unique solution of:

$$D_H(\xi)D_H^T(\xi)y(\xi) = K(\xi)(H_0 - H(\xi)).$$

Let us pause now on the nature of the perturbation term in (2.5). For $x_0 \in \Gamma_\epsilon$, let us note by $\Gamma(x_0)$ the manifold $\{\xi \in \Gamma_\epsilon;\ H(\xi) = H(x_0)\}$. For $\xi \in \Gamma(x_0)$, all vectors orthogonal to $\Gamma(x_0)$ at ξ are of the form $D_H^T(\xi)y$, where $y \in \mathbb{R}^q$. Thus, the perturbation term in (2.5) is orthogonal to the level manifolds $\Gamma(x_0)$. On the other hand it is clear that only such perturbations can be used to improve the reliability of the numerical integration on Γ of the system (2.1) using the observer approach. Now, let us consider for the class of perturbations of the form

$$D_H^T(\xi)Y(\xi)(H_0 - H(\xi)), \tag{2.9}$$

where Y is a $q \times q$ matrix function defined on Γ_ϵ. It is easy to see that the estimation (2.8) can be obtained along the lines of the proof of the Main Lemma precisely in the case when

$$Y(\xi) = \left[D_H(\xi)D_H^T(\xi)\right]^{-1} K(\xi),$$

where $K(\xi)$ is an invertible, not necessarily diagonal, $q \times q$ matrix depending on $\xi \in \Gamma_\epsilon$ satisfying for some $\eta > 0$:

$$\inf_{\xi \in \Gamma_\epsilon} \|K(\xi)y\| \geq \eta\|y\|$$

for every $y \in \mathbb{R}^q$, i.e.

$$\sup_{\xi \in \Gamma_\epsilon} \|K(\xi)^{-1}\| \leq \frac{1}{\eta}.$$

Finally, let us consider the case when all functions under considerations are real analytic. In this case it is easy to see that for every point $x_0 \in \Gamma$ (see (2.3)) there exists a neighbourhood $V \subset \mathbb{R}^p$ of x_0 such that for every vector function $\Phi : V \to \mathbb{R}^p$ vanishing on Γ one has

$$\Phi(\xi) = Y(\xi)(H_0 - H(\xi)),$$

where $\xi \in V$ and Y is a real analytic $q \times q$ matrix function defined on V. Thus, at least, in the real analytic case the form (2.9) of the perturbation is unavoidable.

Now let us drastically simplify the problem (PP) by considering the following initial value problem

$$\left\{ \begin{array}{lcl} \frac{d\xi}{dt} & = & F(\xi) + D_H^T(\xi)K(\xi)(H_0 - H(\xi)) \\ \xi(0) & = & \xi_0 \end{array} \right. \tag{2.10}$$

At first glance, the qualitative features of the orbits (2.10) are similar to those of the problem (PP).

Now, in the same way as in the case of the Main Lemma, one can prove for this initial value problem that for all $t \geq 0$ for which the solution $\xi(t)$ is defined one has

$$\frac{d}{dt}\|\epsilon_t\|^2 = -2\epsilon_t^T D_H(\xi(t))D_H^T(\xi(t))K(\xi(t))\epsilon_t, \tag{2.11}$$

where recall $\epsilon_t = H_0 - H(\xi(t))$. Unfortunately, in full generality, it is not true that the right side of (2.11) is always a strictly negative number. Indeed, the quadratic form $Q(\epsilon) = \epsilon^T AK(\xi)\epsilon$ on $\mathbb{R}^q$, where A is a symmetric positively defined matrix, is not necessarily positively defined for $q \geq 2$. Consequently, for the initial value problem (2.10) with $q > 1$, one cannot prove the statement of the Main Lemma.

Let us consider now the particular case when $K(\xi)$ is the identity matrix. The symmetric matrix $D_H(\xi)D_H^T(\xi)$ is strictly positively defined. Let us denote by $\lambda(\xi)$ its smallest eigenvalue, $\lambda(\xi) > 0$. In this case (2.11) implies that

$$r(\xi(t),\Gamma) \leq r(\xi_0,\Gamma)e^{-\int_0^t \lambda(\xi(s))ds}$$

From the compacity of Γ_ϵ, it follows that $\int_0^\infty \lambda(\xi(s))ds = +\infty$. Now, one can deduce that the solution of the problem (2.10) is defined for all $t \geq 0$ and that for such t, $\xi(t) \in \Gamma_\epsilon$. Thus the similarity in qualitative behaviour of the problem (PP) and of the problem (2.10) is formally established when $K(\xi)$ is the identity matrix. Clearly, the same remains true when $K(\xi)$ is proportional to a sufficiently small perturbation of the identity matrix. Like the problem (PP), the problem (2.10) can also be used for the numerical integration of the problem (P) with controlled errors on first integrals. This is *the penalty method.* Like for the observer method, we will also talk about the simple penalty method and the normalised penalty method.

Let us underline that comparing the observer and penalty methods, the penalty method is numerically simpler than the observer method. But in the penalty method the control of error on the first integrals is not so good as in the observer method. This point will be verified on examples in Sections 5 and 6.

Finally let us discuss shortly the case of partial first integrals. For more details see the forthcoming paper of Maciejewski and Strelcyn ([56]). Many examples of partial first integrals arises in the study of the Euler-Poisson equations of the motion of a heavy rigid body with fixed point (see for example [2, 52, 55]). Many other examples can be found for example in [11–18, 44, 79]. Let us consider the system of ODE's (2.1). Let $h \in C^1(U)$ be a function which is not constant on any open subset of U and such that its zero level $\Gamma_0 = \{x \in U : h(x) = 0\}$ is not empty. A function h is a *partial first integral* of the system (2.1) if there exist a function $l \in C^1(U)$ such that for all $x \in U$ one has

$$D_h(x)F(x) = l(x)h(x).$$

From from our differentiability assumptions one easily deduce that the zero level set Γ_0 is an invariant subset of the system of ODE's (2.1). This means that Γ_0 is foliated by the orbits of the system of ODE's (2.1); i.e. that if at some instance an orbit of the system (2.1) crosses the set Γ_0 then up to its exit from U, its remains in Γ_0. When $l \equiv 0$, we recover the usual notion of first integral. In many examples much more intricate situations occurs when the invariant set is not of codimension one. To describe this let us introduce the notion of *q-partial first integrals.* Let us suppose now that we have q, $1 \leq q < p$, functions $H_1, \ldots, H_q$ which are not constant on any open subset of U and such that its zero level $\Gamma_0 = \{x \in U : H_i(x) = 0, 1 \leq i \leq q\}$ is not empty. We assume that the functions $H_1, \ldots, H_q$ are functionally independent on Γ_0. Let us suppose that there exist a $q \times q$ matrix function L defined on U with entries in $C^1(U)$ such that

$$D_H(x)F(x) = L(x)H(x), \tag{2.12}$$

where $H(x)$ is the column vector $(H_1, \ldots, H_q)$. In this case the vector function H is called *q-partial first integral.* Like in the partial first integrals, here also the zero level set Γ_0 is an invariant subset for the system of ODE's (2.1). Let us underline that in general the functions $H_1, \ldots, H_q$ are not necessary the partial first integrals. Examples of such q-partial first integrals can be found, for example, in the problem of the motion of the rigid body with a fixed point (see [52]). Like in the case of first integrals, let us suppose that the set Γ_0 is compact. We would like to have a reliable procedure to integrate numerically the initial value problem (P) on the invariant set Γ_0. For this aim let us define the initial value problem (PPP):

$$\begin{array}{rcl} \frac{d\xi}{dt} & = & F(\xi) - D_H^T(\xi)\left[D_H(\xi)D_H^T(\xi)\right]^{-1}[K(\xi) - L(\xi)]H(\xi) \\ \xi(t_0) & = & \xi_0 \end{array}$$

where $\xi_0 \in \Gamma_\epsilon$ which is defined by (2.4) with $H_0 = 0$, and where the $q \times q$ matrix K satisfies (2.6) and (2.7). In the case when $H_1, \ldots, H_q$ are the first integrals, i.e. when the matrix $L(x) \equiv 0$, we recover the problem (PP) Taking into account (2.12) it is very easy to see that for the problem (PPP), the Main Lemma and its proof remain valid word for word. Thus also in the case of partial first integrals one can hope to use the perturbed problem (PPP) to the reliable integration of the system (2.1) on the invariant manifold Γ_0.

3. The Observer Method and Numerical Integration of ODE's

To integrate numerically the problem (PP), as the system (2.5) is an autonomous one, we will use an explicit one-step method

$$\xi_{j+1} = \xi_j + h\phi(\xi_j, h) \tag{3.1}$$

with *increment function* ϕ

As usual, we will suppose that this method is at least of *order one*, i.e. that for every $\xi_0 \in U$ and $t \geq 0$ such that the solution ξ of the problem (PP) is defined on $[0, t+1]$, there exists a constant $C(\xi_0, t) > 0$, such that for every $t \geq 0$ and for every h, $0 < h \leq 1$, one has

$$\|\xi(t+h) - \xi(t) - h\phi(\xi(t), h)\| \leq C(\xi_0, t)h. \tag{3.2}$$

We suppose, of course, the method to be *consistent*, i.e. $\phi(\xi, 0) = \hat{F}(\xi)$, where by $\hat{F}$ we denote the right hand side of (2.5) (see [50] for more details).

Now, as it is the case for all standard explicit one-step methods, like Euler and Runge-Kutta method, we will suppose that for any compact subset $Q \subset U$,

$$\sup_{\xi \in Q} \sup_{0 \leq h \leq 1} \|\phi(\xi, h)\| \overset{def}{=} \phi_Q < +\infty. \tag{3.3}$$

Moreover, as this is the case for the above methods, we will suppose that

$$\sup_{\xi_0 \in Q} \sup_{0 \leq h \leq 1} |C(\xi_0, h)| \overset{def}{=} C_Q < +\infty. \tag{3.4}$$

When the step $h > 0$ is fixed, instead of the orbits $x(t) \equiv \xi(t)$ with $x(0) = \xi(0) = \xi_0 \in \Gamma$ of the problem (PP) we compute numerically *the pseudo-orbit* $\{\xi_j\}$. In general, it is a very difficult problem to understand the relation between the true orbits and the related pseudo-orbits (see for example [8, 45]).

The following theorem makes rigorous the main feature of the observer method already indicated in Section 2

Theorem 3.1

Let us consider the initial problem (PP). Then for every $\epsilon > 0$, there exists h, $0 < h < 1$, small enough and η defined by (2.7) big enough such that if $\xi_0 \in \Gamma$ then for every $j \geq 0$, $\xi_j \in \Gamma_\epsilon$, i.e.

$$r(\xi_j, \Gamma) \overset{def}{=} \|H_0 - H(\xi_j)\| \leq \epsilon$$

where $\{\xi_j\}$ is defined by (3.1).

Proof

First, let us fix the notations and the parameters necessary for the proof.

For $a \in \mathbb{R}^p$, let us denote $d(a, \Gamma) = \inf_{b \in \Gamma} \|a - b\|$.

For $r > 0$, let us denote

$$U_r(\Gamma) = \{\xi \in \mathbb{R}^p; d(\xi, \Gamma) \leq r\}.$$

As Γ is compact, $U_r(\Gamma)$ is also compact.

From our assumption on the functional independence of integrals $H_1, \ldots, H_q$, it follows that for every ϵ, $0 < \epsilon < \epsilon_0$ (cf. the definition (2.4) of Γ_ϵ), there exist numbers δ_ϵ and D_ϵ, $0 < \delta_\epsilon \leq D_\epsilon$, such that

$$U_{\delta_\epsilon}(\Gamma) \subset \Gamma_\epsilon \subset U_{D_\epsilon}(\Gamma) \subset U_{2D_\epsilon}(\Gamma) \subset U. \tag{3.5}$$

Let us note (cf. (3.3) and (3.4))

$$\begin{aligned} \phi_\epsilon &= \phi_{\Gamma_\epsilon}, \\ L_\epsilon &= \sup_{\xi \in U_{2D_\epsilon}(\Gamma)} \|D_H(\xi)\|, \\ C_\epsilon &= C_{U_{2D_\epsilon}(\Gamma)}, \\ \hat{F}_\epsilon &= \sup_{\xi \in \Gamma_\epsilon} \|\hat{F}(\xi)\|, \end{aligned} \tag{3.6}$$

where $\hat{F}$ denote the right hand side of (2.5). Now let us choose h such that

$$0 < h \leq \min\left(\frac{\delta_\epsilon}{\phi_\epsilon}, \frac{\delta_\epsilon}{\hat{F}_\epsilon}, \frac{\epsilon}{2L_\epsilon C_\epsilon}\right). \tag{3.7}$$

For fixed h, let us choose η such that

$$e^{\eta h} \geq 2. \tag{3.8}$$

This condition is satisfied for $\eta h \geq 0.69$.

Let us return to the pseudo-orbit $\{\xi_j\}_{j\geq 0}$ defined by (3.1). Let us denote by $\tilde{\xi}_j$ the value at time h of the solution of the system (2.5) satisfying the initial value condition $\xi(0) = \xi_{j-1}$, $j \geq 1$. From the Main Lemma we know that if $\xi_{j-1} \in \Gamma_\epsilon$ then $\tilde{\xi}_j$ is well defined and $\tilde{\xi}_j \in \Gamma_\epsilon$.

The proof is by induction. For $j = 0$ from (3.1) one has that:

$$\xi_1 - \xi_0 = h\phi(\xi_0, h)$$

and thus from (3.7) one obtains that

$$\|\xi_1 - \xi_0\| \leq h\phi_\epsilon \leq \delta_\epsilon.$$

As $\xi_0 \in \Gamma$, from (3.5) one deduces then that $\xi_1 \in \Gamma_\epsilon$.

Now let us suppose that $\xi_k \in \Gamma_\epsilon$, for $0 \leq k \leq j$. By virtue of (3.1), exactly in the same way as for $j = 0$, we obtain that $\| \xi_{j+1} - \xi_j \| \leq \delta_\epsilon$. On the other hand from the mean value theorem and (3.7) one has

$$\| \tilde{\xi}_{j+1} - \xi_j \| \leq h\hat{F}_\epsilon \leq \delta_\epsilon$$

Thus, $\xi_{j+1}, \tilde{\xi}_{j+1} \in \overline{B}(\xi_j, \delta_\epsilon) = \{x \in \mathbb{R}^p; \| x - \xi_j \| \leq \delta_\epsilon\}$. As one supposes that $\xi_j \in \Gamma_\epsilon$, then from (3.5), $\xi_j \in U_{D_\epsilon(\Gamma)}$ and thus $\overline{B}(\xi_j, \delta_\epsilon) \subset U_{2D_\epsilon(\Gamma)} \subset U$. Taking into account the convexity of $\overline{B}(\xi_j, \delta_\epsilon)$ and (3.6) one has

$$\begin{array}{rcl} r(\xi_{j+1}, \Gamma) & = & \|H(\xi_{j+1}) - H_0\| \leq \|H(\tilde{\xi}_{j+1}) - H_0\| + \|H(\xi_{j+1}) - H(\tilde{\xi}_{j+1})\| \\ & \leq & r(\tilde{\xi}_{j+1}, \Gamma) + L_\epsilon\|\xi_{j+1} - \tilde{\xi}_{j+1}\|. \end{array}$$

Now, taking into account the Main Lemma, (3.1) and (3.2) one obtains that

$$\begin{array}{rcl} r(\xi_{j+1}, \Gamma) & \leq & e^{-\eta h} r(\xi_j, \Gamma) + L_\epsilon\|\tilde{\xi}_{j+1} - \xi_j - h\phi(\xi_j, h)\| \\ & \leq & e^{-\eta h} r(\xi_j, \Gamma) + L_\epsilon C_\epsilon h. \end{array}$$

This inequality remains valid if instead of j we write an arbitrary k, $0 \leq k \leq j$, because for such k, our inductive assumption asserts that $\xi_k \in \Gamma_\epsilon$.

