Observer design for nonlinear systems

Pauline Bernard

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à MINES ParisTech

Synthèse d'observateur pour systèmes non linéaires

Observer design for nonlinear systems

École doctorale n°432

SCIENCES ET MÉTIERS DE L'INGÉNIEUR

Spécialité MATHÉMATIQUES ET AUTOMATIQUE

COMPOSITION DU JURY :

M. Philippe Moireau
Inria Saclay – Ile-de-France, Président

M. Alberto Isidori
Université de Rome, Rapporteur

M. Alain Rapaport
Inra Montpellier, Rapporteur

M. Witold Respondek
INSA Rouen, Rapporteur

Mme Hélène Piet-Lahanier
ONERA, Examinateur

M. Vincent Andrieu
LAGEP Lyon, Examinateur

M. Laurent Praly
MINES ParisTech, Examinateur

Soutenue par **Pauline BERNARD**
le 20 novembre 2017

Dirigée par **Laurent PRALY**
Vincent ANDRIEU

MINES ParisTech | PSL RESEARCH UNIVERSITY PARIS

Remerciements

En tout premier lieu, je souhaite remercier mes encadrants de thèse, Laurent Praly et Vincent Andrieu, pour leur confiance, leur enthousiasme et leur soutien permanents. Ce fut un grand honneur et un immense bonheur de travailler avec vous pendant ces trois années. J'ai énormément appris à votre contact, et je garderai de cette thèse un excellent souvenir, tant sur le plan humain que scientifique. Laurent, je me rappellerai toujours de ces heures passionnantes passées en tête à tête à Fontainebleau à gribouiller des pages et des pages de calculs, à tester des milliers d'idées jusqu'à épuisement du stylo ou du tas de brouillons. Merci d'avoir partagé votre passion et (une infime partie de) l'immense étendue de vos connaissances avec tant de générosité et de gentillesse. Quant à Vincent, tu as toujours été présent malgré la distance, toujours prêt à proposer de nouvelles idées, à m'aider et à m'encourager dans mon travail dès que j'en avais besoin. Merci pour ta disponibilité, ton enthousiasme et pour m'avoir toujours chaleureusement accueillie lors de mes visites, en particulier à Wuppertal, dont je garde un excellent souvenir.

J'aimerais aussi exprimer ma reconnaissance à Nicolas Petit pour ses nombreux conseils avisés et l'attention bienveillante qu'il m'a toujours portée, de près ou de loin, depuis mon S3 Recherche. C'est vous qui m'avez si justement orientée vers cette thèse et à qui je dois ces formidables trois années.

Un grand merci à l'ensemble des membres du CAS pour la merveilleuse ambiance que vous savez créer et entretenir. C'est si agréable et inspirant de travailler parmi des passionnés qui font preuve de curiosité et d'enthousiasme pour tout ce qui touche les sciences. Une pensée particulière aux doctorants avec qui j'ai partagé ces années, Charles-Henri, Rémi, Pierre, ainsi que Jean, mon incorrigible compagnon du rez-de-chaussé, que j'étais toujours ravie de retrouver, fidèle à son poste, à mes retours d'exil à Fontainebleau.

Je tiens aussi à remercier Alberto Isidori, Alain Rapaport et Witold Respondek, qui m'ont fait l'honneur d'accepter d'être rapporteurs de ma thèse ainsi que Hélène Piet-Lahanier et Philippe Moireau pour leur participation à mon jury de soutenance. Votre lecture attentive ainsi que vos questions et commentaires constructifs m'ont aidée à améliorer ce manuscrit et m'ont fourni de nouvelles pistes de réflexion.

Une pensée pour Vivien avec qui j'ai vécu et partagé cette aventure au jour le jour, et qui était toujours là pour me demander comment allaient mes difféomorphismes.

Enfin, je voudrais remercier ma famille, pour qui ces pages paraitront aussi énigmatiques qu'une tablette de hiéroglyphes. Vous avez toujours cru en moi, vous m'avez toujours soutenue et encouragée, et c'est grâce à vous si je peux aujourd'hui écrire ces lignes. Merci du fond du coeur.

Contents

Chapter 1

Introduction

Context

In many applications, estimating the current state of a dynamical system is crucial either to build a controller or simply to obtain real time information on the system for decision-making or surveillance. A common way of addressing this problem is to place some sensors on/in the physical system and design an algorithm, called *observer*, whose role is to process the incomplete and imperfect information provided by the sensors and thereby construct a reliable estimate of the whole system state. Of course, such an algorithm can exist only if the measurements from the sensor somehow contain enough information to determine uniquely the state of the system, namely the system is *observable*.

The number and quality of the sensors being often limited in practice due to cost and physical constraints, the observer plays a decisive role in a lot of applications. Many efforts have thus been made in the scientific community to develop universal methods for the construction of observers. Several conceptions of this object exist, but in this thesis, we mean by observer a finite-dimensional dynamical system fed with the measurements, and for which a function of the state must converge in time to the true system state. Although very satisfactory solutions are known for linear systems, nonlinear observer designs still suffer from a significant lack of generality. The very vast literature available on the subject consists of scattered results, each making specific assumptions on the structure and observability of the system. In other words, no unified and systematic method exists for the design of observers for nonlinear systems.

Actually, observer design may be more or less straightforward depending on the coordinates we choose to express the system dynamics. For instance, dynamics which seem nonlinear at first sight could turn out to be linear in other coordinates. In particular, some specific structures, called *normal forms*, have been identified for allowing a direct and easier observer construction. One may cite for instance the state-affine forms with their so-called Luenberger or Kalman observers, or the triangular forms associated to the celebrated high gain design. With this in mind, most solutions available in the literature actually fit in the following three-step methodology :

1. look for a reversible change of coordinates transforming the dynamics of the given nonlinear system into one of the identified normal forms,
2. design an observer in those new coordinates,
3. deduce an estimate for the system state in the initial coordinates via inversion of the transformation.

Of course in order to follow this method one need to know

I. a list of normal forms and their associated observers,

II. under which conditions and thanks to which invertible transformation one can rewrite a dynamical system into one of those forms,

III. how to compute the inverse of this transformation.

When browsing the literature, one discover that the first two points have been extensively studied, although not always under this terminology. In fact, they constitute the core of the observer design problem and they are tightly linked since a particular form is of interest if it admits observers (Point I.) and if a large category of systems can be transformed into that form (Point II.). Therefore, Points I. and II. are often treated simultaneously. On the contrary, very few results concern Point III., mainly because the observer problem is often considered theoretically solved, once an invertible transformation into a normal form has been found.

Problems addressed in this thesis

Actually, in practice, inverting a nonlinear map is far from trivial. Most of the time, the system and the normal form have different dimensions, so that the transformation is at best an injective immersion. Since its inverse is a priori defined only on a submanifold of the observer space, an extension is often necessary. When an explicit expression for a global inverse is not available, numerical inversion usually relies on the resolution of a minimization problem with a heavy computation cost, which thus raises implementation issues. That is why the first goal of this thesis was to develop a methodology to avoid the explicit inversion of the transformation, by bringing the dynamics of the observer (designed in the normal form coordinates) back into the initial system coordinates.

When I started my thesis, some preliminary results in that direction had already been obtained in the case of autonomous systems, but some tools remained to be developed in order to complete the theory and also to make the method implementable in practice. This kept us busy for several months and at the end, we tried to extend our results to time-varying/controlled systems. In doing so, we discovered that surprisingly, the limitation did not come from our method of inversion, but rather from the scarcity of general observer design techniques available for nonlinear controlled systems, namely from Points I.-II. rather than Point III.

In particular, we realized that, in the usual case where the derivatives of the input are unknown, even the widely used high gain design, reputed to be general, had only been proved to work under the assumptions that the system be observable for any input AND that its order of differential observability be equal to the system dimension. In this particular case, the system can indeed be transformed into a triangular normal form with Lipschitz nonlinearities appropriate for the design of a high gain observer. Given the restrictive nature of this framework, we naturally wondered if one of those two assumptions could be relaxed. Actually, the "observable for any input" assumption is necessary to have a triangular form and cannot be altered. However, we discovered that, interestingly, it was often possible to preserve the triangularity of the target form when allowing the order of differential observability to be larger than the dimension of the system, but that the Lipschitzness of the nonlinearities could be lost. This observation led us to address the following two problems : first, what kind of observers can be used for a triangular normal form with continuous (non-Lipschitz) nonlinearities, and second under which conditions a system can be transformed into such a continuous triangular form.

Apart from the high gain paradigm, another general technique for nonlinear observer design had recently been developed, inspired from Luenberger's initial approach to build observers for linear systems. This so-called Kazantzis-Kravaris or Luenberger design consists in transforming the system into a Hurwitz linear form (for which a trivial observer exists) via the resolution of a partial differential equation (PDE). But this approach was only available for autonomous systems and we thus tried to figure out how it could be extended to controlled/time-varying systems, namely how to transform this kind of system into a Hurwitz linear form.

Thesis organization

Summing up, this thesis provides contributions to each of the three points mentioned above :

Contribution 1 observer design for a continuous triangular form (related to Point I.)

Contribution 2 characterization of controlled systems which can be transformed into a continuous triangular form (related to Point II.)

Contribution 3 characterization of controlled systems which can be transformed into a Hurwitz linear form (related to Point II.)

Contribution 4 method to express the dynamics of the observer in the given coordinates to avoid the inversion of the transformation (related to Point III.)

Instead of presenting the results in a chronological way, I thus found clearer to organize my thesis along this three-step methodology and classify the contributions accordingly, namely in three parts :

Part I Normal forms and their observers (with Contribution 1)

Part II Transformation into a normal form (with Contributions 2 and 3)

Part III Observer in given coordinates (with Contribution 4)

Since the topics of Part I and II have been extensively studied in the literature, detailed reviews are provided in each of those parts, so that this thesis finally gives a good overview of the state of the art in terms of observer design for nonlinear systems.

On the other hand, I also had the opportunity to work on applications, in particular the design of observers for permanent magnet synchronous motors (PMSM) without mechanical information (sensorless) and with some unknown parameters. This led to the following contributions :

Contribution 5 gradient observer for the estimation of the rotor position and magnet flux of a PMSM

Contribution 6 observability analysis and observer design for a PMSM with unknown rotor position and unknown resistance.

This work was carried out in parallel to the rest and is detailed in a separate part :

Part IV Observers for PMSMs with unknown parameters (with Contributions 5 and 6).

Publications

The work presented in this thesis has resulted in the following publications :

- Journals

  1. P. Bernard, L. Praly, V. Andrieu, *Observers for a non-Lipschitz triangular form*, Automatica, Vol. 82, p301-313, 2017
  2. P. Bernard, L. Praly, V. Andrieu, *On the triangular normal form for uniformly observable controlled systems*, Automatica, Vol. 85, p293-300, 2017.
  3. P. Bernard, L. Praly, V. Andrieu, *Expressing an observer in given coordinates by augmenting and extending an injective immersion to a surjective diffeomorphism*, Submitted to SIAM

- Conferences

1. P. Bernard, *Luenberger observers for nonlinear controlled systems*, Conference on Decision and Control, 2017 (To appear)
2. P. Bernard, L. Praly, *Robustness of rotor position observer for permanent magnet synchronous motors with unknown magnet flux*, IFAC World Congress, 2017
3. P. Bernard, L. Praly, V. Andrieu, *Non Lipschitz triangular normal form for uniformly observable controlled systems*, IFAC Symposium on Nonlinear Control Systems, 2016
4. P. Bernard, L. Praly, V. Andrieu, *Tools for observers based on coordinate augmentation*, Conference on Decision and Control, 2015

Introduction

Contexte

Dans beaucoup d'applications, l'estimation en temps réel de l'état d'un système dynamique est cruciale, que ce soit pour la synthèse d'un contrôleur ou simplement pour la surveillance et la prise de décision. Une façon usuelle de résoudre ce problème consiste à installer des capteurs sur/dans le système physique et implémenter un algorithme, appelé *observateur*, dont le rôle est de traiter les informations partielles et imparfaites données par les capteurs, et d'en déduire une estimation fiable de l'état complet du système. Bien sûr, un tel algorithme ne peut exister que si les mesures des capteurs contiennent assez d'informations pour déterminer de manière unique l'état du système : le système est alors dit *observable*.

Le nombre et la qualité des capteurs étant souvent limités en pratique en raison de contraintes physiques et de coût, l'observateur est amené à jouer un rôle décisif dans beaucoup d'applications. La communauté scientifique s'est donc efforcée de développer des méthodes aussi universelles que possible pour la synthèse d'observateur. Plusieurs conceptions de cet objet existent, mais dans cette thèse, le terme "observateur" désigne un système dynamique de dimension finie, prenant en entrée les mesures, et dont une fonction de l'état converge en temps vers l'état réel du système. Alors que des solutions satisfaisantes existent pour les systèmes linéaires, les synthèses d'observateurs non linéaires manquent cruellement de généralité. La littérature, par ailleurs très fournie sur le sujet, se compose essentiellement de résultats épars, chacun faisant sa propre hypothèse sur la structure et l'observabilité du système. Autrement dit, il n'existe pas de méthode générale pour la synthèse d'observateur pour système non linéaires.

En fait, il se peut que la synthèse soit plus ou moins facile suivant les coordonnées que l'on a choisies pour exprimer la dynamique du système. Par exemple, une dynamique qui paraît non linéaire au premier abord pourrait s'avérer être linéaire dans d'autres coordonnées. Or, des structures particulières, appelées *formes normales*, ont été identifiées comme permettant la construction facile et directe d'un observateur. Parmi elles, les formes affines en l'état, avec leurs observateurs de Luenberger ou de Kalman, ou les formes triangulaires, associées au célèbre observateur grand gain. A partir de là, la plupart des solutions disponibles dans la littérature s'inscrivent en fait dans une démarche à trois étapes que l'on peut résumer ainsi :

1. chercher un changement de coordonnées réversible qui transforme la dynamique du système non linéaire donné dans l'une des formes normales connues,
2. synthétiser un observateur dans ces coordonnées,
3. en déduire une estimation de l'état du système dans les coordonnées initiales en inversant la transformation.

Bien sûr, pour suivre cette méthode, il est nécessaire de connaître

I. une liste de formes normales et les observateurs associés,

II. sous quelles conditions et grâce à quelle transformation inversible il est possible de réécrire un système dynamique sous l'une de ces formes,

III. comment calculer l'inverse de la transformation.

Il s'avère que les deux premiers points ont beaucoup été étudiés dans la littérature (pas toujours sous cette terminologie). En fait, ils constituent le coeur du problème de synthèse d'observateur et ils sont fortement liés puisqu'une forme particulière n'a d'intérêt que si elle admet un observateur (Point I.) et si une large catégorie de systèmes peuvent être transformés en cette forme (Point II.). Les Points I. et II. sont donc très souvent traités simultanément. Au contraire, très peu de résultats concernent le Point III., principalement parce que le problème d'observateur est souvent considéré comme résolu lorsque une transformation inversible dans une forme normale a été trouvée, c'est-à-dire lorsque les Points I. et II. ont été traités.

Problèmes abordés dans cette thèse

En fait, en pratique, inverser une application non linéaire est loin d'être trivial. La plupart du temps, le système et la forme normale ont des dimensions différentes, et la transformation est donc au mieux une immersion injective. Puisque son inverse n'est a priori définie que sur une sous-variété de l'espace où évolue l'observateur, une extension est souvent nécessaire. En l'absence d'expression explicite et globale de l'inverse, l'inversion numérique repose sur la résolution d'un problème de minimisation coûteux en calcul, ce qui soulève d'importants problèmes d'implémentation. C'est pourquoi le premier objectif de cette thèse était de développer une méthode permettant d'éviter l'inversion explicite de la transformation, en ramenant la dynamique de l'observateur (écrite dans les coordonnées de la forme normale) dans les coordonnées initiales du système.

Lorsque j'ai commencé ma thèse, des résultats préliminaires avaient déjà été obtenus dans cette direction pour les systèmes autonomes, mais il restait à développer certains outils pour compléter la théorie ainsi que pour la rendre implémentable en pratique. Ceci nous a occupés quelques mois, jusqu'à ce que nous essayions d'étendre nos résultats aux systèmes instationnaires/commandés. C'est alors que nous nous rendîmes compte avec surprise que les limitations ne provenaient pas de notre méthode d'inversion, mais plutôt de la rareté des techniques générales de synthèse d'observateurs existant pour les systèmes non linéaires commandés, c'est-à-dire des Points I. et II. plutôt que du Point III.

En particulier, nous réalisâmes que, dans le cas usuel où les dérivées de l'entrée sont inconnues, même la synthèse grand gain, si largement utilisée et réputée générale, ne s'applique théoriquement qu'aux systèmes observables pour toute entrée dont l'ordre d'observabilité différentielle est égal à la dimension du système. Dans ce cas particulier en effet, le système peut être transformé en une forme normale triangulaire avec des non linéarités Lipschitz appropriées à la synthèse d'un observateur grand gain. Vu le caractère restrictif de ce cadre, nous nous demandâmes naturellement si l'une de ces deux hypothèses pouvait être relâchée. Pour ce qui est de la première, l'observabilité "pour toute entrée" est nécessaire pour obtenir une forme triangulaire et ne peut donc être modifiée. Par contre, nous découvrîmes qu'il était souvent possible de préserver la triangularité de la forme cible en autorisant l'ordre d'observabilité différentielle à être supérieur à la dimension du système, mais que le caractère Lipschitz des non linéarités pouvait alors être perdu. Cette observation nous amena naturellement à nous intéresser à deux nouveaux problèmes : d'une part, quels types d'observateurs peuvent être utilisés pour une forme triangulaire avec des non linéarités continues (non-Lipschitz), et d'autre part, sous quelles conditions un système quelconque peut être transformé en une telle forme.

En face de la synthèse grand gain, une autre technique générale de synthèse d'observateurs non linéaires avait été récemment développée, inspirée de l'approche initialement adoptée par Luenberger pour la synthèse d'observateur de systèmes linéaires. Cette synthèse "de Kazantzis-Kravaris" ou "de Luenberger", consiste à transformer le système en une forme linéaire Hurwitz (pour laquelle un observateur trivial existe) via la résolution d'une équation aux dérivées partielles (EDP). Mais cette approche étant disponible seulement pour les systèmes autonomes, nous essayâmes de l'étendre aux systèmes instationnaires/commandés.

Organisation de la thèse

En résumé, cette thèse contribue à chacun des trois points mentionnés plus haut :

Contribution 1 Synthèse d'observateurs pour une forme triangulaire continue (relié au Point I.)

Contribution 2 Caractérisation des systèmes commandés pouvant être transformés en une forme triangulaire continue (relié au Point II.)

Contribution 3 Caractérisation des systèmes commandés pouvant être transformés en une forme linéaire Hurwitz (relié au Point II.)

Contribution 4 Méthode pour exprimer la dynamique de l'observateur directement dans les coordonnées du système pour éviter l'inversion de la transformation (relié au Point III.)

Au lieu de présenter les résultats chronologiquement, j'ai ainsi trouvé plus clair d'organiser ma thèse en suivant cette démarche à trois étapes, et donc de classifier les contributions en trois parties :

Partie I Formes normales et leurs observateurs (avec Contribution 1)

Partie II Transformation dans une forme normale (avec Contributions 2 et 3)

Partie III Expression de l'observateur dans les coordonnées du système (avec Contribution 4)

Les thèmes des Parties I. et II. ayant été intensivement étudiés dans la littérature, ce plan m'a aussi permis de faire apparaître un bilan détaillé des résultats existant en début de ces deux parties. Cette thèse donne donc finalement une bonne vue d'ensemble de l'état de l'art en matière d'observateur pour les systèmes non linéaires.

Enfin, j'ai aussi eu l'opportunité de travailler sur des applications, en particulier sur la synthèse d'observateurs pour moteurs synchrones à aimant permanent (MSAP) en l'absence d'informations mécaniques (sensorless) et avec certains paramètres inconnus. Ce travail a mené aux contributions suivantes :

Contribution 5 Observateur gradient pour l'estimation de la position du rotor et du flux de l'aimant dans un MSAP

Contribution 6 Analyse d'observabilité et synthèse d'observateur pour un MSAP dont la position du rotor et la résistance sont inconnues.

Ceci a été réalisé en parallèle et est donc détaillé dans une partie séparée et indépendante :

Partie IV Observateurs pour MSAPs aux paramètres inconnus (avec Contributions 5 et 6).

Publications

Les travaux présentés dans ce manuscrit ont fait l'objet des publications suivantes :

- Journaux internationaux avec comité de lecture

  1. P. Bernard, L. Praly, V. Andrieu, *Observers for a non-Lipschitz triangular form*, Automatica, Vol. 82, p301-313, 2017
  2. P. Bernard, L. Praly, V. Andrieu, *On the triangular normal form for uniformly observable controlled systems*, Automatica, Vol. 85, p293-300, 2017.
  3. P. Bernard, L. Praly, V. Andrieu, *Expressing an observer in given coordinates by augmenting and extending an injective immersion to a surjective diffeomorphism*, Soumis à SIAM.

- Conférences internationales avec comité de lecture

  1. P. Bernard, *Luenberger observers for non linear controlled systems*, Conference on Decision and Control, 2017 (A paraître)
  2. P. Bernard, L. Praly, *Robustness of rotor position observer for permanent magnet synchronous motors with unknown magnet flux*, IFAC World Congress, 2017
  3. P. Bernard, L. Praly, V. Andrieu, *Non Lipschitz triangular normal form for uniformly observable controlled systems*, IFAC Symposium on Nonlinear Control Systems, 2016
  4. P. Bernard, L. Praly, V. Andrieu, *Tools for observers based on coordinate augmentation*, Conference on Decision and Control, 2015

Chapter 2

Nonlinear observability and observer design problem

Chapitre 2 – Observabilité non-linéaire et synthèse d'observateur. *Ce chapitre présente brièvement la notion d'observabilité pour les systèmes non-linéaires commandés et introduit le problème de la synthèse d'observateur. La méthode introduite en introduction consistant à transformer le système dans une "forme normale" est formalisée et les notations utiles au reste de la thèse sont introduites.*

Contents

This first chapter introduces the problem of observer design for nonlinear controlled systems and presents some basic notions of observability which will be needed throughout the thesis. The subject under discussion here is well-established and widely described in the literature. Our aim is not to provide an exhaustive study on nonlinear observability and observer design, but rather to situate our contribution and introduce the basic tools/notations needed in the rest of this thesis.

2.1 Observation problem

We consider a general system of the form :

$$\dot{x} = f(x,u) \quad , \quad y = h(x,u) \tag{2.1}$$

with x the state in $\mathbb{R}^{d_x}$, u an input function with values in $\mathbb{R}^{d_u}$, y the output (or measurement) with values in $\mathbb{R}^{d_y}$ and f and h sufficiently many times continuously differentiable functions defined on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_u}$. We denote

- $X(x_0, t_0; t; u)$ the solution at time t of (2.1) with input u and passing through x_0 at time t_0. Most of the time, t_0 is the initial time 0 and x_0 the initial condition. In that case, we simply write $X(x_0; t; u)$.

- $Y(x_0, t_0; t; u)$ the output at time t of System (2.1) with input u passing through x_0 at time t_0 i-e :

$$Y(x_0, t_0; t; u) = h(X(x_0, t_0; t; u), u(t)) \ .$$

To alleviate the notations when $t_0 = 0$, we simply note $y_{x_0,u}$, i-e

$$y_{x_0,u}(t) = h(X(x_0; t; u), u(t)) \ .$$

Those notations are used to highlight the dependency of the output on the initial condition (and the input). When this is unnecessary, we simply write $y(t)$.

- $\mathcal{X}_0$ a subset of $\mathbb{R}^{d_x}$ containing the initial conditions that we consider for System (2.1). For any x_0 in $\mathcal{X}_0$, we denote $\sigma^+(x_0; u)$ (resp $\sigma^+_{\mathcal{X}}(x_0; u)$) the maximal time of existence of $X(x_0; \cdot; u)$ in $\mathbb{R}^{d_x}$ (resp in a set $\mathcal{X}$).

- $\mathcal{U}$ the set of all sufficiently many times differentiable inputs $u : [0, +\infty) \to \mathbb{R}^{d_u}$ which the system can be submitted to.

- U a subset of $\mathbb{R}^{d_u}$ containing all the values taken by the inputs $u \in \mathcal{U}$, i-e

$$\bigcup_{u \in \mathcal{U}} u([0, +\infty)) \ \subset U \ .$$

More generally, for an integer m such that any u in $\mathcal{U}$ is m times differentiable, $\overline{U}_m$ denotes a subset of $\mathbb{R}^{d_u(m+1)}$ containing the values taken by the inputs u in $\mathcal{U}$ and its first m derivatives, i-e

$$\bigcup_{u \in \mathcal{U}} \overline{u}_m([0, +\infty)) \ \subset \overline{U}_m \ ,$$

with $\overline{u}_m = (u, \dot{u}, \ldots, u^{(m)})$.

The object of this thesis is to address the following problem :

Observation problem

> For any input u in $\mathcal{U}$, any initial condition x_0 in $\mathcal{X}_0$, find an estimate $\hat{x}(t)$ of $X(x_0; t; u)$ based on the only knowledge of the input and output up to time t, namely $u_{[0,t]}$ and $y_{[0,t]}$, and so that $\hat{x}(t)$ asymptotically approaches $X(x_0; t; u)$, at least when $\hat{x}(t)$ is defined on $[0, +\infty)$.

Note that the solutions are defined from any points in $\mathbb{R}^{d_x}$, but we may choose to restrict our attention to those starting from a subset $\mathcal{X}_0$ of $\mathbb{R}^{d_x}$ (perhaps for physical reasons) and thus, we are only interested in estimating those particular solutions. Otherwise, take $\mathcal{X}_0 = \mathbb{R}^{d_x}$. As for the causality constraint that only the past values of the input $u_{[0,t]}$ can be used at time t, this may be relaxed in the case where the whole trajectory of u is known in advance, namely for a time-varying system.

The continuous differentiability of f says that any solution to System (2.1) is uniquely determined by its initial condition. Thus, the problem could be rephrased as : "given the input, find the only possible initial condition which could have produced the given output up to time t". Of course, this raises the question of uniqueness of the initial condition leading to a given output trajectory, at least after a certain time. This is related to the notion of observability which will be addressed later in this chapter. In any case, one could imagine simulating System (2.1) simultaneously for a set of initial conditions x_0 and progressively removing from the set those producing an output trajectory $Y(x_0; t; u)$ "too far" from $y(t)$ (with the notion of "far" to be defined). However, this method presents several drawbacks : first, one need to have a fairly precise idea of the initial condition to allow a trade off between number of computations and estimation precision, and second, it heavily relies on the model (2.1) which could be imperfect. This path has nevertheless aroused a lot of research :

- either through stochastic approaches, adding random processes to the dynamics (2.1) and to the measurement, and following the probability distribution of the possible values of the state ([Jaz70])
- or in a deterministic way, adding unknown admissible bounded disturbances to the dynamics (2.1) and to the measurement, and producing a "set-valued observer" or "interval observer" such as in [GRHS00, LZA03].

But as far as we know, no viable solution exist for standard nonlinear systems.

Another natural approach is the resolution of the minimization problem ([Zim94])

$$\hat{x}(t) = \mathsf{Argmin}_{\hat{x}} \int_0^t \left| Y(\hat{x}, t; \tau; u) - y(\tau) \right|^2 d\tau$$

or rather with finite memory

$$\hat{x}(t) = \mathsf{Argmin}_{\hat{x}} \int_{t-\bar{t}}^t \left| Y(\hat{x}, t; \tau; u) - y(\tau) \right|^2 d\tau \ .$$

Along this path, a first idea would be to integrate backwards the differential equation (2.1) for a lot of initial conditions $\hat{x}$ at time t until $t - \bar{t}$ and select the "best" one, but this would require a huge number of computations which would be impossible to carry out online and, as before, it would rely too much on the model. Some methods have nonetheless been developed to alleviate the number of computations and solve this optimization problem online, in spite of its non-convexity and the presence of local minima (see [Ala07] for a survey of existing algorithms).

In this thesis, the path we follow is rather to look for a dynamical system using the current value of the input and output and whose state is guaranteed to provide (at least asymptotically) enough information to reconstruct the state of System (2.1). This dynamical system is called an observer. A more rigorous mathematical definition is the following (a sketch is given in Figure 2.1).

Definition 2.1.1.

An *observer* for System (2.1) initialized in $\mathcal{X}_0$ is a couple $(\mathcal{F}, \mathcal{T})$ where

- $\mathcal{F} : \mathbb{R}^{d_z} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$ is continuous
- $\mathcal{T}$ is a family of continuous functions $\mathcal{T}_u : \mathbb{R}^{d_z} \times [0, +\infty) \to \mathbb{R}^{d_x}$, indexed by u in $\mathcal{U}$, which respect the causality[1] condition :

$$\forall \tilde{u} : [0, +\infty) \to \mathbb{R}^{d_u},\ \forall t \in [0, +\infty)\ ,\ u_{[0,t]} = \tilde{u}_{[0,t]} \implies \mathcal{T}_u(\cdot, t) = \mathcal{T}_{\tilde{u}}(\cdot, t)\ .$$

- for any u in $\mathcal{U}$, any z_0 in $\mathbb{R}^{d_z}$ and any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0; u) = +\infty$, any[2] solution $Z(z_0; t; u, y_{x_0,u})$ to

$$\dot{z} = \mathcal{F}(z, u, y_{x_0,u}) \tag{2.2}$$

initialized at z_0 at time 0, with input u and $y_{x_0,u}$, exists on $[0, +\infty)$ and is such that

$$\lim_{t \to +\infty} \left| \hat{X}((x_0, z_0); t; u) - X(x_0; t; u) \right| = 0 \tag{2.3}$$

with

$$\hat{X}((x_0, z_0); t; u) = \mathcal{T}_u\left(Z(z_0; t; u, y_{x_0,u}), t\right)\ .$$

[1] Again, this causality condition may be removed if the whole trajectory of u is explicitly known, for instance in the case of a time-varying system where $u(t) = t$ for all t.

[2] We say "any solution" because $\mathcal{F}$ being only continuous, there may be several solutions. This is not a problem as long as any such solution verifies the required convergence property.

In other words, $\hat{X}((x_0, z_0); t; u)$ is an estimate of the current state of System (2.1) and the error made with this estimation asymptotically converges to 0 as time goes to the infinity.

If $\mathcal{T}_u$ is the same for any u in $\mathcal{U}$ and is defined on $\mathbb{R}^{d_\xi}$ instead of $\mathbb{R}^{d_\xi} \times \mathbb{R}$, i-e is time-independent, $\mathcal{T}$ is said *stationary*. In this case, $\mathcal{T}$ directly refers to this unique function and we may simply say that

$$\dot{z} = \mathcal{F}(z, u, y) \quad , \quad \hat{x} = \mathcal{T}(z)$$

is an observer for System (2.1) initialized in $\mathcal{X}_0$.

In particular, we say that the observer is *in the given coordinates* if $\mathcal{T}$ is stationary and is a projection function from $\mathbb{R}^{d_z}$ to $\mathbb{R}^{d_x}$, namely $\hat{X}((x_0, z_0); t; u)$ can be read directly from d_x components of $Z(z_0; t; u, y_{x_0,u})$. In the particular case where $d_x = d_z$ and $\mathcal{T}$ is the identity function, we may omit to precise $\mathcal{T}$.

Finally, when $\mathcal{X}_0 = \mathbb{R}^{d_x}$, i-e the convergence is achieved for any initial condition of the system, we say "observer" without specifying $\mathcal{X}_0$.

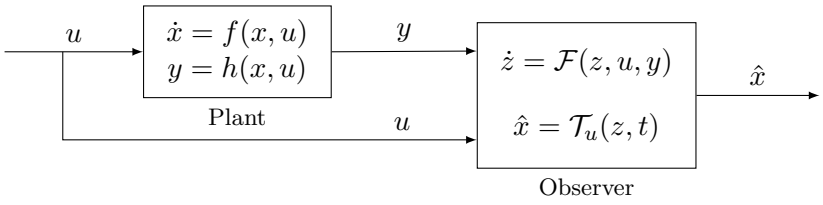


Figure 2.1: Observer : dynamical system estimating the state of a plant from the knowledge of its output and input only.

Remark 1 We will see in Chapter 4 that it is sometimes useful to write the observer dynamics (2.2) as a differential inclusion. In this case, $\mathcal{F}$ is a set-valued map and everything else remains unchanged.

The time-dependence of $\mathcal{T}_u$ enables to cover the case where the knowledge of the input and/or the output is used to build the estimate $\hat{x}$ from the observer state z. For example, using the output sometimes enables to reduce the dimension of the observer state (and thus alleviate the computations). However, for those so-called reduced-order observers, the estimate $\hat{x}$ depends directly on y and is therefore affected by measurement noise. This kind of observer won't be mentioned in this thesis. On the other hand, we will see that it is sometimes necessary to use the input (either implicitly or explicitly) in $\mathcal{T}_u$, but always keeping in mind the causality condition.

The advantage of having an observer in the given coordinates is that the estimate of the system state can directly be read from the observer state. This spares the maybe-complicated computation of $\mathcal{T}_u$. Writing the dynamics of the observer in the given coordinates constitutes one of the goals of this thesis, but we will see that unfortunately, it is not always possible, nor easy.

Anyhow, the role of an observer is to estimate the system state based on the knowledge of the input and output. This means that those signals somehow contain enough information to determine uniquely the whole state of the system. This brings us to the notion of observability.

2.2 Observability and observer design for nonlinear systems

2.2.1 Some notions of observability

In order to have an observer, a detectability property must be satisfied :

Lemma 2.2.1.

Assume there exists an observer for System (2.1). Then, System (2.1) is detectable for any u in $\mathcal{U}$, i-e for any u in $\mathcal{U}$ and for any (x_a, x_b) in $\mathcal{X}_0 \times \mathcal{X}_0$ such that $\sigma^+(x_a, u) = \sigma^+(x_b, u) = +\infty$ and

$$y_{x_a,u}(t) = y_{x_b,u}(t) \quad \forall t \geq 0 \ ,$$

we have

$$\lim_{t\to\infty} |X(x_a; t; u) - X(x_b; t; u)| = 0 \ .$$

The property of detectability says that even if two different initial conditions are not distinguishable with the output, the corresponding system solutions become close asymptotically and thus we still get a "good" estimate no matter which we pick. This is a well-known necessary condition which can be found for instance in [ABS13], and which admits the following straight-forward proof.