Thus for every k, $0 \leq k \leq j$, one has

$$r(\xi_{k+1}, \Gamma) \leq e^{-\eta h} r(\xi_k, \Gamma) + L_\epsilon C_\epsilon h$$

Taking into account this inequality, (3.7) and (3.8) one obtains that

$$\begin{array}{rcl} r(\xi_{j+1}, \Gamma) & \leq & e^{-\eta h}[e^{-\eta h} r(\xi_{j-1}, \Gamma) + L_\epsilon C_\epsilon h] + L_\epsilon C_\epsilon h \\ & = & e^{-2\eta h} r(\xi_{j-1}, \Gamma) + e^{-\eta h} L_\epsilon C_\epsilon h + L_\epsilon C_\epsilon h \;\ldots\ldots\ldots\ldots \\ & \leq & (\sum_{p=0}^{j} e^{-p\eta h}) L_\epsilon C_\epsilon h \\ < \frac{L_\epsilon C_\epsilon h}{1 - e^{-\eta h}} \leq \epsilon. & & \end{array}$$

Thus $\xi_{j+1} \in \Gamma_\epsilon$ and our inductive proof is finish.

Let us note that as the proof of the theorem 3.1 is based only on the Main Lemma, this theorem remains also valid when instead of initial value problem (PP), one considers the problem (PPP) given by (2.14). Although we considered here only the explicit one step method, there is no doubt that similar results hold also for multi-step methods, explicit as well as implicit ones.

Let us note also that the condition (3.8) gives the first restriction on the size of η and thus on the matrix $K(\xi)$. The condition (3.8) is somewhat restrictive. In what follows, when applying the observer method to concrete examples, we will try to choose $\theta_1 \ldots, \theta_q$, in such a way to diminish the possible stiffness ([50]) of problem (PP) and to improve the quality of the numerical computations.

We will see in the examples below that even if ηh is small, the observer method can work very well. Finally, let us note that the explicit exact computations can be given in the case of simple harmonic oscillator defined by the Hamiltonian

$$H(p, q) = \frac{1}{2}(p^2 + q^2),$$

for Euler method with simple observer applied to the first integral H and its value H_0. These computations prove that for the sufficiently small step $h > 0$ and $\theta > 0$ such that $2h < \theta$ following limit:

$$\lim_{t\to\infty} H(\xi_t) = R(h, \theta) > H_0$$

exists and that

$$\lim_{h\to 0} R(h,\theta) = H_0.$$

This means that for such h and θ, the error in H, i.e. $H - H_0$ is tending to some small non-zero value.

4. The Preliminary Tests on the Observer Oethod

In this section we present two simple examples illuminating main features of the observer method: the planar Kepler problem and the double harmonic oscillators.

The computations were performed on a VAX 4000 computer with the double precision of the VAX FORTRAN.

4.1. **The Planar Kepler Problem.** It is well known that the Euler method of numerical integration of ODE's gives, in general, very poor results and cannot be used for a reliable integration except for a very short time interval. This, in particular, is the case for the planar Kepler problem ([45]). But, as it will be seen below, when one uses the Euler method with simple observer, the results of integration are of good quality over a quite long time interval.

In the polar coordinates (r,φ) on the plane, the Hamiltonian H describing the planar Kepler problem is

$$H(r,\varphi,p_r,p_\varphi) = \frac{1}{2}\left(p_r^2 + \frac{p_\varphi^2}{r^2}\right) - \frac{\mu}{r},$$

where μ is a real strictly positive number.

The corresponding Hamilton's equations are

$$\begin{cases} \dot r &=& \frac{\partial H}{\partial p_r} = p_r \\ \dot\varphi &=& \frac{\partial H}{\partial p_\varphi} = \frac{p_\varphi}{r^2} - \frac{\mu}{r^2} \\ \dot p_r &=& -\frac{\partial H}{\partial r} = \frac{p_\varphi^2}{r^3} \\ \dot p_\varphi &=& -\frac{\partial H}{\partial \varphi} = 0 \end{cases}$$

Thus, in particular, p_φ is a first integral of the above system.

For a fixed value of p_φ, the first and the third equations of this system constitute the independent system of two ODE's describing the radial component of the Keplerian motion called the reduced system. For convenience we will write shortly p instead of p_r. The reduced system

$$\begin{cases} \dot r &=& \frac{\partial H_1}{\partial p} = p \\ \dot p &=& -\frac{\partial H_1}{\partial r} = \frac{p_\varphi^2}{r^3} - \frac{\mu}{r^2} \end{cases} \tag{4.1}$$

is a Hamiltonian one, with the Hamiltonian function

$$H_1(r,p) = \frac{1}{2}\left(p^2 + \frac{p_\varphi^2}{r^2}\right) - \frac{\mu}{r}.$$

Thus H_1 is a first integral of the initial system. We will now integrate numerically the elliptic orbit of the reduced system corresponding to $H_1 = -0.4$. We will use three different methods: adaptive Runge-Kutta-Fehlberg method of order 4(5) with control error 10^{-6} (shortly RKF method), the Euler method with a fixed step-size $h = 10^{-2}$ and the Euler method as above together with the simple observer method.

Let us note by $(\tilde r(t_i), \tilde p(t_i))_{i\geq 0}$ the computed orbit of the reduced system (4.1). The integration by RKF method gives figure 1. The error in the computed Hamiltonian $\tilde H_1(t_i) = H_1(\tilde r(t_i), \tilde p(t_i))$ increases linearly with time (see figure 2), but is smaller than 10^{-8} up to integration time $t = 10000$. This will be the reference trajectory because it can be considered as a very good approximation of the true orbit.

Later on, when considering a first integral, its symbol with the tilde will denote the numerically computed value of this integral. Applying the Euler method, one can observe that the computed Hamiltonian $\tilde H_1$ increases very quickly, and that after the approximate value of time $t = 6000$, one observes its apparent stabilisation (see figure 3). Moreover, the observed jumps of the computed Hamiltonian correspond, for the small time, to the passage through the pericenter—the point of the orbit nearest to the origin.

When the time increases, the value of computed Hamiltonian strongly increases and our initially elliptic orbit becomes hyperbolic with $r \to +\infty$ (see figure 4).

Thus, the numerical errors inherent to the Euler method destroy completely the phase portrait of the reduced system (4.1) and this occurs for a very short time interval of integration. On the contrary, when

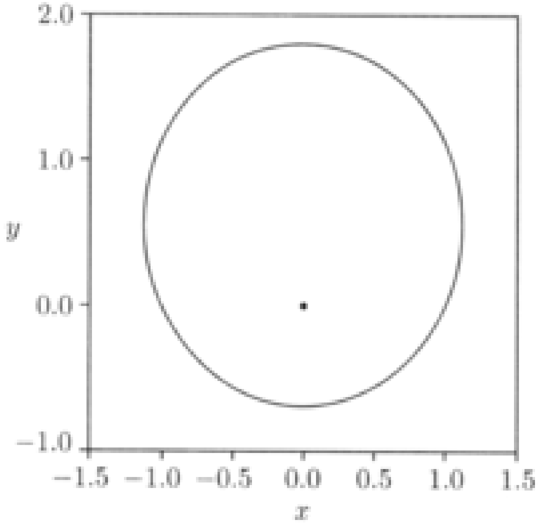


FIGURE 1. The reference orbit for the planar Kepler problem on Cartesian (x, y) coordinates.

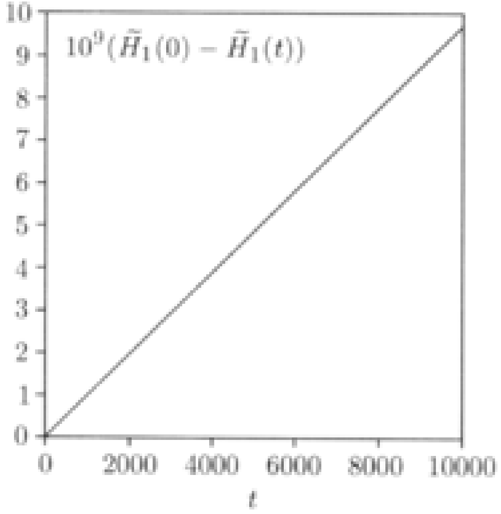


FIGURE 2. Errors of computed reduced Hamiltonian $\tilde{H}_1$ of the planar Kepler problem integrated numerically by the RKAD method.

one uses the same Euler method as above with the simple observer with $H = H_1$ and $\theta = 1$, one obtains quite good results. After the integration time $t = 10000$, the computed orbit is always elliptic, exactly as in figure 1 and the error in computed Hamiltonian is smaller than $5 \cdot 10^{-6}$. In fact, the strong variations of this error are related to the passage through the pericenter of the orbit (see figure 5). Let us emphasise that, in this example, application of the simple observer improves preservation of the integral 10^6 times.

4.2. **The double harmonic oscillator.** Our aim is now to study how different applications of the simple observer method improve results when equations of motion of double harmonic oscillator are integrated by the Euler method. For this aim we will study numerically the system of two uncoupled oscillators described by the Hamiltonian

$$H_2(q,p) = \frac{1}{2}(p_1^2 + p_2^2 + q_1^2 + q_2^2). \tag{4.2}$$

It is obvious that $H_1(q,p) = \frac{1}{2}(p_1^2 + q_1^2)$ is also the first integral of the Hamiltonian system defined by Hamiltonian H_2.

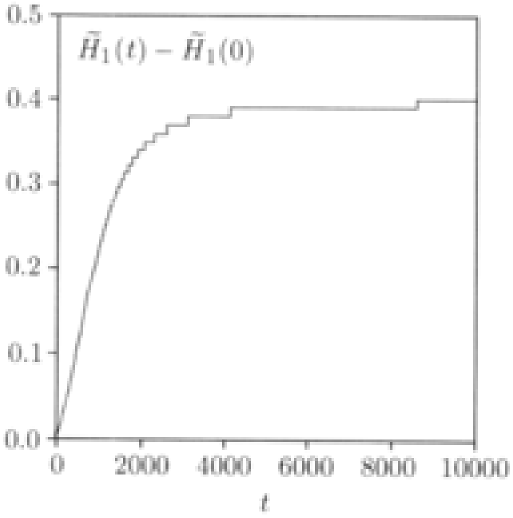


FIGURE 3. Errors of computed reduced Hamiltonian $\tilde{H}_1$ of the planar Kepler problem integrated numerically by the Euler method.

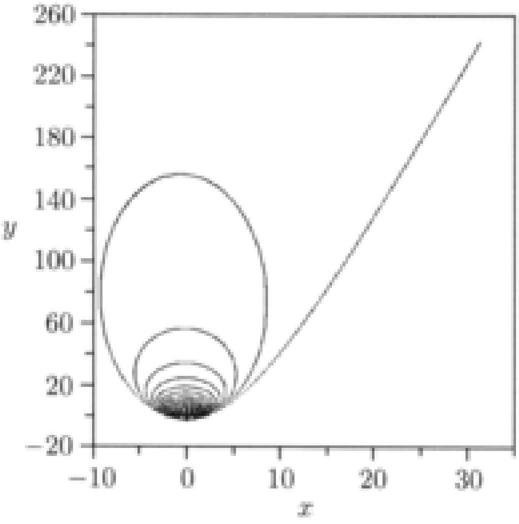


FIGURE 4. The destruction of elliptic orbit computed by the Euler method.

We chose the initial condition for which $H = (H_1, H_2) = (1.5, 1.9)$. Results of computation by the Euler method with the step-size $h = 0.001$ are shown in figure 6a. Integration with the same step-size when the simple observer method is applied for both integrals with $\theta_1 = \theta_2 = 1000$ gives results reported in figure 6b. For the step-size $h = 0.1$ the Euler method is completely out of use for the problem under consideration. After 55 steps absolute values of errors for both integral are bigger than 1. For longer time span of integration these errors grow rapidly. However, for the same step-size when the simple observer method is applied for two integrals and $\theta_1 = \theta_2 = 15$ we obtain quite satisfactory results (see figure 7). Now let us compare the behaviour of the integrals H_1 and H_2 when the Hamiltonian system defined by Hamiltonian H_2 is integrated by the Euler method as above, and by Euler method with the simple observer but applied only to: $H = H_1$ on the level manifold $H_1(q, p) = 1.5$, and $H = H_2$ on the level manifold $H_2(q, p) = 1.9$, respectively. To obtain a reasonable size of error on the first integral H_2 during the integration with the Euler method we chose the step-size $h = 10^{-3}$. To satisfy the inequality (3.8) when the simple observer is used, we chose $\theta = 10^3$. The results presented in figures 8a and 8b indicate that the use of observer method improves the preservation of non-observed first integrals. This phenomenon will be confirmed in all our further computations.

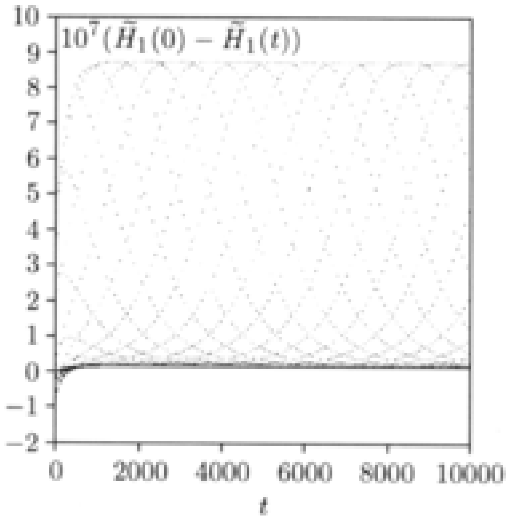


FIGURE 5. Errors of computed reduced Hamiltonian $\tilde{H}_1$ of the planar Kepler problem for the Euler method with the simple observer method.

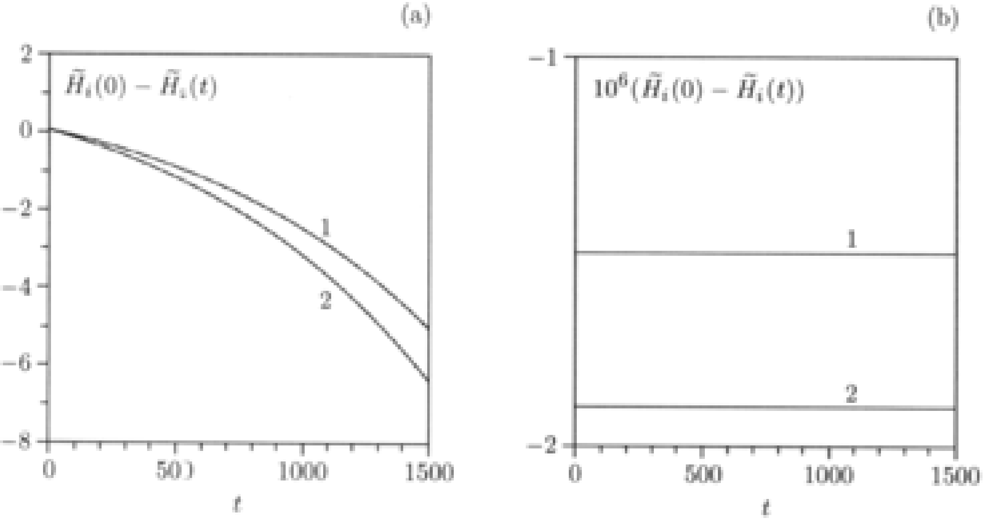


FIGURE 6. Errors of computed first integrals: (1) $\tilde{H}_1$ and (2) $\tilde{H}_2$ of double harmonic oscillator obtained by: (a) the Euler method; (b) the Euler method with the simple observer method applied to both integrals. The step-size was $h = 0.001$.

5. The Spatial Kepler Problem

5.1. **Description of the system.** Let us consider the classical Kepler problem in $\mathbb{R}^3$ ([3,71]). In Cartesian coordinates $(q_1, q_2, q_3, p_1, p_2, p_3)$ of its phase space $\mathbb{R}^6$, this problem is described by the Hamiltonian:

$$H = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) - \frac{\mu}{q},$$

where $q = \sqrt{q_1^2 + q_2^2 + q_3^2}$ and μ is a real, strictly positive number. We will use vectorial notation below, so we put:

$$\mathbf{q} = (\mathbf{q_1}, \mathbf{q_2}, \mathbf{q_3}), \qquad \mathbf{p} = (\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}),$$

and for any vector $\mathbf{v}$ its length will be denoted by v i.e., $v^2 = \mathbf{v} \cdot \mathbf{v}$, where the dot denotes the scalar product. Besides the Hamiltonian, the Kepler system has other first integrals which are the components of the angular momentum $\mathbf{c}$ and the components of the so called Laplace vector $\mathbf{e}$ (see [71]). They are

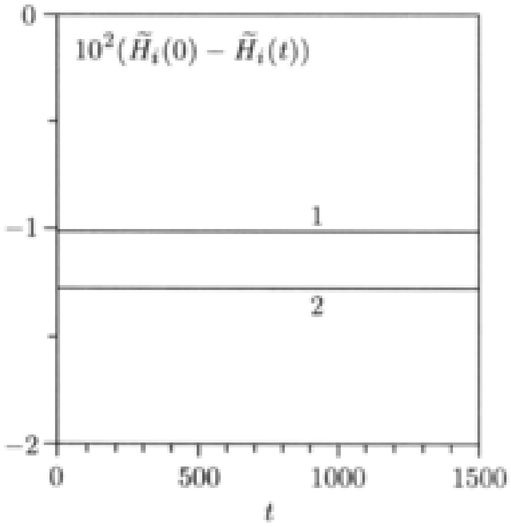


FIGURE 7. The same as in figure 6b for the step-size $h = 0.1$.

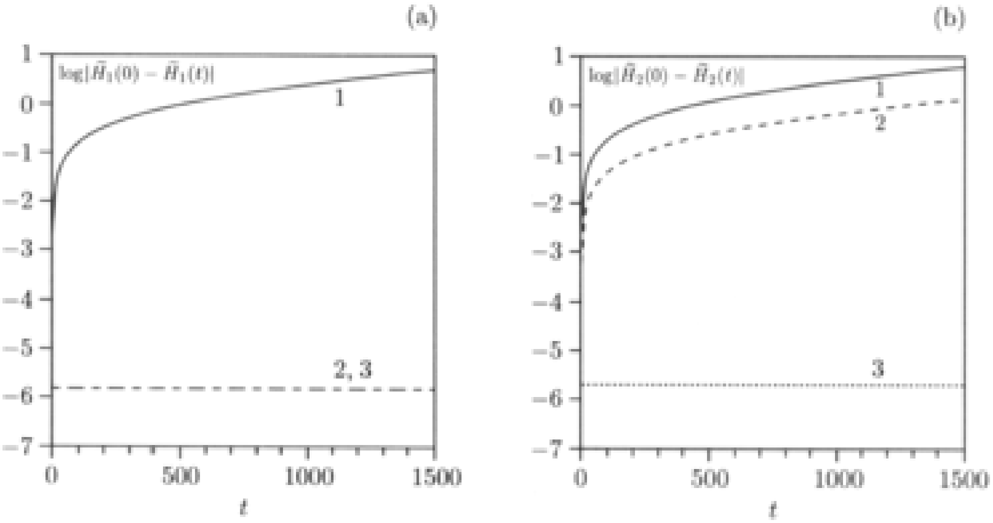


FIGURE 8. Errors in computed first integrals: (a) $\tilde{H}_1$, (b) $\tilde{H}_2$ obtained by: (1) the Euler method, (2) the Euler method with the simple observer method applied to H_1, (3) the Euler method with the simple observer method applied to H_2.

defined as follows:

$$\mathbf{c} = \mathbf{q} \times \mathbf{p}, \qquad \mu\mathbf{e} = \mathbf{p} \times \mathbf{c} - \frac{\mu}{\mathbf{q}}\mathbf{q}. \tag{5.1}$$

Vector $\mathbf{c}$ is perpendicular to the plane of motion. For $c \neq 0$, vector $\mathbf{e}$ is directed from the origin of coordinate system to the pericenter. For $c = 0$, orbits are straight lines passing through the origin and in this case the vector $\mathbf{e}$ is always collinear with the radius vector $\mathbf{q}$ and its length is equal to 1. When $c \neq 0$, the length e of the vector $\mathbf{e}$ is equal to the eccentricity of the orbit. All those seven first integrals of the problem are not functionally independent. In fact, the following relations hold:

$$\mathbf{c} \cdot \mathbf{e} = \mathbf{0}, \qquad \mathbf{Hc^2} = \mu^{\mathbf{2}}(\mathbf{e^2} - \mathbf{1}),$$

where H is the value of the Hamiltonian (energy). In the configuration space $\mathbb{R}^3\{\mathbf{q}\}$, for $c \neq 0$ an orbit lies on the fixed plane passing through the origin and perpendicular to the vector $\mathbf{c}$. For $H < 0$ and $c \neq 0$ all orbits are periodic. In this case an orbit has the form of an ellipse with focus at the origin, semi-major axis $a = -\mu/(2H)$ and the eccentricity e. Orientation of an ellipse in $\mathbb{R}^3\{\mathbf{q}\}$ is traditionally

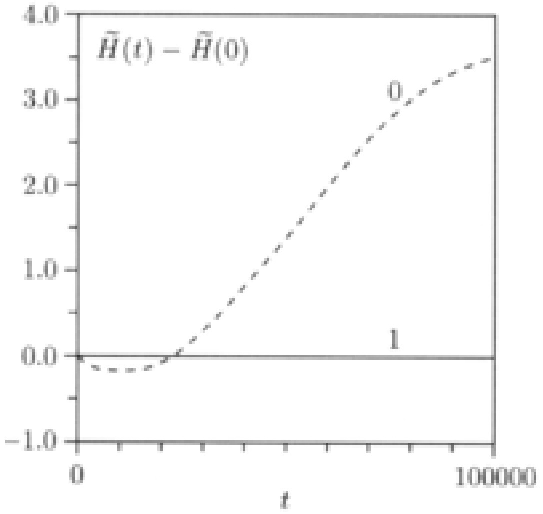


FIGURE 9. Computed energy errors for the spatial Kepler problem over 100000 revolutions by (0) DOPRI integrator and (1) DOPRI integrator when the normalised observer method is applied to integrals (H, c^2).

given by means of the Euler angles (Ω, i, ω). More precisely: Ω—the longitude of ascending node—is the angle between the first axis of $\mathbb{R}^3\{\mathbf{q}\}$ and the line of intersection of the orbital plane and the (q_1, q_2)-plane (this line is called the line of nodes), i—the inclination—is the angle between the vector $\mathbf{c}$ and the q_1-axis, ω—the argument of the pericenter—is the angle between the line of nodes and the vector $\mathbf{e}$. Kepler system appears, among others, in celestial mechanics as unperturbed problem of more complicated systems describing e.g., a model of planetary systems. Its numerical integration shows generally what kind of problems we can meet integrating planetary equations for long time interval. The main problem is the accumulation of energy errors because it causes substantial changes of the angular frequency of the periodic motion and therefore create a rapid growth of the errors of position along the orbit which is also perturbed. This is especially true for highly eccentric orbits. For the Kepler problem we can apply the observer method in many different ways choosing different subsets of integrals. However it is reasonable to consider only H and $\mathbf{c}$ because, generally, in many body problem only these integrals are known. It is well known that levels $H = const < 0$ and $c = const > 0$ are compact. In our tests we have applied the observer and penalty methods with integrals $H_1 = H$ and $H_2 = c^2$. Let us denote $\mathcal{H} = (H_1, H_2)^T$. Let us verify where these integrals are independent. It is easy to show that:

$$D_{\mathcal{H}}(\mathbf{q}, \mathbf{p}) = \begin{bmatrix} \frac{\mu}{q^3}\mathbf{q}, & \mathbf{p} \\ 2\mathbf{p} \times \mathbf{c}, & 2\mathbf{c} \times \mathbf{q} \end{bmatrix} \tag{5.2}$$

Evidently, for $c = 0$, i.e. for straight line orbits, the rank of the above matrix is one. Let us assume now that $\mathbf{c} \neq \mathbf{0}$. The rank of matrix (5.2) will not be maximal iff:

$$\frac{\mu}{q^3}\mathbf{q} = \alpha \mathbf{p} \times \mathbf{c}, \qquad \mathbf{p} = \alpha \mathbf{c} \times \mathbf{q}, \tag{5.3}$$

where $\alpha \neq 0$. Under our assumption vectors $\mathbf{q}$ and $\mathbf{p}$ are not collinear. Taking the scalar product of both sides of equations (5.3) with vectors $\mathbf{q}$ and $\mathbf{p}$, respectively, we obtain:

$$\frac{\mu}{q} = \alpha c^2, \qquad p^2 = \alpha c^2, \qquad \mathbf{q} \cdot \mathbf{p} = \mathbf{0}.$$

The last condition implies that $c = qp$ thus, from the second one we obtain that $1 = \alpha q^2$. Using this, we can rewrite the first of the equations (5.3) in the form:

$$\frac{\mu}{q}\mathbf{q} = \mathbf{p} \times \mathbf{c}.$$

Comparing this with the definition (5.1) of Laplace vector we can see that $\mathbf{e} = \mathbf{0}$, i.e. the orbit is circular. Concluding, integrals H and c^2 are dependent only in the cases of a circular or a straight line orbit.

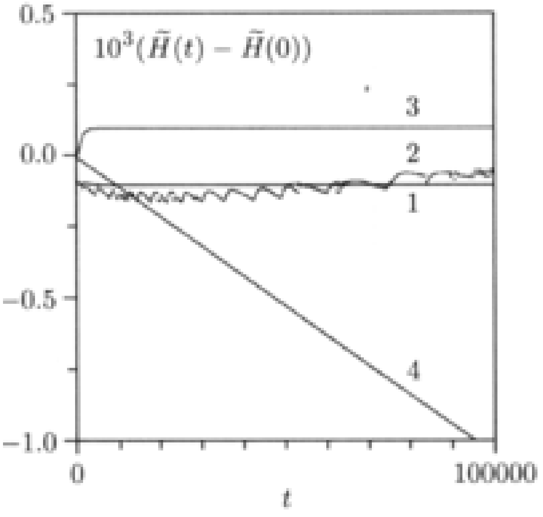


FIGURE 10. The same as in figure 9 obtained by: (1) DOPRI integrator when the normalised observer method is applied to integrals (H, c^2), (2) DOPRI integrator when the normalised penalty method is applied to integrals (H, c^2), (3) fourth order symplectic integrator of Yoshida and (4) DOPRI integrator with K-S regularised problem.

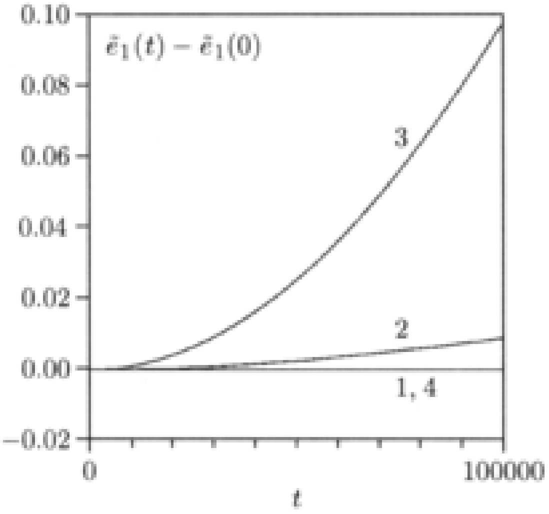


FIGURE 11. Computed errors of the first component of the Laplace vector. Numbers label methods in the same way as in figure 10.

5.2. **Numerical results.** For all tests we choose the orbit with $a = 1$, $e = 0.8$, $\Omega = \omega = \pi/2$, $i = \pi/4$, and we put $\mu = 4\pi^2$. For these values of μ and a the orbital period is equal to one. Initial point was always located at the pericenter. After every hundred of revolutions results were stored. The equations of motion were integrated numerically over the time span of 10^5 by the Runge-Kutta procedure DOPRI with adaptive step-size control. This procedure implements the RK 5(4) algorithm of J.R. Dormand and P.J. Prince and can be found in appendix of the book of Hairer [42]. We translated original FORTRAN code to Pascal. Influence of roundoff errors was minimised because we used nineteen significant digits representation of floating point numbers (we use the extended type of Turbo Pascal). Local precision of integration was chosen 10^{-6}. First tests with the simple observer method and the simple penalty method applied to the integrals H and c^2 with $\theta_1 = 1$ and $\theta_2 = 1$ have shown that they cannot be used effectively for integration of the Kepler system . The step-size chosen by the integration procedure was very small (of order 10^{-10} after several revolutions for the observer method). This shows that magnitudes of terms

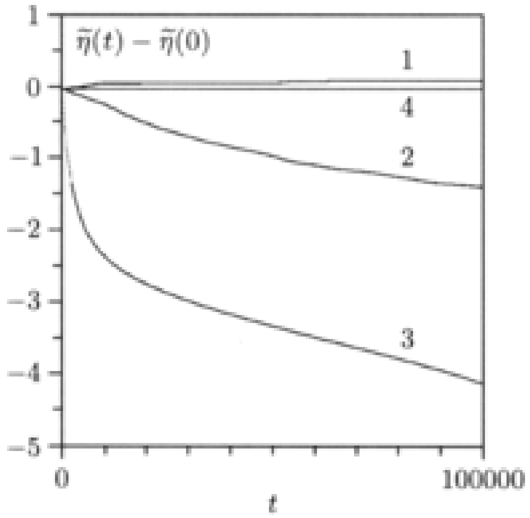


FIGURE 12. Computed errors in position in orbit obtained by four integrators. Numbers label methods in the same way as in figure 10.