Proof : Consider any u in $\mathcal{U}$ and any (x_a, x_b) in $\mathcal{X}_0^2$ such that $\sigma^+(x_a, u) = \sigma^+(x_b, u) = +\infty$ and $y_{x_a,u} = y_{x_b,u}$. Take z_0 in $\mathbb{R}^{d_z}$ and pick a solution $Z(z_0; t; u; y_{x_a,u})$ of (2.2) with input $y_{x_a,u}$. It is also a solution to (2.2) with input $y_{x_b,u}$. Therefore, by denoting $\hat{X}((x_a, z_0); t; u) = \mathcal{T}\left(Z(z_0; t; u, y_{x_a,u}), u(t), y_{x_a,u}(t)\right)$, we have

$$\lim_{t\to\infty} |\hat{X}((x_a, z_0); t; u) - X(x_a; t; u)| = 0$$

and

$$\lim_{t\to\infty} |\hat{X}((x_a, z_0); t; u) - X(x_b; t; u)| = 0 \ .$$

The conclusion follows. ■

This means that detectability at least is necessary to be able to construct an observer. Actually, we often ask for stronger observability properties such as :

Definition 2.2.1.

Consider an open subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$. System 2.1 is

- *distinguishable* on $\mathcal{S}$ for some input $u : \mathbb{R} \to \mathbb{R}^{d_u}$ if
  for all (x_a, x_b) in $\mathcal{S} \times \mathcal{S}$,

$$y_{x_a,u}(t) = y_{x_b,u}(t) \quad \forall t \in [0, \min\{\sigma^+(x_a; u), \sigma^+(x_b; u)\}) \quad \Longrightarrow \quad x_a = x_b \ .$$

- *instantaneously distinguishable* on $\mathcal{S}$ for some input $u : \mathbb{R} \to \mathbb{R}^{d_u}$ if
  for all (x_a, x_b) in $\mathcal{S} \times \mathcal{S}$, for all $\bar{t}$ in $(0, \min\{\sigma^+(x_a; u), \sigma^+(x_b; u)\})$

$$y_{x_a,u}(t) = y_{x_b,u}(t) \quad \forall t \in [0, \bar{t}) \quad \Longrightarrow \quad x_a = x_b \ .$$

- *uniformly observable* on $\mathcal{S}$ if
  it is distinguishable on $\mathcal{S}$ for any input $u : \mathbb{R} \to \mathbb{R}^{d_u}$ (not only for u in $\mathcal{U}$).

- *uniformly instantaneously observable* on $\mathcal{S}$ if
  it is instantaneously distinguishable on $\mathcal{S}$ for any input $u : \mathbb{R} \to \mathbb{R}^{d_u}$ (not only for u in $\mathcal{U}$).

In particular, the notion of instantaneous distinguishability means that the state of the system can be uniquely deduced from the output of the system as quickly as we want. In the particular case where f, h and u are analytical, y is an analytical function of time ([Die60, 10.5.3]) and the notions of distinguishability and instantaneous distinguishability are equivalent

because two analytical functions which are equal on an interval are necessarily equal on their maximal interval of definition. Besides, for any x_0, there exists t_{x_0} such that

$$y_{x_0,u}(t) = \sum_{k=0}^{+\infty} \frac{y_{x_0,u}^{(k)}(0)}{k!}\, t^k \quad , \quad \forall t \in [0, t_{x_0}) \ ,$$

and distinguishability is thus closely related to the important notion of differential observability which will be defined in Chapter 5 and which roughly says that the state of the system at a specific time is uniquely determined by the value of the output and of its derivatives (up to a certain order) at that time.

The notion of uniform observability could appear unnecessary at first sight because it seems sufficient that the system be observable for any u in $\mathcal{U}$, namely for any considered input, rather than for any $u : \mathbb{R} \to \mathbb{R}^{d_u}$. However, we will see that this (strong) observability property infers some structural properties on the system which are useful for the design of certain observers.

In fact, more or less strong observability properties are needed depending on the observer design method and on what is required from the observer (tunability, exponential convergence etc). For example, it is shown in [ABS13], that for autonomous systems, instantaneous distinguishability is necessary to have a tunable observer, i-e an observer giving an arbitrarily small error on the estimate in an arbitrarily short time.

2.2.2 Observer design

It is proved in [ABS13] that if there exists an observer $(\mathcal{F}, \mathcal{T})$ for an autonomous system

$$\dot{x} = f(x) \quad , \quad y = h(x)$$

and a compact subset of $\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}$ which is invariant by the dynamics $(f, \mathcal{F})$, then there exist compact subsets $\mathcal{C}_x$ of $\mathbb{R}^{d_x}$ and $\mathcal{C}_z$ of $\mathbb{R}^{d_z}$, and a closed set-valued map T defined on $\mathcal{C}_x$ such that the set

$$\mathcal{E} = \{(x, z) \in \mathcal{C}_x \times \mathcal{C}_z \ : \ z \in T(x)\}$$

is invariant, attractive, and verifies :

$$\forall (x, z) \in \mathcal{E} \ , \quad \mathcal{T}(z, h(x)) = x \ .$$

In other words, the pair made of the system state x (following the dynamics f) and the observer state z (following the dynamics $\mathcal{F}$) converges necessarily to the graph of some set-valued map T and $\mathcal{T}$ is a left-inverse of this mapping. Note that this injectivity is of a peculiar kind since it is conditional to the knowledge of the output, namely "$x \mapsto T(x)$ is injective knowing $h(x)$". This result justifies the usual methodology of observer design for autonomous systems which consists in transforming, via a function T, the system into a form for which an observer is available, then design the observer in those new coordinates (i-e find $\mathcal{F}$), and finally deduce an estimate in the original coordinates via inversion of T (i-e find $\mathcal{T}$). Note that in practice, we look for a single-valued map T because it is simpler to manipulate than a set-valued map.

When considering a time-varying or controlled system, the same methodology can be used, but two paths are possible :

- either we keep looking for a stationary transformation $x \mapsto T(x)$ like for autonomous systems
- or we look for a time-varying transformation $(x, t) \mapsto T_u(x, t)$ which depends either explicitly or implicitly on the input u.

It is actually interesting to detail what we mean by explicitly/implicitly. In building a time-varying transformation, two approaches exist, each attached to a different vision of controlled systems :

- either we consider, as in System (2.1), that only the current value of the input (or sometimes the extended input $\overline{u}_m = (u, \dot{u}, \dots, u^{(m)})$) is necessary to determine $T_u(\cdot, t)$ at time t, i-e there exists a function $\tilde{T}$ such that for any u in $\mathcal{U}$, $T_u(x,t) = \tilde{T}(x, u(t))$.

- or we consider System (2.1) as a family of systems indexed by u in $\mathcal{U}$, i-e

$$\dot{x} = f_u(x) \quad , \quad y = h_u(x)$$

and we obtain a family of functions T_u, each depending on a whole function u in $\mathcal{U}$. In this case, it is necessary to ensure that $T_u(\cdot, t)$ depends only on the past values of u to guarantee causality.

Along this thesis, we will encounter/develop methods from each of those categories. In any case, here is a sufficient condition to build an observer for System (2.1) :

Theorem 2.2.1.

Consider an integer d_ξ and continuous maps $F : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$, $H : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_y}$ and $\mathcal{F} : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$ such that

$$\dot{\hat{\xi}} = \mathcal{F}(\hat{\xi}, u, \tilde{y}) \tag{2.4}$$

is an observer for[3]

$$\dot{\xi} = F(\xi, u, H(\xi, u)) \quad , \quad \tilde{y} = H(\xi, u) \tag{2.5}$$

i-e for any $(\hat{\xi}_0, \xi_0)$ in $(\mathbb{R}^{d_\xi})^2$ and any u in $\mathcal{U}$, any solution $\hat{\Xi}(\hat{\xi}_0; t; u, \tilde{y}_{\xi_0,u})$ of (2.4) and any solution $\Xi(\xi_0; t; u)$ of (2.5) verify

$$\lim_{t \to +\infty} \left| \hat{\Xi}(\hat{\xi}_0; t; u, \tilde{y}_{\xi_0,u}) - \Xi(\xi_0; t; u) \right| = 0 \ . \tag{2.6}$$

Now suppose that for any u in $\mathcal{U}$, there exists a continuous function $T_u : \mathbb{R}^{d_x} \times \mathbb{R} \to \mathbb{R}^{d_\xi}$ and a subset $\mathcal{X}$ of $\mathbb{R}^{d_x}$ such that :

a) for any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0; u) = +\infty$, $X(x_0; \cdot; u)$ remains in $\mathcal{X}$.

b) there exists a concave $\mathcal{K}$ function ρ and a positive real number $\bar{t}$ such that for all (x_a, x_b) in $\mathcal{X}^2$ and all $t \geq \bar{t}$

$$|x_a - x_b| \leq \rho\Big(|T_u(x_a, t) - T_u(x_b, t)|\Big) \ ,$$

i-e $x \mapsto T_u(x,t)$ becomes injective on $\mathcal{X}$, uniformly in time and in space, after a certain time $\bar{t}$.

c) T_u transforms System (2.1) into System (2.5), i-e for all x in $\mathcal{X}$ and all t in $[0, +\infty)$

$$L_{(f,1)} T_u(x,t) = F(T_u(x,t), u(t), h(x, u(t)) \quad , \quad h(x, u(t)) = H(T_u(x,t), u(t)) \ , \tag{2.7}$$

where $L_{(f,1)} T_u$ is the Lie derivative of T_u along the extended vector field $(f, 1)$, namely

$$L_{(f,1)} T_u(x,t) = \lim_{h \to 0} \frac{T_u(X(x,t;t+h;u), t+h) - T_u(x,t)}{h}$$

d) T_u respects the causality condition

$$\forall \tilde{u} : [0, +\infty) \to \mathbb{R}^{d_u}, \ \forall t \in [0, +\infty) \ , \ u_{[0,t]} = \tilde{u}_{[0,t]} \implies T_u(\cdot, t) = T_{\tilde{u}}(\cdot, t) \ .$$

[3]The expression of the dynamics under the form $F(\xi, u, H(\xi, u))$ can appear strange and abusive at this point because it is highly non unique and we should rather write $F(\xi, u)$. However, we will see in Part I how specific structures of dynamics $F(\xi, u, y)$ allow the design of an observer (2.4).

Then, for any u in $\mathcal{U}$, there exists a function $\mathcal{T}_u : \mathbb{R}^{d_\xi} \times [0, +\infty) \to \mathbb{R}^{d_x}$ such that for each $t \geq \bar{t}$, $\xi \mapsto \mathcal{T}_u(\xi, t)$ is uniformly continuous on $\mathbb{R}^{d_\xi}$ and verifies

$$\mathcal{T}_u(T_u(x,t),t) = x \quad \forall x \in \mathcal{X} \ .$$

Besides, denoting $\mathcal{T}$ the family of functions $\mathcal{T}_u$ for u in $\mathcal{U}$, $(\mathcal{F}, \mathcal{T})$ is an observer for System (2.1) initialized in $\mathcal{X}_0$.

Solving the partial differential equation (2.7) a priori gives a solution T_u depending on the whole trajectory of u and rather situates this result in the last design category presented above. But this formalism actually covers all three approaches and was chosen for its generality. In fact, the dependence of T_u on u may vary, but what is crucial is that they all transform the system into the same target form (2.5) for which an observer (2.4) is known.

Proof : Take u in $\mathcal{U}$. For any $t \geq \bar{t}$, $x \mapsto T_u(x,t)$ is injective on $\mathcal{X}$, thus there exists a function $T_{u,t}^{-1} : T_u(\mathcal{X}, t) \to \mathcal{X}$ such that for all x in $\mathcal{X}$, $T_{u,t}^{-1}(T_u(x,t)) = x$. Taking any $\tilde{u} : [0, +\infty) \to \mathbb{R}^{d_u}$ such that $u_{[0,t]} = \tilde{u}_{[0,t]}$ thus gives $T_{u,t}^{-1} = T_{\tilde{u},t}^{-1}$ on $T_u(\mathcal{X}, t) = T_{\tilde{u}}(\mathcal{X}, t)$ according to d). Besides, with b), for all (ξ_1, ξ_2) in $T_u(\mathcal{X}, t)^2$,

$$|T_{u,t}^{-1}(\xi_1) - T_{u,t}^{-1}(\xi_2)| \leq \rho\Big(|\xi_1 - \xi_2|\Big) \ . \tag{2.8}$$

Applying [McS34, Theorem 2] to each component of $T_{u,t}^{-1}$, there exist $c > 0$ and an extension[4] of $T_{u,t}^{-1}$ on $\mathbb{R}^{d_\xi}$ verifying (2.8) with $\bar{\rho} = c\rho$ for all (ξ_1, ξ_2) in $(\mathbb{R}^{d_\xi})^2$ (i-e $T_{u,t}^{-1}$ is uniformly continuous on $\mathbb{R}^{d_\xi}$) and such that $T_{u,t}^{-1} = T_{\tilde{u},t}^{-1}$ on $\mathbb{R}^{d_\xi}$. Defining $\mathcal{T}$ on $\mathbb{R}^{d_\xi} \times [0, +\infty)$ as

$$\mathcal{T}_u(\xi, t) = \begin{cases} T_{u,t}^{-1}(\xi) & , \text{ if } t \geq \bar{t} \\ 0 & , \text{ otherwise} \end{cases}$$

$\mathcal{T}_u$ verifies the causality condition and we have for all $t \geq \bar{t}$ and all (x, ξ) in $\mathcal{X} \times \mathbb{R}^{d_\xi}$,

$$|\mathcal{T}_u(\xi, t) - x| \leq \bar{\rho}\Big(|\xi - T_u(x,t)|\Big) \ . \tag{2.9}$$

Now consider x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0; u) = +\infty$. Then, from a) and c), since $X(x_0; \cdot; u)$ remains in $\mathcal{X}$ and $T_u(X(x_0; \cdot; u), t)$ is a solution to (2.5) initialized at $\xi_0 = T_u(x_0, 0)$ and for all t, $y_{x_0,u}(t) = \tilde{y}_{\xi_0,u}(t)$. Thus, because of (2.6), for any $\hat{\xi}_0$ in $\mathbb{R}^{d_\xi}$ and any solution $\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0,u})$ of

$$\dot{\hat{\xi}} = \mathcal{F}(\hat{\xi}, u, y_{x_0,u})$$

we have

$$\lim_{t \to +\infty} \Big| \hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0,u}) - T_u(X(x_0; t; u), t) \Big| = 0 \ .$$

If follows from (2.9) that

$$\lim_{t \to +\infty} \Big| \hat{X}((x_0, \hat{\xi}_0); t; u) - X(x_0; t; u) \Big| = 0$$

with $\hat{X}((x_0, \hat{\xi}_0); t; u) = \mathcal{T}_u(\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0,u}), t)$. Thus, $(\mathcal{F}, \mathcal{T})$ is an observer for System (2.1). ■

Remark 2 Without the assumption of concavity of ρ, it is still possible to show that $x \mapsto T_u(x,t)$ admits a continuous left-inverse $\mathcal{T}_u$ defined on $\mathbb{R}^{d_\xi}$. But, as shown in [SL16, Example 4], continuity of $\mathcal{T}$ is not enough to deduce the convergence of $\hat{x}$ from that of $\hat{\xi}$: uniform continuity[5] is necessary. Note that if $\mathcal{X}$ is bounded, the concavity of ρ is no longer a constraint, since a concave upper-approximation can always be obtained by saturation of ρ (see [McS34] for more details).

Besides, if there exists a compact set $\mathcal{C}$ such that $\mathcal{X}$ is contained in $\mathcal{C}$, it is enough to ensure the existence of ρ for (x_a, x_b) in $\mathcal{C}^2$. As long as for all t, $x \mapsto T_u(x,t)$ is injective on $\mathcal{C}$, then for all t, there exists a concave $\mathcal{K}$ function ρ_t verifying the required inequality for all (x_a, x_b) in $\mathcal{C}^2$

[4] Denoting $T_{u,t,j}^{-1}$ the jth component of $T_{u,t}^{-1}$, take $T_{u,t,j}^{-1}(\xi) = \min_{\tilde{\xi} \in T_u(\mathcal{X},t)} \{T_{u,t,j}^{-1}(\tilde{\xi}) + \rho(|\tilde{\xi} - \xi|)\}$ or equivalently $T_{u,t,j}^{-1}(\xi) = \min_{x \in \mathcal{X}} \{x_j + \rho(|T_u(x,t) - \xi|)\}$

[5] A function γ is uniformly continuous if and only if $\lim_{n \to +\infty} |x_n - y_n| = 0$ implies $\lim_{n \to +\infty} |\gamma(x_n) - \gamma(y_n)| = 0$. This property is indeed needed in the context of observer design.

(see Lemma A.3.2). Thus, only uniformity in time should be checked, namely that there exists a concave $\mathcal{K}$ function ρ greater than all the ρ_t, in other words that $x \mapsto T(x,t)$ does not become "less and less injective" with time. Of course, when T_u is time-independent, no such problem exists and it is sufficient to have $x \mapsto T_u(x)$ injective on $\mathcal{C}$. This is made precise in the following corollary.

Corollary 2.2.1.

Consider an integer d_ξ and continuous maps $F : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$, $H : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_y}$ and $\mathcal{F} : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$ such that (2.4) is an observer for (2.5). Suppose there exists a continuous function $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_\xi}$ and a compact set $\mathcal{C}$ of $\mathbb{R}^{d_x}$ such that :

- for any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0, u) = +\infty$, $X(x_0; \cdot; u)$ remains in $\mathcal{C}$.
- $x \mapsto T(x)$ is injective on $\mathcal{C}$.
- T transforms System (2.1) into System (2.5) on $\mathcal{C}$, i-e for all x in $\mathcal{C}$, all u in $\mathcal{U}$, all t in $[0, +\infty)$

$$L_{f(\cdot,u)}T(x) = F(T(x), u(t), h(x, u(t))) \quad , \quad h(x, u(t)) = H(T(x), u(t)) \ .$$

Then, there exists a uniformly continuous function $\mathcal{T} : \mathbb{R}^{d_\xi} \to \mathbb{R}^{d_x}$ such that

$$\mathcal{T}(T(x)) = x \quad \forall x \in \mathcal{C} \ ,$$

and $(\mathcal{F}, \mathcal{T})$ is an observer for System (2.1) initialized in $\mathcal{X}_0$.

Proof : This is a direct consequence of Lemma A.3.2 and Theorem 2.2.1. ■

2.3 Organization of the thesis

As illustrated in Figure 2.3, Theorem 2.2.1 shows that a possible strategy to design an observer is to transform the system into a favorable form (2.5) for which an observer is known, and then bring the estimate back into the initial coordinates by inverting the transformation. This design procedure is widely used in the literature and raises three crucial questions :

1. what favorable forms (2.5) do we know and which observers are they associated to ?
2. how to transform a given nonlinear system into one of those forms ?
3. how to invert the transformation ?

The present thesis contributes to each of those questions and is thus organized accordingly, dedicating one part to each of them. Since the first two have aroused a lot of research, detailed literature reviews are provided in each case to help the reader situate our contributions. As for the third one, it has not received a lot of attention as far as we know, although it constitutes a recurrent problem in practice.

To those contributions, we add in a fourth part the results obtained in parallel concerning observer design for permanent magnet synchronous motors with some unknown parameters.

Here is a more detailed account of the content of this thesis :

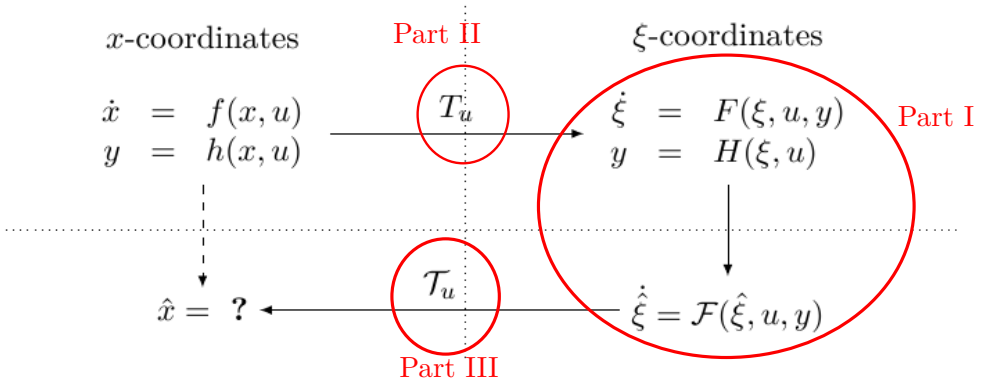


Figure 2.2: Process of observer design suggested by Theorem 2.2.1 and organization of the thesis.

Part I : Normal forms and their observers. We start by making a list of system structures (2.5) for which we know an observer (2.4). We call those favorable structures *normal forms*. Chapter 3 reviews the normal forms existing in the literature and recalls their associated observer : state-affine forms with Luenberger or Kalman observers and triangular forms with high gain, homogeneous or mixed high gain-Kalman observers. Noticing that few observers exist for non-Lipschitz triangular forms, we then fill this gap in Chapter 4, by extending the use of existing homogeneous observers to a broader class of Hölder triangular forms and proposing a new observer for the "only continuous" triangular form.

Part II : Transformation into a normal form. We address the problem of transforming a nonlinear system into one of the previously mentioned normal forms. In each case, sufficient observability conditions on the system are given. A lot of results in this area already exist in the literature and are recalled in Chapter 5. Then, we present in Chapters 6 and 7 our new results concerning the transformation of nonlinear systems into continuous triangular forms and Hurwitz forms.

Part III : Expression of the observer dynamics in the initial system coordinates. Although the observer design problem seems solved with Part I and II according to Theorem 2.2.1, implementation issues may arise such as the computation of the inverse $\mathcal{T}_u$ of the transformation. That is why, we develop in Part III a novel methodology to avoid the inversion of T_u by bringing the dynamics (2.4) back in the x-coordinates, i-e find $\hat{x}$ and obtain an observer *in the given coordinates* as defined in Definition 2.1.1. Although this process is quite common in the case where T_u is a diffeomorphism, completeness of solutions is not always ensured and we show how to solve this problem. Most importantly, we extend this method to the more complex situation where T_u is only an injective immersion, i-e the dimension of the observer state is larger than the one of the system state. This is done by adding some new coordinates to the system.

Part IV : Observers for permanent magnet synchronous motors with unkown parameters. This part gathers results concerning observability and observer design for permanent magnet synchronous motors when some parameters such as the magnet flux or the resistance are unknown. Simulations on real data are provided. This work was carried out in parallel and this part is mostly independent from the rest of the thesis.

Part I

Normal forms and their observers

Chapter 3

Quick review of existing normal forms and their observers

Chapitre 3 – Formes normales existantes et leurs observateurs *Ce chapitre présente les principales formes normales observables qui existent dans la littérature et pour chacune d'entre elles, rappelle la ou les observateurs associés. Deux principales catégories sont dissociées : d'une part les formes affines en l'état pour lesquelles des observateurs de Luenberger ou de Kalman sont utilisés, et d'autre part, les formes triangulaires auxquelles s'appliquent les observateurs de type grand gain.*

Contents

In this chapter we consider systems of the form[1]

$$\dot{\xi} = F(\xi, u, y) \quad , \quad y = H(\xi, u) \tag{3.1}$$

with ξ the state in $\mathbb{R}^{d_\xi}$, u an input with values in $U \subset \mathbb{R}^{d_u}$, y the output with values in $\mathbb{R}^{d_y}$ and F (resp H) a continuous function defined on $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u} \times \mathbb{R}^{d_y}$ (resp $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u}$). We are interested in finding normal forms, namely specific expressions of the functions F and H such that an explicit observer for System (3.1) can be written in the given coordinates[2], i-e the ξ-coordinates. Indeed, an a priori knowledge of such forms is necessary to apply Theorem 2.2.1 and design an observer for a nonlinear system.

We do not claim to be exhaustive, neither about the list of normal forms nor about their history. We select the most popular forms and associated observer, and endeavor to give the most sensible references. Our goal is only to introduce some definitions and results which will be of interest throughout this thesis, and give a starting point to the problem of observer design

[1]The notation $F(\xi, u, y)$ is somehow abusive because y is not an input to the dynamics of ξ. We should rather write $F(\xi, u, H(\xi, u))$ as in (2.5) but this latter notation is less straight-forward. We thus decided to keep the former for clarity.

[2]see Definition 2.1.1

for nonlinear systems. Note that according to Theorem 2.2.1, we are only interested in global observers with guaranteed convergence. This excludes for example the extended Kalman filters, obtained by linearizing the dynamics and the observation along the trajectory of the estimate ([Gel74]). Indeed, their convergence is only local in the sense that the estimate converges to the true state if the initial error is not too large and the linearization does not present any singularity ([BS15] and references therein).

Before giving the results of this chapter, we need the following definition.

Definition 3.0.1.

The *observability grammian* of a linear system of the form

$$\dot{\chi} = A(\nu)\chi \quad , \quad y = C(\nu)\chi$$

with input ν and output y, is the function defined by :

$$\Gamma_\nu(t_0, t_1) = \int_{t_0}^{t_1} \Psi_\nu(\tau, t_0)^\top C(\nu(\tau))^\top C(\nu(\tau))\Psi_\nu(\tau, t_0)\, d\tau$$

where Ψ_ν denotes the transition matrix[3], namely the unique solution to :

$$\begin{aligned} \frac{\partial \Psi_\nu}{\partial \tau}(\tau, t) &= A(\nu(\tau))\Psi_\nu(\tau, t) \\ \Psi_\nu(t, t) &= I \ . \end{aligned}$$

3.1 State-affine normal forms

In this section, we consider a system with dynamics of the form :

$$\dot{\xi} = A(u, y)\,\xi + B(u, y) \quad , \quad y = H(\xi, u) \tag{3.2}$$

where ξ is a vector of $\mathbb{R}^{d_\xi}$, $A : \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi \times d_\xi}$, $B : \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$ and $H : \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$ are continuous functions.

3.1.1 Constant linear part : Luenberger design

In this section, we consider the case where A is constant, with two sub-cases :

- A is Hurwitz and H any continuous function
- A is any matrix but H is linear.

A Hurwitz : Luenberger's original form

We introduce the following definition :

Definition 3.1.1.

We call *Hurwitz form* dynamics of the type:

$$\dot{\xi} = A\,\xi + B(u, y) \quad , \quad y = H(\xi, u) \ . \tag{3.3}$$

where A is a Hurwitz matrix in $\mathbb{R}^{d_\xi \times d_\xi}$ and B and H are continuous functions.

[3] See for instance [Che84]

For a Hurwitz form, a trivial observer is made of a copy of the dynamics of the system :

Theorem 3.1.1.

The system

$$\dot{\hat{\xi}} = A\,\hat{\xi} + B(u,y) \tag{3.4}$$

is an observer for system (3.3).

Indeed, the error $\hat{\xi} - \xi$ decays exponentially according to dynamics

$$\dot{\widehat{\hat{\xi} - \xi}} = A(\hat{\xi} - \xi)\ .$$

We have referred to this form as "Luenberger's original form" because originally in [Lue64], Luenberger's methodology to build observers for linear systems was to look for an invertible transformation which would map the linear system into a Hurwitz one, which admits a very simple observer. We will study in Part II under which condition a standard nonlinear system can be transformed into such a form, namely extend Luenberger's methodology to nonlinear systems.

H linear : $H(\xi,u) = C\xi$ with C constant

We consider now a system of the form[4]

$$\dot{\xi} = A\,\xi + B(u,y) \quad , \quad y = C\,\xi \tag{3.5}$$

where B is a continuous function. The following well-known result can be deduced from [Lue64]:

Theorem 3.1.2.

If the pair (A,C) is observable, there exists a matrix K such that $A - KC$ is Hurwitz. For any such matrix K, the system

$$\dot{\hat{\xi}} = A\hat{\xi} + B(u,y) + K(y - C\hat{\xi}) \tag{3.6}$$

is an observer for system (3.5).

As opposed to Theorem 3.1.1, A is not supposed Hurwitz but H is a linear function.

3.1.2 Time-varying linear part : Kalman design

We suppose in this section that H is linear, but not necessarily constant namely

$$\dot{\xi} = A(u,y)\,\xi + B(u,y) \quad , \quad y = C(u)\xi\ . \tag{3.7}$$

The most famous observer used for this kind of system is the Kalman and Bucy's observer presented in [KB61] for linear time-varying systems, i-e with $A(t)$, $B(t)$ and $C(t)$ replacing $A(u,y)$, $B(u,y)$ and $C(u)$ respectively. Later, a "Kalman-like" design was proposed in [HM90, BBH96] for the case where $A(u,y) = A(u)$. This design can be easily extended to System (3.7) by considering (u,y) as an extended input. The difference with the time-varying case studied by Kalman and Bucy in [KB61] is that every assumption must be verified uniformly for any such

[4] In [AK01], the authors propose an observer for a more general form $\dot{\xi} = A\,\xi + B(u,y) + G\rho(H\xi)$, $y = C\,\xi$, under certain conditions on ρ.

extended input, namely for any input u and for any output function y coming from any initial condition. To highlight this fact more rigorously, we denote

$$y_{\xi_0,u}(t) = C(u(t))\,\Xi(\xi_0;t;u)$$

the output at time t of system (3.7) initialized at ξ_0 at time 0.

Theorem 3.1.3. [HM90, BBH96]

Assume the input u is such that
- for any ξ_0, $t \mapsto A(u(t), y_{\xi_0,u}(t))$ is bounded by A_{max},
- for any ξ_0, the extended input $\nu = (u, y_{\xi_0,u})$ is regularly persistent for the auxiliary dynamics

$$\dot{\chi} = A(u, y_{\xi_0,u})\chi \quad , \quad y = C(u)\chi \tag{3.8}$$

uniformly with respect to ξ_0, i-e there exist strictly positive numbers t_0, $\bar{t}$ and α such that for any ξ_0 and any time $t \geq t_0$,

$$\Gamma_\nu(t, t+\bar{t}) \geq \alpha\, I$$

where Γ_ν is the observability grammian (see Definition 3.0.1) associated to System (3.8). Then, for any $\gamma > 2A_{max}$, there exist strictly positive numbers α_1 and α_2 such that the matrix differential equation

$$\dot{P} = -\gamma P - PA(u,y) - A(u,y)^\top P + C(u)^\top C(u) \tag{3.9}$$

initialized at $P(t_0) = P(t_0)^\top > 0$, admits a unique solution verifying for all $t \geq t_0$,

$$P^\top(t) = P(t) \quad , \quad \alpha_1 I \leq P(t) \leq \alpha_2 I\ .$$

Besides, the system

$$\dot{\hat{\xi}} = A(u,y)\,\hat{\xi} + B(u,y) + K\left(y - C(u)\hat{\xi}\right) \tag{3.10}$$

with the gain

$$K = P^{-1}C(u)^\top \tag{3.11}$$

is an observer for the state-affine system (3.7).

Remark 3

- It is important to note that K is time-varying and depends on the functions $t \mapsto u(t)$ and $t \mapsto y_{\xi_0,u}(t)$ and thus on ξ_0.
- The assumptions of boundedness of A and regular persistence are mainly to ensure that the solution to (3.9) is uniformly bounded from below and above, namely that P (and thus the gain K) neither goes to 0 nor to infinity.
- An equivalent way of writing (3.9) and (3.11) is with

$$\begin{aligned} \dot{P} &= A(u,y)P + PA(u,y)^\top - PC(u)^\top C(u)P + \gamma P \\ K &= P\,C(u)^\top \end{aligned}$$

(i-e P is replaced by P^{-1}). This implementation does not require the computation of the inverse of $P(t)$ at each step.
- Following Kalman and Bucy's original paper [KB61], the gain K can also be computed with

$\dot{P}(t) = A(u(t),y(t))P(t) + P(t)A(u(t),y(t))^\top$

$$-P(t)C(u(t))^\top R^{-1}(t)C(u(t))P(t) + D(t)Q(t)D(t)^\top$$
$$K(t) = P(t)C(u(t))^\top R^{-1}(t)$$

where $R(t)$ (resp $Q(t)$) is a positive definite matrix representing the covariance at time t of the noise which enters the measurement (resp the dynamics) and $D(t)$ describes how the noise enters the dynamics. In the case where those noises are independent white noise processes, this observer solves the following optimal problem : given the values of u and y up to time t, find an estimate $\hat{\xi}(t)$ of $\xi(t)$ which minimizes the conditional expectation $\mathsf{E}\left(|\hat{\xi}(t) - \xi(t)|^2 \,|\, y_{[t_0,t]}, u_{[t_0,t]}\right)$. In order to ensure asymptotic convergence of the observer, according to [KB61, Theorem 4], the following assumptions are needed :

- boundedness of A
- uniform complete observability of (A, C) : this corresponds to the regular persistence condition of Theorem 3.1.3 when A and C depend on an input u and A is bounded (see [Kal60])
- uniform complete controllability of (A, D) : this is the dual of uniform complete observability, namely uniform complete observability of $(A^\top, D^\top)$ (see [Kal60])
- R and Q are uniformly lower and upper-bounded in time.

Only the first two assumptions depend on the system and they are the same as in Theorem 3.1.3 ; the other two must be satisfied by an appropriate choice of the design parameters R and Q.

3.2 Triangular normal forms

3.2.1 Nominal form : high-gain designs

Triangular forms became of interest when [GB81] related their structure to uniformly observable systems, and when [Zei84] introduced the phase-variable form for differentially observable systems. The celebrated high gain observer proposed in [Tor89, EKNN89] for phase variable forms and later in [BH91, GHO92] for triangular forms, have been extensively studied ever since. It would be too long for the interest of this thesis to provide a thorough review of this literature, but we refer the interested reader to [KP13] and the references therein for a detailed analysis of the high gain design.

Definition 3.2.1.

We call *continuous triangular form* dynamics of the form:

$$\left\{\begin{array}{rcll} \dot{\xi}_1 &=& \xi_2 + \Phi_1(u, \xi_1) & \\ & \vdots & & \\ \dot{\xi}_i &=& \xi_{i+1} + \Phi_i(u, \xi_1, \ldots, \xi_i) & , \quad y = \xi_1 \\ & \vdots & & \\ \dot{\xi}_m &=& \Phi_m(u, \xi) & \end{array}\right. \tag{3.12}$$

where for all i in $\{1, ..., m\}$, ξ_i is in $\mathbb{R}^{d_y}$, $\xi = (\xi_1, \ldots, \xi_m)$ is in $\mathbb{R}^{d_\xi}$, with $d_\xi = m d_y$, $\Phi_i : \mathbb{R}^{d_u} \times \mathbb{R}^{i d_y} \to \mathbb{R}^{d_y}$ are continuous functions. In the particular case where only Φ_m is nonzero, we say *continuous phase-variable form.*

If now the functions $\Phi_i(u, \cdot)$ are globally Lipschitz on $\mathbb{R}^{i d_y}$ uniformly in u, namely there exists $\mathfrak{a}$ in $\mathbb{R}$ such that for all u in U, all (ξ_a, ξ_b) in $(\mathbb{R}^{d_\xi})^2$ and for all i in $\{1, ..., m\}$

$$|\Phi_i(u, \xi_{1a}, \ldots, \xi_{ia}) - \Phi_i(u, \xi_{1b}, \ldots, \xi_{ib})| \leq \mathfrak{a} \sum_{j=1}^{i} |\xi_{ja} - \xi_{jb}| \ ,$$

we say *Lipschitz triangular form* and *Lipschitz phase-variable form.*

Lipschitz triangular form

The Lipschitz triangular form is well-known because it allows the design of a high gain observer :

Theorem 3.2.1.