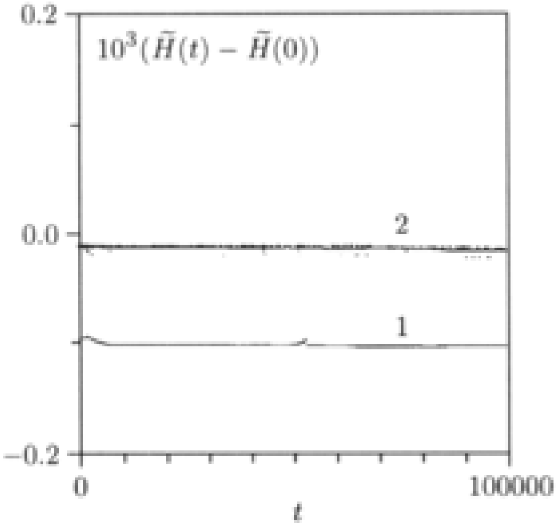


FIGURE 13. Computed energy errors for the spatial Kepler problem over 100000 revolutions by DOPRI integrator with (1) the normalised observer method applied to integrals (H, c^2) and (2) the normalised observer applied to integrals (H, c_1, c_2, c_3).

introduced by the simple observer and the simple penalty methods are too big. There are two ways to overcome this difficulty. One can decrease values of θ_1 and θ_2 or, one can take the normalised observer and penalty methods. We chose the second possibility. We used the normalised observer method with $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = \beta_2 = 1$ (see the definition of the normalised observer). In our first test we compared results obtained from the DOPRI integration of Hamilton's equations for the Kepler system with results obtained by application of the normalised observer method to DOPRI integration of this problem. In figure 9 errors of the computed Hamiltonian as function of time are presented. Because the standard integration gives so bad results in further figures we do not include them. Next, we compare the normalised observer method with: the normalised penalty method, symplectic integration scheme of Yoshida (see [46, 83]), and finally with integration of equations of motion regularised by Kustaanheimo-Stiefel method which will be called shortly the K-S method (see [77]). For the last method the original

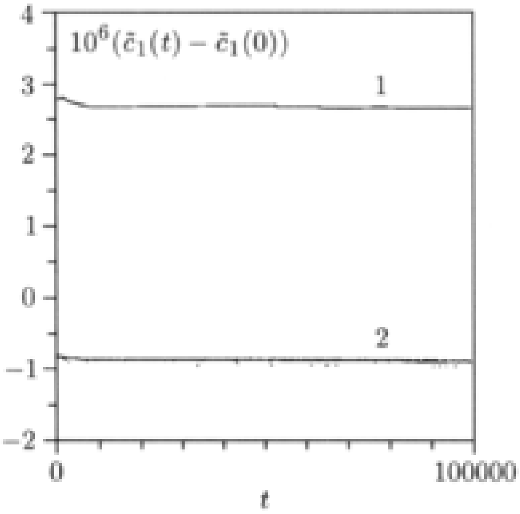


FIGURE 14. Computed errors of the first component of angular momentum for the spatial Kepler problem over 100000 revolutions by DOPRI integrator with (1) the normalised observer method applied to integrals (H, c^2) and (2) the normalised observer applied to integrals (H, c_1, c_2, c_3).

initial value problem is transformed to the form:

$$\left\{ \begin{array}{lclrclrcl} \frac{d\mathbf{u}}{ds} & = & \mathbf{v}, & \frac{d\mathbf{v}}{ds} & = & \frac{H}{2}\mathbf{u}, & \frac{dt}{ds} & = & \mathbf{u}\cdot\mathbf{u}, \\ \mathbf{u(0)} & = & \mathbf{u_0}, & \mathbf{v(0)} & = & \mathbf{v_0}, & t(0) = 0 & & \end{array} \right. \tag{5.4}$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\mathbf{4}$, t is the physical time, and

$$H = \frac{2\mathbf{v_0}\cdot\mathbf{v_0} - \mu}{\mathbf{u_0}\cdot\mathbf{u_0}}.$$

The Cartesian coordinates $\mathbf{q}$ can be expressed in terms of $\mathbf{u}$ by the following formula

$$\begin{matrix} q_1 \\ q_2 \\ q_3 \end{matrix} = \begin{matrix} u_1 & -u_2 & -u_3 & u_4 \\ u_2 & u_1 & -u_4 & -u_3 \\ u_3 & u_4 & u_1 & u_2 \end{matrix} \begin{matrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{matrix}.$$

For more details see [77]. Integration of the K-S system (5.4) shows that the DOPRI procedure chooses, for this system, the step-size too big for good preservation of integrals. Thus, for this integration the maximal step-size was limited to 0.01 while for other integrations it was equal 1. The step-size for the symplectic integrator was taken equal to 0.095 in order to have the computed energy error on the same level as in the normalised observer method. For this integration results were stored after every 95 revolutions. In figure 10 the computed energy errors for all four methods are presented. Figures 11 and 12 show errors in the computed first component of the Laplace vector and errors in the position in the orbit, respectively. The last quantity is defined as the angle between the initial radius vector and the radius vector after some number of full revolutions:

$$\eta(nT) = \arccos\frac{\mathbf{q(0)}\cdot\mathbf{q(nT)}}{q(0)q(nT)}$$

where T denotes the period of the orbit and n is an integer. Although we do not report all obtained results, from our computations follow that the K-S method gives better results than the observer method only in the position in orbit and in the argument of pericenter. However, it should be noted that in all cases the K-S method has linear growth of errors. Comparison of all results obtained for the symplectic integrator and the normalised observer shows that the last one is much more precise although the symplectic integrator is better in preservation of the components of the angular momentum. However it should be noted that we used the observer method with only two integrals H and c^2. Thus, one can expect that the use of the observer method applied to four integrals H and the components of the Laplace vector $\mathbf{e}$ should improve precision of obtained results. In our last test we checked the above hypothesis. We applied the normalised observer method to the integrals H and $\mathbf{c} = (\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3})$ with $\alpha_i = 10, 1 \leq i \leq 4$, and $\beta_1 = \beta_3 = 10$, $\beta_2 = \beta_4 = 1$. Let us note here that it is very important for the observer method to choose appropriate values of constants: θ_i in the case of the simple observer and α_i, β_i in the case of the normalised observer. Especially in this last test it was difficult to find the 'optimal' ones. Figures 13 and 14 show errors in energy and first component of the angular momentum, respectively. Improvement of preservation of angular momentum is not substantial and still the symplectic integrator gives better results for these quantities.

6. Gavrilov-Shil'nikov System

6.1. **Description of the system.** In this section we will consider two parameter family of Hamiltonian systems introduced by N.K. Gavrilov and L.P. Shil'nikov in [34].

These systems are defined on $\mathbb{R}^4$ and are given by following Hamiltonian functions:

$$H = H(x, \omega, \epsilon), \qquad x = (q_1, q_2, p_1, p_2) \in \mathbb{R}^4, \qquad \omega, \epsilon \in \mathbb{R};$$

$$H = \omega(q_1 p_2 - q_2 p_1) - \frac{\epsilon}{2}(p_1^2 + p_2^2) + \frac{\epsilon}{2}(q_1^2 + q_2^2) + \frac{\epsilon}{4}(p_1^2 + p_2^2)^2. \tag{6.1}$$

$x_0 = 0$ is the equilibrium point for this system. The type of this equilibrium point depends on the sign of ϵ. For $\epsilon < 0$, $\omega \neq 0$ the equilibrium point is center-center (i.e., matrix of linearised system possesses four purely imaginary eigenvalues). When $\epsilon = 0$, $\omega \neq 0$ we have the non-semisimple resonance of second order in our system (i.e., the matrix of linearised system possesses two indentical pairs of purely imaginary eigenvalues and it is not diagonalisable). Finally, for $\epsilon > 0$, $\omega \neq 0$ it is the saddle-focus and thus unstable (i.e., the matrix of linearized system possesses four complex eigenvalues with non-zero real and imaginary parts). The systems (6.1) are completely integrable. The second first integral has the form:

$$K = q_1 p_2 - q_2 p_1.$$

Let us fix now $\epsilon = 1$. The obtained system will be shortly called the G-S system. For better understanding of the nature of the phase flow of our system let us introduce new canonical variables (with respect to the standard symplectic structure on $\mathbb{R}^4$):

$$\begin{cases} (q_1, q_2, p_1, p_2) \to (r, \phi, P, I), & \\ q_1 = -P\cos(\phi) + \frac{I\sin(\phi)}{r}, & p_1 = r\cos(\phi), \\ q_2 = -P\sin(\phi) - \frac{I\cos(\phi)}{r}, & p_2 = r\sin(\phi). \end{cases} \tag{6.2}$$

Let us note here that similar variables were introduced by Kovalev and Chudnenko ([47]) while they studied stability conditions of an equilibrium point in a Hamiltonian system with two degrees of freedom in the case of non-semisimple resonance of second order (see also [76] and the formula (11) on p.262 in [2]). In new variables the Hamiltonian (6.1) has the form:

$$H(r, \phi, P, I) = \omega I - \frac{1}{2}r^2 + \frac{1}{2}\left(P^2 + \frac{I^2}{r^2}\right) + \frac{1}{4}r^4.$$

The coordinate ϕ is cyclic and, as a consequence $I = K$ is the first integral. This allows us to reduce our system to one degree of freedom. Hamiltonian of the reduced system is:

$$H_R(r, P) = \frac{1}{2}\left(P^2 - r^2 + \frac{I^2}{r^2}\right) + \frac{1}{4}r^4. \tag{6.3}$$

Levels $H_R = const$ define phase curves of the reduced system on the (r, P)-plane. Because

$$\frac{d\phi}{dt} = \frac{\partial H}{\partial I} = \omega + \frac{I}{r^2},$$

when $\omega \neq 0$, $\frac{d\phi}{dt}$ has a constant sign for sufficiently small $|I|$, these level curves can be interpreted as the closure of the traces of orbits of original system defined by the Hamiltonian (6.3) on the Poincaré surface of section $\phi = 0 \,(\mathrm{mod}\, 2\pi)$, at least when these orbits are not periodic ones. Figures 15a and 15b show these levels for $I = 0$ and $I = 0.1$, respectively. In the first of these figures we can see that there exists the 'figure 8' homoclinic loop for the reduced system. This can be easily shown analytically. Thus in the G-S system, the stable and unstable manifolds of the equilibrium have a common point. Because $\omega \neq 0$, the equilibrium point is of saddle-focus type. Consequently, as the system is integrable, these manifolds coincide. N.K. Gavrilov and L.P. Shil'nikov introduced systems described above during their investigations of the phenomenon of changing stability type of an isolated equilibrium in general one parameter family of Hamitonian systems with two degrees of freedom. We found this system challenging

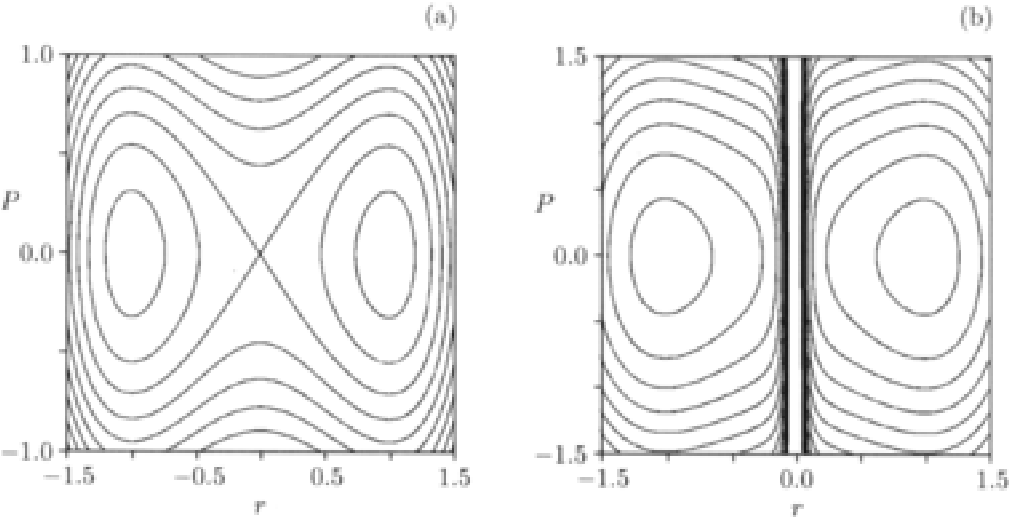


FIGURE 15. Constant values levels of the G-S reduced Hamiltonian (6.3) for (a) $I = 0$ (contours correspond to values of $H_R = -0.3 + 0.1k$, $k = 1, 2\ldots$; the smallest ovals correspond to the minimal value of H_R) and (b) $I = 0.1$ (contours correspond to values of $H_R = -0.35 + 0.2k$, $k = 1, 2, \ldots$).

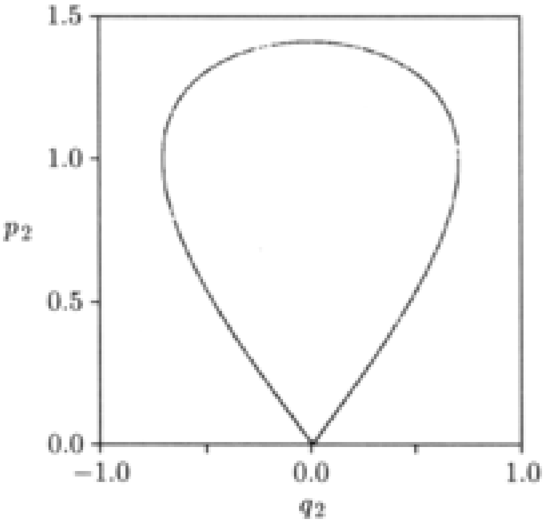


FIGURE 16. The trace on the surface of section of the orbit homoclinic to the equilibrium point $x_0 = 0$ of the G-S system computed without the observer method.

for testing precision of numerical integration. Every numerical procedure 'disturbs' the original system in some way. As it was shown by L.M. Lerman and Ya.L. Umanski in [53] small perturbations of an integrable two degrees of freedom Hamiltonian system with a saddle-focus equilibrium point cause, typically, splitting of the asymptotic surfaces, appearance of transversal homoclinic orbit, and thus nonintegrability (see [25]). Thus, one can expect that any numerical procedure should produce 'numerical chaos' for the G-S system in a neighbourhood of the equilibrium point $x_0 = 0$ on the $H = K = 0$ level. Let us note that this level is compact and that on it the integrals H and K are dependent at the point $x_0 = 0$.

6.2. **Numerical results.** For numerical experiments we put $\omega = 2\pi$. In all cases presented below equations of motion corresponding to the Hamiltonian (6.1) were integrated numerically. For this purpose we applied general extrapolating code for solving ODE's from ([42]). Original FORTRAN procedure ODEX was translated into Pascal. Influence of roundoff errors on our results is minimal because we used nineteen significant digits floating numbers representation (the extended type of the Turbo Pascal v.6.0 of

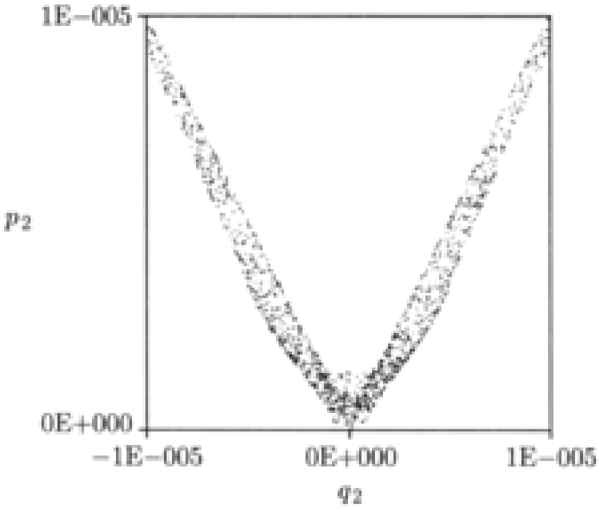


FIGURE 17. The magnification of the neighbourhood of the equilibrium point from figure 16.

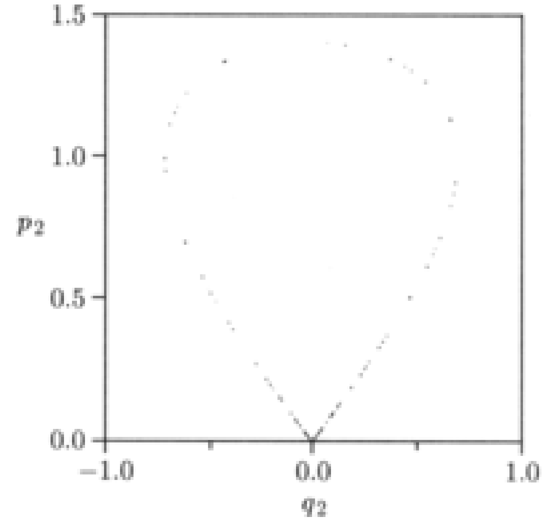


FIGURE 18. The same as in figure 16 when the simple observer method is applied to integrals (H, K).