Suppose the functions $\Phi_i(u,\cdot)$ are globally Lipschitz on $\mathbb{R}^{id_y}$, uniformly in u. For any $(k_1,...,k_m)$ in $\mathbb{R}^m$ such that the roots of the polynomial

$$s^m + k_m s^{m-1} + \ldots + k_2 s + k_1$$

have strictly negative real parts, there exists L^* in $\mathbb{R}^+$ such that for any input function u with values in U, for any $L \geq L^*$, the system

$$\begin{cases} \dot{\hat{\xi}}_1 &= \hat{\xi}_2 + \Phi_1(u,\hat{\xi}_1) - L\, k_1\, (\hat{\xi}_1 - y) \\ \dot{\hat{\xi}}_2 &= \hat{\xi}_3 + \Phi_2(u,\hat{\xi}_1,\hat{\xi}_2) - L^2\, k_2\, (\hat{\xi}_1 - y) \\ &\vdots \\ \dot{\hat{\xi}}_m &= \Phi_m(u,\hat{\xi}) - L^m\, k_m\, (\hat{\xi}_1 - y) \end{cases} \tag{3.13}$$

is an observer for the Lipschitz triangular form (3.12).

Actually, extensions of this high gain observer exist for more complex triangular forms, in particular when each block does not have the same dimension, but extra assumptions on the dependence of the function Φ_i must be made to ensure convergence (see [BH91] or later [HBB10] for instance). We omit these here because they are of no use for this thesis.

In any cases, the standard implementation of a high gain observer necessitates the global Lipschitzness of the nonlinearities Φ_i. In the case where they are only locally Lipschitz, it is still possible to use observer (3.13) if the trajectories of the system evolve in a compact set, by saturating Φ_i outside this compact set (see Section 4.4). Otherwise, several researchers have tried to adapt the high gain L online by "following" the Lipschitz constant of Φ_i when it is observable from the output ([PJ04, AP05, APA09, SP11] and references therein).

Unfortunately, when the nonlinearities are only continuous, we will see in the next Chapter 4 that the convergence of the high gain observer can be lost, but that, under specific Hölder-like conditions, it still provides arbitrary small errors (by taking a sufficiently large gain). In particular, it has been known for a long time, mostly in the context of dirty-derivatives and output differentiation, that a high gain observer can provide an arbitrary small error for a phase-variable form as long as Φ_m is bounded ([Tor89] among many others).

Hölder continuous triangular form

Fortunately, moving to a generalization of high gain observers exploiting homogeneity makes it possible to achieve convergence in the case of non-Lipschitz nonlinearities verifying some Hölder conditions. It is at the beginning of the century that researchers started to consider homogeneous observers with various motivations: exact differentiators ([Lev b, Lev03, Lev05]), domination as a tool for designing stabilizing output feedback ([YL04], [Qia05], [QL06], [APA08] and references therein (in particular [APA06])), ... The advantage of this type of observers is their ability to face Hölder nonlinearities. In [Qia05], or in more general context in [APA08], the following observer design is used :

Theorem 3.2.2. [Qia05]

Consider a continuous triangular form (3.12). Assume there exists d_0 in $(-1, 0]$ and $\mathfrak{a}$ in $\mathbb{R}_+$ such that for all i in $\{1, \dots, m\}$, for all ξ_a and ξ_b in $\mathbb{R}^{d_\xi}$ and u in U

$$|\Phi_i(u, \xi_{1a}, \dots, \xi_{ia}) - \Phi_i(u, \xi_{1b}, \dots, \xi_{ib})| \leq \mathfrak{a} \sum_{j=1}^{i} |\xi_{ja} - \xi_{jb}|^{\frac{r_{i+1}}{r_j}} , \tag{3.14}$$

where r is a vector in $\mathbb{R}^{m+1}$, called weight vector, the components of which, called weights, are defined by

$$r_i = 1 - d_0(m - i) . \tag{3.15}$$

There exist $(k_1, \dots, k_m)$ and $L^* \geq 1$ such that for all $L \geq L^*$, the system[5]

$$\begin{cases} \dot{\hat{\xi}}_1 &= \hat{\xi}_2 + \Phi_1(u, \hat{\xi}_1) - L\, k_1 \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_2}{r_1}} \\ \dot{\hat{\xi}}_2 &= \hat{\xi}_3 + \Phi_2(u, \hat{\xi}_1, \hat{\xi}_2) - L^2\, k_2 \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_3}{r_1}} \\ &\vdots \\ \dot{\hat{\xi}}_m &= \Phi_m(u, \hat{\xi}) - L^m\, k_m \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_{m+1}}{r_1}} \end{cases} \tag{3.16}$$

is an observer for the continuous triangular form (3.12).

d_0 is called degree of the observer. When $d_0 = 0$, all the weights r_i are equal to 1, the nonlinearities are Lipschitz and we recover the high gain observer (3.13). In that sense, we can say that the homogeneous observer (3.16) is an extension of (3.13). Noticing that the Hölder constraints (3.14) become less and less restrictive as d_0 goes to -1, it is interesting to wonder what happens in the limit case where $d_0 = -1$. In that case, $r_{m+1} = 0$, which makes the last correction term of (3.16) equal to $\left\lfloor \hat{\xi}_1 - y \right\rceil^0 = \mathtt{sign}(\hat{\xi}_1 - y)$. This function being discontinuous at 0, the system becomes a differential inclusion when defining the sign function as the set valued map[6] :

$$\mathtt{S}(a) = \begin{cases} \{1\} & \text{if } a > 0\,, \\ [-1, 1] & \text{if } a = 0\,, \\ \{-1\} & \text{if } a < 0\,. \end{cases} \tag{3.17}$$

Note that this set valued map is upper semi-continuous with nonempty, compact and convex values, namely it verifies the usual basic conditions for existence of absolutely continuous solutions for differential inclusions given in [Fil88, Smi01].

Actually, when $d_0 = -1$, we recover the same correction terms as in the exact differentiator presented in [Lev b], where finite-time convergence is established for a phase-variable form with Φ_m is bounded. Quite naturally, this boundedness condition on Φ_m is exactly the condition we obtain when taking $d_0 = -1$ in the Hölder constraint (3.14). Actually, we will show in the next Chapter 4 that Theorem 3.2.2 still holds when allowing the degree to be -1, i-e that the exact differentiator presented in [Lev b] can also be used in presence of continuous nonlinearities on every line, provided they verify the Hölder constraint (3.14) with $d_0 = -1$.

Note that a generalization of observer (3.16) was presented in [APA08] in the context of "bi-limit" homogeneity, i-e for nonlinearities having two homogeneity degrees (around the origin and around infinity), namely

$$|\Phi_i(u, \xi_{1a}, \dots, \xi_{ia}) - \Phi_i(u, \xi_{1b}, \dots, \xi_{ib})| \leq \mathfrak{a}_0 \sum_{j=1}^{i} |\xi_{ja} - \xi_{jb}|^{\frac{r_{0,i+1}}{r_{0,j}}} + \mathfrak{a}_\infty \sum_{j=1}^{i} |\xi_{ja} - \xi_{jb}|^{\frac{r_{\infty,i+1}}{r_{\infty,j}}} ,$$

[5] We denote the signed power function as $\lfloor a \rceil^b = \mathtt{sign}(a)\,|a|^b$, for $b > 0$.

[6] Writing $c = \lfloor a \rceil^0$ will mean $c \in \mathtt{S}(a)$.

with

$$r_{0,i} = 1 - d_0(m-i) \quad , \quad r_{\infty,i} = 1 - d_\infty(m-i)$$

and $-1 < d_0 \leq d_\infty < \frac{1}{m+1}$. It would also be interesting to see if this design is still valid when $d_0 = -1$.

Continuous triangular form ?

The only existing observer we are aware of able to cope with Φ no more than continuous is the one presented in [BBD96]. Its dynamics are described by a differential inclusion[7]

$$\dot{\hat{\xi}} \in \mathcal{F}(\hat{\xi}, y, u)$$

where $(\hat{\xi}, y, u) \mapsto \mathcal{F}(\hat{\xi}, y, u)$ is a set valued map defined by : $(v_1, \ldots, v_m)$ is in $\mathcal{F}(\hat{\xi}, y, u)$ if there exists $(\tilde{\xi}_2, ..., \tilde{\xi}_m)$ in $\mathbb{R}^{m-1}$ such that

$$\begin{array}{rcl} v_1 & = & \tilde{\xi}_2 + \Phi_1(u, y) \\ \tilde{\xi}_2 & \in & \mathsf{sat}_{M_2}(\hat{\xi}_2) - k_1\, \mathtt{S}(y - \hat{\xi}_1) \\ & \vdots & \\ v_i & = & \tilde{\xi}_{i+1} + \Phi_i(u, y, \tilde{\xi}_2, \ldots, \tilde{\xi}_i) \\ \tilde{\xi}_{i+1} & \in & \mathsf{sat}_{M_{i+1}}(\hat{\xi}_{i+1}) - k_i\, \mathtt{S}(\hat{\xi}_i - \tilde{\xi}_i) \\ & \vdots & \\ v_m & \in & \Phi_m(u, y, \tilde{\xi}_2, \ldots, \tilde{\xi}_m) - k_m\, \mathtt{S}(\hat{\xi}_m - \tilde{\xi}_m) \end{array}$$

where sat is the saturation function

$$\mathsf{sat}_a(x) = \max\{\min\{x, a\}, -a\} \tag{3.18}$$

and M_i are known bounds for the solution. It can be shown that any absolutely continuous solution gives in finite time an estimate of ξ under the only assumption of boundedness of the input and of the state trajectory. But the set valued map $\mathcal{F}$ above does not satisfy the usual basic assumptions given in [Fil88, Smi01] (upper semi-continuous with non-empty, compact and convex values). It follows that we are not guaranteed of existence of absolutely continuous solutions nor of possible sequential compactness of such solutions and therefore of possibilities of approximations of $\mathcal{F}$.

That is why we dedicate the next Chapter 4 to the problem of designing observers for the continuous triangular forms. In particular, we propose a novel cascade of homogeneous observers whose convergence is established without requiring anything but the continuity of the nonlinearities and boundedness of trajectories.

3.2.2 General form : High gain-Kalman design

A more general triangular form is the following :

Definition 3.2.2.

[7] See Remark 1.

We call *general continuous triangular form* dynamics of the form

$$\left\{\begin{array}{rcl} \dot{\xi}_1 &=& A_1(u,y)\,\xi_2 + \Phi_1(u,\xi_1) \\ &\vdots& \\ \dot{\xi}_i &=& A_i(u,y)\,\xi_{i+1} + \Phi_i(u,\xi_1,\ldots,\xi_i) \\ &\vdots& \\ \dot{\xi}_m &=& \Phi_m(u,\xi) \end{array}\right. \quad , \quad y = C_1(u)\xi_1 \tag{3.19}$$

where for all i in $\{1,\ldots,m\}$, ξ_i is in $\mathbb{R}^{N_i}$, $\sum_{j=1}^{m} N_j = d_\xi$, $A_i : \mathbb{R}^{d_u} \times \mathbb{R}^{d_y} \to \mathbb{R}^{N_i \times N_{i+1}}$, $C_1 : \mathbb{R}^{d_u} \to \mathbb{R}^{d_y \times N_1}$, and $\Phi_i : \mathbb{R}^{d_u} \times \mathbb{R}^{\sum_{j=1}^{i} N_j} \to \mathbb{R}^{N_i}$ are continuous functions.

If besides the functions $\Phi_i(u,\cdot)$ are globally Lipschitz on $\mathbb{R}^i$ uniformly in u, then we will say *general Lipschitz triangular form.*

Note that when the values of the functions A_i are constant full-column rank matrices and $C_1(u)$ is the identity function, this form covers the standard triangular form (3.12) if $N_i = N_j$ for all (i,j), and also the forms studied in [BH91] or [HBB10]. In those cases, a high gain observer is possible because the system is observable for any input and the functions Φ are triangular and Lipschitz. When the dependence on the input and output is allowed in A_i however, the observability of the system depends on those signals and a high gain is no longer sufficient. In fact, System (3.19) is a combination of both (3.2) and (3.12). It is thus quite natural to combine both Kalman and high gain designs, as proposed in [Bes99] for the case where $N_i = 1$ for all i, and then in [BT07] for the general case.

In the following, we denote $y_{\xi_0,u}$ the output at time t of system (3.19) initialized at ξ_0 at time 0, and

$$A(u,y) = \begin{pmatrix} 0 & A_1(u,y) & 0 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ & & & & 0 \\ & & & & A_{m-1}(u,y) \\ 0 & & \ldots & & 0 \end{pmatrix} \quad , \quad C(u) = (C_1(u)\,,\,0\,,\ldots,\,0)$$

$$\Phi(u,\xi) = \begin{pmatrix} \Phi_1(u,\xi_1) \\ \vdots \\ \Phi_i(u,\xi_1,\ldots,\xi_i) \\ \vdots \\ \Phi_m(u,\xi_1,\ldots,\xi_m) \end{pmatrix} \quad , \quad \Lambda(L) = \begin{pmatrix} L\,I_{N_1} & 0 & \ldots & & 0 \\ 0 & \ddots & & & \\ \vdots & & L^i\,I_{N_i} & & \vdots \\ & & & \ddots & 0 \\ 0 & & \ldots & 0 & L^m\,I_{N_m} \end{pmatrix}$$

Theorem 3.2.3. [BT07]

Assume the input u is such that

a) For any ξ_0, $t \mapsto A(u(t), y_{\xi_0,u}(t))$ is bounded by A_{max},

b) for any ξ_0, the extended input $\nu = (u, y_{\xi_0,u})$ is locally regular for the dynamics

$$\dot{\chi} = A(u, y_{\xi_0,u})\chi \quad , \quad y = C(u)\chi \tag{3.20}$$

uniformly with respect to ξ_0, i-e there exist strictly positive real numbers α and L_0 such that for any ξ, any $L \geq L_0$ and any $t \geq \frac{1}{L}$,

$$\Gamma_\nu\left(t - \frac{1}{L}, t\right) \geq \alpha L\,\Lambda(L)^{-2}$$

where Γ_ν is the observability grammian (see Definition 3.0.1) associated to System (3.20).

c) the functions $\Phi_i(u,\cdot)$ are globally Lipschitz on $\mathbb{R}^{\sum_{j=1}^{i} N_j}$ uniformly in u,
Then, there exists a strictly positive real gain L^* such that for any $L \geq L^*$ and any $\gamma \geq 2A_{max}$, there exist strictly positive real numbers α_1 and α_2 such that the matrix differential equation

$$\dot{P} = L\left(-\gamma P - A(u,y)^\top P - PA(u,y) + C(u)^\top C(u)\right)$$

initialized at $P(0) = P(0)^\top > 0$ admits a unique solution verifying for all t

$$P(t)^\top = P(t) \quad , \quad \alpha_1 I \leq P(t) \leq \alpha_2 I \ .$$

Besides, the system

$$\dot{\hat{\xi}} = A(u,y)\,\hat{\xi} + \Phi(u,\hat{\xi}) + K\left(y - C(u)\hat{\xi}\right) \tag{3.21}$$

with gain

$$K = \Lambda(L)P^{-1}C(u)^\top$$

is an observer for the general Lipschitz triangular form (3.19).

As opposed to the classical Kalman observer (3.10), the input needs to be more than regularly persistent, namely to be locally regular. This is because in a high gain design, observability at arbitrarily short times is necessary. Note that in the case where the matrices A_i are of dimension one, [GK01, Lemma 2.1] shows that the gain K can be taken constant under the only condition that there exists A_{min} and A_{max} such that for any ξ_0,

$$0 < A_{min} < A_i(u(t), y_{\xi_0,u}(t)) < A_{max} \ .$$

3.3 Conclusion

We have introduced in this chapter the main normal forms and their associated observer design with guaranteed global convergence. They are summed up in Table 3.1.

Although the Lipschitz triangular form and its high gain observer have been widely studied, its continuous version has received little attention. This is quite unfortunate because in Part II, we will show that a large category of nonlinear systems can be transformed into this form, and not in the Lipschitz one. Partial solutions exist nevertheless, such as the homogeneous observer (3.16) when the nonlinearities verify some Hölder conditions. In the next Chapter 4, we show that the use of this type of observer can be extended to a broader class of Hölder nonlinearities and present a novel observer made of a cascade of homogeneous observers which requires only continuity of the nonlinearities and boundedness of trajectories : we are thus going to fill lines 6-7 of Table 3.1 which for now are empty.

Note that we concentrate our efforts on the continuous triangular form (3.12) because it is of special interest for Part II. But many of the techniques used in the following chapter should also be applicable to the general continuous triangular form (3.19) (lines 9 of Table 3.1).

Structure			Observability assumption	Observer design
State-affine forms	H nonlinear	A constant Hurwitz	$\emptyset$	copy of the dynamics
	H linear	A and C constant	(A, C) observable	Luenberger
		A or C non constant A bounded	(u, y) regularly persistent	Kalman
Triangular forms	Nominal (3.12)	Φ_i Lipschitz	$\emptyset$	High-gain
		Φ_i Hölder (3.14), $d_0 \in (-1, 0]$	$\emptyset$	Homogeneous of degree d_0
		Φ_i Hölder (3.14), $d_0 = -1$	?	?
		Φ_i continuous	?	?
	General (3.19)	Φ_i Lipschitz A_i bounded	(u, y) locally regular	High gain-Kalman
		Φ_i continuous ?	?	?

Table 3.1: Normal forms and their associated observer design

Chapter 4

Observers for the continuous triangular form

Chapitre 4 – Observateurs pour la forme triangulaire continue. *Dans ce chapitre, nous montrons qu'en l'absence de caractère Lipschitz et sous une condition de type Hölder, le grand gain usuel donne au mieux une convergence pratique, c'est -à-dire avec une erreur finale arbitrairement faible. Lorsque cette condition n'est pas satisfaite, nous proposons un nouvel observateur grand gain en cascade. Cependant, cette convergence pratique peut nécessiter l'emploi de très grands gains, ce qui devient problématique en présence de bruit de mesure. Sous une hypothèse un peu plus restrictive, nous montrons que des observateurs homogènes donnent par contre une convergence asymptotique. Comme pour le grand gain, nous proposons une cascade d'observateurs homogènes pour le cas où cette condition ne serait pas respectée. La convergence asymptotique est alors prouvée sous la seule hypothèse de continuité. Dans un souci de complétude, pour chaque observateur, des perturbations sur la dynamique et sur la mesure sont prises en compte, et les résultats sont énoncés sous la forme stabilité entrée-sortie.*

Contents

In this chapter, we address the problem of designing observers for the continuous triangular normal form (3.12). We will see in Chapter 6 that this form is useful for a certain category of systems, namely those which are uniformly observable and differentially observable at an order greater than the dimension of the system. Indeed, those systems may be transformed in a triangular form but with nonlinearities which may not be locally Lipschitz.

The content of this chapter has been published in [BPA17a].

In order to present results which are as complete as possible, we consider the continuous triangular form (3.12), but unlike in the rest of this thesis, we add some disturbances on the dynamics and on the measurement, namely[1]

$$\left\{\begin{array}{rcl} \dot{\xi}_1 &=& \xi_2 + \Phi_1(u,\xi_1) + w_1 \\ &\vdots& \\ \dot{\xi}_i &=& \xi_{i+1} + \Phi_i(u,\xi_1,\ldots,\xi_i) + w_i \quad , \quad y = \xi_1 + v \\ &\vdots& \\ \dot{\xi}_{d_\xi} &=& \Phi_{d_\xi}(u,\xi) + w_{d_\xi} \end{array}\right. \tag{4.1}$$

where ξ is the state in $\mathbb{R}^{d_\xi}$, y is a measured output in $\mathbb{R}$, Φ is a continuous function which is not assumed to be locally Lipschitz and (v,w) are time-functions which verify the Caratheodory conditions. w can model either a known or an unknown disturbance on the dynamics and v is an unknown disturbance.

We show in Section 4.1 that the classical high gain observer may still be used when the nonlinearities Φ_i verify some Hölder-type condition. Nevertheless, the asymptotic convergence is lost and only a convergence with an arbitrary small error remains.

On the other hand, according to Theorem 3.2.2, homogeneous observers enable to ensure asymptotic convergence in presence of Hölder nonlinearities. In particular, the homogeneous observer (3.16) with degree d_0 in $(-1,0]$ is built in [APA08] following a Lyapunov design. We show in Section 4.2 that the same Lyapunov design can be extended to the case where the degree of homogeneity is $d_0 = -1$. This is interesting since the Hölder constraints (3.14) on the nonlinearities become less and less restrictive as the degree gets closer to -1. It turns out that we recover with this method the exact differentiator presented in [Lev b] and which is defined by an homogeneous differential inclusion. As opposed to [Lev b] where convergence is established only for a phase-variable form via a solution-based analysis, in our case, convergence is guaranteed by construction for the triangular form since the Lyapunov design provides a strict homogeneous Lyapunov function which allows the presence of homogeneous disturbances on the dynamics. Actually, many efforts have been made to get expressions of Lyapunov functions for the output differentiator from [Lev b]. First limited to small dimensions (see [ORSM15]), it was only recently achieved (simultaneously to our work) at any dimension in [CZM16]. This approach is much harder since the authors look for a Lyapunov function for an already existing observer (Lyapunov analysis), while in our work, the observer and the Lyapunov function are built at the same time (Lyapunov design).

To face the unfortunate situation where the nonlinearities verify none of the above mentioned Hölder type conditions, we propose novel observers made of a cascade of high gain observers in Section 4.3.1 and of homogeneous observers in Section 4.3.2 of dimension less or equal to $\frac{d_\xi(d_\xi+1)}{2}$. We prove that the high gain version converges with an arbitrary small error, and the homogeneous version converges asymptotically, all this without requiring anything but continuity of the nonlinearities in the case where the system trajectories and the input are bounded.

All along this chapter, we sometimes use stronger assumptions than necessary in order to simplify the presentation of our results. We signal them to the reader with a (☼) symbol as in *"the trajectories are complete (☼)"*. We discuss how they can be relaxed later in Section 4.4, in particular when we restrict our attention to compact sets.

Finally, we illustrate our observers with an example in Section 4.5.

Notations

[1] To simplify the computations in this chapter, we consider the case $d_y = 1$, i-e each ξ_i is of dimension 1, but everything still holds for a block triangular form (3.12) with ξ_i of dimension $\mathbb{R}^{d_y}$.

For $(\xi_1,\dots,\xi_i)$ and $(\hat\xi_1,\dots,\hat\xi_i)$ (resp. $(\hat\xi_{i1},\dots,\hat\xi_{ii})$) in $\mathbb{R}^i$, we denote

$$\begin{array}{c}\boldsymbol{\xi}_i=(\xi_1,\dots,\xi_i)\quad,\quad \hat{\boldsymbol{\xi}}_i=(\hat\xi_1,\dots,\hat\xi_i)\quad(\text{resp.}\quad \hat{\boldsymbol{\xi}}_i=(\hat\xi_{i1},\dots,\hat\xi_{ii}))\\ e_{ij}=\hat\xi_{ij}-\xi_j\ ,\quad e_j=\hat\xi_j-\xi_j\ ,\ \mathbf{e}_i=\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i\ .\end{array}\tag{4.2}$$

To simplify the presentation, we assume that the solutions to (4.1) are defined for all $t\geq 0$ (i.e. the trajectories are complete (☼)). Besides, wanting to present the results in a unified and concise way, we will say that the function Φ verifies the property $\mathcal{H}(\alpha,\mathfrak{a})$ or a positive real number $\mathfrak{a}$, and a vector α in $[0,1]^{\frac{d_\xi(d_\xi+1)}{2}}$, if :

Property $\mathcal{H}(\boldsymbol{\alpha},\mathfrak{a})$ (☼)

For all i in $\{1,\dots,d_\xi\}$, for all ξ_a and ξ_b in $\mathbb{R}^{d_\xi}$ and u in U, we have[2] :

$$|\Phi_i(u,\boldsymbol{\xi}_{ia})-\Phi_i(u,\boldsymbol{\xi}_{ib})|\leq\mathfrak{a}\sum_{j=1}^{i}|\xi_{ja}-\xi_{jb}|^{\alpha_{ij}}\ .\tag{4.3}$$

This property captures many possible contexts. In the case in which $\alpha_{ij}>0$, it implies that the function Φ is Hölder with power α_{ij}. When the $\alpha_{ij}=0$, it simply implies that the function Φ is bounded.

It is possible to employ the degree of freedom given in (4.1) by the time functions w to deal with the case in which the given function $\Phi(u,\xi)$ doesn't satisfy $\mathcal{H}(\mathfrak{a},\alpha)$. In this case, an approximation procedure can be carried out to get a function $\hat\Phi$ satisfying $\mathcal{H}(\mathfrak{a},\alpha)$ and selecting $w=\Phi(u,\xi)-\hat\Phi(u,\xi)$ which is an unknown disturbance. The quality of the estimates obtained from the observer will then depend on the quality of the approximation (i-e the norm of w). This is what is done for example in [MV00] when dealing with locally Lipschitz approximations. We will further discuss in Section 4.4 how to relax assumption $\mathcal{H}(\mathfrak{a},\alpha)$.

4.1 High gain observer ?

We consider in this section the standard high gain observer already presented in the previous chapter

$$\left\{\begin{array}{rcl}\dot{\hat\xi}_1&=&\hat\xi_2+\Phi_1(u,\hat\xi_1)+\hat w_1-L\,k_1\,(\hat\xi_1-y)\\ \dot{\hat\xi}_2&=&\hat\xi_3+\Phi_2(u,\hat\xi_1,\hat\xi_2)+\hat w_2-L^2\,k_2\,(\hat\xi_1-y)\\ &\vdots&\\ \dot{\hat\xi}_{d_\xi}&=&\Phi_{d_\xi}(u,\hat\xi)+\hat w_{d_\xi}-L^{d_\xi}\,k_{d_\xi}\,(\hat\xi_1-y)\end{array}\right.\tag{4.4}$$

where L and the k_i's are gains to be tuned, y is the measurement. The $\hat w_i$ are approximations of the w_i. In particular, when w_i represents unknown disturbances, the corresponding $\hat w_i$ is simply set to 0. In the following, we denote

$$\Delta w=\hat w-w\ .$$

When Φ satisfies the property $\mathcal{H}(\alpha,\mathfrak{a})$ with $\alpha_{ij}=1$ for all $1\leq j\leq i\leq d_\xi$, we recognize the usual triangular Lipschitz property for which the nominal high-gain observer gives an input to state stability (ISS) property with respect to the measurement disturbance v and dynamics disturbance w. Specifically, we have the following well known result (see for instance [KP13] for a proof).

[2] Actually Φ_i can depend also on ξ_{i+1} to ξ_m as long as (4.3) holds. It can also depend on time requiring some uniform property (see Section 4.4).

Theorem 4.1.1. Nominal high-gain

There exist real numbers $k_1, \ldots, k_{d_\xi}$, L^*, λ, β and γ such that,
a) for all functions Φ satisfying$^{(\divideontimes)}$ for all i and for all $\boldsymbol{\xi}_{ia}$ and $\boldsymbol{\xi}_{ib}$ in $\mathbb{R}^i$

$$|\Phi_i(u, \boldsymbol{\xi}_{ia}) - \Phi_i(u, \boldsymbol{\xi}_{ib})| \leq \mathfrak{a} \sum_{j=1}^{i} |\xi_{ja} - \xi_{jb}| + \mathfrak{b}_i \tag{4.5}$$

b) for all $L \geq \max\{\mathfrak{a}L^*, 1\}$,
c) for all locally bounded time function $(u, v, w, \hat{w})$, all $(\xi_0, \hat{\xi}_0)$ in $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_\xi}$,
any solution $\hat{\Xi}(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})$ of (4.4) verifies, for all t_0 and t such that $t \geq t_0 \geq 0$, and for all i in $\{1, ..., d_\xi\}$,

$$\left|\hat{\Xi}_i(t) - \Xi_i(t)\right| \leq \max\left\{ L^{i-1}\beta \left|\hat{\Xi}_i(t_0) - \Xi_i(t_0)\right| e^{-\lambda L(t-t_0)}, \gamma \sup_{\substack{1\leq j\leq m \\ s\in[t_0,t]}} \left\{ L^{i-1} |v(s)|, \frac{|\Delta w_j(s)| + \mathfrak{b}_j}{L^{j-i+1}} \right\} \right\} \tag{4.6}$$

where we have used the abbreviations $\Xi(t) = \Xi(\xi_0; t; u, w)$ and $\hat{\Xi}(t) = \hat{\Xi}(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})$.

Since the nominal high-gain observer gives asymptotic convergence for Lipschitz nonlinearities, we may wonder what type of property is preserved when the nonlinearities are only Hölder. In the following theorem, we show that the usual high-gain observer can provide an arbitrary small error on the estimate provided the Hölder orders α_{ij} satisfy the restrictions given in Table 4.1 or Equation (4.7).

i \ j		1	2	...	$d_\xi-2$	$d_\xi-1$	d_ξ
1		$\frac{d_\xi-2}{d_\xi-1}$					
2		$\frac{d_\xi-3}{d_\xi-2}$	$\frac{d_\xi-3}{d_\xi-2}$				
$\vdots$	$\alpha_{ij} >$	$\vdots$	$\vdots$	$\ddots$			
$d_\xi-2$		$\frac{1}{2}$	$\frac{1}{2}$	...	$\frac{1}{2}$		
$d_\xi-1$		0	0	...	...	0	
d_ξ	$\alpha_{d_\xi j} \geq$	0	0	...	...	...	0

Table 4.1 : Hölder restrictions on Φ for arbitrarily small errors with a high gain observer.

Theorem 4.1.2.

Assume the function Φ verifies $\mathcal{H}(\alpha, \mathfrak{a})$ for some $(\alpha, \mathfrak{a})$ in $[0,1]^{\frac{d_\xi(d_\xi+1)}{2}} \times \mathbb{R}_+$ satisfying, for $1 \leq j \leq i$

$$\begin{array}{c} \frac{d_\xi-i-1}{d_\xi-i} < \alpha_{ij} \leq 1 \quad \text{for} \quad i = 1 \ldots, d_\xi - 1 \ , \\ 0 \leq \alpha_{d_\xi j} \leq 1 \end{array} \tag{4.7}$$

Then, there exist real numbers $k_1, \ldots, k_{d_\xi}$, such that, for all $\epsilon > 0$ we can find positive real numbers λ, β, γ, and L^* such that, for all $L \geq L^*$, for all locally bounded time function $(u, v, w, \hat{w})$ and all $(\xi_0, \hat{\xi}_0)$ in $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_\xi}$, any solution $\hat{\Xi}(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})$ of (4.4) verifies, for

all t_0 and t such that $t \geq t_0 \geq 0$, and for all i in $\{1, ..., d_\xi\}$,

$$\left|\hat{\Xi}_i(t) - \Xi_i(t)\right| \leq \max\left\{ \epsilon\,,\, L^{i-1}\beta \left|\hat{\Xi}_i(t_0) - \Xi_i(t_0))\right| e^{-\lambda L(t-t_0)}, \right.$$
$$\left. \gamma \sup_{\substack{1\leq j\leq d_\xi \\ s\in[t_0,t]}} \left\{ L^{i-1}\,|v(s)|, \frac{|\Delta w_j(s)|}{L^{j-i+1}} \right\} \right\}$$

where we have used the abbreviation $\Xi(t) = \Xi(\xi_0; t; u, w)$ and $\hat{\Xi}(t) = \hat{\Xi}(\hat{\xi}_0\xi_0; t; u, v, w, \hat{w})$.

Comparing this inequality with (4.6), we have now the arbitrarily small non zero ε in the right hand side but this is obtained under the Hölder condition instead of the Lipschitz one.

Proof : With Young's inequality, we obtain from (4.3) that, for all σ_{ij} in $\mathbb{R}_+$ and all $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ in $\mathbb{R}^{d_\xi}$

$$\left|\Phi_i(u, \hat{\boldsymbol{\xi}}_i) - \Phi_i(u, \boldsymbol{\xi}_i)\right| \;\leq\; \sum_{j=1}^{i} \mathfrak{a}_{ij}|\hat{\xi}_j - \xi_j| + \mathfrak{b}_{ij}\;, \tag{4.8}$$

with $\mathfrak{a}_{ij}$ and $\mathfrak{b}_{ij}$ defined as

$$\begin{cases} \mathfrak{a}_{ij} = 0\;,\; \mathfrak{b}_{ij} = \mathfrak{a}\;, & \text{if } \alpha_{ij} = 0 \\ \mathfrak{a}_{ij} = \mathfrak{a}^{\frac{1}{\alpha_{ij}}}\alpha_{ij}\sigma_{ij}^{\frac{1}{\alpha_{ij}}}\;,\; \mathfrak{b}_{ij} = \frac{1-\alpha_{ij}}{\sigma_{ij}^{\frac{1}{1-\alpha_{ij}}}} & \text{if } 0 < \alpha_{ij} < 1 \\ \mathfrak{a}_{ij} = \mathfrak{a}\;,\; \mathfrak{b}_{ij} = 0 & \text{if } \alpha_{ij} = 1 \end{cases} \tag{4.9}$$

With (4.8), the assumptions of Theorem 4.1.1 are satisfied with $\mathfrak{b}_i = \sum_{j=1}^{i}\mathfrak{b}_{ij}$. It gives $k_1, \ldots, k_{d_\xi}$, L^*, λ, β and γ and, if $L > \max_{i\geq j}\{\mathfrak{a}_{ij}L^*, 1\}$, the solution satisfies the ISS inequality (4.6). The result will follow if there exist L and σ_{ij} such that

$$L > \max_{i\geq j}\{\mathfrak{a}_{ij}L^*, 1\}\;,\quad \max_{i,j}\sum_{\ell=1}^{j}\gamma\mathfrak{b}_{j\ell}L^{i-j-1} \leq \epsilon\;. \tag{4.10}$$

At this point, we have to work with the expressions of $\mathfrak{a}_{ij}$ and $\mathfrak{b}_{j\ell}$ given in (4.9). From (4.7), α_{ij} can be zero only if $i = d_\xi$. And, when $\alpha_{d_\xi\ell} = 0$, we get

$$\gamma\mathfrak{b}_{d_\xi\ell}L^{i-d_\xi-1} \;=\; \gamma\mathfrak{a}L^{i-d_\xi-1} \;\leq\; \frac{\gamma\mathfrak{a}}{L}$$

Say that we pick $\sigma_{d_\xi\ell} = 1$ in this case. For all the other cases, we choose

$$\sigma_{j\ell} = \left(\frac{2j\gamma}{\epsilon}(1-\alpha_{j\ell})L^{(d_\xi-j-1)}\right)^{1-\alpha_{j\ell}}\;,$$

to obtain from (4.9)

$$\gamma\mathfrak{b}_{j\ell}L^{i-j-1} \;\leq\; \epsilon\frac{1}{j}\frac{1}{2L^{d_\xi-i}}\;.$$

So, with this selection of the $\sigma_{j\ell}$, the right inequality in (4.10) is satisfied for L sufficiently large. Then, according to (4.9), the a_{ij} are independent of L or proportional to $L^{(d_\xi-i-1)\frac{1-\alpha_{ij}}{\alpha_{ij}}}$. But with (4.7) we have

$$0 < (d_\xi - i - 1)\frac{1-\alpha_{ij}}{\alpha_{ij}} < 1\;.$$

This implies that $\frac{\mathfrak{a}_{ij}}{L}$ tends to 0 as L tends to $+\infty$. We conclude that (4.10) holds if we pick L sufficiently large. ■

It is interesting to remark the weakness of the assumptions imposed on the last two components of the function Φ. Indeed, (4.7) only imposes that $\Phi_{d_\xi-1}$ be Hölder without any restriction on the order, and that Φ_{d_ξ} be bounded (☼).