Borland International). Local precision of integration was 10^{-14}. Results are presented on the Poincaré surface of section $\{p_1 = 0, p_2 > 0\}$, on the constant energy surface $H = 0$, which contains our equilibrium point. Position of points on the section was determined with the precision higher than 10^{-15}. We always stopped integrations after obtaining 10000 points. In every test we integrated equations of motions with the initial condition $(0, 0, 0, \sqrt{2})$. This point lies on the homoclinic loop. Thus, theoretically we should obtain a sequence of points lying on one half of the homoclinic loop and approaching asymptotically the equilibrium point. However, in practice, because of numerical errors, the computed orbit passes near the equilibrium and we obtain a sequence of points that lie near the entire homoclinic loop. First, original Hamilton's equations of motion corresponding to the Hamiltonian (6.1) were integrated. Figure 16 shows, on the surface of section, the obtained homoclinic loop. In figure 17 the magnification of the neighbourhood of the equilibrium is presented. Our suggestion that in the G-S system the numerical chaos is generated is fully confirmed. We obtained qualitatively similar results when for integration the Merson's fourth order Runge-Kutta method (see chapter 2 of [42]) was used. Next, we integrated the G-S system equations of motion with the simple observer applied to both integrals H and K and with $\theta_1 = \theta_2 = 10$. The obtained homoclinic loop is shown in figure 18. Notice the difference between

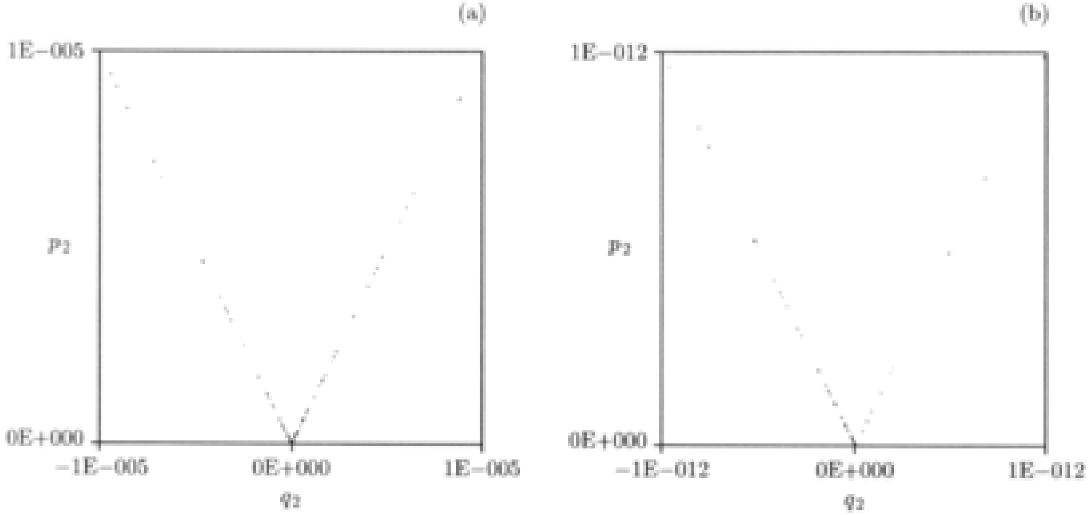


FIGURE 19. Two magnification of the neighbourhood of the equilibrium point from figure 18.

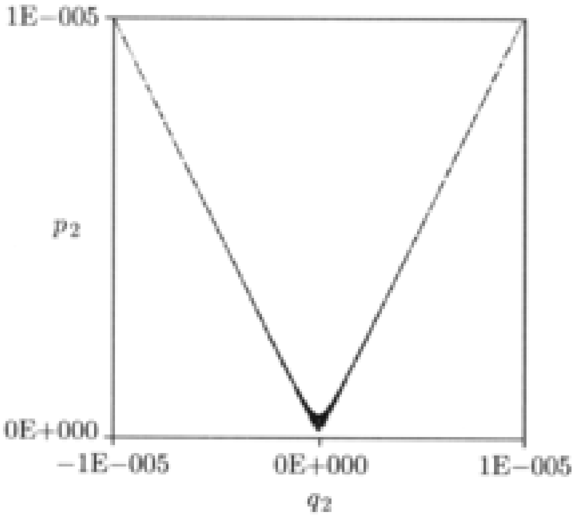


FIGURE 20. The same as in figure 17 when the simple observer method is applied only to the integral H.

figures 16 and 18. During the integration with the observer perturbations the computed orbit goes only few times along the homoclinic loop, and after that it oscillates very closely near the equilibrium point. As a result almost 90% points in figure 18 lie in a very small neighbourhood of the equilibrium point. Moreover, two different magnifications of the neighbourhood presented in figures 19 a and 19b show that numerical chaos disappeared. We applied also the simple observer method to the integral H only with $\theta = 0.2$. In figure 20 obtained results are presented. Comparison with figure 19a clearly shows that chaotic region does not disappear. However, it is confined in the small vicinity of the homoclinic loop. Finally, the simple penalty method was applied to both integrals H and K with $\theta_1 = \theta_2 = 10$. In figures 21a and 21b, as in figures 19 and 20, magnifications of the neighbourhood of the equilibrium point are presented. Let us notice that chaotic region also appears but it is very thin. In our next tests, we compared the simple observer method with a symplectic integrator. For symplectic integration we chose Butcher's fourth order implicit Runge-Kutta method ([21]). The fact that the Butcher method is symplectic one is discussed in [72]. First, we integrated the original equations of motion of the G-S system with the simple observer applied to both integrals using Merson's fourth order Runge-Kutta method. The fixed step-size was chosen equal to 0.01 and we put $\theta_1 = \theta_2 = 10$. Initial condition was

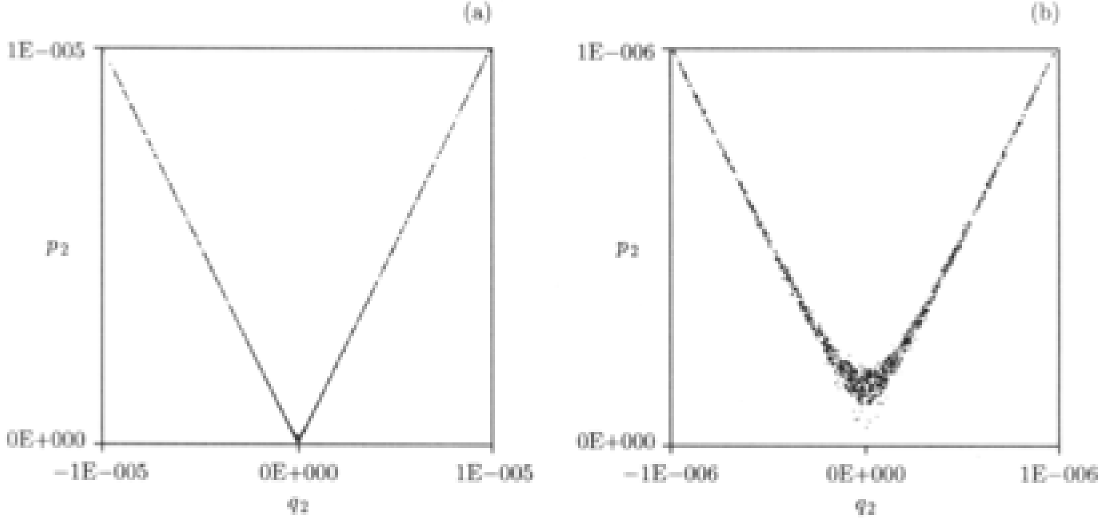


Figure 21. Two magnification of the neighbourhood of the equilibrium point when the simple penalty method is applied to two integrals (H, K).

the same as in previous tests. Results are presented in figures 22a–d. In figure 22a we can see that numerically computed orbit approaches asymptotically the equilibrium point. This is fully confirmed in figures 22b, c, d where magnifications of the neighbourhood of the equilibrium point are shown. In these figures one side of the neighbourhood is only shown because there are no points with $q_2 < 0$). Thus, these results coincide with the theoretical predictions. We noticed that during numerical computation when orbit approaches the equilibrium, errors of both integrals decreases rapidly. After obtaining 100 points on the cross section, errors of both integrals are practically zero. Here it should be explained why, potentially better, integrator ODEX gave worse results than the fixed step-size Runge-Kutta method (compare figures 18 and 22). It was checked that the procedure ODEX is too 'liberal' in choosing a step of integration. When maximal allowed step-size was limited to 0.1 then the integration with ODEX procedure repeated last results. Next, we integrated the G-S system using Butcher's symplectic method with the step-size 0.01. Results are presented in figure 23. Notice that computed orbit goes quite far from the equilibrium point. However, Butcher's method has an excellent property: it does not produce the 'numerical chaos' in the G-S system. We checked it choosing different initial conditions on the homoclinic loop. Even if we start integration from a point lying in a distance of order 10^{-17} from the equilibrium point the computed orbit lies in the close vicinity of the homoclinic loop, but the 'numerical chaos' is undetectable. This very good behaviour of the symplectic integrator can be perhaps explained by the fact that the G-S system is relatively simple: one of its integrals is quadratic and among equations of motion two are linear (see e.g. [72]). In order to compare the observer method and the symplectic integrator on a more complicated system we introduced modified G-S system with the following Hamiltonian:

$$\begin{aligned} H &= \omega(q_1p_2 - q_2p_1) - \tfrac{1}{2}(p_1^2 + p_2^2) + \tfrac{1}{2}(q_1^2 + q_2^2) \\ &\quad + \tfrac{1}{2}(p_1^2 + p_2^2)\left\{\tfrac{1}{2}(p_1^2 + p_2^2) + A(q_1p_2 - q_2p_1) + B(q_1^2 + q_2^2)\right\} \end{aligned}$$

where ω, A, B are real parameters. Note, that we add to the Hamiltonian (6.1) of the G-S system two fourth order terms, thus the type of the equilibrium point in the new system is the same as in the G-S system. Moreover, using canonical transformation (6.2) we can easily prove that the modified G-S system is integrable with the same second integral K and that it possesses, for $B > -1$, the 'figure 8' homoclinic loop on the (r, P) plane. For tests we chose $A = 0.2$, $B = -0.2$, $\omega = 2\pi$ and the step-size $h = 0.01$. Initial condition was $(0, -P, 0, r)$ with $r = 10^{-10}$ and

$$P = r\sqrt{\frac{2 - r^2}{2(1 + Br^2)}}.$$

This point lies on the homoclinic loop in the distance, approximately 10^{-10} from the equilibrium point. The orbit with this initial condition for $t \to +\infty$ goes along the whole loop and asymptotically approaches

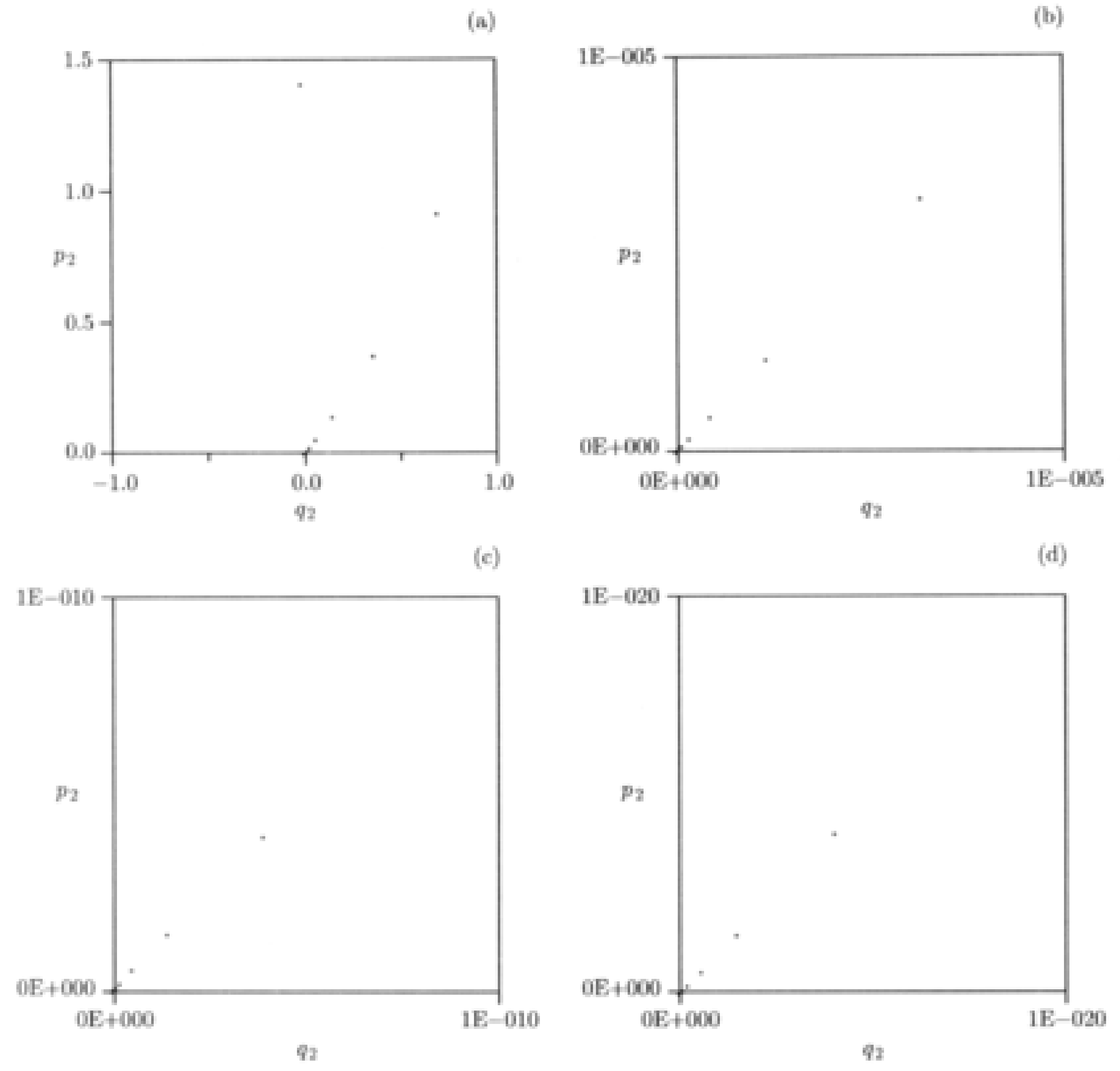


FIGURE 22. (a) the trace on the surface of section of the orbit homoclinic to the equilibrium point $x_0 = 0$ of the G-S system computed by the Merson's RK method with the simple observer method applied to integrals (H, K). (b-d) successive magnifications of the neighbourhood of the equilibrium point for (a).

the equilibrium point from the positive side of the q_2-axis. Using Merson's Runge-Kutta method we integrated equations of the modified G-S system with the simple observer method applied to both integrals, and $\theta_1 = \theta_2 = 10$. Results are presented in figure 24. The series of magnifications of the neighbourhood of the equilibrium point (see figures 24b, c, d) show that computed orbit behaves exactly as the theory predicts. Symplectic integration with Butcher's method gives results presented in figure 25. We can see that for the modified G-S system the symplectic integrator generate the 'numerical chaos' although it is located in the very thin layer around the homoclinic loop. The extremally good coincidence between the theoretically predicted and the numerically computed orbits of our system, clearly indicates that here the use of the observer method increases the reliability of the numerical computations.

7. THE EULER EQUATIONS ON LIE ALGEBRA SO(4)

7.1. **Description of the system.** The general theory of Euler equations on Lie algebra can be found in [11, 32, 33, 63, 65–70, 80]. The specific case of the Euler equations on Lie algebra $so(4)$ is studied in more details in [1, 40, 80, 82]. Because these equations are used here exclusively as an interesting and non trivial example for the application of the observer method, we write them down directly without any explanation of their Lie algebraic origin.

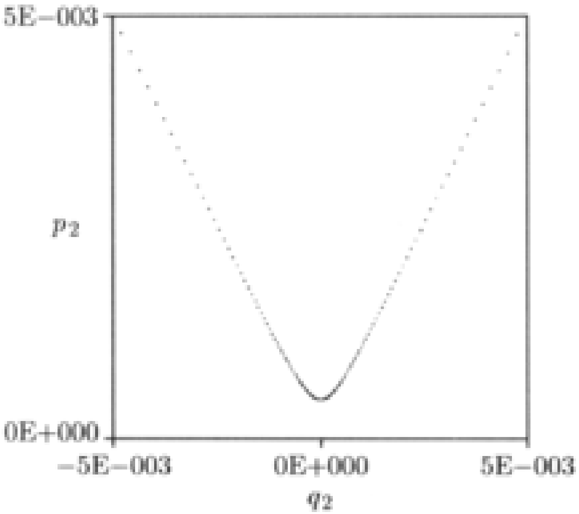


FIGURE 23. The magnifications of the neighbourhood of the equilibrium point when the homoclinic orbit was computed by fourth order Butcher's symplectic integrator.