We have shown with Theorem 4.1.2 that one can hope to obtain an arbitrarily small error when taking the high gain L sufficiently large. In the next section, we show that actually asymptotic convergence can be achieved when considering homogeneous observers.

i \ j	1	2	...	$d_\xi-2$	$d_\xi-1$	d_ξ
1	$\frac{d_\xi-1}{d_\xi}$					
2	$\frac{d_\xi-2}{d_\xi}$	$\frac{d_\xi-2}{d_\xi-1}$				
$\vdots$	$\vdots$	$\vdots$	$\ddots$			
$d_\xi-2$	$\frac{2}{d_\xi}$	$\frac{2}{d_\xi-1}$	...	$\frac{2}{3}$		
$d_\xi-1$	$\frac{1}{d_\xi}$	$\frac{1}{d_\xi-1}$	...	...	$\frac{1}{2}$	
d_ξ	0	0	...	...	...	0

($\alpha_{ij} =$)

Table 4.2 : Hölder restrictions on Φ for a homogeneous observer with $d_0 = -1$.

4.2 Homogeneous observer

4.2.1 Main result

In this section, we consider the homogeneous observer (3.16) to which we add the estimation of the perturbations $\hat{w}_i$:

$$\begin{cases} \dot{\hat{\xi}}_1 &= \hat{\xi}_2 + \Phi_1(u,\hat{\xi}_1) + \hat{w}_1 - L\,k_1 \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_2}{r_1}} \\ \dot{\hat{\xi}}_2 &= \hat{\xi}_3 + \Phi_2(u,\hat{\xi}_1,\hat{\xi}_2) + \hat{w}_2 - L^2\,k_2 \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_3}{r_1}} \\ &\vdots \\ \dot{\hat{\xi}}_{d_\xi} &= \Phi_{d_\xi}(u,\hat{\xi}) + \hat{w}_{d_\xi} - L^{d_\xi}\,k_{d_\xi} \left\lfloor \hat{\xi}_1 - y \right\rceil^{\frac{r_{d_\xi+1}}{r_1}} \end{cases} \tag{4.11}$$

where r is the weight vector in $\mathbb{R}^{d_\xi+1}$ defined by

$$r_i = 1 - d_0(d_\xi - i)\ , \tag{4.12}$$

and where L and the k_i's are gains to be tuned, d_0 the degree to be chosen in $[-1,0]$. We have seen in Theorem 4.1.2 that the usual high-gain observer can provide an estimation with an arbitrary small error provided the nonlinearity satisfies the property $\mathcal{H}(\alpha,\mathfrak{a})$ with the α_{ij} verifying (4.7). But since [APA08] (see Theorem 3.2.2), we know that asymptotic estimation may be obtained with homogeneous correction terms and when considering nonlinearities which satisfies $\mathcal{H}(\alpha,\mathfrak{a})$ with the α_{ij} verifying

$$\alpha_{ij} = \frac{1 - d_0(d_\xi - i - 1)}{1 - d_0(d_\xi - j)} = \frac{r_{i+1}}{r_j}\ ,\ 1 \le j \le i \le d_\xi\ . \tag{4.13}$$

for some d_0 in $(-1,0]$. As announced in the introduction, we want to extend this result to the extreme case where $d_0 = -1$ i-e for nonlinearities satisfying $\mathcal{H}(\alpha,\mathfrak{a})$ with α_{ij} given in Table 4.2.

Theorem 4.2.1.

Assume that there exist d_0 in $[-1,0]$ and $\mathfrak{a}$ in $\mathbb{R}_+$ such that Φ satisfies $\mathcal{H}(\alpha,\mathfrak{a})$ with α verifying (4.13) (☼) . There exist $(k_1,\ldots,k_{d_\xi})$, such that for all $\overline{\mathrm{w}}_{d_\xi} > 0$ there exist $L^* \ge 1$ and a positive constant γ such that, for all $L \ge L^*$ there exists a class $\mathcal{KL}$ function β such that for all locally bounded time function $(u,v,w,\hat{w})$, and all $(\xi_0,\hat{\xi}_0)$ in $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_\xi}$ system (4.11) admits absolutely continuous solutions $\hat{\Xi}(\hat{\xi}_0,\xi_0;t;u,v,w,\hat{w})$ defined on $\mathbb{R}_+$ and for any such

solution the following implications hold for all t_0 and t such that $t \geq t_0 \geq 0$, and for all i in $\{1, ..., d_\xi\}$:

If $d_0 > -1$:

$$|\hat{\Xi}_i(t) - \Xi_i(t)| \leq \max \left\{ \beta(|\hat{\Xi}(t_0) - \Xi(t_0)|, t - t_0), \right. \\ \left. \gamma \sup_{\substack{1 \leq j \leq i \\ s \in [t_0, t]}} \left\{ L^{i-1} |v(s)|^{\frac{r_i}{r_1}}, \frac{|\Delta w_j(s)|^{\frac{r_i}{r_{j+1}}}}{L^{\mu_{ij}}} \right\} \right\} \quad (4.14)$$

where $\mu_{ij} = (j - i + 1)\frac{r_1}{r_{j+1}}$, and we have used the abbreviation $\Xi(t) = \Xi(\xi_0; t; u, w)$ and $\hat{\Xi}(t) = \hat{\Xi}(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})$.
Moreover, when $d_0 < 0$ and $v(t) = w_j(t) = 0$ for all t and $j = 1, \ldots, d_\xi$, there exists $\bar{t}$ such that $\hat{\Xi}(t) = \Xi(t)$ for all $t \geq \bar{t}$.

If $d_0 = -1$ and $|\Delta w_{d_\xi}(t)| \leq \overline{\mathrm{w}}_{d_\xi}$:

$$|\hat{\Xi}_i(t) - \Xi_i(t)| \leq \max \left\{ \beta(|\hat{\Xi}(t_0) - \Xi(t_0)|, t - t_0), \right. \\ \left. \gamma \sup_{\substack{1 \leq j \leq i-1 \\ s \in [t_0, t]}} \left\{ L^{i-1} |v(s)|^{\frac{r_i}{r_1}}, \frac{|\Delta w_j(s)|^{\frac{r_i}{r_{j+1}}}}{L^{\mu_{ij}}} \right\} \right\} \quad (4.15)$$

where μ_{ij}, $\Xi(t)$ and $\hat{\Xi}(t)$ are defined above.
Moreover, when $v(t) = w_j(t) = 0$ for all t and $j = 1, \ldots, d_\xi$, there exists $\bar{t}$ such that $\hat{\Xi}(t) = \Xi(t)$ for all $t \geq \bar{t}$.

Note that j is in $\{1, \ldots, i\}$ in (4.14) whereas it is in $\{1, \ldots, i-1\}$ in (4.15).

The proof of Theorem 4.2.1 for the case $d_0 \in (-1, 0]$ and without disturbances is given for example in [APA08]. Actually [APA08] gives a Lyapunov design of a generalized version of observer (4.11) with a recursive construction of both Lyapunov function and observer. Here we are concerned with the case $d_0 = -1$. In this limit case, observer (4.11) is a differential inclusion corresponding to the exact differentiator studied in [Lev b], where convergence is established in the particular case in which $\Phi_i = 0$ for $j = 1, \ldots, d_\xi - 1$ and Φ_{d_ξ} is bounded. We prove in Lemma 4.2.2 that the Lyapunov design of [APA08] can be extended to this case. This allows us to show that observer (4.11) still converges if, for each i, Φ_i is Hölder with order α_{ij} equal to the values given in Table 4.2, where i is the index of Φ_i and j is the index of e_j. We also recover the same bound in presence of a noise v as the one given in [Lev b]. Note that knowing the convergence of the exact differentiator from [Lev b], we could also have deduced the existence of such a Lyapunov function via a converse theorem as in [NYN04]. But with only existence, quantifying of the effect of the disturbances is nearly impossible.

Finally, it is interesting to remark that in the case $d_0 = -1$, the ISS property between the disturbance w_{d_ξ} and the estimation error is with restrictions as defined in [Tee96, Definition 3.1]. If $|\Delta w_{d_\xi}(t)| \leq \overline{\mathrm{w}}_{d_\xi}$ and L is chosen sufficiently large, then asymptotic convergence is obtained. However, nothing can be said when $|\Delta w_{d_\xi}| > \overline{\mathrm{w}}_{d_\xi}$. Moreover, it may be possible for a bounded large disturbance to induce a norm of the estimation error which goes to infinity. We believe that this problem could be solved employing homogeneous in the bi-limit observer as in [APA08]. It is shown to be doable in dimension 2 in [CZMF11].

Proof : The set-valued function $e_1 \mapsto \lfloor e_1 \rceil^0 = \mathtt{S}(e_1)$ defined in (3.17) is upper semi-continuous and has

convex and compact values. Thus, according to [Fil88], there exist absolutely continuous solutions to (4.11).

Let $\mathcal{L} = \mathsf{diag}(1, L, ..., L^{d_\xi - 1})$. The error $e = \hat{\xi} - \xi$ produced by the observer (4.11) satisfies

$$\dot{e} \in L A_{d_\xi} e + \delta + L\mathcal{L}\mathfrak{K}(e_1 + v) \tag{4.16}$$

where A_{d_ξ} is the shifting matrix of order d_ξ,

$$\delta = \Phi(u, \hat{\xi}) + \hat{w} - \Phi(u, \xi) - w \ ,$$

and $\mathfrak{K}$ is the homogeneous correction term the components of which are defined as

$$(\mathfrak{K}(e_1))_i = -k_i \lfloor e_1 \rceil^{\frac{r_{i+1}}{r_1}}$$

where $(k_1, \ldots, k_{d_\xi})$ are positive real number and r_i is defined in (4.12). In the scaled error coordinates $\varepsilon = \mathcal{L}^{-1} e$, those error dynamics read

$$\frac{1}{L}\dot{\varepsilon} \in A_{d_\xi}\varepsilon + \mathcal{D}_L + \mathfrak{K}(\varepsilon_1 + v) \tag{4.17}$$

with $\mathcal{D}_L = \mathcal{L}^{-1}\delta$. With this mind, the proof consists in finding an ISS homogeneous Lyapunov function for the L independent auxiliary system

$$\dot{\bar{e}} \in A_{d_\xi}\bar{e} + \mathfrak{K}(\bar{e}_1) \tag{4.18}$$

with state $\bar{e}$ in $\mathbb{R}^{d_\xi}$, then extending it to (4.17) by a robustness analysis, and finally deducing the result on (4.16).

Let $V : \mathbb{R}^{d_\xi} \to \mathbb{R}_+$ be the function defined as

$$V(\bar{e}) = \sum_{i=1}^{d_\xi - 1} \int_{\lfloor \bar{e}_{i+1} \rceil^{\frac{r_i}{r_{i+1}}}}^{\ell_i \bar{e}_i} \left[\lfloor \tau \rceil^{\frac{d_V - r_i}{r_i}} - \lfloor \bar{e}_{i+1} \rceil^{\frac{d_V - r_i}{r_{i+1}}} \right] d\tau + \frac{|\bar{e}_{d_\xi}|^{d_V}}{d_V}, \tag{4.19}$$

where d_V and ℓ_i are positive real numbers such that $d_V > 2d_\xi - 1$. It is shown in [APA08, Theorem 3.1] that, in the case where d_0 is in $(-1, 0]$, and by appropriately selecting the parameters ℓ_i and k_i, V is a strict C^1 Lyapunov function for the auxiliary system (4.18) and is homogeneous of degree d_V with weight vector r. In fact, the same construction is still valid for the case $d_0 = -1$ as stated in the following technical result, which is proved in the next subsection to ease the reading.

Lemma 4.2.1.

For all d_0 in $[-1, 0]$, the function V defined in (4.19) is positive definite and there exist positive real numbers $k_1, \ldots k_{d_\xi}$, $\ell_1, \ldots \ell_{d_\xi}$, λ, c_δ and c_v such that for all $\bar{e}$ in $\mathbb{R}^{d_\xi}$, $\bar{\delta}$ in $\mathbb{R}^{d_\xi}$ and $\bar{v}$ in $\mathbb{R}$ the following implication holds :

$$|\bar{\delta}_i| \le c_\delta V(\bar{e})^{\frac{r_{i+1}}{d_V}}, \quad \forall i\,, \quad \text{and} \quad |\bar{v}| \le c_v V(\bar{e})^{\frac{r_1}{d_V}}$$
$$\Longrightarrow {}^{3} \quad \max\left\{ \frac{\partial V}{\partial \bar{e}}(\bar{e})(A_{d_\xi}\bar{e} + \bar{\delta} + \mathfrak{K}(\bar{e}_1 + \bar{v})) \right\} \le -\lambda V(\bar{e})^{\frac{d_V + d_0}{d_V}}.$$

This Lemma says V is a ISS Lyapunov function for the auxiliary system (4.18). See [SW95, Proof of Lemma 2.14] for instance. With this result in hand a robustness analysis can be carried out on a system of the form (4.17).

Indeed, since Φ satisfies $\mathcal{H}(\alpha, \mathfrak{a})$, with (4.13) and $\frac{r_{i+1}}{r_j} \le 1$, we obtain, for all $L \ge 1$

$$\begin{aligned} |\mathcal{D}_{L,i}| &\le \frac{\mathfrak{a}}{L} \sum_{j=1}^{i} L^{(j-1)\frac{r_{i+1}}{r_j} - i + 1} |\varepsilon_j|^{\frac{r_{i+1}}{r_j}} + \frac{|\Delta w_i|}{L^i} \ , \\ &\le \frac{\mathfrak{a}}{L} \sum_{j=1}^{i} |\varepsilon_j|^{\frac{r_{i+1}}{r_j}} + \frac{|\Delta w_i|}{L^i} \ , \\ &\le \frac{c}{L} V(\varepsilon)^{\frac{r_{i+1}}{d_V}} + \frac{|\Delta w_i|}{L^i} \ , \end{aligned}$$

[3] Here the max is with respect to s in $\lfloor \bar{e}_1 + \bar{v} \rceil^0 = \mathtt{S}(\bar{e}_1 + \bar{v})$ appearing in the d_ξth component of $\mathfrak{K}(\bar{e}_1 + \bar{v})$ when $d_0 = -1$.

where c is a positive real number obtained from Lemma A.1.2 in Appendix A.1. With Lemma 4.2.1, where $\bar{\delta}_i$ plays the role of $\mathcal{D}_{L,i}$, $\bar{v}$ the role of v and $\bar{e}$ the role of ε, we obtain that, by picking L^* sufficiently large such that $\frac{c}{L^*} \leq \frac{c_\delta}{2}$, we have, for all $L > L^*$,

$$\left\{ \begin{array}{rl} \frac{|\Delta w_i|}{L^i} & \leq \frac{c_\delta}{4} V(\varepsilon)^{\frac{r_{i+1}}{d_V}} \ , \ \forall i \\ |v| & \leq c_v V(\varepsilon)^{\frac{r_1}{d_V}} \end{array} \right. \quad \Longrightarrow \quad \frac{1}{L} \max\left\{ \frac{\partial V}{\partial e}(\varepsilon)\dot{\varepsilon} \right\} \leq -\lambda V(\varepsilon)^{\frac{d_V+d_0}{d_V}} \ . \tag{4.20}$$

Now, when evaluated along a solution, ε gives rise to an absolutely continuous function $t \mapsto \varepsilon(t)$. Similarly the function defined by $t \mapsto \nu(t) = V(\varepsilon(t))$ is absolutely continuous. It follows that its time derivative is defined for almost all t and, according to [Smi01, p174], (4.20) implies, for almost all t,

$$\left\{ \begin{array}{rl} \frac{|\Delta w_i|}{L^i} & \leq \frac{c_\delta}{4} \nu(t)^{\frac{r_{i+1}}{d_V}} \ , \ \forall i \\ |v| & \leq c_v \nu(t)^{\frac{r_1}{d_V}} \end{array} \right. \quad \Longrightarrow \quad \frac{1}{L}\dot{\nu}(t) \leq -\lambda \nu(t)^{\frac{d_V+d_0}{d_V}} \ . \tag{4.21}$$

Here two cases have to be distinguished.

1. If d_0 is in $]-1, 0]$, with Lemma A.1.4 in Appendix A.1 (see also [SW95]), we get the existence of a class $\mathcal{KL}$ function β_V such that[4]

$$V(\varepsilon(t)) \leq \max_{i \in \{1,\ldots,d_\xi\}} \left\{ \beta_V(V(\varepsilon(0)), \lambda L t), \sup_{s \in [0,t]} \left\{ \left(\frac{4|\Delta w_i(s)|}{L^i c_\delta} \right)^{\frac{d_V}{r_{i+1}}}, \frac{|v(s)|^{\frac{d_V}{r_1}}}{c_v} \right\} \right\} \ .$$

The result holds since with Lemma A.1.2 there exist a positive real number c_1 such that

$$\left| \frac{e_i}{L^{i-1}} \right| \leq c_1 V(\epsilon)^{\frac{r_i}{d_V}} \ .$$

Moreover, when $v(t) = \Delta w_j(t) = \mathfrak{b}_j = 0$ for $j = 1, \ldots, d_\xi$, (4.21) implies finite time convergence in the case in which $d_0 < 0$.

2. If $d_0 = -1$, then $r_{d_\xi+1} = 0$. We choose L^* sufficiently large to satisfy

$$\frac{\bar{w}_{d_\xi}}{(L^*)^{d_\xi}} \leq \frac{c_\delta}{4} \ .$$

We obtain that the first condition in (4.21) is satisfied for $i = d_\xi$ when $L \geq L^*$. With Lemma A.1.4 in Appendix A.1 (see also [SW95]), the implication (4.21) implies the existence of a class $\mathcal{KL}$ function β_V such that[4]

$$V(\varepsilon(t)) \leq \max_{i \in \{1,\ldots,d_\xi-1\}} \left\{ \beta_V(V(\varepsilon(0)), \lambda L t), \sup_{s \in [0,t]} \left\{ \left(\frac{4|\Delta w_i(s)|}{L^i c_\delta} \right)^{\frac{d_V}{r_{i+1}}}, \frac{|v(s)|^{\frac{d_V}{r_1}}}{c_v} \right\} \right\} \ .$$

And the result holds as in the previous case.

■

4.2.2 Proof of Lemma 4.2.1

The proof is based on the following Lemma (4.2.2) which establishes that for a chain of integrator it is possible to construct homogeneous correction terms which provide an observer and that it is possible to construct a smooth strict homogeneous Lyapunov function.

Lemma 4.2.2.

For all d_0 in $[-1, 0]$, the function V defined in (4.19) is positive definite and there exists positive real numbers $k_1, \ldots k_{d_\xi}$, $\ell_1, \ldots \ell_{d_\xi}$, $\tilde{\lambda}$ such that for all $\bar{e}$ in $\mathbb{R}^{d_\xi}$, the following holds :

$$\max\left\{ \frac{\partial V}{\partial \bar{e}}(\bar{e}) \left(A_{d_\xi} \bar{e} + \mathfrak{K}(\bar{e}_1) \right) \right\} \ \leq \ -\tilde{\lambda} V(\bar{e})^{\frac{d_V+d_0}{d_V}} . \tag{4.22}$$

[4] according to Lemma A.1.4, $\beta_V(s,t) = \max\{0, s^{\frac{-d_0}{d_V}} - t\}^{\frac{d_V}{-d_0}}$

Proof : [Case $\mathbf{d_0 = -1}$ (see [APA08] otherwise)]

We denote $E_i = (\bar{e}_i, ..., \bar{e}_{d_\xi})$. Let d_V be an integer such that $d_V > 2d_\xi - 1$ and the functions $\mathfrak{K}_i$ recursively defined by :

$$\mathfrak{K}_{d_\xi}(\bar{e}_{d_\xi}) = -\left\lfloor \bar{e}_{d_\xi} \right\rceil^0 = -\mathtt{S}(\bar{e}_{d_\xi}) \quad , \quad \mathfrak{K}_i(\bar{e}_i) = \begin{pmatrix} -\left\lfloor \ell_i \bar{e}_i \right\rceil^{\frac{r_{i+1}}{r_i}} \\ \mathfrak{K}_{i+1}\left(\left\lfloor \ell_i \bar{e}_i \right\rceil^{\frac{r_{i+1}}{r_i}}\right) \end{pmatrix} .$$

Note that the jth component of $\mathfrak{K}_i$ is homogeneous of degree $r_{j+1} = d_\xi - j$ and, for any $\bar{e}_i$ in $\mathbb{R}$, the set $\mathfrak{K}_i(\bar{e}_i)$ can be expressed as

$$\mathfrak{K}_i(\bar{e}_i) = \{\tilde{\mathfrak{K}}_i(\bar{e}_i, s)\,, \quad s \in \mathtt{S}(\bar{e}_i)\} \ ,$$

where $\tilde{\mathfrak{K}}_i : \mathbb{R} \times [-1, 1] \to \mathbb{R}$ is a continuous (single valued) function.

Let $V_{d_\xi}(\bar{e}_{d_\xi}) = \frac{|\bar{e}_{d_\xi}|^{d_V}}{d_V}$ and for all i in $\{1, \ldots, d_\xi - 1\}$, let also $\bar{V}_i : \mathbb{R}^2 \to \mathbb{R}$ and $V_i : \mathbb{R}^{d_\xi - i + 1} \to \mathbb{R}$ be the functions defined by

$$\bar{V}_i(\nu, \bar{e}_{i+1}) = \int_{\left\lfloor \bar{e}_{i+1} \right\rceil^{\frac{r_i}{r_{i+1}}}}^{\nu} \left\lfloor x \right\rceil^{\frac{d_V - r_i}{r_i}} - \left\lfloor \bar{e}_{i+1} \right\rceil^{\frac{d_V - r_i}{r_{i+1}}} dx \ ,$$

$$V_i(E_i) = \sum_{j=d_\xi - 1}^{i} \bar{V}_j(\ell_j \bar{e}_j, \bar{e}_{j+1}) + V_{d_\xi}(\bar{e}_{d_\xi}) \ .$$

With these definitions, the Lyapunov function V defined in (4.19) is simply $V(e) = V_1(e)$ and the homogeneous vector field $\mathfrak{K}(\bar{e}_1) = \mathfrak{K}_1(\bar{e}_1)$ with

$$k_i = \ell_i^{\frac{r_{i+1}}{r_i}} \ell_{i-1}^{\frac{r_{i+1}}{r_{i-1}}} ... \ell_2^{\frac{r_{i+1}}{r_2}} \ell_1^{\frac{r_{i+1}}{r_1}} \ .$$

The proof of Proposition 4.2.2 is made iteratively from $i = d_\xi$ toward 1. At each step, we show that V_i is positive definite and we look for a positive real number ℓ_i, such that for all E_i in $\mathbb{R}^{d_\xi - i + 1}$

$$\max_{s \in \mathtt{S}(\bar{e}_i)} \left\{ \frac{\partial V_i}{\partial E_i}(E_i)(A_{d_\xi - i + 1} E_i + \tilde{\mathfrak{K}}_i(\bar{e}_i, s)) \right\} \leq -c_i V_i(E_i)^{\frac{d_V - 1}{d_V}} \ , \tag{4.23}$$

where c_i is a positive real number. The lemma will be proved once we have shown that the former inequality holds for $i = 1$.

Step $i = d_\xi$: At this step, $E_{d_\xi} = \bar{e}_{d_\xi}$. Note that we have

$$\max_{s \in \mathtt{S}(\bar{e}_{d_\xi})} \left\{ \frac{\partial V_{d_\xi}}{\partial E_{d_\xi}}(E_{d_\xi}) \tilde{\mathfrak{K}}_{d_\xi}(\bar{e}_{d_\xi}, s) \right\} = -|E_{d_\xi}|^{d_V - 1} = -c_{d_\xi} V_{d_\xi}(E_{d_\xi})^{\frac{d_V - 1}{d_V}} \ ,$$

with $c_{d_\xi} = d_V^{\frac{d_V - 1}{d_V}}$. Hence, equation (4.23) holds for $i = d_\xi$.

Step $i = j$: Assume V_{j+1} is positive definite and assume there exists $(\ell_{j+1}, \ldots, \ell_{d_\xi})$ such that (4.23) holds for $j = i - 1$. Note that the function $x \mapsto \left\lfloor x \right\rceil^{\frac{d_V - r_j}{r_j}} - \left\lfloor \bar{e}_{i+1} \right\rceil^{\frac{d_V - r_j}{r_{j+1}}}$ is strictly increasing, is zero if and only if $x = \left\lfloor \bar{e}_{j+1} \right\rceil^{\frac{r_j}{r_{j+1}}}$, and therefore has the same sign as $x - \left\lfloor \bar{e}_{j+1} \right\rceil^{\frac{r_j}{r_{j+1}}}$. Thus, for any $\bar{e}_{j+1}$ fixed in $\mathbb{R}$, the function $\nu \mapsto \overline{V}_j(\nu, \bar{e}_{j+1})$ is non negative and is zero only for $v = \left\lfloor \bar{e}_{j+1} \right\rceil^{\frac{r_j}{r_{j+1}}}$. Thus, $\bar{V}_j$ is positive and we have

$$V_j(E_j) = 0 \iff \begin{cases} V_{j+1}(E_{j+1}) = 0 \\ \overline{V}_j(\ell_j \bar{e}_j, \bar{e}_{j+1}) = 0 \end{cases} \iff \begin{cases} E_{j+1} = 0 \\ \ell_j \bar{e}_j = \left\lfloor \bar{e}_{j+1} \right\rceil^{\frac{r_j}{r_{j+1}}} = 0 \end{cases}$$

so that V_j is positive definite.

On another hand, let $\tilde{V}_j(\nu, E_{j+1}) = V_{j+1}(E_{j+1}) + \bar{V}_j(\nu, \bar{e}_{j+1})$ and let T_1 be the function defined

$$T_1(\nu, E_{j+1}) = \max_{s \in \mathtt{S}(\nu)} \left\{ \tilde{T}_1(\nu, E_{j+1}, s) \right\}$$

with $\tilde{T}_1$ continuous and defined by

$$\tilde{T}_1(\nu, E_{j+1}, s) = \frac{\partial \tilde{V}_j}{\partial E_{j+1}}(E_{j+1})(A_{d_\xi - i - 1} E_{i+1} + \tilde{\mathfrak{K}}_{j+1}(\left\lfloor \nu \right\rceil^{\frac{r_{j+1}}{r_j}}, s)) + \frac{c_{j+1}}{2} \tilde{V}_j(\nu, E_{j+1})^{\frac{d_V - 1}{d_V}} \ .$$

Let also T_2 be the continuous real-valued function defined by

$$T_2(v, E_{j+1}) = -\frac{\partial \tilde{V}_j}{\partial \nu}(\nu, E_{i+1})(\bar{e}_{j+1} - \lfloor \nu \rceil^{\frac{r_{j+1}}{r_j}}) \ .$$

Note that T_1 and T_2 are homogeneous with weight r_j for ν and r_i for $\bar{e}_i$ and degree $d_V - 1$. Besides, they verify the following two properties :

-for all E_{j+1} in $\mathbb{R}^{d_\xi - j}$, ν in $\mathbb{R}$

$$T_2(\nu, E_{j+1}) \geq 0$$

(since $(\lfloor \nu \rceil^{\frac{r_{j+1}}{r_j}} - \bar{e}_{j+1})$ and $(\lfloor \nu \rceil^{\frac{d_V - r_j}{r_j}} - \lfloor \bar{e}_{j+1} \rceil^{\frac{d_V - r_j}{r_{j+1}}})$ have the same sign)

-for all (ν, E_{j+1}) in $\mathbb{R}^{d_\xi - j+1} \setminus \{0\}$, and s in $\mathtt{S}(\nu)$, we have the implication

$$T_2(\nu, E_{j+1}) = 0 \quad \Longrightarrow \quad \tilde{T}_1(\nu, E_{j+1}, s) < 0$$

since T_2 is zero only when $\lfloor \nu \rceil^{\frac{r_{j+1}}{r_j}} = \bar{e}_{j+1}$ and

$$\tilde{T}_1(\lfloor \bar{e}_{j+1} \rceil^{\frac{r_{j+1}}{r_j}}, E_{j+1}, s) = \frac{\partial V_{j+1}}{\partial E_{j+1}}(E_{j+1})(A_{n-i}E_{j+1} + \tilde{\mathfrak{K}}_{j+1}(\bar{e}_{j+1}, s)) \\ + \frac{c_{j+1}}{2} V_{j+1}(E_{j+1})^{\frac{d_V - 1}{d_V}} \leq -\frac{c_{j+1}}{2} V_{j+1}(E_{j+1})^{\frac{d_V - 1}{d_V}} \ ,$$

where we have employed (4.23) for $i = j - 1$.

Using Lemma A.1.3 in Appendix A.1, there exists ℓ_j such that

$$T_1(\nu, E_{j+1}) - \ell_j T_2(\nu, E_{j+1}) \leq 0 \ , \ \forall \ (\nu, E_{j+1}) \ .$$

Finally, note that

$$\max_{s \in \mathtt{S}(\bar{e}_i)} \left\{ \frac{\partial V_j}{\partial E_j}(E_j)(A_{m-j+1}E_j + \tilde{\mathfrak{K}}_j(\bar{e}_j, s)) \right\} = T_1(\ell_j \bar{e}_j) - \ell_j T_2(\ell_j \bar{e}_j, E_{j+1}) - \frac{c_{j+1}}{2} V_j(E_j)^{\frac{d_V - 1}{d_V}}$$

Hence, (4.23) holds for $i = j$. ■

We are now ready to finish the proof of Lemma 4.2.1. Let $\tilde{\mathfrak{K}}(\bar{e}_1, s)$ be the function defined as

$$\left(\tilde{\mathfrak{K}}(\bar{e}_1, s)\right)_i = (\mathfrak{K}(\bar{e}_1))_i \ , \ i \in [1, d_\xi - 1] \ ,$$

and,

$$\left(\tilde{\mathfrak{K}}(\bar{e}_1, s)\right)_{d_\xi} = \begin{cases} k_{d_\xi} s \ , & \text{when } d_0 = -1 \\ (\mathfrak{K}(\bar{e}_1))_{d_\xi} \ , & \text{when } d_0 > -1 \end{cases} \ .$$

$\tilde{\mathfrak{K}}$ is a continuous (single) real-valued function which satisfies for all $\bar{e}_1$ in $\mathbb{R}$

$$\mathfrak{K}(\bar{e}_1) = \{\tilde{\mathfrak{K}}(\bar{e}_1, s) \, , \quad s \in \mathtt{S}(\bar{e}_1)\} \ .$$

Consider also the functions

$$\tilde{\eta}(\bar{e}, \bar{\delta}, \bar{v}, s) = \frac{\partial V}{\partial \bar{e}}(\bar{e})(A_{d_\xi}\bar{e} + \bar{\delta} + \tilde{\mathfrak{K}}(\bar{e}_1 + \bar{v}, s)) + \frac{\tilde{\lambda}}{2} V(\bar{e})^{\frac{d_V + d_0}{d_V}} ,$$

and

$$\gamma(\bar{\delta}, v) = \sum_{i=1}^{d_\xi} |\bar{\delta}_i|^{\frac{d_V + d_0}{r_{i+1}}} + |\bar{v}|^{\frac{d_V + d_0}{r_1}} .$$

With (4.22), we invoke Lemma A.1.3 to get the existence of a positive real number c_1 satisfying for all s in $\mathtt{S}(\bar{e}_1 + \bar{v})$:

$$\frac{\partial V}{\partial \bar{e}}(\bar{e})(A_{d_\xi}\bar{e} + \bar{\delta} + \tilde{\mathfrak{K}}(\bar{e}_1 + \bar{v}, s)) \leq -\frac{\tilde{\lambda}}{2} V(\bar{e})^{\frac{d_V + d_0}{d_V}} + c_1 \sum_{i=1}^{d_\xi} \bar{\delta}_i^{\frac{d_V + d_0}{r_{i+1}}} + c_1 |\bar{v}|^{\frac{d_V + d_0}{r_1}} \ .$$

This can be rewritten,

$$\frac{\partial V}{\partial \bar{e}}(\bar{e})(A_{d_\xi}\bar{e}+\bar{\delta}+\tilde{\mathfrak{K}}(\bar{e}_1+v,s)) \leq -\frac{\tilde{\lambda}}{2(d_\xi+2)}V(\bar{e})^{\frac{d_V+d_0}{d_V}} + \sum_{i=1}^{d_\xi}\left(c_1|\bar{\delta}_i|^{\frac{d_V+d_0}{r_{i+1}}} - \frac{\tilde{\lambda}}{2(d_\xi+2)}V(\bar{e})^{\frac{d_V+d_0}{d_V}}\right) + c_1|\bar{v}|^{\frac{d_V+d_0}{r_1}} - \frac{\tilde{\lambda}}{2(d_\xi+2)}V(\bar{e})^{\frac{d_V+d_0}{d_V}}.$$

Consequently, the result of Lemma 4.2.1 holds with $\lambda = \frac{\tilde{\lambda}}{2(d_\xi+2)}$, $c_\delta = c_v = \left(\frac{\lambda}{c_1}\right)^{\frac{r_1}{d_V+d_0}}$.

4.3 Cascade of observers

4.3.1 High gain cascade

According to Theorem 4.1.2, the classical high gain observer can provide an arbitrary small error when the last nonlinearity is only bounded and when there is no disturbance. We exploit here this observation by proposing the following cascaded high gain observer to deal with the case where the functions Φ_i do not satisfy (4.7):

$$\begin{cases} \dot{\hat{\xi}}_{11} = \hat{w}_1 - L_1\, k_{11}\,(\hat{\xi}_{11}-y) \\ \cdots\cdots \\ \dot{\hat{\xi}}_{21} = \hat{\xi}_{22}+\Phi_1(u,\hat{\xi}_{11}) + \hat{w}_1 - L_2\, k_{21}\,(\hat{\xi}_{21}-y) \\ \dot{\hat{\xi}}_{22} = \hat{w}_2 - L_2^2\, k_{22}\,(\hat{\xi}_{21}-y) \\ \cdots\cdots \\ \quad\vdots \\ \cdots\cdots \\ \dot{\hat{\xi}}_{d_\xi 1} = \hat{\xi}_{d_\xi 2}+\Phi_1(u,\hat{\xi}_{(d_\xi-1)1}) + \hat{w}_1 - L_{d_\xi}\, k_{d_\xi 1}\,(\hat{\xi}_{d_\xi 1}-y) \\ \dot{\hat{\xi}}_{d_\xi 2} = \hat{\xi}_{d_\xi 3}+\Phi_2(u,\hat{\xi}_{(d_\xi-1)1},\hat{\xi}_{(d_\xi-1)2}) + \hat{w}_2 - L_{d_\xi}^2\, k_{d_\xi 2}\,(\hat{\xi}_{d_\xi 1}-y) \\ \quad\vdots \\ \dot{\hat{\xi}}_{d_\xi d_\xi} = \hat{w}_{d_\xi} - L_{d_\xi}^{d_\xi}\, k_{d_\xi d_\xi}\,(\hat{\xi}_{d_\xi 1}-y) \end{cases} \tag{4.24}$$

with the gain k_{ij} chosen as in a classical high gain observer of dimension i, $\hat{w}_i$ are estimations of w_i and L_i are the high gains parameters to be chosen. It is important to notice that the arguments of all the nonlinearities Φ_j in block i come from the block $i-1$ (thanks to triangularity) and that Φ_i is not present (because we saw that a bounded error is allowed on the last line of a high gain observer).

Assuming the input function and the system solution are bounded, it is shown in the following that estimation with an arbitrary small error can be achieved by the cascaded high-gain observer (4.24).

Theorem 4.3.1.

Assume Φ is continuous. For any positive real numbers $\overline{\xi}$ and $\overline{\mathsf{u}}$, for any strictly positive real number ϵ, there exist a choice of $(k_{11},\ldots,k_{d_\xi d_\xi})$ and of $(L_1,\ldots,L_{d_\xi})$, a class $\mathcal{KL}$ function β and two class $\mathcal{K}_\infty$ functions γ_1 and γ_2 such that, for all locally bounded time function $(u,v,w,\hat{w})$, for all $(\xi_0,\hat{\xi}_0)$ in $\mathbb{R}^{d_\xi}\times\mathbb{R}^{d_\xi}$ and for all t such that $|\Xi(\xi_0;s;u,w)| \leq \overline{\xi}$ and $|u(s)| \leq \overline{\mathsf{u}}$ for all $0 \leq s \leq t$, any solution $\left(\hat{\mathbf{\Xi}}_1(\hat{\xi}_0,\xi_0;t;u,v,w,\hat{w}),\ldots,\hat{\mathbf{\Xi}}_{d_\xi}(\hat{\xi}_0,\xi_0;t;u,v,w,\hat{w})\right)$

of (4.24) verifies, for all i in $\{1,\dots,d_\xi\}$,

$$|\hat{\boldsymbol{\Xi}}_i(t)-\boldsymbol{\Xi}_i(t)|\le \max\left\{\varepsilon\,,\ \beta\left(\sum_{j=1}^{i}|\hat{\xi}_j-\xi_j|,t\right)\,,\ \sup_{s\in[0,t]}\left\{\gamma_1(|v(s)|),\gamma_2(|\Delta w(s)|)\right\}\right\}$$

where $\hat{\boldsymbol{\Xi}}_i$ is the state of the ith block (see Notation (4.2)) and we have used the abbreviation $\hat{\boldsymbol{\Xi}}_i(t)=\hat{\boldsymbol{\Xi}}_i(\hat{\xi}_0,\xi_0;t;u,v,w,\hat{w})$ and $\boldsymbol{\Xi}_i(t)=\boldsymbol{\Xi}_i(\xi_0;t;u,w)$.

Proof : This result is nothing but a straightforward consequence of the fact that a cascade of ISS systems is ISS.

Specifically the error system attached to the high gain observer in block i has state $\mathbf{e}_i$ (see Notation (4.2)) and input v and $\delta_{i\ell}$ defined as[5]

$$\begin{array}{rcl}\delta_{i\ell} &=& \left[\Phi_\ell(u,\hat{\boldsymbol{\xi}}_{(i-1)})-\Phi_\ell(u,\boldsymbol{\xi}_{(i-1)})\right]+[\hat{w}_\ell-w_\ell]\\ \delta_{ii} &=& -\xi_{i+1}-\Phi_i(u,\boldsymbol{\xi}_i)+\hat{w}_i-w_i\end{array}$$

with $\xi_{d_\xi+1}=0$. With Theorem 4.1.1, we have the existence of $k_{i1},\dots,k_{ii}$, λ_i, β_i and γ_i such that we have, for all $L_i\ge 1$, all $t\ge t_i\ge 0$, all j in $\{1,\dots,i\}$ and with $e_{ij}(t)$ denoting the jth error in the ith block evaluated along the solution at time t,

$$|e_{ij}(t)|\le \max\left\{L_i^{j-1}\beta_i\,|\mathbf{e}_i(t_i)|\,e^{-\lambda_i L_i(t-t_i)},\gamma_i\sup_{\substack{1\le\ell\le j\\ s\in[t_i,t]}}\left\{L_i^{j-1}\,|v(s)|,\frac{|\delta_{i\ell}(s)|}{L_i^{\ell-j+1}}\right\}\right\}\ .$$

But according to Lemma A.2.1, the continuity of the Φ_ℓ implies the existence of a function ρ of class $\mathcal{K}$ such that, for all ℓ in $\{1,\dots,d_\xi\}$ and for all $(\boldsymbol{\xi}_{(i-1)},\hat{\boldsymbol{\xi}}_{(i-1)},u)$ in $\mathbb{R}^{i-1}\times\mathbb{R}^{i-1}\times U$ satisfying $|\boldsymbol{\xi}_{(i-1)}|\le\overline{\xi}$ and $|u|\le\overline{\mathrm{u}}$,

$$|\Phi_\ell(u,\hat{\boldsymbol{\xi}}_{(i-1)})-\Phi_\ell(u,\boldsymbol{\xi}_{(i-1)})|\le\rho\left(|\mathbf{e}_{(i-1)}|\right)\ .$$

This implies

$$\begin{array}{rcl}|\delta_{i\ell}(s)| &\le& \rho(|\mathbf{e}_{i-1}(s)|)+|\Delta w_\ell(s)|\ ,\ \ell=1,\dots,j-1\ ,\\ |\delta_{ii}(s)| &\le& \overline{\xi}_{i+1}+\overline{\Phi}_i+|\Delta w_i(s)|\ ,\end{array}$$

where $\overline{\Phi}_i=\max_{|u|\le\overline{\mathrm{u}},|\boldsymbol{\xi}_i|\le\overline{\xi}}|\Phi_i(u,\boldsymbol{\xi}_i)|$ and $\overline{\xi}_i$ is a bound for $|\Xi_i(\xi,s;u,w)|$ (which is less than $\overline{\xi}$). Hence, we have the existence of c_i independent of L_i such that

$$|\mathbf{e}_i(t)|\le c_i\max\left\{L_i^{i-1}\,|\mathbf{e}_i(t_i)|\,e^{-\lambda_i L_i(t-t_i)}\,,\ \sup_{s\in[t_i,t]}L_i^{i-1}\,|v(s)|\,,\right.$$
$$\left.\sup_{s\in[t_i,t]}\frac{\rho(|\mathbf{e}_{i-1}(s)|)}{L_i^{2-i}}\,,\ \sup_{\substack{1\le\ell\le i\\ s\in[t_i,t]}}\frac{|\Delta w_\ell(s)|}{L_i^{\ell-i+1}}\,,\ \frac{\overline{\xi}_{i+1}+\overline{\Phi}_i}{L_i}\right\}\ .$$

This makes precise what we wrote above that we have a cascade of ISS systems. Hence (see [Son89, Prop. 7.2]), for each i in $\{1,\dots,m\}$, there exist a class $\mathcal{KL}$ function $\bar{\beta}_i$ and class $\mathcal{K}$ functions γ_{vi} and γ_{wi}, each depending on L_1 to L_i and such that we have, for all $t\ge 0$,

$$|\mathbf{e}_i(t)|\le\max\left\{\bar{\beta}_i\left(\max_{j\in\{1,\dots,i\}}\{|\mathbf{e}_j(0)|\},t\right),\varpi_i\,,\ \sup_{s\in[0,t]}\{\gamma_{vi}(|v(s)|),\gamma_{wi}(|\Delta w(s)|)\}\right\}\ .$$

where ϖ_i is a positive real number defined by the sequences

$$\varpi_1=c_1\frac{\overline{\xi}_2+\overline{\Phi}_1}{L_1}\ ,\ \varpi_i=c_i\max\left\{\frac{\overline{\xi}_{i+1}+\overline{\Phi}_i}{L_i},\frac{\rho(\varpi_{i-1})}{L_i^{2-i}}\right\}\ .$$

Then by picking $L_i\ge L_i^*$ where L_i^* is defined recursively as :

$$\epsilon_{d_\xi}=\epsilon\ ,\quad \epsilon_i=\min\left(\epsilon,\rho^{-1}\left(\frac{\epsilon_{i+1}}{c_{i+1}L_{i+1}^{i-2}}\right)\right)$$
$$L_{d_\xi}^*=\frac{c_{d_\xi}\overline{\Phi}_{d_\xi}}{\varepsilon_{d_\xi}}\ ,\quad L_i^*=\frac{c_i[\overline{\xi}_{i+1}+\overline{\Phi}_i]}{\varepsilon_i}$$

[5] We write $\Phi_\ell(u,\boldsymbol{\xi}_{(i-1)})$ although $\boldsymbol{\xi}_{(i-1)}$ is the state of the previous block of dimension $i-1$, which can be larger than ℓ. We should rather have introduced a symbol for the ℓ-first coordinates of $\boldsymbol{\xi}_{(i-1)}$, but we thought this would unnecessarily complexify the notations. Indeed, for the present proof, we only need to know that those variables come from the previous block of dimension $i-1$.

we obtain $\varpi_i \leq \epsilon$ for all i, hence the result. ■

Note that unlike for the previous observers, we cannot state the result with "there exists $(L_1^*, \ldots, L_{d_\xi}^*)$ such that for any $(L_1, \ldots, L_{d_\xi})$ satisfying $L_i \geq L_i^*$ for all i, [...]" because L_i^* depends on $(L_{i+1}, \ldots, L_{d_\xi})$ and not $(L_{i+1}^*, \ldots, L_{d_\xi}^*)$.

This observer has the advantage of working without any assumption on the nonlinearities besides their continuity. Note however that it requires the knowledge of a bound on the system solution and on the input. Also we may not need to build d_ξ blocks, since according to Theorem 4.1.2, we need to create a new block only for the indexes i where Φ_i does not verify Property $\mathcal{H}(\alpha, \mathfrak{a})$ for any $\mathfrak{a} \geq 0$ and with α satisfying (4.7). Unfortunately, as it appears from the proof of Theorem 4.3.1, the choice of $(L_1, ..., L_{d_\xi})$ can be complicated. Besides, only a convergence with an arbitrary small error is obtained. It may thus be necessary to take very large gains which is problematic in terms of peaking (see [Kha02, Section 14.5] for instance) and most importantly in presence of noise (see Section 4.5). In the following section, we design a similar cascade observer, but with homogeneous correction terms, and show that it enables to obtain asymptotic convergence.

4.3.2 Homogeneous cascade

When we cannot find d_0 in $[-1, 0]$ and $\mathfrak{a}$ such that the nonlinearities satisfy $\mathcal{H}(\alpha, \mathfrak{a})$, with α defined in (4.13), we may lose the convergence of observer (4.11), or the possibility of making the final error arbitrarily small. In such a bad case, we can still take advantage of the fact that, for α verifying (4.13) with $d_0 = -1$, $\mathcal{H}(\alpha, \mathfrak{a})$ does not impose any restriction besides boundedness of the last functions Φ_{d_ξ} (see Table 4.2).

From the remark that observer (4.11)

1. can be used for the system

$$\begin{aligned} \dot{\xi}_1 &= \xi_2 + \psi_1(t) \\ &\vdots \\ \dot{\xi}_{k-1} &= \xi_k + \psi_{k-1}(t) \\ \dot{\xi}_k &= \varphi_k(t) \end{aligned}$$

   provided the functions ψ_i are known and the function φ_k is unknown but bounded, with known bound.

2. gives estimates of the ξ_i's in finite time,

we see that it can be used as a preliminary step to deal with the system

$$\begin{aligned} \dot{\xi}_1 &= \xi_2 + \psi_1(t) \\ &\vdots \\ \dot{\xi}_{k-1} &= \xi_k + \psi_{k-1}(t) \\ \dot{\xi}_k &= \xi_{k+1} + \Phi_k(u, \xi_1, \ldots, \xi_k) \\ \dot{\xi}_{k+1} &= \varphi_{k+1}(u, \xi_1, \ldots, \xi_{k+1}) \end{aligned}$$

Indeed, thanks to the above observer we know in finite time the values of $\xi_1, \ldots, \xi_k$, so that the function $\Phi_k(u, \xi_1, \ldots, \xi_k)$ becomes a known signal $\psi_k(t)$.

From this, we can propose the following observer made of a cascade of homogeneous ob-

servers :

$$\left\{\begin{array}{l}
\dot{\hat{\xi}}_{11} \in \hat{w}_1 - L_1\, k_{11}\, \mathtt{S}(\hat{\xi}_{11} - y) \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\dot{\hat{\xi}}_{21} = \hat{\xi}_{22} + \Phi_1(u, \hat{\xi}_{11}) + \hat{w}_1 - L_2\, k_{21}\, \left\lfloor \hat{\xi}_{21} - y \right\rceil^{\frac{1}{2}} \\
\dot{\hat{\xi}}_{22} \in \hat{w}_2 - L_2^2\, k_{22}\, \mathtt{S}(\hat{\xi}_{21} - y) \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\quad \vdots \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\dot{\hat{\xi}}_{d_\xi 1} = \hat{\xi}_{d_\xi 2} + \Phi_1(u, \hat{\xi}_{11}) + \hat{w}_1 - L_{d_\xi}\, k_{d_\xi 1}\, \left\lfloor \hat{\xi}_{d_\xi 1} - y \right\rceil^{\frac{d_\xi - 1}{d_\xi}} \\
\quad \vdots \\
\dot{\hat{\xi}}_{d_\xi (d_\xi - 1)} = \hat{\xi}_{d_\xi d_\xi} + \Phi_{d_\xi - 1}(u, \hat{\xi}_{(d_\xi - 1)1}, \ldots, \hat{\xi}_{(d_\xi - 1)(d_\xi - 1)}) + \hat{w}_{d_\xi - 1} - L_{d_\xi}^{d_\xi - 1}\, k_{d_\xi (d_\xi - 1)}\, \left\lfloor \hat{\xi}_{d_\xi 1} - y \right\rceil^{\frac{1}{d_\xi}} \\
\dot{\hat{\xi}}_{d_\xi d_\xi} \in \hat{w}_{d_\xi} - L_{d_\xi}^{d_\xi}\, k_{d_\xi d_\xi}\, \mathtt{S}(\hat{\xi}_{d_\xi 1} - y)
\end{array}\right. \tag{4.25}$$

where the k_{ij} and L_i are positive real numbers to be tuned.

As a direct consequence of Theorem 4.2.1 and following the same steps as in the proof of Theorem 4.3.1, we have

Theorem 4.3.2.

Assume Φ is continuous. For any positive real numbers $\bar{\xi}$, $\bar{\mathtt{u}}$, $\bar{\mathtt{w}}$, we can find positive real numbers k_{ij} and L_i^*, two class $\mathcal{K}$ functions γ_1 and γ_2 and a class $\mathcal{KL}$ function β such that, for all $(L_1, \ldots, L_{d_\xi})$ verifying $L_i \geq L_i^*$, for all locally bounded time function $(u, v, w, \hat{w})$, and all $(\xi_0, \hat{\xi}_0)$ in $\mathbb{R}^{d_\xi} \times \mathbb{R}^{d_\xi}$, the observer (4.25) admits absolutely continuous solutions $\left(\hat{\boldsymbol{\Xi}}_1(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w}), ..., \hat{\boldsymbol{\Xi}}_{d_\xi}(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})\right)$ which are defined on $\mathbb{R}_+$ and for any such solution we have for all i in $\{1, ..., d_\xi\}$ and for all t such that $|\Xi(\xi_0, s; u, w)| \leq \bar{\xi}$, $|u(s)| \leq \bar{\mathtt{u}}$ and $|\Delta w(s)| \leq \bar{\mathtt{w}}$ for all $0 \leq s \leq t$:

$$|\hat{\boldsymbol{\Xi}}_i(t) - \boldsymbol{\Xi}_i(t)| \leq \max\left\{\beta(|\xi_0 - \hat{\xi}_0|, t), \sup_{\substack{1 \leq j \leq i-1 \\ s \in [t_0, t]}} \{\gamma_1(|v(s)|), \gamma_2(|\Delta w_j(s)|)\}\right\}.$$

where $\hat{\boldsymbol{\Xi}}_i$ is the state of the ith block (see Notation (4.2)) and we have used the abbreviation $\hat{\boldsymbol{\Xi}}_i(t) = \hat{\boldsymbol{\Xi}}_i(\hat{\xi}_0, \xi_0; t; u, v, w, \hat{w})$ and $\boldsymbol{\Xi}_i(t) = \boldsymbol{\Xi}_i(\xi_0; t; u, w)$.
Moreover, when $v(t) = \Delta w_j = 0$, there exists $\bar{t}$ such that $\hat{\boldsymbol{\Xi}}_i(t) = \boldsymbol{\Xi}_i(t)$ for all $t \geq \bar{t}$.

Proof : This proof is very similar to that of Theorem 4.3.1 (but also much simpler). We give it all the same for the sake of completeness. The error system attached to the homogeneous observer in block i has state $\mathbf{e}_i$ and input v and $\delta_{i\ell}$ defined as

$$\begin{array}{rcl}
\delta_{i\ell} & = & \left[\hat{\Phi}_\ell(u, \hat{\boldsymbol{\xi}}_{(i-1)}) - \Phi_\ell(u, \boldsymbol{\xi}_{(i-1)})\right] + [\hat{w}_\ell - w_\ell] \\
\delta_{ii} & = & -z_{i+1} - \Phi_i(u, \boldsymbol{\xi}_i) + \hat{w}_i - w_i
\end{array}$$

with $z_{m+1} = 0$. δ_{ii} is bounded, thus Theorem 4.2.1 gives the existence of $k_{i1}, \ldots, k_{ii}$, L_i^*, λ_i, β_i and γ_i such that we have, for all $L_i \geq L_i^*$, all $t \geq 0$, all j in $\{1, \ldots, i\}$ and with $e_{ij}(t)$ denoting the jth error in the ith block evaluated along the solution at time t,

$$|e_{ij}(t)| \leq \max\left\{\beta_i\left(|\mathbf{e}_i(0)|, t\right), \gamma_i \sup_{\substack{1 \leq \ell \leq j-1 \\ s \in [0,t]}} \left\{L^{j-1}|v(s)|^{\frac{r_j}{r_1}}, \frac{|\delta_{i\ell}(s)|^{\frac{r_j}{r_{\ell+1}}}}{L^{\mu_{j\ell}}}\right\}\right\}.$$

But the continuity of the Φ_ℓ implies the existence of a function ρ of class $\mathcal{K}$ such that, for all ℓ in $\{1, \ldots, i\}$ and for all $(\boldsymbol{\xi}_{(i-1)}, \hat{\boldsymbol{\xi}}_{(i-1)}, u)$ in $\mathbb{R}^{i-1} \times \mathbb{R}^{i-1} \times U$ satisfying $|\boldsymbol{\xi}_{(i-1)}| \leq \bar{\xi}$ and $|u| \leq \bar{\mathtt{u}}$,

$$|\Phi_\ell(u, \hat{\boldsymbol{\xi}}_{(i-1)}) - \Phi_\ell(u, \boldsymbol{\xi}_{(i-1)})| \leq \rho\left(|\mathbf{e}_{(i-1)}|\right).$$

This implies

$$|\delta_{i\ell}(s)| \leq \rho(|\mathbf{e}_{i-1}(s)|) + |\Delta w_\ell(s)| \ , \ \ell = 1, \ldots, j-1 \ .$$

Hence, we have the existence of two class $\mathcal{K}$ functions γ_{iv}, γ_{iw} such that

$$|\mathbf{e}_i(t)| \leq \max \left\{ \beta_i \left(|\mathbf{e}_i(0)|, t \right) , \sup_{s\in[0,t]} \gamma_{iv}(|v(s)|) , \sup_{s\in[0,t]} \frac{\rho(|\mathbf{e}_{i-1}(s)|)}{L_i^{2-i}} , \sup_{\substack{1\leq\ell\leq i-1 \\ s\in[0,t]}} \gamma_{iv}(|\Delta w_\ell(s)|) \right\} .$$

Hence, by recursion, for each i in $\{1, \ldots, d_\xi\}$, there exist a class $\mathcal{KL}$ function $\bar{\beta}_i$ and class $\mathcal{K}$ functions γ_{vi} and γ_{wi}, each depending on L_1 to L_i and such that we have, for all $t \geq 0$,

$$|\mathbf{e}_i(t)| \leq \max \left\{ \bar{\beta}_i \left(\max_{j\in\{1,\ldots,i\}} \{|\mathbf{e}_j(0)|\}, t \right) , \sup_{s\in[0,t]} \{\gamma_{vi}(|v(s)|), \gamma_{wi}(|\Delta w(s)|)\} \right\} .$$

■

Observer (4.25) is an extension of the cascaded high gain observer (4.24) presented in Section 4.3.1. The use of homogeneity enables here to obtain asymptotic convergence without demanding anything but the knowledge of a bound on the input and on the system solution. A drawback of a cascade of observers is that it gives an observer with dimension $\frac{d_\xi(d_\xi+1)}{2}$ in general. However, as seen in Section 4.3.1, it may be possible to reduce this dimension since, for each new block, one may increase the dimension by more than one, when the corresponding added functions Φ_i satisfy $\mathcal{H}(\alpha, \mathfrak{a})$ for some α verifying (4.13) with $d_0 = -1$ and for some $\mathfrak{a}$.

Finally, note that the result of Theorem 4.3.2 does not mean that the observer is ISS with respect to Δw. Indeed, Δw must be bounded to obtain this ISS-like inequality : the system is ISS with restrictions. Again, we believe that this problem could be solved employing homogeneous in the bi-limit observer as in [APA08].

4.4 Relaxing the assumptions marked with (☼)

First, if System (4.1) is not complete, every ISS inequalities still holds for any solution $\Xi(\xi_0; t; u, w)$, but only on $[0, \sigma^+(\xi_0, u)[$ where $\sigma^+(\xi_0, u)$ is its maximal time of existence in $\mathbb{R}^{d_\xi}$.

The global aspect of boundedness, Hölder, $\mathcal{H}(\alpha, \mathfrak{a})$, ..., can be relaxed as follows. Let U be bounded and let $\mathcal{M}$ be a given compact set. We define $\hat{\Phi}$ as[6]

$$\hat{\Phi}_i(u, \xi_1, ..., \xi_i) = \mathsf{sat}_{\overline{\Phi}_i}(\Phi_i(u, \xi_1, ..., \xi_i)) \tag{4.26}$$

where

$$\overline{\Phi}_i = \max_{u\in U, \xi\in\mathcal{M}} |\Phi_i(u, \xi_1, ..., \xi_i)| \ .$$

Now consider any compact set $\tilde{\mathcal{M}}$ strictly contained[7] in $\mathcal{M}$. We have $\hat{\Phi} = \Phi$ on $\tilde{\mathcal{M}}$, so that if the system trajectories remain in $\tilde{\mathcal{M}}$, the model (4.1) with $\hat{\Phi}$ replacing Φ is still valid. Besides, according to Lemma A.2.2 in Appendix A.2, there exists $\tilde{\mathfrak{a}}$ such that (4.3) holds for $\hat{\Phi}$ for all $(\boldsymbol{\xi}_a, \boldsymbol{\xi}_b)$ in $\mathbb{R}^{d_\xi} \times \tilde{\mathcal{M}}$. Then, by taking $\hat{\Phi}$ instead of Φ in the observers, we can modify the assumptions

- in Theorem 4.1.1, so that (4.5) holds only on the compact set $\mathcal{M}$;
- in Theorems 4.1.2 and 4.2.1, so that Φ verifies $\mathcal{H}(\alpha, \mathfrak{a})$ only on the compact set $\mathcal{M}$;
- Theorems 4.3.1 and 4.3.2 remain unchanged.

In this case, the results hold for the particular system solutions $\Xi(\xi_0; t; u, w)$ which are in the compact set $\tilde{\mathcal{M}}$ for t in $[0, \sigma^+_{\tilde{\mathcal{M}}}(\xi_0, u))$. Precisely, for these solutions, the bounds on $\hat{\Xi}_i(t) - \Xi_i(t)$ given in these theorems hold for all t in $[0, \sigma^+_{\tilde{\mathcal{M}}}(\xi_0, u))$.

[6] The saturation function is defined on $\mathbb{R}$ by $\mathsf{sat}_M(x) = \max\{\min\{x, M\}, -M\}$.

[7] By strictly contained, we mean that $\tilde{\mathcal{M}}$ is contained in the interior of $\mathcal{M}$.

Note also that if $\mathcal{H}(\alpha, \mathfrak{a})$ holds on a compact set, then for any $\tilde{\alpha}$ such that $\tilde{\alpha}_{ij} \leq \alpha_{ij}$ for all (i,j), there exists $\tilde{\mathfrak{a}}$ such that $\mathcal{H}(\tilde{\alpha}, \tilde{\mathfrak{a}})$ also holds on this compact set. It follows that the constraints given by (4.13) or Table 4.2 in Theorem 4.2.1 can be relaxed to $\alpha_{ij} \geq \frac{1-d_0(d_\xi-i-1)}{1-d_0(d_\xi-j)}$, and the less restrictive conditions one may ask for are obtained for $d_0 = -1$.

Finally, in Theorems 4.1.1, 4.1.2 and 4.2.1, it is possible to consider the case where Φ depends also on time as long as any assumption made on Φ is satisfied uniformly with respect to time.

4.5 Illustrative example

As an example, we consider the triangular normal form of dimension 4 defined by

$$\left\{\begin{array}{rcl} \dot{\xi}_1 &=& \xi_2 \\ \dot{\xi}_2 &=& \xi_3 \\ \dot{\xi}_3 &=& \xi_4 + \Phi_3(u,\xi_1,\xi_2,\xi_3) \\ \dot{\xi}_4 &=& \Phi_4(u,\xi) \end{array}\right. \quad , \quad y = \xi_1 \ , \tag{4.27}$$

where

$$\Phi_3(u,\xi_1,\xi_2,\xi_3) = 5u|\xi_3+\xi_1|^{\frac{4}{5}} \lfloor \xi_1 \rceil^{\frac{1}{5}} \quad , \quad \Phi_4(u,\xi) = \Psi(u,\psi(\xi))$$

with $\Psi : \mathbb{R}^{d_u} \times \mathbb{R}^3 \to \mathbb{R}$ and $\psi : \mathbb{R}^4 \to \mathbb{R}^3$ continuous function defined by :

$$\Psi(u,\psi) = \psi_1 - 2\psi_1\psi_3^5 + 20\psi_3^3\psi_1^3\psi_2^2 - 15\psi_3^4\psi_2^2\psi_1 + 5\psi_3^4\psi_1^3 - 5\psi_3^9\psi_1^3 + \psi_3^{10}\psi_1 + u(-20\psi_3^3\psi_1^2\psi_2 + 5\psi_3^4\psi_2)$$

$$\psi(\xi) = \left(\xi_1 \ , \ \xi_2 \ , \ \left(\frac{(\xi_3+\xi_1)\xi_1 + \left[(\xi_4+\xi_2) + 3|(\xi_3+\xi_1)\lfloor\xi_1\rceil^{\frac{3}{2}}|^{\frac{4}{5}}\xi_2 \right] \xi_2}{\xi_1^2+\xi_2^2} \right)^{\frac{1}{5}} \right) .$$

Those seemingly mysterious expressions do not make a lot of sense for now. We shall see how they appear in an example in Chapter 6. In fact, they are given here for the sake of completeness but only the expression of Φ_3 and the fact that Φ_4 is continuous matter here. We are interested in estimating trajectories remaining in a given compact set which will be defined in Chapter 6.

The function Φ_3 is not Lipschitz at the points on the hyperplanes $\xi_3 = -\xi_1$ and $\xi_1 = 0$. The function Φ_4 is continuous and therefore bounded on any compact set. Besides, for $\hat{\boldsymbol{\xi}}_3$ and $\boldsymbol{\xi}_3$ in a compact set, and u bounded there exist[8]

$$|\Phi_3(u,\hat{\xi}_1,\hat{\xi}_2,\hat{\xi}_3) - \Phi_3(u,\xi_1,\xi_2,\xi_3)| \leq c_1|\hat{\xi}_1-\xi_1|^{\frac{1}{5}} + c_3|\hat{\xi}_3-\xi_3|^{\frac{4}{5}} \ .$$

This implies that Φ_3 is Hölder with order $\frac{1}{5}$.

Hence the nonlinearities Φ_3 and Φ_4 verify the conditions of Table 4.1. It follows that for L sufficiently large, convergence with an arbitrary small error can be achieved with the high gain observer (4.4). However, Φ_3 does not verify the conditions of Table 4.2. Thus, there is no theoretical guarantee that the homogeneous observer (4.11) with $d_0 = -1$ will provide exact convergence.

4.5.1 An observer of dimension 4 ?

We consider a solution to System (4.27) which regularly crosses the Lipschitzness singularities $\xi_3 = -\xi_1$ or $\xi_1 = 0$, as illustrated in Figure 4.1. In the following, we use the same noisy

[8] Let $\Delta\Phi_3(\xi_1,\xi_3,e_1,e_3) = |\xi_3+e_3+\xi_1+e_1|^{\frac{4}{5}}\lfloor\xi_1+e_1\rceil^{\frac{1}{5}} - |\xi_3+\xi_1|^{\frac{4}{5}}\lfloor\xi_1\rceil^{\frac{1}{5}} = |\xi_3+\xi_1|^{\frac{4}{5}}\left(\lfloor\xi_1+e_1\rceil^{\frac{1}{5}} - \lfloor\xi_1\rceil^{\frac{1}{5}}\right) + \lfloor\xi_1+e_1\rceil^{\frac{1}{5}}\left(|\xi_3+\xi_1+e_3+e_1|^{\frac{4}{5}} - |\xi_3+\xi_1|^{\frac{4}{5}}\right)$. By Lemma A.1.5, we have $\left|\lfloor\xi_1+e_1\rceil^{\frac{1}{5}} - \lfloor\xi_1\rceil^{\frac{1}{5}}\right| \leq 2^{\frac{4}{5}}|e_1|^{\frac{1}{5}}$ and $\left||\xi_3+\xi_1+e_3+e_1|^{\frac{4}{5}} - |\xi_3+\xi_1|^{\frac{4}{5}}\right| \leq 2^{\frac{1}{5}}(|e_3|+|e_1|)^{\frac{4}{5}} \leq 2^{\frac{1}{5}}(|e_3|^{\frac{4}{5}}+|e_1|^{\frac{4}{5}})$. Besides, $|\xi_1+e_1|^{\frac{1}{5}} \leq |\xi_1|^{\frac{1}{5}} + |e_1|^{\frac{1}{5}}$, so that for ξ_1 and ξ_3 in compact sets, $|\Delta\Phi_3(\xi_1,\xi_3,e_1,e_3)| \leq c_1|e_1|^{\frac{1}{5}} + c_2|e_3|^{\frac{4}{5}} + c_3|e_1|^{\frac{4}{5}} + c_4|e_1|^{\frac{1}{5}}|e_3|^{\frac{4}{5}} + c_5|e_1|$. By Young's inequality, $|e_1|^{\frac{1}{5}}|e_3|^{\frac{4}{5}} \leq \frac{1}{5}|e_1| + \frac{4}{5}|e_3|$, and finally, for e_1 and e_3 in compact sets, $|\Delta\Phi_3(\xi_1,\xi_3,e_1,e_3)| \leq \tilde{c}_1|e_1|^{\frac{1}{5}} + \tilde{c}_3|e_3|^{\frac{4}{5}}$.