Let us fix the real numbers $\{\lambda_i\}_{1\le i\le 6}$ and let us consider the system of six ODE's:

$$\left\{\begin{array}{rclcl} \dfrac{dx_1}{dt} &=& (\lambda_3-\lambda_2)x_2x_3 &+& (\lambda_6-\lambda_5)x_5x_6 \\ \dfrac{dx_2}{dt} &=& (\lambda_1-\lambda_3)x_1x_3 &+& (\lambda_4-\lambda_6)x_4x_6 \\ \dfrac{dx_3}{dt} &=& (\lambda_2-\lambda_1)x_1x_2 &+& (\lambda_5-\lambda_4)x_4x_5 \\ \dfrac{dx_4}{dt} &=& (\lambda_3-\lambda_5)x_3x_5 &+& (\lambda_6-\lambda_2)x_2x_6 \\ \dfrac{dx_5}{dt} &=& (\lambda_4-\lambda_3)x_3x_4 &+& (\lambda_1-\lambda_6)x_1x_6 \\ \dfrac{dx_6}{dt} &=& (\lambda_2-\lambda_4)x_2x_4 &+& (\lambda_5-\lambda_1)x_1x_5 \end{array}\right. \tag{7.1}$$

The above system is the system of Euler equations on Lie algebra $so(4)$ corresponding to the 'Hamiltonian' $\dfrac{1}{2}\sum_{i=1}^{6}\lambda_i x_i^2$.

It always admits the following three first integrals:

$$\left\{\begin{array}{rcl} H_1 &=& x_1x_4+x_2x_5+x_3x_6 \\ H_2 &=& \sum_{i=1}^{6} x_i^2 \\ H_3 &=& \sum_{i=1}^{6} \lambda_i x_i^2 \end{array}\right. \tag{7.2}$$

When numbers $\{\lambda_i\}_{1\le i\le 6}$ are not all equal to each other, these three integrals are functionally independent. The integrals H_1 and H_2 are intimately related to the Lie algebra $so(4)$, they represent the 'Casimir functions' of it.

To be integrable (see Section 28 of [3]), the system (7.1) needs to have a supplementary fourth first integral H_4, functionally independent of H_1, H_2, H_3. Let us suppose that among the $\{\lambda_i\}_{1\le i\le 6}$, the equality $\lambda_i=\lambda_j$ for $i\ne j$ occurs at most two times. Then the unique known case when such a fourth first integral exists is the so called *Manakov case*, defined by the condition

$$\lambda_1\lambda_4(\lambda_2+\lambda_5-\lambda_3-\lambda_6)+\lambda_2\lambda_5(\lambda_3+\lambda_6-\lambda_1-\lambda_4)+\lambda_3\lambda_6(\lambda_1+\lambda_4-\lambda_2-\lambda_5)=0$$

In this case as a fourth functionally independent first integral one can take

$$H_4=a_1x_1^2+a_2x_2^2+a_3x_3^2+a_4x_4^2+a_5x_5^2+a_6x_6^2,$$

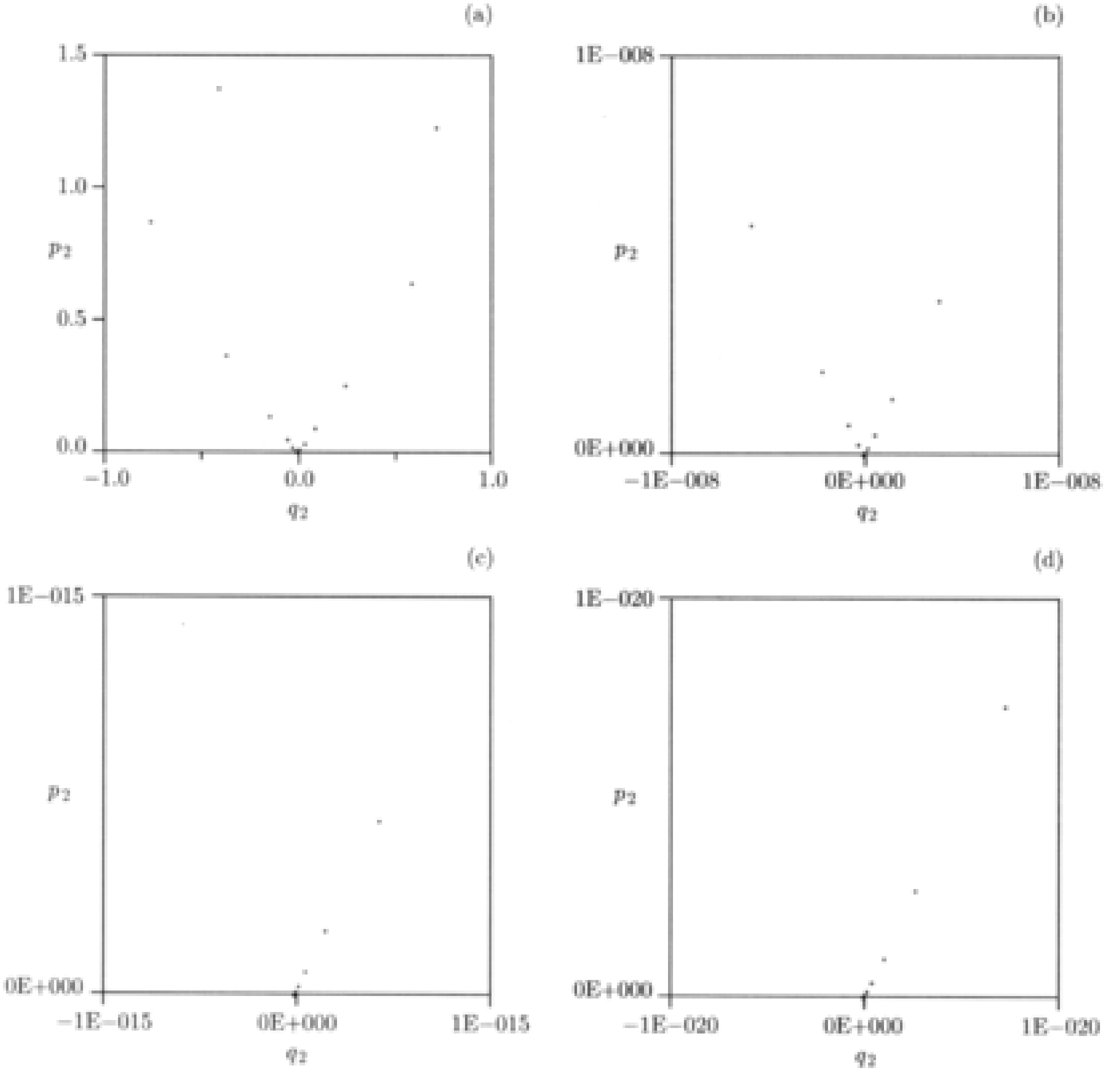


FIGURE 24. The same as in figure 22 for the modified G-S system.

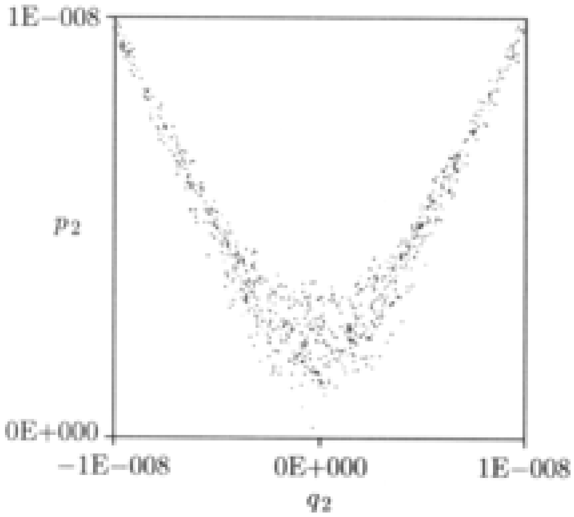


FIGURE 25. The same as in figures 23 for the modified G-S system.

with appropriate constants $\{a_i\}_{1\leq i\leq 6}$ (see [1]). One can prove that, under our assumption concerning the $\{\lambda_i\}_{1\leq i\leq 6}$, except for the Manakov case, the system (7.1) is never algebraically completely integrable (see [1,40,41]. Our main aim is to obtain a reliable graphical representation of the behaviour of the orbits of (7.1). As usual, we will use the Poincaré surface of section.

For $\{\lambda_i\}_{1\leq i\leq 6}$ fixed, let us denote

$$M(h_1,h_2,h_3)=\{x\in\mathbb{R}^6; H_1(x)=h_1, H_2(x)=h_2, H_3(x)=h_3\},$$

where $x=(x_1,x_2,\dots,x_6)$.

Typically $M(a,b,c)$, when non-empty, is a compact, smooth, three dimensional submanifold of $\mathbb{R}^6$, filled with orbits of the system (7.1).

In what follows we will concentrate on the particular non-Manakov case:

$$\lambda_1=1,\quad \lambda_2=5,\quad \lambda_3=2,\quad \lambda_4=1,\quad \lambda_5=3,\quad \lambda_6=4, \tag{7.3}$$

and

$$h_1=1,\quad h_2=15,\quad h_3=25. \tag{7.4}$$

Let us underline that this choice of values of parameters is largely fortuitous except for the fact that $\lambda_1=\lambda_4=1$ and $h_2<h_3$; indeed, these two last conditions will play an essential role in the determination of our surface of section.

Now, let us describe it. Let us consider the two dimensional manifold

$$\tilde{M}(h_1,h_2,h_3)=\{x\in M(h_1,h_2,h_3); x_3=0\}.$$

We will like to choose $\tilde{M}(h_1,h_2,h_3)$ as our surface of section. Unfortunately it is not clear how to choose the global coordinate system on it. Thus we will proceed as follows.

Let us consider a point $X=(x_1,0,x_3,x_4,x_5,x_6)\in\tilde{M}(h_1,h_2,h_3)$ and let us count how many points of $\tilde{M}(h_1,h_2,h_3)$ have the same fifth and sixth coordinates as X.

From (7.2), (7.3) and (7.4) one gets

$$x_3=\epsilon\sqrt{h_3-h_2-2x_5^2-3x_6^2}, \tag{7.5}$$

where $\epsilon=\pm1$. On the other hand from (7.2) one obtains that

$$x_1x_4=h_1-x_3x_6\overset{def}{=}P_\epsilon, \tag{7.6}$$

and that

$$x_1^2+x_4^2=h_2-x_3^2-x_5^2-x_6^2\overset{def}{=}S, \tag{7.7}$$

where x_3 is the same as in (7.5) and the sign of ϵ_i in (7.6) is the same as in (7.5). Finally from (7.6) and (7.7) one deduces that

$$x_1+x_4=\eta\sqrt{S+2P_\epsilon},$$

where $\eta=\pm1$, and that

$$x_1-x_4=\theta\sqrt{S-2P_\epsilon},$$

where $\theta=\pm1$.

Thus, when x_5 and x_6 are fixed, we have at most eight different points in $\tilde{M}(h_1,h_2,h_3)$ with these particular x_5 and x_6 as fifth and sixth coordinate. Consequently, we can cartography $\tilde{M}(h_1,h_2,h_3)$ with eight charts defined by choice of ϵ, η and θ, i.e. by fixing the sign of x_3, x_1+x_4 and x_1-x_4. In any of such charts, (x_5,x_6) is the global system of coordinates.

These charts, noted $\{\Gamma_i\}_{1\leq i\leq 8}$, are defined as follows

	Γ_1	Γ_2	Γ_3	Γ_4	Γ_5	Γ_6	Γ_7	Γ_8
ε	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$
η	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
θ	$-$	$+$	$-$	$+$	$-$	$+$	$-$	$+$

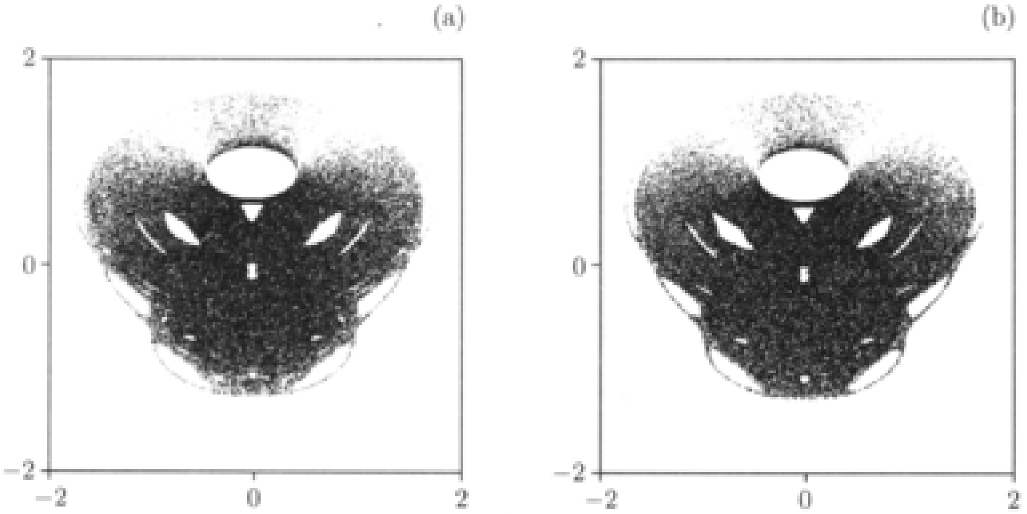


FIGURE 26. The chaotic region on the chart Γ_1 obtained as the trace of one chaotic orbit computed by (a) RKAD method with the simple observer method applied to the first integrals (H_1, H_2, H_3) and (b) the Lie-Poisson integrator. There are 140858 and 146221 points on figure (a) and (b) respectively.

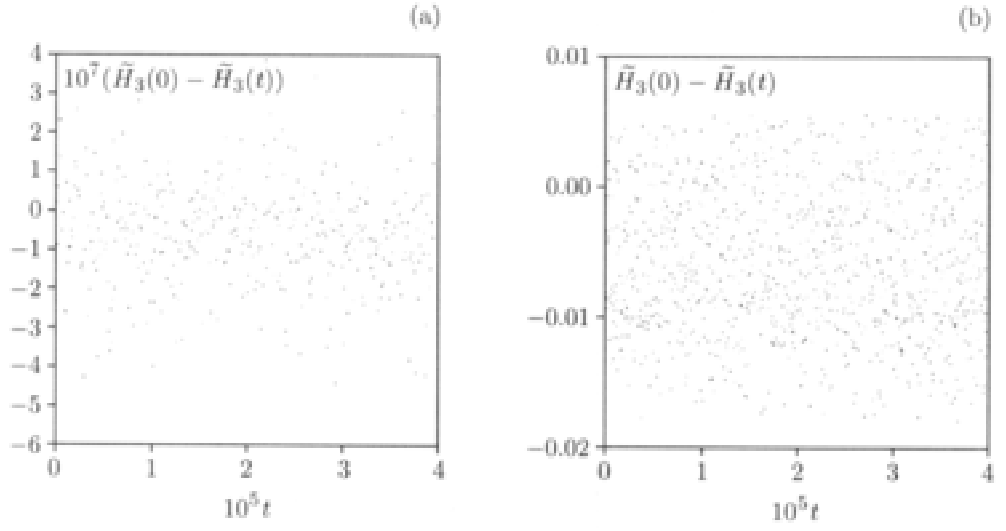


FIGURE 27. The error in computed first integral $\tilde{H}_3$ obtained by (a) RKAD method with the simple observer method applied to the first integrals (H_1, H_2, H_3) and (b) the Lie-Poisson integrator.

7.2. **Numerical results.** First we will study the non-Manakov case defined by (7.3). In this case the system (7.1) was integrated by RKF method (see Section 4.1) with the simple observer, where $\theta_1 = \theta_2 = \theta_3 = 10$. The control error was 5.10^{-8}. The integration time was approximately 400000. All computations in this section were done on the same computer as in the Section 4.