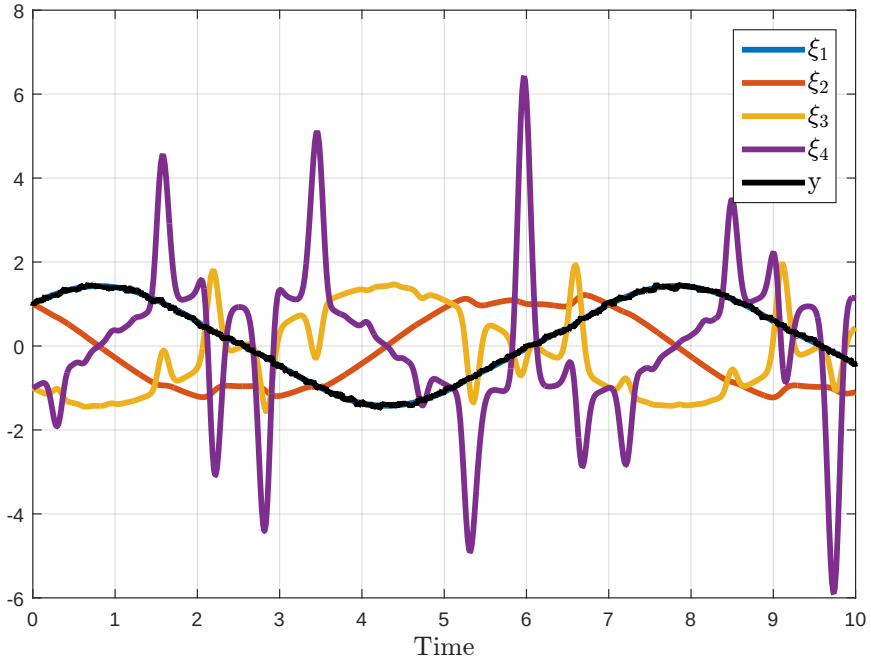


Figure 4.1: Trajectory of System (4.27), with initial condition $\xi = (1, 1, -1, -1)$, with input $u = 5\sin(10\,t)$ and with the noisy measurement y (filtered gaussian noise with standard deviation $\sigma = 0.03$ and 1st order filtering parameter $a = 50$), used to test the observers.

measurement y, shown on Figure 4.1, in every simulation with noise.

We first implement a high gain observer of dimension 4, in the absence of noise, initialized at $\hat{\xi} = (0.1, 0.1, -0.1, -0.1)$, and with the gains $k_1 = 14$, $k_2 = 99$, $k_3 = 408$, $k_4 = 833$. As an illustration of Theorem 4.1.2, the convergence with an arbitrary small error is achieved and is illustrated in Table 4.3. However, we observe that the decrease of the errors, especially for e_4, is very slow compared to the increase of the peaking and a very high gain is needed to obtain "acceptable" final errors. In presence of noise, the tradeoff between final error and noise amplification becomes impossible : with the noisy measurement of Figure 4.1, the smallest final error e_4 is 200, achieved for $L = 2$. Of course, there might exist a choice of the gains k_i giving better results. But overall a high gain observer may not be a systematic solution in practice for non-Lipschitz triangular systems, especially when the solution regularly crosses the Lipschitz-singularities.

L	e_1	e_2	e_3	e_4	$\max \lvert e \rvert$
2	0.15	4	60	200	300
5	$6\,.\,10^{-4}$	0.04	1.5	30	4000
8	$5\,.\,10^{-5}$	$4\,.\,10^{-3}$	0.25	7	$1.5\,.\,10^{4}$
10	$8\,.\,10^{-6}$	$1\,.\,10^{-3}$	0.1	4	$3.5\,.\,10^{4}$
15	$1.5\,.\,10^{-6}$	$3\,.\,10^{-4}$	0.03	2	$1.2\,.\,10^{5}$

Table 4.3: Decrease of the final error ($e_i = \hat{\xi}_i - \xi_i$) with the gain L, with a high gain observer and in the absence of noise. The last columns shows however that the peaking increases, i-e the errors reach higher and higher values during the transient before converging.

Let us now implement an homogeneous observer of dimension 4 with an explicit Euler method with fixed measurement and integration steps equaling 10^{-5}, and with the Matlab sign function. The degree is $d_0 = -1$, and the gains are chosen according to [Lev05], i-e $k_1 = 5$, $k_2 = 8.77$,

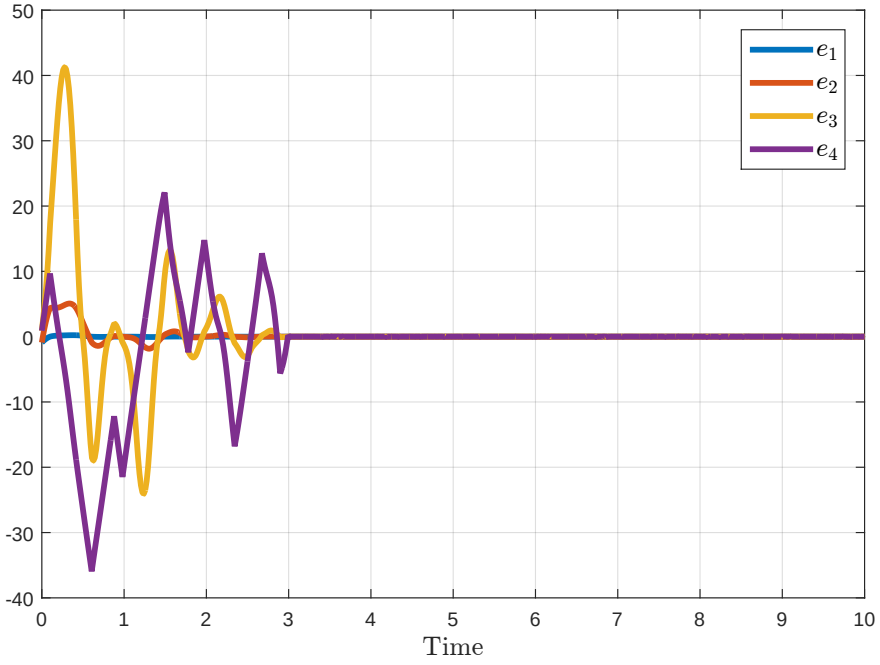


Figure 4.2: Convergence of an homogeneous observer with degree -1 in the absence of noise ($e_i = \hat{\xi}_i - \xi_i$)

$k_3 = 4.44$, $k_4 = 1.1$. For a gain $L = 3$, the convergence is achieved with a final error of $|e_4| = 8\,.\,10^{-4}$, even though the Hölder restriction of Theorem 4.2.1 is a priori not satisfied around $\xi_1 = 0$. The results are given in Figure 4.2. Unfortunately, the final errors are heavily impacted in presence of noise, as illustrated in Table 4.4. This may also come from a lack of ISS property. Notice that the amplification of the noise by the gain L is not as rapid as expected from the bound in Theorem 4.2.1. The final errors remain nonetheless too large, although, once again, we did not optimize our choice of gains k_i.

L	e_1	e_2	e_3	e_4
2.5	0.15	3.5	30	18
3	0.15	3	35	25
4	0.1	2	25	50
5	0.1	2	30	80
6	0.1	2	35	120

Table 4.4: Final errors given by a homogeneous observer of degree -1 in presence of noise.

4.5.2 Cascaded observers

In the absence of noise, the cascaded observers presented in Sections 4.3.1 and 4.3.2 give similar results to the corresponding observers in dimension 4, i-e arbitrary small asymptotic error and finite time convergence respectively. However, they seem to provide better accuracies in presence of noise.

In the case of a high gain cascade observer, the errors, although smaller than in the high gain observer of dimension 4, remain too large to consider it a viable solution. On the other

hand, the homogeneous cascade observer :

$$\begin{cases} \dot{\hat{\xi}}_{11} = \hat{\xi}_{12} - L_1\, k_{11}\, \left\lfloor \hat{\xi}_{11} - y \right\rceil^{\frac{2}{3}} \\ \dot{\hat{\xi}}_{12} = \hat{\xi}_{13} - L_1^2\, k_{12}\, \left\lfloor \hat{\xi}_{11} - y \right\rceil^{\frac{1}{3}} \\ \dot{\hat{\xi}}_{13} \in -L_1^3\, k_{13}\, \mathtt{S}(\hat{\xi}_{11} - y) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \dot{\hat{\xi}}_{21} = \hat{\xi}_{22} - L_2\, k_{21}\, \left\lfloor \hat{\xi}_{21} - y \right\rceil^{\frac{3}{4}} \\ \dot{\hat{\xi}}_{22} = \hat{\xi}_{23} - L_2^2\, k_{22}\, \left\lfloor \hat{\xi}_{21} - y \right\rceil^{\frac{1}{2}} \\ \dot{\hat{\xi}}_{23} = \hat{\xi}_{24} + \mathsf{sat}(\mathfrak{g}_3(\hat{\xi}_{11}, \hat{\xi}_{12}, \hat{\xi}_{13}))u - L_2^3\, k_{23}\, \left\lfloor \hat{\xi}_{21} - y \right\rceil^{\frac{1}{4}} \\ \dot{\hat{\xi}}_{24} \in -L_2^4\, k_{24}\, \mathtt{S}(\hat{\xi}_{21} - y) \end{cases}$$

with the coefficients k_{1j} chosen, according to [Lev05], as $k_{11} = 3$, $k_{12} = 2.6$, $k_{13} = 1.1$, and k_{2j} as above, and with the gains $L_1 = 2.5$ and $L_2 = 3$, gives the following final errors :

$$e_{11} = 0.05, \quad e_{12} = 0.4, \quad e_{13} = 2.5, \quad e_{24} = 12$$

Comparing to Table 4.4, we see that implementing an intermediate homogeneous observer of dimension 3 enables to obtain much better estimates of the first three states ξ_i, which are then used in the nonlinearity of the second block, thus giving a better estimate of ξ_4.

4.6 Conclusion

To summarize the most important ideas, we provide in Table 4.6 a synthetic comparison of the four observers proposed in this chapter, in the case where the system trajectories and the input are bounded.

We have shown the convergence with an arbitrary small error of the classical high gain observer (4.4) in presence of nonlinearities verifying some Hölder-like condition. Also, for the case when this Hölder condition is not verified, we have proposed a novel cascaded high gain observer (4.24). On the other hand, under slightly more restrictive Hölder-like conditions, we have made a Lyapunov design of the homogeneous observer (4.11) and proved its asymptotic convergence with the help of an explicit Lyapunov function. As for its cascaded version (4.25), asymptotic convergence has been established under the only condition of continuity of the nonlinearities and the fact that the trajectories (and input) are bounded. We conclude that a global observer exists for the continuous triangular form (3.12).

Although it is an extremely important aspect, we have had no time to devote much attention to the impact of disturbances on the behavior of the observers. Nevertheless, we have established an ISS property with respect to dynamics and measurement noises. Our numerical experience seems to indicate that it is very difficult to tune the gains of both high gain and homogeneous observers in presence of measurement noise, although it is slightly simpler for the latter since smaller gains are sufficient to ensure convergence. Simulations on our example suggest that the situation may be more favorable with the cascaded homogeneous observer. Anyway, the presented results are still unsatisfactory in presence of noise, and the question of the construction of robust observers for non-Lipschitz triangular forms remains unanswered. Our theoretical ISS bounds being far too conservative, it would be necessary to carry out a finer study if we wanted to optimally tune the gains of the observers. It may also be appropriate to use on-line gain adaptation techniques since large gains should be necessary only around the points where the nonlinearities are not Lipschitz. About these two aspects, we refer the reader to the survey in [KP13, Sections 3.2.2 and 3.2.3] and the references therein.

	High gain (4.4)	High gain cascade (4.24)	Homogeneous (4.11)	Homogeneous cascade (4.25)
Assumption on $\mathfrak{g}_i$	Hölder with order greater than in Table 4.1	Continuous	Hölder with order greater than in (4.13) or Table 4.2 for $d_0 = -1$	Continuous
Convergence	Arbitrary small error	Arbitrary small error	Asymptotic convergence	Asymptotic convergence
Advantages	Easy choice of gains	No constraint on Φ_i	Not necessarily large gains because convergence	No constraint on Φ_i, convergence, apparently better in terms of noise
Drawbacks	Large gains necessary to obtain small error $\Rightarrow$ numerical problems (peaking) and sensitivity to noise	Same as for high gain, but also gains difficult to choose and large dimension	Implementation of the sign function if $d_0 = -1$ (chatter etc)	Large dimension and a lot of gains to choose

Table 4.5: Comparison between observers of Chapter 4 for a continuous triangular forms when the system state and the input are bounded.

Part II

Transformation into a normal form

Chapter 5

Review of existing transformations

Chapitre 5 – Bilan des transformations existantes. *Tout au long de Partie I, nous avons listé un certain nombre de formes normales pour lesquelles un observateur est connu. Afin d'appliquer Théorème 2.2.1, il nous faut maintenant étudier comment transformer un système non linéaire quelconque en l'une de ces formes. C'est l'objet de Partie II. Comme dans la précédente, nous commencons par faire un rapide bilan des résultats existant dans la littérature concernant ce problème en soulignant les points qui n'ont pas encore été étudiés. Ceci nous permet de situer nos contributions qui seront ensuire détaillées dans les chapitres suivants.*

Contents

Throughout Part I, we have given a list of normal forms and their associated observers. We now have to study how a nonlinear system can be transformed into one of those forms to apply Theorem 2.2.1. This is the object of Part II. Using the same methodology as in the previous part, we start by reviewing the literature concerning this problem in order to highlight along the way the points which have not been addressed yet, and we situate our contributions which will be detailed in the next chapters.

More precisely, we consider a general nonlinear system of the form

$$\dot{x} = f(x,u) \quad , \quad y = h(x,u) \tag{5.1}$$

with x the state in $\mathbb{R}^{d_x}$, u an input function in $\mathcal{U}$ with values in $U \subset \mathbb{R}^{d_u}$, y the output in $\mathbb{R}^{d_y}$. For each normal form presented in Part I of the form

$$\dot{\xi} = F(\xi, u, H(\xi,u)) \quad , \quad y = H(\xi,u)\ , \tag{5.2}$$

we look for sufficient conditions on System (5.1) for the existence of a subset $\mathcal{X}$ and functions $T_u : \mathcal{X} \times [0,+\infty[\to \mathbb{R}^{d_\xi}$ for each u in $\mathcal{U}$ which transforms System (5.1) into the normal form (5.2) in the sense of Theorem 2.2.1, i-e for all x in $\mathcal{X}$ and all t in $[0,+\infty)$

$$L_{(f,1)}T_u(x,t) = F(T_u(x,t),u(t),h(x,u(t))) \quad , \quad h(x,u(t)) = H(T_u(x,t),u(t))\ .$$

Indeed, according to Theorem 2.2.1 and Corollary 2.2.1, the observer design problem is then solved for System (5.1) if the solutions of System (5.1) which are of interest remain in $\mathcal{X}$ and

- *either* for any u in $\mathcal{U}$, $x \mapsto T_u(x,t)$ becomes injective on $\mathcal{X}$ uniformly in space and in time after a certain time ;
- *or* $\mathcal{C} = \mathcal{X}$ is a compact set, and for any u in $\mathcal{U}$, T_u is a same stationary transformation T injective on $\mathcal{C}$.

5.1 Transformations into a state-affine normal forms

5.1.1 Linearization by output injection

Constant linear part

The problem of transforming a nonlinear system into a linear one of the form (3.5) i-e

$$\dot{\xi} = A\,\xi + B(u,\tilde{y}) \quad , \quad \tilde{y} = \psi(y) = C\,\xi \tag{5.3}$$

with the pair (A,C) observable and $\psi : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$ a possible change of output, has a very long history. The first results appeared in [KI83, BZ83] for autonomous systems and were then extended by [KR85] to multi-input multi-output systems. In those papers, the authors looked for necessary and sufficient conditions on the functions f and h for the existence of a local change of coordinates (and possibly change of output) which brings the system into the form (5.3), which they called "observer form". [BRG89] then gave conditions for the existence of a local (and global) immersion[1] (instead of diffeomorphism) in the particular case of control affine systems. A vast literature followed on the subject, either developing algebraic algorithms to check the existence of a transformation or tools to explicitly find the transformation.

In [Jou03], the general problem of finding an immersion (rather than a diffeomorphism) which transforms a nonlinear system of the form (5.3) is addressed. If such a transformation exists, the system is said linearizable by output injection. The following result is proved :

Theorem 5.1.1. [Jou03, Theorem 2.3]

A system of the form

$$\dot{x} = f(x,u) \quad , \quad y = h(x)$$

is linearizable by output injection if and only if there exist a $C^+\infty$ function T and a diffeomorphism $\psi : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$ transforming the system into the particular triangular form

$$\left\{ \begin{array}{rcl} \dot{\xi}_1 &=& \xi_2 + \Phi_1(u,\xi_1) \\ &\vdots& \\ \dot{\xi}_i &=& \xi_{i+1} + \Phi_i(u,\xi_1) \\ &\vdots& \\ \dot{\xi}_{d_\xi} &=& \Phi_{d_\xi}(u,\xi_1) \end{array} \right. \quad , \quad \tilde{y} = \psi(y) = \xi_1 \ .$$

Thus, the linearization problem reduces to the existence of a transformation into this latter observable form. Note that if besides this transformation is required to be injective (like in our context of observer design), then the system is necessarily uniformly observable[2]. Actually, the class of systems considered here is even strictly smaller because for a uniformly observable system, the functions Φ_i would be allowed to depend on $\xi_1, \ldots, \xi_i$, and not only on ξ_1.

From this, it is possible to deduce :

[1] $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_\xi}$ is an immersion if the rank of $\frac{\partial T}{\partial x}$ is d_x. Contrary to a diffeomorphism, this allows to take $d_\xi \geq d_x$.

[2] See Definition 2.2.1.

Theorem 5.1.2. [Jou03, Theorem 2.6]

An autonomous system

$$\dot{x} = f(x) \quad , \quad y = h(x)$$

is linearizable by output injection if and only if there exists a C^∞ function ψ, an integer d_ξ and d_ξ C^∞ functions Φ_1, ..., Φ_{d_ξ} such that

$$L_f^{d_\xi}\tilde{h}(x) = L_f^{d_\xi-1}\Phi_1 \circ \tilde{h} + L_f^{d_\xi-2}\Phi_2 \circ \tilde{h} + L_f\Phi_{d_\xi-1} \circ \tilde{h} + \Phi_{d_\xi} \circ \tilde{h}$$

with $\tilde{h} = \psi \circ h$.

This is the so-called characteristic equation which extends the same notion for linear systems and was introduced in [Kel87] originally with $d_\xi = d_x$. This partial differential equation (PDE) is important in practice because several results show that the linearization of a controlled system first necessitates the linearization of its uncontrolled parts or drift dynamics[3] ([KR85, BRG89, Jou03] among others). A first difficulty thus lies in solving this PDE, which does not always admit solutions ([Jou03, BS04]).

Along the history of linearization, we must also mention some generalizations such as [Kel87], where the function B is allowed to depend on the derivatives of the input and later on the derivatives of the output in [GMP96, PG97], or [Gua02, RPN04] where it is proposed to use an output-depending time-scale transformation.

We conclude that linearizing both the dynamics and the output function is very demanding and requires some very restrictive conditions on the system. The existence of the transformation is difficult to check and involves quite tedious symbolic calculations which do not always provide the transformation itself, and even when they do, its validity is often only local.

Time-varying linear part

In parallel, others allowed the linear part A to depend on the input/output, i-e looked for conditions to transform the system in the state-affine form (3.7)

$$\dot{\xi} = A(u,y)\,\xi + B(u,y) \quad , \quad y = C(u)\xi \ .$$

The first to address this problem were [Fli82, FK83] but without allowing output injection in the dynamics, namely requiring $A(u)$ and $B(u)$. This led to the very restrictive finiteness criterion of the observation space, which roughly says that the linear space containing the successive derivatives of the output along any vector field of the type $f(\cdot, u)$ is finite. Later, [HK96, HC91, BB97] allowed A and B to depend on the output to broaden the class of concerned systems. But it remains difficult to characterize those systems because there are often many possible ways to parametrize the system via the output. Besides, even when the transformation exists and is known, the input must satisfy an extra excitation condition to allow the design of a Kalman observer (see Chapter 3).

5.1.2 Transformation into Hurwitz form

In a completely independent line of research, some researchers have tried to reproduce Luenberger's original methodology presented in [Lue64] for linear systems on nonlinear systems. It consists in finding a transformation into a Hurwitz form (3.3)

$$\dot{\xi} = A\,\xi + B(u,y) \quad , \quad y = H(\xi,u)$$

[3]Dynamics with u equal to a constant

with A Hurwitz, for which a trivial observer (3.4) (made of a copy of the dynamics) exists. Note that unlike in the previous section, this procedure is not a linearization of the system, since the output function H can be any nonlinear function (see [KK98, Remark 4]). It is not even necessary to know its expression since it is not needed in the observer. This crucial difference leads to far less restrictive conditions on the system.

The extension to autonomous nonlinear systems of Luenberger's original methodology ([Lue64]) was proposed and analyzed in a general context by [Sho92]. It was rediscovered later by [KK98] who gave a local analysis close to an equilibrium point under conditions relaxed later on in [KX03]. The localness as well as most of the restrictive assumptions were then by-passed in [AP06]. As noticed in [KX06] and [AP06], this nonlinear Luenberger observer is also strongly related to the observer proposed in [KE03].

In [AP06], the authors investigate the possibility of transforming an autonomous system

$$\dot{x} = f(x) \quad , \quad y = h(x)$$

into a Hurwitz autonomous form

$$\dot{\xi} = A\,\xi + B(y) \ .$$

This raises the question of finding, for some integer d_ξ, a continuous function $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_\xi}$ verifying

$$L_f T(x) = A\,T(x) + B(h(x)) \quad , \quad \forall x \in \mathcal{X} \tag{5.4}$$

with A some Hurwitz matrix of dimension d_ξ and $B : \mathbb{R}^{d_y} \to \mathbb{R}^{d_\xi}$ some continuous function. The existence of such a transformation is shown for any Hurwitz matrix A and for some well-chosen functions B under the only assumption that the system is backward-complete[4] in $\mathcal{X}$ ([AP06, Theorem 2]). Of course, this is not enough since, as we saw in the introduction, it is required that T be uniformly injective on $\mathcal{X}$ to deduce from the estimate of $T(x)$ an estimate of x. The authors show in [AP06, Theorem 3] that injectivity of T is achieved for almost any diagonal complex Hurwitz matrix A of dimension[5] $(d_x+1)d_y$ on $\mathbb{C}$ and for any B verifying some growth condition under the assumption that the system is backward $\mathcal{S}$-distinguishable[6] on $\mathcal{X}$ for some open set $\mathcal{S}$ containing $\mathcal{X}$, i-e for any (x_a, x_b) in $\mathcal{X}^2$ such that $x_a \neq x_b$, there exists t in $\left(\max\left\{\sigma^-_{\mathcal{S}}(x_a), \sigma^-_{\mathcal{S}}(x_b)\right\}, 0\right]$ such that $y_{x_a}(t) \neq y_{x_b}(t)$.

In the case where $\mathcal{X}$ is bounded, the result simplifies into :

Theorem 5.1.3. [AP06]

Assume that $\mathcal{X}$ and $\mathcal{S}$ are open bounded subset of $\mathbb{R}^{d_x}$, such that $\mathtt{cl}(\mathcal{X})$ is contained in $\mathcal{S}$ and System (5.1) is backward $\mathcal{S}$-distinguishable on $\mathcal{X}$. There exists a strictly positive number ℓ and a set $\mathcal{R}$ of zero Lebesgue measure in $\mathbb{C}^{d_x+1}$ such that denoting $\Omega = \{\lambda \in \mathbb{C} \ : \ \Re(\lambda) < -\ell\}$, for any $(\lambda_1, \ldots, \lambda_{d_x+1})$ in $\Omega^{d_x+1} \setminus \mathcal{R}$, there exists a function $T : \mathbb{R}^{d_x} \to \mathbb{R}^{(d_x+1)\times d_y}$ uniformly injective on $\mathcal{X}$ and verifying (5.4) with

$$A = \begin{pmatrix} \tilde{A} & & & \\ & \ddots & & \\ & & \tilde{A} & \\ & & & \ddots \\ & & & & \tilde{A} \end{pmatrix} \quad , \quad B(y) = \begin{pmatrix} \tilde{B} & & & \\ & \ddots & & \\ & & \tilde{B} & \\ & & & \ddots \\ & & & & \tilde{B} \end{pmatrix} y$$

[4] Any solution exiting $\mathcal{X}$ in finite time must cross the boundary of $\mathcal{X}$. See [AP06, Definition 1].

[5] Separating the real/imaginary parts, the observer is thus of dimension $2(d_x+1)d_y$ on $\mathbb{R}$.

[6] This notion is similar to the distinguishability defined in Definition 2.2.1 but in negative time and with the constraint that $\bar{t}$ occurs when both solutions are still in $\mathcal{S}$.

and

$$\tilde{A} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{d_x+1} \end{pmatrix} \quad , \quad \tilde{B} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} .$$

Besides, if $\mathcal{X}$ is backward-invariant, the function T is unique on $\mathcal{X}$ and defined by:

$$T(x) = \int_{-\infty}^{0} e^{-A\tau} B(h(X(x,\tau)))\, d\tau \ . \tag{5.5}$$

We conclude from this result that it is possible to design an observer for an autonomous nonlinear system under the weak assumption of backward-distinguishability. Note that with a stronger assumption of strong differential observability[7] of order m, and still in a bounded set, it is also proved in [AP06, Theorem 4] that injectivity of (5.5) is ensured for any choice of m real strictly negative λ_i smaller that $-\ell$ with ℓ sufficiently large.

The difficulty lies in the computation of the function T, let alone its inverse. Even when $\mathcal{X}$ is bounded and backward-invariant, the use of its explicit expression (5.5) is not easy since it necessitates to integrate backwards the differential equation at each time step. Several examples will be given in Chapter 7 or Chapter 11 to show how the function T can be computed without relying on this formula. In particular, we will see in Chapter 7 how this task can sometimes be made easier by allowing T to be time-varying.

The extension of this Luenberger methodology to controlled systems is not straight-forward. First steps in this direction were made in [RZ13, Tru07] for linear time-varying systems, in [Ham08] for nonlinear time-varying systems, and in [Eng07] for nonlinear controlled systems. This is the object of Chapter 7.

5.2 Transformations into triangular normal forms

5.2.1 Lipschitz triangular form

The Lipschitz triangular form[8] (3.12)

$$\left\{ \begin{array}{rcl} \dot{\xi}_1 &=& \xi_2 + \Phi_1(\tilde{u}, \xi_1) \\ & \vdots & \\ \dot{\xi}_i &=& \xi_{i+1} + \Phi_i(\tilde{u}, \xi_1, \ldots, \xi_i) \\ & \vdots & \\ \dot{\xi}_m &=& \Phi_m(\tilde{u}, \xi) \end{array} \right. \quad , \quad y = \xi_1$$

is well-known because it is associated to the classical high gain observer (3.13). The idea of transforming a nonlinear system into a phase-variable form[9] (i-e with $\Phi_i = 0$ except Φ_m) appeared in [Zei84]. For an autonomous system,

$$\dot{x} = f(x) \quad , \quad y = h(x)$$

the function $\mathbf{H}_m$ defined by the output and its $m-1$ first derivatives, namely

$$\mathbf{H}_{d_\xi}(x) = (h(x), L_f h(x), \ldots, L_f^{d_\xi - 1} h(x)) \ ,$$

transforms the system into

$$\dot{\xi}_1 = \xi_2 \quad , \quad \ldots \quad , \quad \dot{\xi}_i = \xi_{i+1} \quad , \quad \ldots \quad , \quad \dot{\xi}_m = L_f^m h(x) \quad , \quad y = \xi_1 \ .$$

[7] See Definition 5.2.2 in the autonomous case.

[8] It is useful to denote here the input $\tilde{u}$ because we will see that it can be for instance $\tilde{u} = (u, \dot{u}, \ddot{u}, \ldots)$.

[9] See Definition 3.2.1.

This is a Lipschitz phase-variable form if and only if there exists a function Φ_m Lipschitz on $\mathbb{R}^{d_\xi}$ such that

$$\forall x \in \mathcal{X} \quad , \quad L_f^m h(x) = \Phi_m(\mathbf{H}_m(x)) \ ,$$

i-e the mth-derivative of the output can be expressed "in a Lipschitz way" in terms of its $m-1$ first derivatives. This is possible for example if $\mathcal{X}$ is bounded and $\mathbf{H}_m$ is an injective immersion[10] on some open set $\mathcal{S}$ containing $\mathtt{cl}(\mathcal{X})$ (see Theorem 5.2.1 below for this result in the general controlled case).

In the remaining of this section, we review the existing results in terms of transformation of general controlled systems into a Lipschitz triangular form.

Time varying transformation

The first natural idea introduced in [Zei84] is to keep considering the transformation made of the output and its $m-1$ first derivatives, despite the presence of the input, and transform the system into a phase-variable form in the same way as for autonomous systems. In order to properly define this transformation, we need the following definition.

Definition 5.2.1.

Given an integer m, and using the notation

$$\overline{\nu}_m = (\nu_0, \dots, \nu_m) \ ,$$

we call *dynamic extension of order* m of System (5.1) the extended dynamical system

$$\dot{\overline{x}} = \overline{f}(\overline{x}, u^{(m+1)}) \qquad , \qquad y = \overline{h}(\overline{x}) \tag{5.6}$$

with input $u^{(m+1)}$ in $\mathbb{R}^{d_u}$, extended state $\overline{x} = (x, \overline{\nu}_m)$ in $\mathbb{R}^{d_x} \times \mathbb{R}^{d_u(m+1)}$, extended vector field $\overline{f}$ defined by

$$\overline{f}(\overline{x}, u^{(m+1)}) = \left(f(x, \nu_0) \, , \, \nu_1 \, , \, \dots \, , \, \nu_m \, , \, u^{(m+1)} \right)$$

and extended measurement function $\overline{h}$ defined by

$$\overline{h}(\overline{x}) = h(x, \nu_0) \ .$$

Note that for any solution x to System (5.1) with some input u, $(x, \overline{u}_m)$ is solution to the dynamic extension (5.6), with the notation $\overline{u}_m = (u, \dot{u}, \dots, u^{(m)})$. While $\overline{\nu}_m$ is an element of $\mathbb{R}^{d_u(m+1)}$, $\overline{u}_m$ is a function defined on $[0, +\infty)$ such that $\overline{u}_m(s) = (u(t), \dot{u}(t), ..., u^{(m)}(t))$ is in $\overline{U}_m \subset \mathbb{R}^{d_u(m+1)}$. The successive time derivatives of the output y are related to the Lie derivatives of $\overline{h}$ along the vector fields $\overline{f}$, namely for any $j \leq m$ and any (x_0, t_0) in $\mathcal{X} \times [0, +\infty)$

$$\frac{\partial^j Y}{\partial t^j}(x_0, t_0; t; u) = L_{\overline{f}}^j \overline{h}(X(x_0, t_0; t; u), \overline{u}_m(t)) \ .$$

We are now ready to define the notion of differential observability.

Definition 5.2.2.

Consider the function $\overline{\mathbf{H}}_m$ on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_u(m+1)}$ defined by

$$\overline{\mathbf{H}}_m(x, \overline{\nu}_m) = \left(\overline{h}(x, \overline{\nu}_m) \, , \, L_{\overline{f}} \overline{h}(x, \overline{\nu}_m) \, , \, \dots \, , \, L_{\overline{f}}^{m-1} \overline{h}(x, \overline{\nu}_m) \right) . \tag{5.7}$$

[10] $\mathbf{H}_m$ is injective and $\frac{\partial \mathbf{H}_m}{\partial x}(x)$ has full-rank on $\mathcal{X}$

- *weakly differentially observable of order m on $\mathcal{S}$* if for any $\overline{\nu}_m$ in $\overline{U}_m$, the function $x \mapsto \overline{\mathbf{H}}_m(x, \overline{\nu}_m)$ is injective on $\mathcal{S}$.
- *strongly differentially observable of order m on $\mathcal{S}$* if for any $\overline{\nu}_m$ in $\overline{U}_m$, $x \mapsto \overline{\mathbf{H}}_m(x, \overline{\nu}_m)$ is an injective immersion on $\mathcal{S}$.

The function $\overline{\mathbf{H}}_m(\cdot, \overline{\nu}_m)$ is equivalent to $\mathbf{H}_m$ for autonomous systems since it is made of the successive derivatives of the output, but it now depends on the input and its derivatives. The notion of differential observability of order m thus means that when knowing the current input and its derivatives, the current state is uniquely determined by the current output and its first $m-1$ derivatives. With this in hand, a straightforward extension of the stationary case along the idea presented in [Zei84] is :

Theorem 5.2.1.

If $\overline{U}_m$ is a compact subset of $\mathbb{R}^{d_u(m+1)}$ and there exists an integer m and a subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$ such that System (5.1) is weakly (resp strongly) differentially observable of order m on $\mathcal{S}$, then, for any compact subset $\mathcal{C}$ of $\mathcal{S}$ and any u in $\mathcal{U}$, the function defined by

$$T(x,t) = \overline{\mathbf{H}}_m(x, \overline{u}_m(t))$$

transforms System (5.1) into a continuous (resp Lipschitz) phase-variable form of dimension $d_\xi = m d_y$ on $\mathcal{C}$ and with input $\tilde{u} = \overline{u}_m$. Besides, $x \mapsto T(x,t)$ is uniformly injective in space and in time on $\mathcal{C}$.

Proof : Assume first that the system is weakly differentially observable of order m, i-e for all $\overline{\nu}_m$ in $\overline{U}_m$, $x \mapsto \overline{\mathbf{H}}_m(x, \overline{\nu}_m)$ is injective on $\mathcal{C}$. According to Lemma A.3.5, it is uniformly injective in space and in time on $\mathcal{C}$ and for any $\overline{\nu}_m$, it admits a uniformly continuous left inverse i-e there exists a function $\overline{\mathbf{H}}_m^{-1} : \mathbb{R}^{d_\xi} \times \mathbb{R}^{d_u(m+1)} \to \mathbb{R}^{d_x}$ such that for all $\overline{\nu}_m$ in $\overline{U}_m$ and all x in $\mathcal{C}$

$$x = \overline{\mathbf{H}}_m^{-1}(\overline{\mathbf{H}}_m(x, \overline{\nu}_m), \overline{\nu}_m) \ .$$

Now define

$$\Phi_m(\xi, \overline{\nu}_m) = L_f^m \overline{h}(\overline{\mathbf{H}}_m^{-1}(\xi, \overline{\nu}_m), \overline{\nu}_m) \ .$$

T transforms System (5.1) into the continuous phase-variable form

$$\left\{ \begin{array}{rcl} \dot{\xi}_1 & = & \xi_2 \\ & \vdots & \\ \dot{\xi}_{m-1} & = & \xi_m \\ \dot{\xi}_m & = & \Phi_m(\xi, \overline{u}_m(t)) \end{array} \right.$$

Assume now the system is strongly observable. Still with Lemma A.3.5, $\xi \mapsto \overline{\mathbf{H}}_m^{-1}(\xi, \overline{\nu}_m)$ can now be taken Lipschitz on $\mathbb{R}^{d_\xi}$, with the same Lipschitz constant for all $\overline{\nu}_m$ in $\overline{U}_m$. It follows that $\xi \mapsto \Phi_m(\xi, \overline{\nu}_m)$ is Lipschitz on any compact set of $\mathbb{R}^{d_\xi}$ containing the compact set of interest $\overline{\mathbf{H}}_m(\mathcal{C} \times \overline{U}_m)$, with the same Lipschitz constant for all $\overline{\nu}_m$ in $\overline{U}_m$. According to Kirszbraun-Valentine theorem [Kir34, Val45], it can be extended to a Lipschitz function on $\mathbb{R}^{d_\xi}$ with still the same Lipschitz constant. This new extended function $\xi \mapsto \Phi_m(\xi, \overline{\nu}_m)$ is globally Lipschitz uniformly in $\overline{\nu}_m$ and has not changed on $\overline{\mathbf{H}}_m(\mathcal{C} \times \overline{U}_m)$ where the system solutions evolve, thus we have a Lipschitz phase-variable form. ■

The assumptions given in Theorem 5.2.1 are sufficient to ensure the existence of the function Φ_m in the phase-variable variable form. But they are not necessary. The possibility of finding such a function, namely to express $L_f^m \overline{h}$ (the mth derivative of the output) in terms of $h, L_f \overline{h}, \ldots, L_f^{m-1} \overline{h}$ (the output and its $m-1$ first derivatives) and $\overline{u}_m$ (the input and its m first derivatives) is thoroughly studied in [JG96] through the so-called "ACP(m) condition". We refer the reader to [JG96] (or [GK01]) for a more complete analysis of those matters.

Remark 4 Note that as we saw in Chapter 4, for a high gain design, it is not necessary to have global Lipschitzness of the function Φ_m with respect to ξ. It is sufficient to have

$$|\Phi_m(\xi, \overline{\nu}_m) - \Phi_m(\hat{\xi}, \overline{\nu}_m)| \leq \mathfrak{a}|\xi - \hat{\xi}|$$

for all $\hat{\xi}$ in $\mathbb{R}^{d_\xi}$, all $\overline{\nu}_m$ in $\overline{U}_m$ and ξ in a compact set containing $\overline{\mathbf{H}}_m(\mathcal{C} \times \overline{U}_m)$ where the system solutions evolve. Thus, the Lipschitz extensions made in the proof of Theorem 5.2.1 are not necessary in practice : as suggested in Chapter 4, it is sufficient to take[11]

$$\Phi_m(\xi, \overline{\nu}_m) = \mathsf{sat}_M(L_f^m \overline{h}(\overline{\mathbf{H}}_m^{-1}(\xi, \overline{\nu}_m), \overline{\nu}_m)) \tag{5.8}$$

where M is a bound for $|L_f^m \overline{h}|$ on $\mathcal{C} \times \overline{U}_m$ and $\overline{\mathbf{H}}_m^{-1}$ is any locally Lipschitz function defined on $\mathbb{R}^{d_\xi} \times \overline{U}_m$ which is a left-inverse of $\overline{\mathbf{H}}_m$ on $\overline{\mathbf{H}}_m(\mathcal{C} \times \overline{U}_m)$. It follows that the only difficulty is the computation of a globally defined left-inverse for $\overline{\mathbf{H}}_m$, which is needed anyway to deduce an estimate $\hat{x}$ from $\hat{\xi}$ (see [RM04]).

Stationary transformation

We have seen that under an appropriate injectivity assumption, the function made of the successive derivatives of the output transforms the system into a Lipschitz phase-variable form. The drawback of this design is that the transformation depends on the derivatives of the input, which we may not have access to, in particular if we are not in an output feedback configuration. It turns out that under appropriate assumptions involving uniform observability, a control-affine multi-input single-output system

$$\dot{x} = f(x) + g(x)u \quad , \quad y = h(x) \in \mathbb{R} \tag{5.9}$$

can be transformed into a Lipschitz triangular form (3.12) by a stationary transformation. This famous result was first proved in [GB81] and then in a simpler way in [GHO92]. Before stating the result, we need the following definition.

Definition 5.2.3.

We call *drift system* of System (5.9) the dynamics with $u \equiv 0$, namely

$$\dot{x} = f(x) \quad , \quad y = h(x) \ .$$

Applying Definition 5.2.2, we say that the drift system of System (5.9) is weakly (resp strongly) differentially observable of order m on $\mathcal{S}$ if the function

$$\mathbf{H}_m(x) = (h(x), L_f h(x), \ldots, L_f^{m-1} h(x))$$

is injective (resp an injective immersion) on $\mathcal{S}$.

Differential observability of the drift system is weaker than differential observability of the system since it is only for $u \equiv 0$[12]. In order to obtain a triangular form, it is necessary to add an assumption of uniform observability :

Theorem 5.2.2. [GB81, GHO92]

Assume that there exists an open subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$ such that

[11] The saturation function is defined by $\mathsf{sat}_M(s) = \min\{M, \max\{s, -M\}\}$.

[12] or any other constant value

- System (5.9) is uniformly instantaneously observable[13] on $\mathcal{S}$;

- The drift system of System (5.9) is strongly differentially observable of order d_x on $\mathcal{S}$.

Then, $\mathbf{H}_{d_x}$ defined by

$$\mathbf{H}_{d_x}(x) = (h(x), L_f h(x), \dots, L_f^{d_x-1} h(x)) \tag{5.10}$$

which is a diffeomorphism on $\mathcal{S}$ by assumption, transforms System (5.9) into a Lipschitz triangular form (3.12) of dimension $d_\xi = d_x$ on $\mathcal{S}$.

Triangularity makes the form (3.12) instantaneously observable for any input. Since the transformation $\mathbf{H}_{d_x}$ itself is independent from the input and injective, this observability property must necessarily be verified by the original System (5.9). Thus, the first assumption is necessary. A usual case where this property is verified is when there exists an order p such that the system is weakly differentially observable of order p.

It is crucial that the order of strong differential observability of the drift system be d_x (the dimension of the state) to ensure the Lipschitzness of the triangular form in order to use a high gain observer. When this order is larger than d_x, we will see in Chapter 6 that triangularity is often preserved but the Lipschitzness is lost : the triangular form is only continuous and observers from Chapter 4 must be used.

5.2.2 General Lipschitz triangular form

Consider a general multi-input single-output control-affine system

$$\dot{x} = f(x) + g(x)u \quad , \quad y = h(x) + h^u(x)u \ \in \mathbb{R} \tag{5.11}$$

where g and h^u are matrix fields with values in $\mathbb{R}^{d_x \times d_u}$ and $\mathbb{R}^{1 \times d_u}$ such that for any $u = (u_1, ..., u_{d_u})$ in $\mathbb{R}^{d_u}$,

$$g(x)u = \sum_{k=1}^{d_u} g_k(x)u_k \quad , \quad h^u(x)u = \sum_{k=1}^{d_u} h_k^u(x)u_k$$

with g_k vector fields of $\mathbb{R}^{d_x}$ and h_k^u real-valued functions. We want to know under which conditions this system can be transformed into a general Lipschitz triangular form (3.19)

$$\left\{ \begin{array}{rcll} \dot{\xi}_1 & = & A_1(u,y)\,\xi_2 + \Phi_1(u,\xi_1) & \\ & \vdots & & \\ \dot{\xi}_i & = & A_i(u,y)\,\xi_{i+1} + \Phi_i(u,\xi_1,\dots,\xi_i) & , \quad y = C_1(u)\xi_1 \\ & \vdots & & \\ \dot{\xi}_m & = & \Phi_m(u,\xi) & \end{array} \right.$$

for which a Kalman-High gain observer (3.21) may exist[14]. Before stating the main result, we need some definitions introduced in [HK77].

Definition 5.2.4.

The *observation space* of System 5.11, denoted $\mathcal{O}$, is the smallest real vector space such that

- $x \mapsto h(x)$ and $x \mapsto h_k^u(x)$ for any k in $\{1, \dots, d_u\}$ are in $\mathcal{O}$;

- $\mathcal{O}$ is stable under the Lie derivative along the vector fields f, g_1, ... , g_{d_u}, i-e for any

[13] see Definition 2.2.1.

[14] An additional excitation condition on the input is needed, see Chapter 3.

element ϕ of $\mathcal{O}$, $L_f\phi$ and $L_{g_k}\phi$ for all k in $\{1,\dots,d_u\}$ are in $\mathcal{O}$.

We denote $d\mathcal{O}$ the codistribution of $\mathbb{R}^{d_x}$ defined by

$$d\mathcal{O}(x) = \left\{ d\phi(x) \ , \quad \phi \in \mathcal{O} \right\} .$$

This leads to the following observability notion.

Definition 5.2.5.

System 5.11 is said to satisfy the *observability rank condition* at a point x in $\mathbb{R}^{d_x}$ (resp on $\mathcal{S}$) if

$$\dim(d\mathcal{O}(x)) = d_x \quad (\text{resp } \forall x \in \mathcal{S}) \ .$$

It is proved in [HK77] that the observability rank condition is sufficient to ensure the so-called "local weak observability", which roughly means that any point can be instantaneously distinguished from its neighbors via the output. In fact, this property is also necessary on a dense subset of $\mathcal{X}$. We refer the interested reader to [HK77] for a more precise account of those notions.

In [BT07], the authors relate the observability rank condition to the ability of transforming (at least locally) a system into a general Lipschitz triangular form.

Theorem 5.2.3. [BT07]

If System (5.11) satisfies the observability rank condition at x_0 then there exists a neighborhood $\mathcal{V}$ of x_0 and an injective immersion T on $\mathcal{V}$ which transforms System (5.11) into a general Lipschitz triangular form (3.19) on $\mathcal{V}$ with the linear parts A_i independent from the output i-e $A_i(u,y) = A_i(u)$.

This result is local because the rank condition is of local nature and does not say that we can select the same immersion T around every point of $\mathcal{X}$, let alone that this function is injective on $\mathcal{X}$. However, we give this result all the same because the idea of the construction of the function T is the same whether we look for a global immersion or a local one. Here is the algorithm presented in [BT07]:

1. Take $T^1(x) = (h(x), h_1^u(x), \dots, h_{d_u}^u(x))$ of dimension $N_1 = d_u + 1$.
2. Suppose $T^1, \dots, T^i$ have been constructed in the previous steps, of dimension $N_1, \dots, N_i$. Pick among their $N_1+...+N_i$ differentials a maximum number ν_i of differentials $d\phi_1, \dots, d\phi_{\nu_i}$ which generate a regular codistribution around x_0, i-e there exists a neighborhood of x_0 where $\dim(\mathsf{span}\{d\phi_1(x), \dots, d\phi_{\nu_i}(x)\})$ is constant and equal to ν_i.
   - if $\nu_i = d_x$ stop ;
   - otherwise build T^{i+1} with every functions $L_f T_j^i$ and $L_{g_k} T_j^i$, with j in $\{1,\dots,N_i\}$ and k in $\{1,\dots,d_u\}$, except those whose differential already belongs to $\mathsf{span}\{d\phi_1(x), \dots, d\phi_{\nu_i}(x)\}$ in a neighborhood of x_0.

   Finally, denoting m the number of iterations, take $T(x) = (T^1(x), \dots, T^m(x))$.

The observability rank condition ensures that the algorithm stops at some time because computing the successive T^i comes back to progressively generating all $\mathcal{O}$ which is of dimension d_x around x_0. Besides, it is shown in [BT07], that when the differential $d\phi$ of some real valued function ϕ is such that, in a neighborhood of x_0, $d\phi(x)$ belongs to $\mathsf{span}\{d\phi_1(x), \dots, d\phi_{\nu_i}(x)\}$ with $d\phi_1(x), \dots, d\phi_{\nu_i}(x)$ independent, then ϕ can be locally expressed in a Lipschitz way in

terms of $\phi_1, \dots, \phi_{\nu_i}$. Therefore, either the derivatives of the elements of T^i are in T^{i+1} or they can be expressed in terms of the previous $T^1, \dots, T^i$. It follows that for any i, there exist a matrix $A_i(u)$ and a function Φ_i (linear in u and with $\Phi(u, \cdot)$ Lipschitz) such that

$$\dot{\overbrace{T^i(x)}} = L_f T^i(x) + \sum_{k=1}^{d_u} u_k \, L_{g_k} T^i(x) = A_i(u) T^{i+1}(x) + \Phi_i(u, T^1(x), \dots, T^i(x)) \ ,$$

which gives the general triangular form (3.19).

Note that the transformation T thus obtained is a local immersion. If we are interested in a global transformation, the same algorithm can be applied but everything must be checked globally (and not in a neighborhood of x_0) and we need to go on with this algorithm until obtaining a global injective immersion. But there is no guarantee that this will be possible, unless a stronger assumption is made. In particular, if the drift system (i-e with $u \equiv 0$) is strongly differentially observable of some order p, the algorithm provides a global injective immersion in a maximum of p iterations. Beware however, that it still remains to check that the functions Φ_i exist globally. If this is not the case, it is always possible to put the corresponding $L_f T^i_j(x)$ or $L_{g_k} T^i_j(x)$ in T^{i+1}, but this is bound to considerably increase the dimension of T (and thus of the observer).

Finally, it is important to remark that this design enables to avoid the strong assumption of uniform observability needed for the classical triangular form, by stuffing the $L_{g_k} T^i_j(x)$ which do not verify the triangularity constraint into the state. The first obvious setback is that it often leads to observers of very large dimension. But mostly, unlike the classical Lipschitz triangular form which admits a high gain observer without further assumption, the possibility of observer design for the general Lipschitz triangular form is not automatically achieved as seen in Chapter 3 : building the transformation is not enough, one need to check an additional excitation condition on the input.

5.3 Conclusion

A lot of results exist in the literature concerning the characterization of systems which can be transformed into a normal form and we have tried to give in this chapter as thorough an account as possible. Those results are summed up in Table 5.1. However, some cases have not been addressed yet. They are highlighted in the table with the sign **?** and will be studied in the following chapters :

- Chapter 6 : transformation into a continuous triangular form. We study what becomes of Theorem 5.2.2 when the system has an order of differential observability larger than d_x. We show that using the same transformation, triangularity may be preserved but not its Lipschitzness, i-e the system may be transformed into a continuous triangular form, instead of a Lipschitz triangular form.

- Chapter 7 : transformation of time-varying/controlled systems into a Hurwitz form. We extend the results presented in Section 5.1.2 for autonomous systems to controlled systems. We show that similar results can be obtained under the assumption of backward distinguishability in finite time, or strong differential observability.

Normal form		Type of system	Observability assumption	Transformation T	T time-varying ?	Validity
State-affine form (3.5) or (3.7)		/	diverse but very restrictive	variable	no	often only local
Hurwitz form (3.3)		Autonomous	Backward-distinguishability	solution to PDE (5.4)	no	global
		Time-varying	?	?	?	?
		Controlled	?	?	?	?
Triangular form (3.12)	Continuous phase-variable	/	weakly observable of order m	output and its derivatives up to order $m-1$	yes, with derivatives of input	at least on any compact set
	Lipschitz phase-variable	/	strongly observable of order m	output and its derivatives up to order $m-1$	yes, with derivatives of input	at least on any compact set
	Continuous triangular	?	?	?	?	?
	Lipschitz triangular	control-affine single-output	uniformly observable and strongly observable of order d_x	output and its derivatives along drift vector field up to order d_x-1	no	global
General Lipschitz triangular form (3.19)		control-affine single-output	observability rank condition	output and its derivatives along each vector field	no	at least locally

Table 5.1: Which type of nonlinear system, under which observability condition and with which transformation and domain of validity can be transformed into each of the normal forms presented in Part I. It also says if the transformation in each case is time-varying.

Chapter 6

Transformation into a continuous triangular form

Chapitre 6 – Transformation dans une forme triangulaire continue. *Ce chapitre étend le résultat présenté dans [GB81, GHO92] et rappelé dans Theorem 5.2.2, selon lequel tout système instantanément uniformément observable et, pour $u \equiv 0$, fortement différentiellemment observable d'ordre sa dimension d_x, peut être transformé en une forme triangulaire Lipschitz* (3.12)*. En particulier, nous étudions le cas plus général où l'ordre d'observabilité différentielle est quelconque, c'est-à-dire éventuellemment supérieur à la dimension du système. Nous montrons que dans ce cas, la dynamique du système peut encore (au moins partiellemment) être décrite par une forme triangulaire continue mais que cette forme n'est plus nécessairement Lipschitz. Des conditions nécessaires et suffisantes pour que le caractère Lipschitz soit assuré sont établies, et en particulier un lien étroit avec l'observabilité infinitésimale uniforme est mis en évidence.*

Contents

This chapter extends the result presented in [GB81, GHO92] and recalled in Theorem 5.2.2 which says that any uniformly instantaneously observable[1] single-output control-affine system whose drift system is strongly differentially observable[2] of order its dimension d_x, can be transformed into a Lipschitz triangular form (3.12). In particular, we investigate what happens in the more general case where the drift system is weakly differentially observable of some order, namely of an order larger or equal to the dimension of the system. We shall see that, in this case, the system dynamics may still be described by a (partial) continuous triangular form but with nonlinear functions Φ_i which may not be locally Lipschitz. As we saw in Chapter 4,

[1] See Definition 2.2.1
[2] See Definition 5.2.3.

this loss of Lipschitzness can prevent the use of a high gain observer, and we establish in this chapter necessary and sufficient conditions on the system for the Lipschitzness to be ensured. In particular, a tight link with uniform infinitesimal observability is revealed. The results presented in this chapter have been published in [BPA17b].

6.1 Presentation of the problem

As in Section 5.2.1, we consider a single-output control-affine system of the form :

$$\dot{x} = f(x) + g(x)u \quad , \qquad y = h(x) \tag{6.1}$$

where x is the state in $\mathbb{R}^{d_x}$, u is an input in $\mathbb{R}^{d_u}$, y is a measured output in $\mathbb{R}$ and the functions f, g and h are sufficiently many times differentiable. As in the previous chapter, we go on denoting

$$\mathbf{H}_i(x) = (h(x), L_f h(x), ..., L_f^{i-1} h(x)) \in \mathbb{R}^i \ . \tag{6.2}$$

According to Definition 5.2.3, we say that the drift system of System (6.1) is weakly (resp. strongly) differentially observable of order m on $\mathcal{S}$ if $\mathbf{H}_m$ is injective (resp. an injective immersion) on $\mathcal{S}$.

We are interested in solving :

Problem $\mathfrak{T}$

Given a compact subset $\mathcal{C}$ of $\mathbb{R}^{d_x}$, under which condition do there exist integers $\mathfrak{r}$ and d_ξ, a continuous injective function $T : \mathcal{C} \to \mathbb{R}^{d_\xi}$, and continuous functions $\varphi_{d_\xi} : \mathbb{R}^{d_\xi} \to \mathbb{R}$ and $\mathfrak{g}_i : \mathbb{R}^i (\text{or } \mathbb{R}^{d_\xi}) \to \mathbb{R}^{d_u}$ such that T transforms System (6.1) into the up-to-$\mathfrak{r}$-triangular form

$$\left\{ \begin{array}{rcl} \dot{\xi}_1 &=& \xi_2 + \mathfrak{g}_1(\xi_1)\, u \\ &\vdots& \\ \dot{\xi}_\mathfrak{r} &=& \xi_{\mathfrak{r}+1} + \mathfrak{g}_\mathfrak{r}(\xi_1, \ldots, \xi_\mathfrak{r})\, u \\ \dot{\xi}_{\mathfrak{r}+1} &=& \xi_{\mathfrak{r}+2} + \mathfrak{g}_{\mathfrak{r}+1}(\xi)\, u \\ &\vdots& \\ \dot{\xi}_{d_\xi} &=& \varphi_{d_\xi}(\xi) + \mathfrak{g}_{d_\xi}(\xi)\, u \end{array} \right. \quad , \quad y = x_1 \tag{6.3}$$

on $\mathcal{C}$.

Because $\mathfrak{g}_i$ depends only on ξ_1 to ξ_i, for $i \leq \mathfrak{r}$, but potentially on all the components of ξ for $i > \mathfrak{r}$, we call this particular form *up-to-$\mathfrak{r}$-triangular normal form* and $\mathfrak{r}$ is called the order of triangularity. When $d_\xi = \mathfrak{r} + 1$, we say full triangular normal form. When the functions φ_{d_ξ} and $\mathfrak{g}_j$ are locally Lipschitz we say Lipschitz up-to-$\mathfrak{r}$-triangular normal form.

According to Theorem 5.2.2, in the case where System (6.1) is instantaneously uniformly observable and $\mathbf{H}_{d_x}$ is a diffeomorphism on an open set $\mathcal{S}$ containing the given compact set, $T = \mathbf{H}_{d_x}$ transforms the system on $\mathcal{C}$ into a full Lipschitz triangular normal form of dimension $d_\xi = d_x$. However, in general, it is possible for the system not to be strongly differentially observable of order d_x everywhere. This motivates our interest in the case where the drift system is strongly differentially observable of order $m > d_x$, i-e $\mathbf{H}_m$ is an injective immersion but not a diffeomorphism.

The specificity of the triangular normal form (6.3) is not so much in its structure but more in the dependence of its functions $\mathfrak{g}_i$ and φ_{d_ξ}. Indeed, by choosing $T = \mathbf{H}_{d_\xi}$, we obtain in general:

$$\dot{\overbrace{\mathbf{H}_{d_\xi}(x)}} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{pmatrix} \mathbf{H}_{d_\xi}(x) + \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ L_f^{d_\xi} h(x) \end{pmatrix} + L_g \mathbf{H}_{d_\xi}(x) u$$

But, to get (6.3), we need further the existence of functions φ_{d_ξ} and $\mathfrak{g}_i$ satisfying, for $i > \mathfrak{r}$,

$$L_f^{d_\xi} h(x) \;=\; \varphi_{d_\xi}(\mathbf{H}_{d_\xi}(x)) \quad , \quad L_g L_f^{i-1}(x) \;=\; \mathfrak{g}_i(\mathbf{H}_{d_\xi}(x)) \quad \forall x \in \mathcal{C} \tag{6.4}$$

and, for $i \leq \mathfrak{r}$,

$$L_g L_f^{i-1}(x) \;=\; \mathfrak{g}_i(h(x), \ldots, L_f^{i-1} h(x)) \quad \forall x \in \mathcal{C} \; . \tag{6.5}$$

Let us illustrate via the following elementary example what can occur.

Example 6.1.1 Consider the system defined as

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& x_2 \\ \dot{x}_2 &=& x_3^3 \\ \dot{x}_3 &=& 1+u \end{array} \right. \quad , \quad y \;=\; x_1$$

We get

$$\mathbf{H}_3(x) = \left(\begin{array}{c} h(x) \\ L_f h(x) \\ L_f^2 h(x) \end{array} \right) = \left(\begin{array}{c} x_1 \\ x_2 \\ x_3^3 \end{array} \right) \quad , \quad \mathbf{H}_5(x) = \left(\begin{array}{c} \mathbf{H}_3(x) \\ L_f^3 h(x) \\ L_f^4 h(x) \end{array} \right) = \left(\begin{array}{c} \mathbf{H}_3(x) \\ 3x_3^2 \\ 6x_3 \end{array} \right)$$

Hence $\mathbf{H}_3$ is a bijection and $\mathbf{H}_5$ is an injective immersion on $\mathbb{R}^3$. So the drift system is weakly differentially observable of order 3 on $\mathbb{R}^3$ and strongly differentially observable of order 5 on $\mathbb{R}^3$. Also the function $(x_1, x_2, x_3) \mapsto (y, \dot{y}, \ddot{y})$ being injective for all u, it is uniformly instantaneously observable on $\mathbb{R}^3$. From this we could be tempted to pick $d_\xi = 3$ or 5 and the compact set $\mathcal{C}$ arbitrary in $\mathbb{R}^3$. Unfortunately, if we choose $d_\xi = 3$, we must have

$$\varphi_3(\mathbf{H}_3(x)) = L_f^3 h(x) \;=\; 3x_3^2 \;=\; 3(L_f^2 h(x))^{2/3}$$

and there is no locally Lipschitz function φ_3 satisfying (6.4) if the given compact set $\mathcal{C}$ contains a point satisfying $x_3 = 0$. If we choose $d_\xi = 5$, we must have

$$\mathfrak{g}_3(\mathbf{H}_3(x)) = L_g L_f^2 h(x) = 3x_3^2 = L_f^3 h(x) = 3(L_f^2 h(x))^{2/3}$$

and there is no locally Lipschitz function $\mathfrak{g}_3$ satisfying (6.5) if the given compact set $\mathcal{C}$ contains a point satisfying $x_3 = 0$. ▲

Following this example, we leave aside the Lipschitzness requirement for the time being, and focus on the existence of continuous functions φ_{d_ξ} and $\mathfrak{g}_i$ verifying (6.4) and (6.5). It turns out that (6.4) is easily satisfied as soon as $\mathbf{H}_{d_\xi}$ is injective :

Theorem 6.1.1.

Suppose the drift system of System (6.1) is weakly (resp. strongly) differentially observable of order m on an open set $\mathcal{S}$ containing the given compact set $\mathcal{C}$. For any $d_\xi \geq m$, there exist continuous (resp. Lipschitz) functions $\varphi_{d_\xi} : \mathbb{R}^{d_\xi} \to \mathbb{R}$, $\mathfrak{g}_i : \mathbb{R}^{d_\xi} \to \mathbb{R}$ satisfying (6.4).

Proof : There is nothing really new in this result. It is a direct consequence of the fact that a continuous injective function, like $\mathbf{H}_m$, defined on a compact set admits a continuous left inverse defined on $\mathbb{R}^{d_\xi}$ (see Lemma A.3.3), and that when it is also an immersion, its left-inverse can be chosen Lipschitz on $\mathbb{R}^{d_\xi}$ (see Lemma A.3.5 or [RM04]). ■

We conclude that the real difficulty lies in finding triangular functions $\mathfrak{g}_i$ satisfying (6.5).

6.2 Existence of $\mathfrak{g}_i$ satisfying (6.5)

6.2.1 Main result

We will prove the following result :

Theorem 6.2.1.

Suppose System (6.1) is uniformly instantaneously observable on an open set $\mathcal{S}$ containing the given compact set $\mathcal{C}$. Then,

- there exists a continuous function $\mathfrak{g}_1 : \mathbb{R} \to \mathbb{R}^{d_u}$ satisfying (6.5).
- if, for some i in $\{2, \dots, d_x\}$, $\mathbf{H}_2, \dots, \mathbf{H}_i$ defined in (6.2) are open maps, then, for all $j \leq i$, there exists a continuous function $\mathfrak{g}_j : \mathbb{R}^j \to \mathbb{R}^{d_u}$ satisfying (6.5).

The rest of this section is dedicated to the proof of this result through a series of lemmas. Note that, in the case where the drift system is strongly differentially observable of order d_x, $\mathbf{H}_i$ is a submersion and thus open for all $i \leq d_x$, and the result holds.

A first important thing to notice is that the following property must be satisfied for the identity (6.5) to be satisfied (on $\mathcal{S}$).

Property $\mathcal{A}(i)$

$$L_g L_f^{i-1} h(x_a) = L_g L_f^{i-1} h(x_b) \qquad \forall (x_a, x_b) \in \mathcal{S}^2 \, : \, \mathbf{H}_i(x_a) = \mathbf{H}_i(x_b)$$

Actually the converse is true and is a direct consequence of Lemma A.3.3 :

Lemma 6.2.1.

If Property $\mathcal{A}(i)$ is satisfied with $\mathcal{S}$ containing the given compact set $\mathcal{C}$, then there exists a continuous function $\mathfrak{g}_i : \mathbb{R}^i \to \mathbb{R}^{d_u}$ satisfying (6.5).

Property $\mathcal{A}(i)$ being sufficient to obtain the existence of a function $\mathfrak{g}_i$ satisfying (6.5), we study now under which conditions it holds. Clearly $\mathcal{A}(i)$ is satisfied for all $i \geq m$ if $\mathbf{H}_m$ is injective. If we do not have this injectivity property the situation is more complex. To overcome the difficulty we introduce the following property for $2 \leq i \leq d_x + 1$.

Property $\mathcal{B}(i)$

For any (x_a, x_b) in $\mathcal{S}^2$ such that $x_a \neq x_b$ and $\mathbf{H}_i(x_a) = \mathbf{H}_i(x_b)$, there exists a sequence $(x_{a,k}, x_{b,k})$ of points in $\mathcal{S}^2$ converging to (x_a, x_b) such that for all k, $\mathbf{H}_i(x_{a,k}) = \mathbf{H}_i(x_{b,k})$ and $\dfrac{\partial \mathbf{H}_{i-1}}{\partial x}$ is full-rank at $x_{a,k}$ or $x_{b,k}$.

As in this property, let $x_a \neq x_b$ be such that $\mathbf{H}_i(x_a) = \mathbf{H}_i(x_b)$. If $\frac{\partial \mathbf{H}_{i-1}}{\partial x}$ is full-rank at either x_a or x_b, then we can take $(x_{a,k}, x_{b,k})$ constant equal to (x_a, x_b). Thus, it is sufficient to check $\mathcal{B}(i)$ around points where neither $\frac{\partial \mathbf{H}_{i-1}}{\partial x}(x_a)$ nor $\frac{\partial \mathbf{H}_{i-1}}{\partial x}(x_b)$ is full-rank. But according to [GK01, Theorem 4.1], the set of points where $\frac{\partial \mathbf{H}_{d_x}}{\partial x}$ is not full-rank is of codimension at least one for a uniformly observable system. Thus, it is always possible to find points $x_{a,k}$ as close to x_a as we want such that $\frac{\partial \mathbf{H}_{i-1}}{\partial x}(x_{a,k})$ is full-rank. The difficulty of $\mathcal{B}(i)$ thus rather lies in ensuring that we have also $\mathbf{H}_i(x_{a,k}) = \mathbf{H}_i(x_{b,k})$.

In Section 6.2.2, we prove :

Lemma 6.2.2.

Suppose System (6.1) is uniformly instantaneously observable on $\mathcal{S}$.

- Property $\mathcal{A}(1)$ is satisfied.
- If, for some i in $\{2, \dots, d_x + 1\}$, Property $\mathcal{B}(i)$ holds and Property $\mathcal{A}(j)$ is satisfied for

all j in $\{1,\dots,i-1\}$, then Property $\mathcal{A}(i)$ holds.

Thus, the first point in Theorem 6.2.1 is proved. Besides, a direct consequence of Lemmas 6.2.1 and 6.2.2 is that a sufficient condition to have the existence of the functions $\mathfrak{g}_i$ for i in $\{2,\dots,d_x+1\}$ is to have $\mathcal{B}(j)$ for j in $\{2,\dots,i\}$. The following lemma finishes the proof of Theorem 6.2.1 by showing that $\mathcal{B}(j)$ is in fact satisfied when $\mathbf{H}_j$ is an open map.

Lemma 6.2.3.

Suppose that for some j in $\{2,\dots,d_x\}$, $\mathbf{H}_j$ is an open map on $\mathcal{S}$. Then, $\mathcal{B}(j)$ is satisfied.

Proof : Take (x_a,x_b) in $\mathcal{S}^2$ such that $x_a \neq x_b$ and $\mathbf{H}_j(x_a)=\mathbf{H}_j(x_b)=y_0$. Let Π be the set of points of $\mathcal{S}$ such that $\frac{\partial \mathbf{H}_j}{\partial x}$ is not full-rank. According to Sard's theorem, $\mathbf{H}_j(\Pi)$ is of measure zero in $\mathbb{R}^j$. Now, take $p>0$ and consider $B_p(x_a)$ and $B_p(x_b)$ the open balls of radius $\frac{1}{p}$ centered at x_a and x_b respectively. Since $\mathbf{H}_j$ is open, $\mathbf{H}_j(B_p(x_a))$ and $\mathbf{H}_j(B_p(x_b))$ are open sets, both containing y_0. Thus, $\mathbf{H}_j(B_p(x_a)) \cap \mathbf{H}_j(B_p(x_b))$ is a non-empty open set. It follows that $(\mathbf{H}_j(B_p(x_a)) \cap \mathbf{H}_j(B_p(x_b))) \setminus \mathbf{H}_j(\Pi)$ is non-empty and contains a point y_p. We conclude that there exist $(x_{a,p},x_{b,p})$ in $B_p(x_a)\times B_p(x_b)$ such that $\mathbf{H}_j(x_{a,p})=\mathbf{H}_j(x_{b,p})=y_p$ and $\frac{\partial \mathbf{H}_j}{\partial x}$ (and thus $\frac{\partial \mathbf{H}_{j-1}}{\partial x}$) is full-rank at $x_{a,p}$ and $x_{b,p}$. Besides $(x_{a,p},x_{b,p})$ converges to (x_a,x_b), and $\mathcal{B}(j)$ is satisfied. ■

Note that the assumption that $\mathbf{H}_j$ is an open map is stronger than $\mathcal{B}(j)$ since it leads to the full rank of $\frac{\partial \mathbf{H}_j}{\partial x}$, while, in $\mathcal{B}(j)$, we only need the full-rank for $\frac{\partial \mathbf{H}_{j-1}}{\partial x}$. We show in the following example that the openness of $\mathbf{H}_j$ is not necessary.

Example 6.2.1 Consider the system defined as

$$\left\{\begin{array}{rcl} \dot{x}_1 &=& x_2 \\ \dot{x}_2 &=& x_3^3 x_1 \\ \dot{x}_3 &=& 1+u \end{array}\right. \quad , \quad y \;=\; x_1 \tag{6.6}$$

On $\mathcal{S}=\{x\in\mathbb{R}^3 : x_1^2+x_2^2\neq 0\}$, and whatever u is, the knowledge of the function $t\mapsto y(t)=X_1(x,t)$ and therefore of its three first derivatives

$$\dot{y} \;=\; x_2 \;, \quad \ddot{y} \;=\; x_3^3 x_1 \;, \quad \dddot{y} \;=\; 3x_3^2 x_1(1+u)+x_3^3 x_2$$

gives us x_1, x_2 and x_3. Thus, the system is uniformly instantaneously observable on $\mathcal{S}$. Besides, the function

$$\mathbf{H}_4(x)=\begin{pmatrix} x_1 \\ x_2 \\ x_3^3 x_1 \\ 3x_3^2 x_1 + x_3^3 x_2 \end{pmatrix}$$

is injective on $\mathcal{S}$, thus the system is weakly differentially observable of order 4 on $\mathcal{S}$. Now, although $\mathbf{H}_2$ is trivially an open map on $\mathcal{S}$, $\mathbf{H}_3$ is not. Indeed, consider for instance the open ball[3] $B_{\frac{1}{2}}(0,x_2,0)$ in $\mathbb{R}^3$ for some x_2 such that $|x_2|>\frac{1}{2}$. $B_{\frac{1}{2}}(0,x_2,0)$ is contained in $\mathcal{S}$. Suppose its image by $\mathbf{H}_3$ is an open set of $\mathbb{R}^3$. It contains $\mathbf{H}_3(0,x_2,0)=(0,x_2,0)$ and thus $(\varepsilon,x_2,\varepsilon)$ for any sufficiently small ε. This means that there exist x in $B_{\frac{1}{2}}(0,x_2,0)$ such that $(\varepsilon,x_2,\varepsilon)=\mathbf{H}_3(x)$, i-e necessarily $x_1=\varepsilon$ and $x_3=1$. But this point is not in $B_{\frac{1}{2}}(0,x_2,0)$, and we have a contradiction. Therefore, $\mathbf{H}_3$ is not open. However, $\mathcal{B}(3)$ trivially holds because $\mathbf{H}_2$ is full-rank everywhere. ▲

6.2.2 Proof of Lemma 6.2.2

Lemma 6.2.2 is fundamental for the main result of this chapter. That is why we dedicate a whole section to its proof. It is built in the same spirit as the one in [GHO92] but in a more

[3] $B_r(x)$ denotes the open ball centered at x and with radius r.

detailed and complete way so that the reader can understand how the fact that $\mathbf{H}_{d_\xi}$ is no longer a diffeomorphism makes a great difference.