In figure 26a one can find the trace of one chaotic orbit on Γ_1, computed by the above method. We verified that the shape of the obtained chaotic region is independent of the chosen initial orbit, by tracing the similar figures for many different orbits passing by chaotic region of figure 26a.

The system (7.1) can also be integrated by the so called Lie-Poisson integrators ([35]) which generalise the symplectic methods to the case of Hamiltonian equations on Lie algebras.

In [62] the Lie-Poisson integration scheme of Ge Zhong and J.E. Marsden ([35]) in the form given by P.J. Channell and J.C. Scovel ([22]) was applied to integrate the system (7.1) with $\{\lambda_i\}_{1\leq i\leq 6}$ as above.

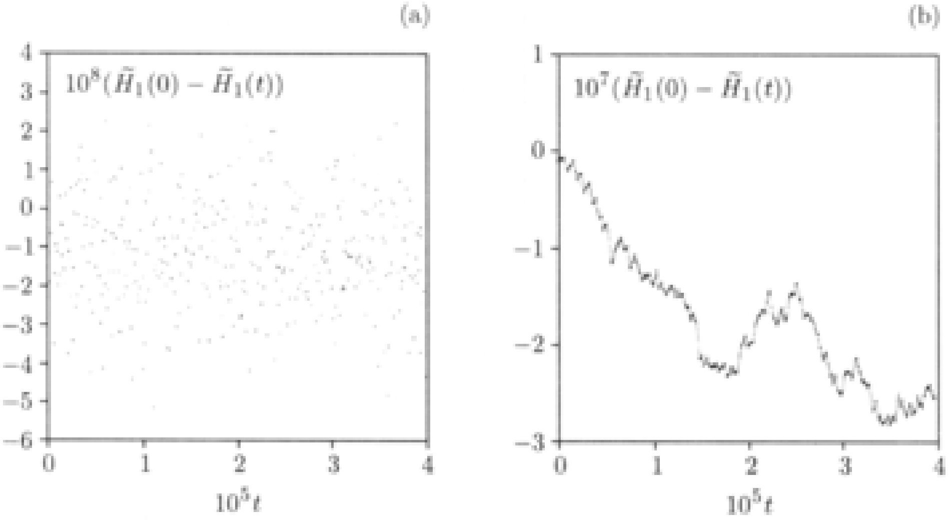


FIGURE 28. The same as in figure 27 but for the computed first integral $\tilde{H}_1$.

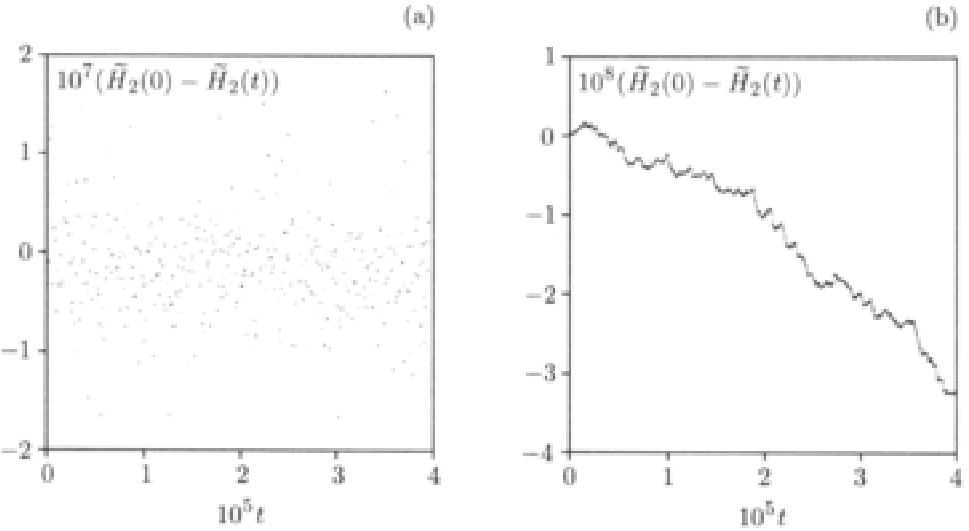


FIGURE 29. The same as in figure 27 but for the computed first integral $\tilde{H}_2$.

More precisely, integration step 1/400 and precision 10^{-13} were used. The integration time was 400000. The figure 26b (from [62]) was obtained in exactly the same way as figure 26a but with the use of this integration scheme (see [62] for more details).

The resemblance between both figures is exceptionally good. This is an argument in what concerns the reliability of our computations.

Let us see now how the computed first integrals H_1, H_2 and H_3 are preserved.

One can see (figures 27a, b) that the Hamiltonian H_3 is much better preserved by the RKF method with the simple observer than by the Lie-Poisson integration scheme. The errors on the remaining integrals H_1 and H_2 are comparable (see figures 28a, b and figures 29a, b). Nevertheless, the error in RKF method with the simple observer is oscillating around a stable non-zero value (see the remark at the end of Section 3) while the error in Lie-Poisson integration scheme increases for H_1 and H_2. Figures 27b, 28b and 29b are taken from the paper of Neron de Surgy *et al.*

Finally, let us compare the behaviour of the computed first integrals H_1, H_2 and H_3 when the observer method is applied to some of them. More precisely, we use the simple observer method only in order to preserve H_3 (the Hamiltonian of the system) and we plot the behaviour of the error for all first integrals. As one can see in figure 30a, the Hamiltonian H_3 is well preserved but moreover, H_1 and H_2 are quite

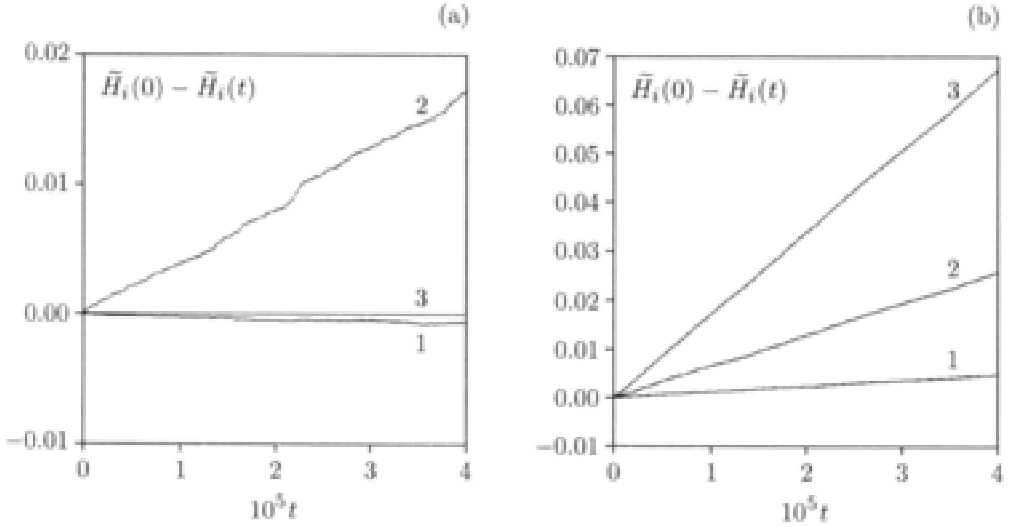


FIGURE 30. The errors in computed first integrals $(\tilde{H}_1, \tilde{H}_2, \tilde{H}_3)$. obtained by (a) RKAD method with the simple observer method applied only to the Hamiltonian H_3 and (b) RKAD method without observer method. The three curves correspond to $i = 1$, 2 and 3, respectively.

well preserved too. Moreover, H_1 and H_2 are better preserved than without the use of the observer method on H_3 (see figure 30b). This confirms the fact that the use of the observer method increases the reliability of computations.

8. An Example of Anosov Flow

8.1. **Description of the system.** This Section is inspired by Section 3.2 of [22]. It is well known that among the smooth dynamical systems with continuous time and compact phase space, the Anosov flows have the most chaotic behaviour, enjoying the strongest possible ergodic properties ([73]). The simplest and the oldest examples of Anosov flows are provided by the geodesic flows on compact orientable surfaces M^2 having constant negative curvature equal -1 (see [24,73]). Let M^2 be a compact orientable two dimensional Riemannian manifold (surface) of constant negative curvature -1. Let us consider the open unit disc $\mathbf{D^2} = \{(\mathbf{q_1}, \mathbf{q_2}) \in \mathbb{R}^2; \mathbf{q_1^2} + \mathbf{q_2^2} < \mathbf{1}\}$ equipped with non-Euclidean Riemannian metric

$$\frac{dq_1^2 + dq_2^2}{(1 - q_1^2 - q_2^2)^2} \tag{8.1}$$

of constant negative curvature -1 (see Chapter 2 of [26]). It is well known that any surface M^2, as above, is isometric to the quotient space $\mathbf{D^2/\Gamma}$, where Γ is a discrete subgroup of the homographic transformations group acting on the unit disc $\mathbf{D^2}$ (see Chapter 4 of [27]). The geodesics on M^2 can be obtained as projections on $\mathbf{D^2/\Gamma}$ of the geodesics on $\mathbf{D^2}$ corresponding to the Riemannian metric (8.1). The geodesic flow on $\mathbf{D^2}$ is governed by the Hamiltonian

$$H(q,p) = \frac{(1 - q_1^2 - q_2^2)^2 (p_1^2 + p_2^2)}{4}, \tag{8.2}$$

where $q = (q_1, q_2) \in \mathbf{D^2}$, $p = (p_1, p_2) \in \mathbb{R}^2$. From now on let us consider the particular case when M^2 is the doughnut with two holes, i.e. a compact orientable surface of genus 2 equipped with the non-Euclidean metric of constant curvature -1. In [27] one can find the explicit description of the discrete groups Γ of homographies of the unit disc $\mathbf{D^2}$ such that, $M^2 = \mathbf{D^2/\Gamma}$, as well as, of the fundamental region $\Omega \subset \mathbf{D^2}$ of the subgroup Γ which was used by Channell and Scovel ([22]) and which we will also use here. The boundary of the fundamental region Ω is the piecewise smooth curve L defined as follows. Let

$$R_0 = \frac{\sin\beta}{\cos\beta + \sqrt{\cos 2\beta}} \approx 0.21684534\ldots$$

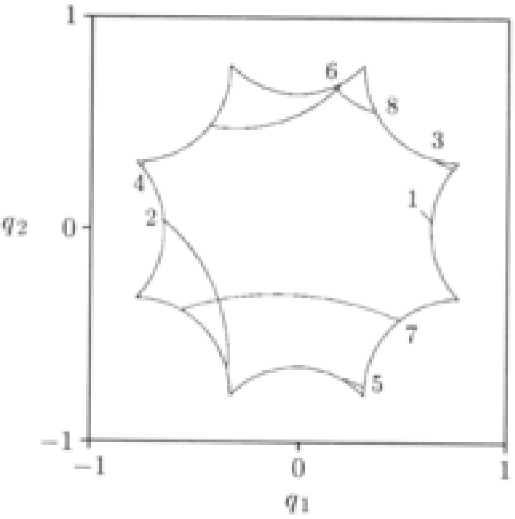


FIGURE 31. A piece of a geodesic on M^2. Numbers mark successive origins of smooth pieces of the geodesic.

with $\beta = \pi/8$. Inside the octant $\Pi_1 = \{(q_1, q_2) \in \mathbf{D^2}; \mathbf{0} \leq |\mathbf{q_2}| \leq \mathbf{q_1}\mathrm{tg}\beta\}$ the boundary curve is defined by the equation

$$1 + q_1^2 + q_2^2 - 2q_1\frac{1+R_0^2}{1-R_0^2} = 0.$$

The boundary in the other octants is obtained by its rotation by a multiple of $\pi/4$ (see figure 31). For every point $q \in \mathbf{D^2}$ there exists a homographic transformation $\gamma \in \Gamma$ such that $\gamma(q) \in \Omega$. When $\gamma(q) \in \Omega \backslash L$, then such γ is unique. For $q \in \Pi_1 \backslash \Omega$, but belonging to a thin layer around L, the corresponding homographic transformation γ, $\gamma(q_1 + iq_2) = \bar{q}_1 + i\bar{q}_2$, is of the form

$$(8.3) \qquad \begin{cases} \bar{q}_1 & = & \frac{ab(q_1^2+q_2^2)+q_1(a^2+b^2)+ab}{b^2(q_1^2+q_2^2)+2abq_1+a^2} \\ \bar{q}_2 & = & \frac{q_2(a^2-b^2)}{b^2(q_1^2+q_2^2)+2abq_1+a^2} \end{cases}$$

where $a = 1 + 1/R_0^2$ and $b = 1 - 1/R_0^2$. The similar transformations in the other octants are obtained from the formulae (8.3) by rotations by a multiple of $\pi/4$. Any such mapping induces in a natural way the mapping of the momenta by

$$(8.4) \qquad \bar{p} = (D_\gamma^T(q))^{-1}p$$

where by $D_\gamma^\star(q)$ we denote the transposed of the Jacobian matrix of the mapping γ at q.

8.2. **Numerical results.** The Hamiltonian (8.2) written down in Section 3.2 of [22] was used by them for numerical study of the geodesic flow on doughnut with two holes. This study was done with the use of the symplectic integrator of the fourth order. But the final result concerning the preservation of the computed Hamiltonian $\tilde{H}$ was, in the opinion of their authors, not completely satisfactory. They suggested that the very strong ergodic properties of the system under consideration prevent really good numerical results, independently of the method used. We will now show, applying exactly the same geometrical machinery as in [22], that the use of a standard Runge-Kutta integration procedure with the simple observer method allows to obtain results substantially better than those reported by Channell and Scovel. We choose at random a point $q_0 \in \Omega$ and a vector $p_0 \in \mathbb{R}^2$. We compute numerically the piece of the orbit of the phase flow governed by the Hamiltonian (8.2) up to the first computed point (q, p) such that $q \notin \Omega$. Then, using the formulae (8.3) and (8.4) we obtain the point $(\bar{q}, \bar{p})$ where $\bar{q} \in \Omega$ and we continue our computations as before. Let us note that on M^2 the projections of the point q and $\bar{q}$ are almost the same. The pieces of orbits in Ω obtained in this way represent the computed geodesic line on M^2 passing through q_0 in the direction p_0. In figure 31 one can find a piece of such a geodesic on M^2 represented in this way inside Ω. In figure 32 the errors of the computed Hamiltonian for four integration with the fixed step-size $h = 0.01$ are presented. We integrated equations of motion using Merson's Runge-Kutta method, Merson's Runge-Kutta method with the simple observer, Butcher's

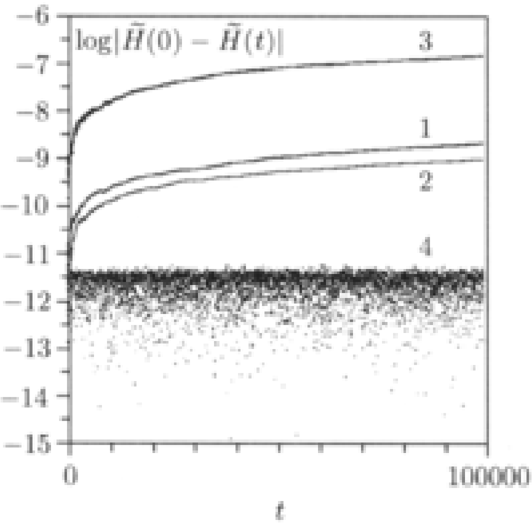


Figure 32. Errors in computed Hamiltonian for geodesic flow on M^2 obtained by (1) Merson's RK method, (2) fourth order Butcher's symplectic method, (3) fourth order generating function symplectic integrator and (4) Merson's RK Method with the simple observer method

fourth order symplectic Runge-Kutta method and fourth order generating function integrator described in [22], respectively. The advantage of the application of the observer method clearly appears.

Acknowledgement 1. *The fourth author thanks Dr. S. Wojciechowski from Department of Mathematics of Linköping Institute of Technology (Sweden), who many years ago introduced him to the world of Euler equations and Dr. M. Balabane from Department of Mathematics of University Paris-Nord (France) for the discussion on the observer method and his idea of the penalty method. We thank O. Neron de Surgy from l'Observatoire de Paris for his kind permission for the reproduction of figures from [62]. We thank also Dr. T. Nadzieja from Department of Mathematics of University of Wroclaw (Poland), Dr. C. Simó from Department of Applied Mathematics of University of Barcelona (Spain) for their interest in this work, Dr. J.-G. Caputo from Institut National des Sciences Appliquées in Rouen for improving English of the manuscript as well as K. Goździewski from N. Copernicus University (Poland) for his computing and technical help. We thank O. Benois and G. Grancher from Departement of Mathematics of University of Rouen and Dr. H. Ghidouche from Departement of Mathematics of University of Paris-Nord for their computer assistance during the edition of the present paper. The first, third and fourth authors want to thank their wives Prisque, Zbroja and Helena for their patience. The third and fourth authors thank also Rouen and Toruń Universities for their excellent working conditions during authors respective visits in August and in December 1992. Last but not least we thank the organisers of conference Equadiff 91 where the observer method was presented for the first time. For the third author this work was supported by the grant KBN 22149203. During the final stage of elaboration of this paper he was supported by the fellowship ERB3510PL9210479 from the European Community's Action for Cooperation in Science and Technology with Central and Eastern European Countries.*

References

[1] M. Adler and P. van Moerbeke The algebraic integrability of geodesic flow on SO(4) Invent. Math 67 1982 297–331

[2] V. I. Arnold, V. V. Kozlov and A. I. Neishtadt Mathematical aspects of classical and celestial mechanics Arnold V I (Edit.) *Dynamical Systems III, Encyclopaedia of Mathematical Sciences 3*, (Springer Verlag) Berlin 1988

[3] V. I. Arnold Mathematical Methods of Classical Mechanics Graduate Texts in Math., Springer Verlag 60 Berlin 1989

[4] M. Artigue, V. Gautheron and E. Isambert Ensemble de bifurcation et topologie des variétés intégrales dans le problème du solide pesant Journal de mécanique théorique et appliquée 5 1986 429–469

[5] M. Audin and R. Shilol Variétés abéliennes réelles et toupie de Kovalevski Public. de l'IRMA (preprint) Strasbourg 1992

[6] J. Baumgarte Stabiliztion of the differential equations of Keplerian motion Tapley B D and Szebehely V (Edit.) *Recent Advances in Dynamical Astronomy*, D. Reidel Publ Comp, Dodrecht 1973 38–44

[7] J. Baumgarte Stabiliztion, manipulation and analytic step adaptation Szebehely V and Tapley B D (Edit.) *Long Time Predictions in Dynamics*, D. Reidel Publ Comp, Dodrecht 1976 153–163

[8] G. Benettin, M. Casartelli, L. Galgani, A. Giorgilli and J.-M. Strelcyn On the reliability of numerical study of stochasticity Part I: Existence of times averages Il Nuovo Cimento 44B 1978 183-195

[9] G. Benettin, M. Casartelli, L. Galgani, A. Giorgilli and J.-M. Strelcyn On the reliability of numerical study of stochasticity Part II: Identification of time averages Il Nuovo Cimento 50B 1979 211–232

[10] D. G. Bettis (Edit.) Proceedings of the conference on the numerical solution of ordinary differential equations 19 20 october *1972* the University of Texas at Austin, Lecture Notes in Mathem. 362, Springer Verlag Berlin 1974

[11] O. I. Bogoyavlensky Integrable Euler equations on Lie algebras arising in problems of mathematical physics Math. USSR Izvestya 25 207–257 1984

[12] O. I. Bogoyavlensky Integrable Euler equations on SO(4) and their physical applications Commmun. Math. Phys. 93 1984 417–436

[13] O. I. Bogoyavlensky New integrable problems of classical mechanics Commun. Math. Phys. 94 1984 225–269

[14] O. I. Bogoyavlensky Euler equations on finite dimensional Lie algebras arising in physical problems Commun. Math. Phys. 95 1984 307–315

[15] O. I. Bogoyavlensky Periodic solutions in a model of pulsar rotation Commun. Math. Phys. 102 1985 349–359

[16] O. I. Bogoyavlensky Integrable cases of rigid body dynamics and integrable systems on the ellipsoids Commun. Math. Phys. 103 1986 305–322

[17] O. I. Bogoyavlensky Some integrable cases of Euler equations I (in Russian) Dokl. Akad. NauK SSSR 287 1986 1105–1108

[18] O. I. Bogoyavlensky Some integrable cases of Euler equations II (in Russian) Dokl. Akad. Nauk SSSR 292 1987 318–322

[19] O. I. Bogoyavlensky Boundary value problems of mathematicals physics Proceeding of the Steklov Institute of Math., Publ. by Amer. Math. Soc. 95 1988

[20] O. I. Bogoyavlensky Euler equations on finite dimensional Lie coalgebras (in Russian) Uspekhi Math. Nauk. 47 (1) 1992 107–146 English translation to appear in Russian Math. Surveys **47 (1)**

[21] J. C. Butcher Implicit Runge-Kutta processes Math. Comput. 18 1964 50–64

[22] P. J. Channell and J. C. Scovel Symplectic integrators of Hamiltonian systems Nonlinearity 3 1990 231–259

[23] P. J. Channell and J. C. Scovel Integrators for Lie-Poisson dynamical systems Physica D 50 1991 80–88

[24] I. P. Cornfeld, S. V. Fomin and Ya G. Sinai Ergodic Theory Springer Verlag Berlin 1982

[25] R. L. Devaney Homoclinic orbits in Hamiltonian systems Journal of Diff. Equat. 21 1976 431–438

[26] B. A. Dubrovin, A. T. Fomenko and S. P. Novikov Modern Geometry— Methods and Applications Part I: The Geometry of Surfaces. Transformation Groups and Fields Graduate Texts in Math. **93** Springer Verlag Berlin 1984

[27] B. A. Dubrovin, A. T. Fomenko and S. P. Novikov Modern Geometry— Methods and Applications Part II: The Geometry and Topology of Manifolds Graduate Texts in Math. **104** Springer Verlag Berlin 1985

[28] T. Eirola and J. M. Sanz-Serna Conservation of integrals and symplectic structure in the integration of differential equation by multistep methods Numer. Math. 61 1992 281–290

[29] El Hamidi Propriétés stochastiques d'un système non-linéaire en dimension finie Thèse de physique théorique Université de Pau 1989

[30] Feng Kang and Qin Meng-Zhao The symplectic methods for the computation of hamiltonian equations in Zhu You-Wan Guo Ben-Yu (Edit.) *Numerical Methods for Partial Differential Equations,* Lectures Notes in Math. **1297**, Springer Verlag, Berlin 1987 1-37

[31] A. F. Filippov Differential eqution with discontinuous right side (in Russian) Nauka Scientific Publ. Moscow 1985

[32] A. T. Fomenko Integrability and nonintegrability in geometry and mechanics Kluwer academic publishers Dodrecht 1988

[33] A. T. Fomenko and V. V. Trofimov Integrable system on Lie algebras and symmetric spaces Gordon and Breach Science Publishers New York 1988

[34] N. K. Gavrilov and L. P. Shil'nikov On bifurcations of the equilibrium states of hamiltonian system with two degrees of freedom Selecta Mathematica Sovietica 10 (1) 1991 61–68

[35] Z. G. Ge and J. E. Marsden Lie-Poisson-Hamilton-Jacobi theory and Lie-Poisson integrators Phys. Letters A 133 (3) 1988 134–139

[36] C. W. Gear Invariants and numerical methods for ODEs Physica D 60 1992 303–310

[37] B. Gladman and M. Dunca Symplectic Integrators for long-term integrations in celestial mechanics Celestial Mech. 52 1991 221–240

[38] D. Greespan Arithmetic Applied mathematics Pergamon Press Oxford 1980

[39] J. W. Grizzle, P. E. Moraal Newton, observers and nonlinear discrete-time control Proceed. of the *29*th Conference on Decision and Control, Honolulu Hawaii 1990 760–767

[40] L. Haine Geodesic flow on SO(4) and abelian surfaces Math. Ann. 263 1983 435–472

[41] L. Haine The algebraic complete integrability of geodesic flow on SO(N) Commun. Math. Phys. 94 1984 271–287

[42] E. Hairer, S. P. Norsett and G. Wanner Ordinary differential equations I, Nonstiff problem Springer Verlag Berlin N.Y. 1987

[43] P. Hartman Ordinary differential equations Wiley J and Sons N. Y. London 1964

[44] W. D. Irtegov Invariant manifolds of stationary motions and their stability Nauka Scientific Publ. - Siberian branch Novosibirsk 1985

[45] K. Hockett Chaotic numerics from an integrable hamiltonian system Proc. of Amer. Math. Soc. 108 1990 271–281

[46] H. Kinoshita, H. Yoshida and H. Nakai Symplectic integrators and their application to dynamical astronomy Celestial Meech. 50 1991 59–71

[47] A. M. Kovalev and A. N. Chudnenko On stability of equilibrium point of Hamiltonian system with two degree of freedom in the case of equal frequencies (in Russian) Doklady of Ukrainian Acad. of Sciences A 11 1977 1010–1013

[48] V. V. Kozlov Integrability and nonintegrability in hamiltonian mechanics (in Russian) Uspekhi Math. Nauk 38 (1) 1983 3–67

[49] V. V. Kozlov and D. A. Onishenko Nonintegrability of Kirchoff equations (in Russian) Doklady Akad. Nauk USSR 266 1982 271–281

[50] J. D. Lambert Computational methods in ordinary differential equations J Wiley New York 1973

[51] M. Lecar A comparison of eleven numerical integrations of the same gravitational 25-body problem Bulletin Astronomique Série 3 vol. III fascicule 1 1968 91–104

[52] E. Leimanis The general problem of the motion of coupled rigid bodies about a fixed point Springer Verlag Berlin 1965

[53] L. M. Lerman and Ya L. Umanski On the existence of separatrix loops in four-dimensional systems similar to the integrable hamiltonian systems Journal of Applied Math. and Mech. 47 (3) 1983 335–340

[54] A. J. Lichtenberg and M. A. Lieberman Regular and stochastic motion Applied Math. Sciences **38** Springer Verlag Berlin 1983

[55] T. Levi-Civita and G. Amaldi Lezioni di meccanica razionale Vol II Part 2 Nicola Zanichelli Bologna 1927

[56] A. J. Maciejewski and J.-M. Strelcyn Numerical integration of differential equations in presence of first integrals: partial first integrals to be published

[57] A. Marciniak Numerical solutions of the n-body problem D, Reided Publ. Comp. Dordrecht 1985

[58] A. Marciniak The selected numerical methods for solving the n-body problem Ed. by Technical School at Poznań 1989

[59] A. Marciniak and D. Greenspan Energy conserving numerical solutions of simplified turbulence equations Dept. Math. the University of Texas at Arlington, Technical Report **269**, Preprint 1990

[60] A. Marciniak and D. Greenspan Arbitrary order Hamiltonian conserving numerical solutions of Calogero and Toda systems Computers Math. Applic. 22 (7) 1991 11–35

[61] P. E. Nacozy The use of integrals in numerical integrations of the n-body problem Astroph. and Space Science 14 1971 40–51

[62] O. Neron de Surgy Sur les intégrateurs symplectiques pour les systèmes Hamiltoniens standards et les systèmes de Lie-Poisson; une application aux équations d'Euler sur SO(4) Rapport de stage fait à Electricité de France 1991

[63] P. J. Olver Applications of Lie groups to differential equations Graduate Texts in Math. **107**, Springer Verlag Berlin 1986

[64] S. A. Orszag and J. B. Mc Laughlin Evidence that random behavior is generic for nonlinear differential equations Physica 1D 1980 68-79

[65] A. M. Perelomov Dynamical systems of classical mechanics with hidden symmetry (in Russian) Preprints of Inst. of Theor. Exper. Phys.Moscow **82** 1981

[66] A. M. Perelomov Integrable systems of classical mechanics and Lie algebras. The simplest systems (in Russian) Preprints of Inst. of Theor. Exper. Phys. Moscow **149** 1981

[67] A. M. Perelomov Integrable systems of classical mechanics and Lie algebras. Constrained systems (in Russian) Preprints of Inst. of Theor. Exper. Phys. Moscow **116** 1983

[68] A. M. Perelomov Integrable systems of classical mechanics and Lie algebras. The motion of rigid body with fixed point (in Russian) Preprints of Inst. of Theor. Exper. Phys. Moscow **147** 1983

[69] A. M. Perelomov Integrable systems of classical mechanics and Lie algebras. The motion of rigid body in the ideal fluid (in Russian) Preprints of Inst. of Theor. Exper. Phys. Moscow **151** 1984

[70] A. M. Perelomov Integrable systems of classical mechanics and Lie algebra Vol. I Birkhauser Verlag Basel 1990

[71] H. Pollard Celestial Mechanics The Carus Math. Monogr. **18** edited by the Math. Assoc. of America 1976

[72] J. M. Sanz-Serna Symplectic integrators for Hamiltonian systems: an overview Acta Numerica, Annual volume published by Cambridge University Press. 1 1991 243–286

[73] Ya G. Sinai & al Ergodic Theory with applications to Dynamical Systems and Statistical Mechanics in Ya G. Sinai (Edit.) Dynamical Systems II, *Encyclopaedia of Mathematical Sciences***2**, Springer Verlag Berlin 1989

[74] S. Smale Topology and mechanics Part I Invent. Math. 10 1970 305–333

[75] S. Smale Topology and mechanics Part II Invent. Math. 11 1970 45–64

[76] A. G. Sokol'sky On the stability of an autonomous hamiltonian system with two degree of freedom in the case of equal frequencies Journal of Applied Math. and Mech. 38 (5) 1974 741–749

[77] E. Stiefel and G. Schiefele Linear and regular celestial mechanics Springer Verlag Berlin 1971

[78] J.-M. Strelcyn The 'Coexistence Problem' for Conservative Dynamical Systems: A Review Colloq.Math. 62(2) 1991 331–345

[79] W. W. Strygin and W. A. Sobolev Separation of motions by the method of integral manifolds (in Russian) Nauka Scientific Publ. Moscow 1988

[80] V. V. Trofimov Introduction to the geometry of manifolds with symmetries Moscow University Press Moscow 1989

[81] S. Ushiki Central difference schema and chaos Physica 4D 1982 407–424

[82] A. P. Veselov On condition of integrability of Euler equations on SO(4) (in Russian) Doklady Akad. Nauk SSSR 270 1983 1298–1300

[83] H. Yoshida Construction of higher order symplectic integrators Phys. Lett. A 150 1990 262–268

Authors

Eric Busvelle, Centre de Recherche Shell S.A., Boite Postale 20, 76530 Grand-Couronne, France

Rachid Kharab, Université de Rouen, Département de Mathématiques, URA CNRS 1378, Boite postale 118, 76134 Mont-Saint Aignan Cedex, France (e-mail: kharab@univ-rouen.fr)

A. J. Maciejewski, Nicolaus Copernicus University, Institute of Astronomy, ul. Chopina 12–18, 87-100 Toruń, Poland (e-mail: maciejka@mat.torun.edu.pl)

Jean-Marie Strelcyn, Université de Rouen, Département de Mathématiques, URA CNRS 1378, Boite postale 118, 76134 Mont-Saint Aignan Cedex, France (e-mail: strelcyn@univ-rouen.fr)

and

Laboratoire Analyse, Géométrie et Applications, URA 742, Université Paris-Nord, Institut Galilée, Département de Mathématiques, Avenue J.-B. Clément, 93430 Villetaneuse, France (e-mail: strelcyn@math.univ-paris13.fr)