Assume the system is uniformly instantaneously observable on $\mathcal{S}$. We first show that property $\mathcal{A}(1)$ holds. Suppose there exists (x_a^*, x_b^*) in $\mathcal{S}^2$ and k in $\{1, ..., d_u\}$ such that $x_a^* \neq x_b^*$ and

$$h(x_a^*) = h(x_b^*) \qquad , \qquad L_{g_k} h(x_a^*) \neq L_{g_k} h(x_b^*) \ .$$

Then, the control law u with all its components zero except its kth one which is

$$u_k \ = \ -\frac{L_f h(x_a) - L_f h(x_b)}{L_{g_k} h(x_a) - L_{g_k} h(x_b)} \ .$$

is defined on a neighborhood of (x_a^*, x_b^*). The corresponding solutions $X(x_a^*; t; u)$ and $X(x_b^*; t; u)$ are defined on some time interval $[0, \bar{t})$ and satisfy

$$h(X(x_a^*; t; u)) = h(X(x_b^*; t; u)) \qquad \forall t \in [0, \bar{t}) \ .$$

Since x_a^* is different from x_b^*, this contradicts the instantaneous observability. Thus $\mathcal{A}(1)$ holds.

Let now i in $\{2, \ldots, d_x + 1\}$ be such that Property $\mathcal{B}(i)$ holds and $\mathcal{A}(j)$ is satisfied for all j in $\{1, \ldots, i-1\}$. To establish by contradiction that $\mathcal{A}(i)$ holds, we assume this is not the case. This means that there exists $(x_{a,0}^*, x_{b,0}^*)$ in $\mathcal{S}^2$ and k in $\{1, ..., d_u\}$ such that $\mathbf{H}_i(x_{a,0}^*) = \mathbf{H}_i(x_{b,0}^*)$ but $L_{g_k} L_f^{i-1}(x_{a,0}^*) \neq L_{g_k} L_f^{i-1}(x_{b,0}^*)$. This implies $x_{a,0}^* \neq x_{b,0}^*$. By continuity of $L_{g_k} L_f^{i-1}$ and according to $\mathcal{B}(i)$, there exists x_a^* (resp x_b^*) in $\mathcal{S}$ sufficiently close to $x_{a,0}^*$ (resp $x_{b,0}^*$) satisfying $x_a^* \neq x_b^*$,

$$\mathbf{H}_i(x_a^*) = \mathbf{H}_i(x_b^*) \ , \quad L_{g_k} L_f^{i-1}(x_a^*) \neq L_{g_k} L_f^{i-1}(x_b^*) \ ,$$

and $\frac{\partial \mathbf{H}_{i-1}}{\partial x}$ is full-rank at x_a^* or x_b^*. Without loss of generality, we suppose it is full-rank at x_a^*. Thus, $\frac{\partial \mathbf{H}_j}{\partial x}$ is full-rank at x_a^* for all $j < i \leq d_x + 1$. We deduce that there exists an open neighborhood $\mathcal{V}_a$ of x_a^* such that for all $j < i$, $\frac{\partial \mathbf{H}_j}{\partial x}$ is full-rank on $\mathcal{V}_a$. Since $\mathcal{A}(j)$ holds for all $j < i$, according to Lemma A.3.4, $\mathbf{H}_j(\mathcal{V}_a)$ is open for all $j < i$ and there exist locally Lipschitz functions $\mathfrak{g}_j : \mathbf{H}_j(\mathcal{V}_a) \to \mathbb{R}^{d_u}$ such that, for all x_α in $\mathcal{V}_a$,

$$\mathfrak{g}_j(\mathbf{H}_j(x_\alpha)) = L_g L_f^{j-1} h(x_\alpha) \ . \tag{6.7}$$

Also, $\mathbf{H}_j(x_a^*) = \mathbf{H}_j(x_b^*)$ implies that $\mathbf{H}_j(x_b^*)$ is in the open set $\mathbf{H}_j(\mathcal{V}_a)$. Continuity of each $\mathbf{H}_j$ implies the existence of an open neighborhood $\mathcal{V}_b$ of x_b^* such that $\mathbf{H}_j(\mathcal{V}_b)$ is contained in $\mathbf{H}_j(\mathcal{V}_a)$ for all $j < i$. Thus, for any x_β in $\mathcal{V}_b$, $\mathbf{H}_j(x_\beta)$ is in $\mathbf{H}_j(\mathcal{V}_a)$, and there exists x_α in $\mathcal{V}_a$ such that $\mathbf{H}_j(x_\alpha) = \mathbf{H}_j(x_\beta)$. According to $\mathcal{A}(j)$ this implies that $L_g L_f^{j-1} h(x_\beta) = L_g L_f^{j-1} h(x_\alpha)$ and with (6.7),

$$L_g L_f^{j-1} h(x_\beta) = L_g L_f^{j-1} h(x_\alpha) = \mathfrak{g}_j(\mathbf{H}_j(x_\alpha)) = \mathfrak{g}_j(\mathbf{H}_j(x_\beta)) \ .$$

Therefore, (6.7) holds on $\mathcal{V}_a$ and $\mathcal{V}_b$.

Then, the control law u with all its components zero except its kth one which is

$$u_k \ = \ -\frac{L_f^i h(x_a) - L_f^i h(x_b)}{L_{g_k} L_f^{i-1} h(x_a) - L_{g_k} L_f^{i-1} h(x_b)}$$

is defined on a neighborhood of (x_a^*, x_b^*). The corresponding solutions $X(x_a^*; t; u)$ and $X(x_b^*; t; u)$ are defined on some time interval $[0, \bar{t})$ where they remain in $\mathcal{V}_a$ and $\mathcal{V}_b$ respectively. Let $Z_a(t) = \mathbf{H}_i(X(x_a^*; t; u))$, $Z_b(t) = \mathbf{H}_i(X(x_b^*; t; u))$ and $W(t) = Z_a(t) - Z_b(t)$ on $[0, \bar{t})$. Since, for

all $j < i$, (6.7) holds on $\mathcal{V}_a$ and $\mathcal{V}_b$, (W, Z_a) is solution to the system :

$$\begin{cases} \dot{w}_1 &= w_2 + (\mathfrak{g}_1(\xi_{a,1}) - \mathfrak{g}_1(\xi_{a,1} - w_1))\, u \\ &\vdots \\ \dot{w}_j &= w_{j+1} + (\mathfrak{g}_j(\xi_{a,1}, ..., \xi_{a,j}) - \mathfrak{g}_j(\xi_{a,1} - w_1, ..., \xi_{a,j} - w_j))\, u \\ &\vdots \\ \dot{w}_i &= 0 \\ \dot{\xi}_{a,1} &= \xi_2 + \mathfrak{g}_1(\xi_{a,1})\, u \\ &\vdots \\ \dot{\xi}_{a,j} &= \xi_{j+1} + \mathfrak{g}_j(\xi_{a,1}, ..., \xi_{a,j})\, u \\ &\vdots \\ \dot{\xi}_{a,i} &= \tilde{u} \end{cases}$$

with initial condition $(0, \mathbf{H}_i(x_a^*))$, where $\tilde{u}$ is the time derivative of $Z_{a,i}(t)$. Note that the function $(0, Z_a)$ is also a solution to this system with the same initial condition. Since the functions involved in this system are locally Lipschitz, it admits a unique solution. Hence, for all t in $[0, \bar{t}[$, $W(t) = 0$, and thus $Z_a(t) = Z_b(t)$, which implies $h(X(x_a^*, t)) = h(X(x_b^*, t))$. Since x_a^* is different from x_b^*, this contradicts the uniform observability. Thus $\mathcal{A}(i)$ holds.

6.2.3 A solution to Problem $\mathfrak{T}$

With Theorems 6.1.1 and 6.2.1, we have the following solution to Problem $\mathfrak{T}$.

Theorem 6.2.2.

Let $\mathcal{S}$ be an open set containing the given compact set $\mathcal{C}$. Suppose

- System (6.1) is uniformly instantaneously observable on $\mathcal{S}$
- the drift system of System (6.1) is weakly differentially observable of order m on $\mathcal{S}$.

With selecting $T = \mathbf{H_m}$ and $d_\xi = m$, we have a solution to Problem $\mathfrak{T}$ if we pick either $\mathfrak{r} = 1$, or $\mathfrak{r} = i$ when $\mathbf{H}_j$ is an open map for any j in $\{2, \ldots, i\}$ with $i \leq d_x$.

Remark 5

- As seen in Example 6.2.1, the openness of the functions $\mathbf{H}_j$ is sufficient but not necessary. We may ask only for $\mathcal{B}(j)$ for any j in $\{2, \ldots, i\}$ with $i \leq d_x + 1$. Besides, this weaker assumption allows to obtain the existence of $\mathfrak{g}_i$ up to the order $d_x + 1$.

- Consider the case where $\mathcal{B}(j)$ is satisfied for all $j \leq d_x + 1$ and $m = d_x + 2$. Then we have $\mathfrak{r} = d_x + 1$ and it is possible to obtain a full triangular form of dimension $d_\xi = \mathfrak{r} + 1 = m = d_x + 2$. Actually, we still have a full triangular form if we choose $d_\xi > m$. Indeed, $\mathbf{H}_m$ being injective, $\mathcal{A}(i)$ is satisfied for all i larger than m, thus there also exist continuous functions $\mathfrak{g}_i : \mathbb{R}^i \to \mathbb{R}^{d_u}$ satisfying (6.5) for all $i \geq m$. It follows that $\mathfrak{r}$ can be taken larger than $d_x + 1$ and $d_\xi = \mathfrak{r} + 1$ larger than m.

- If Problem $\mathfrak{T}$ is solved with $d_\xi = \mathfrak{r} + 1$, we have a full triangular normal form of dimension d_ξ. But, at this point we know nothing about the regularity of the functions $\mathfrak{g}_i$, besides continuity. As we saw in Example 6.1.1, even the usual assumption of strong differential observability is not sufficient to make it Lipschitz everywhere. As studied in Chapter 4, this may impede the convergence of a high gain observer. That is why, in the next section, we look for conditions under which the Lipschitzness is ensured.

- As explained in Section 5.2.1, another way of solving Problem $\mathfrak{T}$ is to allow the transformation T to depend on the control u and its derivatives. In particular, if $d_\xi > \mathfrak{r}+1$, a full triangular form may still be obtained with $T = (\mathbf{H}_\mathfrak{r}, \tilde{T})$ where the components $\tilde{T}_i$ of $\tilde{T}$ are defined recursively as

$$\tilde{T}_1 = L_f^\mathfrak{r} h \quad , \quad \tilde{T}_{i+1} = L_{f+gu}\tilde{T}_i + \sum_{j=0}^{i-2} \frac{\partial \tilde{T}_i}{\partial u^{(j)}} u^{(j+1)}$$

  until (if possible) the map $x \mapsto T(x, u, \dot{u}, ...)$ becomes injective for all $(u, \dot{u}, ...)$. The interest of this approach is to ensure triangularity while reducing the order of differentiation of u compared to Theorem 5.2.1.

Example 6.2.2 Coming back to Example 6.2.1, we have seen that $\mathbf{H}_2$ is open and that $\mathbf{H}_3$ is not but $\mathcal{B}(3)$ is satisfied. Besides, the system is weakly differentially observable of order 4. We deduce that there exists a full-triangular form of order 4. Indeed, we have $L_g h(x) = L_g L_f h(x) = 0$ and

$$L_g L_f^2 h(x) = 3x_3^2 x_1 = 3(L_f^2 h(x))^{\frac{2}{3}} (h(x))^{\frac{1}{3}}$$

so that we can take

$$\mathfrak{g}_1 = \mathfrak{g}_2 = 0 \quad , \quad \mathfrak{g}_3(\xi_1, \xi_2, \xi_3) = 3\xi_3^{\frac{2}{3}} \xi_1^{\frac{1}{3}}.$$

As for φ_4 and $\mathfrak{g}_4$, they are obtained via inversion of $\mathbf{H}_4$ i-e for instance on $\mathbb{R}^4 \setminus \{(0, 0, \xi_3),\ \xi_3 \in \mathbb{R}\}$

$$\mathbf{H}_4^{-1}(\xi) = \left(\xi_1, \xi_2, \left(\frac{(\xi_4 - 3\xi_3^{\frac{2}{3}} \xi_1^{\frac{1}{3}})^2 + \xi_3^2}{\xi_1^2 + \xi_2^2} \right)^{\frac{1}{6}} \right) .$$

6.3 Lipschitzness of the triangular form

6.3.1 A sufficient condition

We saw with Examples 6.1.1 and 6.2.1 that uniform instantaneous observability is not sufficient for the functions $\mathfrak{g}_i$ to be Lipschitz. Nevertheless, we are going to show in this section that it is sufficient except maybe around the image of points where $\dfrac{\partial \mathbf{H}_i}{\partial x}$ is not full-rank ($x_1 = 0$ or $x_3 = 0$ in Example 6.2.1).

Consider the open set $\mathcal{R}_i$ of points in $\mathcal{S}$ where $\dfrac{\partial \mathbf{H}_i}{\partial x}$ has full rank. According to [Leb82, Corollaire p68-69], if $\mathbf{H}_i$ is an open map, $\mathcal{R}_i$ is an open dense set. Anyway, assume $\mathcal{R}_{d_x} \cap \mathcal{C}$ is non empty. Then there exists $\varepsilon_0 > 0$ such that, for all ε in $(0, \varepsilon_0]$, the set

$$K_{i,\varepsilon} = \left\{ x \in \mathcal{R}_i \cap \mathcal{C} \ , \quad d(x\,, \mathbb{R}^{d_x} \setminus \mathcal{R}_i) \geq \varepsilon \right\} .$$

is non-empty and compact, and such that its points are (ε)-away from singular points. The next theorem shows that the functions $\mathfrak{g}_i$ can be taken Lipschitz on the image of $K_{i,\varepsilon}$, i-e everywhere except arbitrary close to the image of points where the rank of the Jacobian of $\mathbf{H}_i$ drops.

Theorem 6.3.1.

Assume System (6.1) is uniformly instantaneously observable on an open set $\mathcal{S}$ containing the compact set $\mathcal{C}$. For all i in $\{1, ..., d_x\}$ and for any ε in $(0, \varepsilon_0]$, there exists a Lipschitz function $\mathfrak{g}_i : \mathbb{R}^i \to \mathbb{R}^{d_u}$ satisfying (6.5) for all x in $K_{i,\varepsilon}$.

Proof : As noticed after the statement of Property $\mathcal{B}(i)$, since $\dfrac{\partial \mathbf{H}_i}{\partial x}$ has full rank in the open set $\mathcal{R}_i$, Property $\mathcal{B}(i)$ holds on $\mathcal{R}_i$ (i-e with $\mathcal{R}_i$ replacing $\mathcal{S}$ in its statement). It follows from Lemma 6.2.2 that $\mathcal{A}(i)$ is satisfied on $\mathcal{R}_i$. Besides, according to Lemma A.3.4, $\mathbf{H}_i(\mathcal{R}_i)$ is open and there exists a C^1

function $\mathfrak{g}_i$ defined on $\mathbf{H}_i(\mathcal{R}_i)$ such that for all x in $\mathcal{R}_i$, $\mathfrak{g}_i(\mathbf{H}_i(x)) = L_g L_f^{i-1} h(x)$. Now, $K_{i,\varepsilon}$ being a compact set contained in $\mathcal{R}_i$, and $\mathbf{H}_i$ being continuous, $\mathbf{H}_i(K_{i,\varepsilon})$ is a compact set contained in $\mathbf{H}_i(\mathcal{R}_i)$. Thus, $\mathfrak{g}_i$ is Lipschitz on $\mathbf{H}_i(K_{i,\varepsilon})$. According to [McS34], there exists a Lipschitz extension of $\mathfrak{g}_i$ to $\mathbb{R}^i$ coinciding with $\mathfrak{g}_i$ on $\mathbf{H}_i(K_{i,\varepsilon})$, and thus verifying (6.5) for all x in $K_{i,\varepsilon}$. ■

For a strongly differentially observable system of order $m = d_x$ on $\mathcal{S}$, the Jacobian of $\mathbf{H}_i$ for any i in $\{1, ..., d_x\}$ has full rank on $\mathcal{S}$. Thus, taking $d_\xi = \mathfrak{r} + 1 = m = d_x$ a full Lipschitz triangular form of dimension d_x exists, i.e. we recover the result of Theorem 5.2.2.

Example 6.3.1 In Example 6.2.1, $\mathbf{H}_3$ is full rank on $\mathcal{S} \setminus \{x \in \mathbb{R}^3 \mid x_1 = 0 \text{ or } x_3 = 0\}$. Thus, according to Theorem 6.3.1, the only points where $\mathfrak{g}_3$ may not be Lipschitz, are the image of points where $x_1 = 0$ or $x_3 = 0$. Let us study more precisely what happens around those points. Take $x_a = (x_{1,a}, x_{2,a}, 0)$ in $\mathcal{S}$. If there existed a locally Lipschitz function $\mathfrak{g}_3$ verifying (6.5) around x_a, there would exist $\mathfrak{a} > 0$ such that for any $x_b = (x_{1,b}, x_{2,a}, x_{3,b})$ sufficiently close to x_a with $x_{1,b} \neq 0$, $|3x_{3,b}^2| \leq \mathfrak{a}|x_{3,b}^3|$, which we know is impossible. Therefore, there does not exist a function $\mathfrak{g}_3$ which is Lipschitz around the image of points where $x_3 = 0$. Let us now study what happens elsewhere, namely on $\tilde{\mathcal{S}} = \mathcal{S} \setminus \{x \in \mathbb{R}^3 \ | x_3 = 0\}$. It turns out that, on any compact set $\mathcal{C}$ of $\tilde{\mathcal{S}}$, there exists[4] $\mathfrak{a}$ such that we have for all (x_a, x_b) in $\mathcal{C}^2$,

$$|x_{3,a}^2 x_{1,a} - x_{3,b}^2 x_{1,b}| \leq \mathfrak{a}(|x_{1,a} - x_{1,b}| + |x_{3,a}^3 x_{1,a} - x_{3,b}^3 x_{1,b}|)$$

Therefore, the continuous function $\mathfrak{g}_3$ found earlier in Example 6.2.2 such that $\mathfrak{g}_3(\mathbf{H}_3(x)) = L_g L_f^2(x) = 3x_3^2 x_1$ on $\mathcal{S}$ (and thus on $\mathcal{C}$) verifies in fact

$$|\mathfrak{g}_3(\xi_a) - \mathfrak{g}_3(\xi_b)| \leq \mathfrak{a}|\xi_a - \xi_b|$$

on $\mathbf{H}_3(\mathcal{C})$ and can be extended to a Lipschitz function on $\mathbb{R}^3$ according to [McS34]. We conclude that although $\mathbf{H}_3$ does not have a full-rank Jacobian everywhere on $\mathcal{C}$ (singularities at $x_1 = 0$), it is possible to find a Lipschitz function $\mathfrak{g}_3$ solution to our problem on this set. ▲

6.3.2 A necessary condition

We have just seen that the condition in Theorem 6.3.1 that the Jacobian of $\mathbf{H}_i$ be full-rank, is sufficient but not necessary. In order to have locally Lipschitz functions $\mathfrak{g}_i$ satisfying (6.5), there must exist for all x a strictly positive number $\mathfrak{a}$ such that for all (x_a, x_b) in a neighborhood of x,

$$|L_g L_f^{i-1} h(x_a) - L_g L_f^{i-1} h(x_b)| \leq \mathfrak{a} \, |\mathbf{H}_i(x_a) - \mathbf{H}_i(x_b)| \ . \tag{6.8}$$

We have the following necessary condition :

Lemma 6.3.1.

Consider x in $\mathcal{S}$ such that (6.8) is satisfied in a neighborhood of x. Then, for any non zero vector v in $\mathbb{R}^{d_x}$, and any k in $\{1, \ldots, d_u\}$, we have :

$$\frac{\partial \mathbf{H}_i}{\partial x}(x)\, v = 0 \quad \Rightarrow \quad \frac{\partial L_{g_k} L_f^{i-1} h}{\partial x}(x)\, v = 0 \ . \tag{6.9}$$

[4] If $x_{1,a}$ and $x_{1,b}$ are both zero, the inequality is trivial. Suppose $|x_{1,a}| > |x_{1,b}|$ and denote $\rho = \frac{x_{1,b}}{x_{1,a}}$. If $\rho < 0$, we have directly $|x_{3,a}^2 - \rho\, x_{3,b}^2| \leq \max\{x_{3,a}^2, x_{3,b}^2\}|1 - \rho|$. If now $\rho > 0$, $x_{3,a}^2 - \rho\, x_{3,b}^2 = \frac{(x_{3,a}^3 - \rho^{\frac{3}{2}}\, x_{3,b}^3)(x_{3,a} + \sqrt{\rho} x_{3,b})}{x_{3,a}^2 + \sqrt{\rho} x_{3,a} x_{3,b} + \rho x_{3,b}^2}$ and thus $|x_{3,a}^2 - \rho\, x_{3,b}^2| \leq \frac{2\sqrt{2}}{\sqrt{x_{3,a}^2 + \rho x_{3,a}^2}} |x_{3,a}^3 - \rho^{\frac{2}{3}}\, x_{3,b}^3|$. Besides, $|x_{3,a}^3 - \rho^{\frac{3}{2}}\, x_{3,b}^3| = |x_{3,a}^3 - \rho x_{3,b}^3 + \rho(1 - \sqrt{\rho})\, x_{3,b}^3| \leq$ $|x_{3,a}^3 - \rho x_{3,b}^3| + \frac{\rho |x_{3,b}^3|}{1 + \sqrt{\rho}} |1 - \rho|$ which gives $\mathfrak{a}$ on compact sets.

Proof : Assume there exists a non-zero vector v in $\mathbb{R}^{d_x}$ such that $\frac{\partial \mathbf{H}_i}{\partial x}(x)\, v = 0$. Choose $r > 0$ such that Inequality (6.8) holds on $B_r(x)$, the ball centered at x and of radius r. Consider for any integer p the vector x_p in $B_r(x)$ defined by $x_p = x - \frac{1}{p}\frac{1}{|v|}v$. This gives a sequence converging to x when p tends to infinity. We have

$$0 \;\leq\; \frac{|L_{g_k} L_f^{i-1} h(x) - L_{g_k} L_f^{i-1} h(x_p)|}{|x - x_p|} \;\leq\; \mathfrak{a}\, \frac{|\mathbf{H}_i(x) - \mathbf{H}_i(x_p)|}{|x - x_p|} \tag{6.10}$$

But, $\frac{\mathbf{H}_i(x) - \mathbf{H}_i(x_p)}{|x - x_p|}$ tends to $\frac{\partial \mathbf{H}_i}{\partial x}(x)\, v$ which by assumption is 0 . Similarly $\frac{1}{|x - x_p|}(L_{g_k} L_f^{i-1} h(x) - L_{g_k} L_f^{i-1} h(x_p))$ tends to $\frac{\partial L_{g_k} L_f^{i-1} h}{\partial x}(x)\, v$ which is also 0 according to (6.10). ■

We conclude that when $\mathbf{H}_i$ does not have a full-rank Jacobian, it must satisfy condition (6.9) to allow the existence of locally Lipschitz triangular functions $\mathfrak{g}_i$. This condition is in fact about uniform infinitesimal observability.

Definition 6.3.1.

See [GK01, Definition I.2.1.3]. Consider the system lifted to the tangent bundle ([GK01, page 10])

$$\left\{ \begin{array}{rcl} \dot{x} &=& f(x) + g(x)u \\ \dot{v} &=& \left[\frac{\partial f}{\partial x}(x) + \frac{\partial g u}{\partial x}(x)\right] v \end{array} \right. , \quad \left\{ \begin{array}{rcl} y &=& h(x) \\ w &=& \frac{\partial h}{\partial x}(x) v \end{array} \right. \tag{6.11}$$

with v in $\mathbb{R}^{d_x}$ and w in $\mathbb{R}$ and the solutions of which are denoted $(X(x;t;u), V((x,v);t;u))$. System (6.1) is *uniformly instantaneously infinitesimally observable* on $\mathcal{S}$ if, for any pair (x,v) in $\mathcal{S} \times \mathbb{R}^{d_x} \setminus \{0\}$, any strictly positive number $\bar{t}$, and any C^1 function u defined on an interval $[0, \bar{t})$, there exists a time $t < \bar{t}$ such that $\frac{\partial h}{\partial x}(X(x;t;u))\, V((x,v);t;u) \neq 0$ and such that $X(x;s;u) \in \mathcal{S}$ for all $s \leq t$.

We have the following result.

Theorem 6.3.2.

Suppose that System (6.1) is strongly differentially observable of order m (or at least that $\mathbf{H}_m$ is an immersion on $\mathcal{S}$) and that Inequality (6.8) is verified at least locally around any point x in $\mathcal{S}$ for any i in $\{1, \ldots, m\}$. Then the system is uniformly infinitesimally observable on $\mathcal{S}$.

Proof : According to Lemma 6.3.1, we have (6.9). Now take x in $\mathcal{S}$ and a non-zero vector v and suppose that there exists $\bar{t} > 0$ such that for all t in $[0, \bar{t})$, $X(x;t;u)$ is in $\mathcal{S}$ and $w(t) = \frac{\partial h}{\partial x}(X(x;t;u))V((x,v);t;u) = 0$. To simplify the notations, we denote $X(t) = X(x;t;u)$ and $V(t) = V((x,v);t;u)$. For all integer i, we denote

$$w_i(t) = \frac{\partial L_f^{i-1} h}{\partial x}(X(t))V(t) \;.$$

We note that for any function $\psi : \mathbb{R}^n \to \mathbb{R}$, we have

$$\dot{\frac{\partial \psi}{\partial x}(X(t))V(t)} = \frac{\partial L_f \psi}{\partial x}(X(t))V(t) + \sum_{k=1}^{d_u} u_k \frac{\partial L_{g_k} \psi}{\partial x}(X(t))V(t) \;.$$

We deduce for all integer i and all t in $[0, \bar{t})$

$$\dot{w}_i(t) = w_{i+1}(t) + \sum_{k=1}^{d_u} u_k \frac{\partial L_{g_k} L_f^{i-1} h}{\partial x}(X(t))V(t) \;.$$

Let us show by induction that $w_i(t) = 0$ for all integer i and all t in $[0, \bar{t})$. It is true for $i = 1$ by assumption. Now, take an integer $i > 1$, and suppose $w_j(t) = 0$ for all t in $[0, \bar{t})$ and all $j \leq i$, i-e $\frac{\partial \mathbf{H}_i}{\partial x}(X(x;t;u))V((x,v);t;u) = 0$ for all $t < \bar{t}$. In particular, $\dot{w}_i(t) = 0$ for all $t < \bar{t}$. Besides,

according to (6.9), $\frac{\partial L_{g_k} L_f^{i-1} h}{\partial x}(X(x;t;u))V((x,v);t;u) = 0$ for all k in $\{1,\dots,d_u\}$ and for all $t < \bar{t}$. Thus, $w_{i+1}(t) = 0$ for all $t < \bar{t}$. We conclude that w_i is zero on $[0,\bar{t}[$ for all i and in particular at time 0, $\frac{\partial \mathbf{H}_m}{\partial x}(x)v = (w_1(0),\dots,w_m(0)) = 0$. But $\mathbf{H}_m$ is an immersion on $\mathcal{S}$, thus, necessarily $v = 0$ and we have a contradiction. ■

Example 6.3.2 We go on with Example 6.2.1. The linearization of the dynamics (6.6) yields

$$\left\{\begin{array}{rcl} \dot{v}_1 &=& v_2 \\ \dot{v}_2 &=& x_3^3 v_1 + 3x_3^2 x_1 v_3 \\ \dot{v}_3 &=& 0 \end{array}\right. , \quad w = v_1 \tag{6.12}$$

Consider $x_0 = (x_1, x_2, 0)$ in $\mathcal{S}$ and $v_0 = (0,0,v_3)$ with v_3 a nonzero real number. The solution to (6.6)-(6.12) initialized at (x_0, v_0) and with a constant input $u = -1$ is such that $X(x_0;t;u)$ remains in $\mathcal{S}$ in $[0,\bar{t})$ for some strictly positive $\bar{t}$ and $w(t) = 0$ for all t in $[0,\bar{t})$. Since v_0 is nonzero, System (6.6) is not uniformly instantaneously infinitesimally observable on $\mathcal{S}$. But, for System (6.6), $\mathbf{H}_7$ is an immersion on $\mathcal{S}$. We deduce from Theorem 6.3.2 that Inequality (6.8) is not satisfied for all i, i-e there does not exist Lipschitz triangular functions $\mathfrak{g}_i$ for all i on $\mathcal{S}$. This is consistent with the conclusion of Example 6.3.1. However, on $\tilde{\mathcal{S}}$, i-e when we remove the points where $x_3 = 0$, the system becomes uniformly instantaneously infinitesimally observable. Indeed, it can easily be checked that for x in $\tilde{\mathcal{S}}$, $w = \dot{w} = \ddot{w} = w^{(3)} = 0$, implies necessarily $v = 0$. Unfortunately, from our results, we cannot infer from this that the functions $\mathfrak{g}_i$ can be taken Lipschitz on $\tilde{\mathcal{S}}$. Nevertheless, the conclusion of Example 6.3.1 is that $\mathfrak{g}_3$ can be taken Lipschitz even around points with $x_1 = 0$. All this suggests a possible tighter link between uniform instantaneous infinitesimal observability and Lipschitzness of the triangular form. ▲

We conclude from this section that uniform instantaneous infinitesimal observability is required to have the Lipschitzness of the functions $\mathfrak{g}_i$ when they exist. However, we don't know if it is sufficient yet.

6.4 Back to Example 4.5 in Chapter 4

Consider the system

$$\left\{\begin{array}{rcl} \dot{x}_1 &=& x_2 \\ \dot{x}_2 &=& -x_1 + x_3^5 x_1 \\ \dot{x}_3 &=& -x_1 x_2 + u \end{array}\right. , \quad y = x_1 \,. \tag{6.13}$$

It would lead us too far from the main subject of this thesis to study here the solutions behavior of this system. We note however that, when u is zero, they evolve in the 2-dimensional surface $\{x \in \mathbb{R}^3 : 3x_1^2 + 3x_2^2 + x_3^6 = c^6\}$. The equilibrium $(0,0,x_3)$ being unstable at least for $c > 1$, we can hope for the existence of solutions remaining in the compact set

$$\mathcal{C}_{r,\epsilon} = \left\{x \in \mathbb{R}^3 : x_1^2 + x_2^2 \geq \epsilon \,,\ 3x_1^2 + 3x_2^2 + x_3^6 \leq r\right\}$$

for instance when u is a small periodic time function, except maybe for pairs of input u and initial condition (x_1,x_2,x_3) for which resonance could occur. An example is given in Figure 6.1.

On $\mathcal{S} = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 \neq 0\}$, and whatever u is, the knowledge of the function $t \mapsto y(t) = X_1(x,t)$ and therefore of its three first derivatives

$$\begin{array}{rcl} \dot{y} &=& x_2 \\ \ddot{y} &=& -x_1 + x_3^5 x_1 \\ \dddot{y} &=& -x_2 - 5x_3^4 x_1^2 x_2 + x_3^5 x_2 + 5x_3^4 x_1 u \end{array}$$

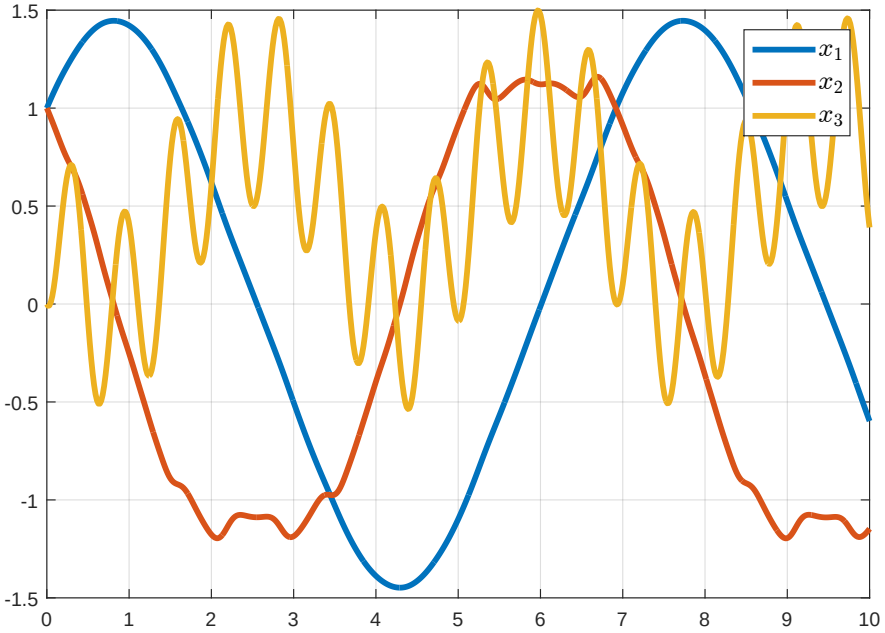


Figure 6.1: Solution of System (6.13) with initial condition $x = (1, 1, 0)$ and input $u = 5\sin(t/10)$.

gives us x_1, x_2 and x_3. Thus, System (6.13) is uniformly instantaneously observable on $\mathcal{S}$. Besides, the function

$$\mathbf{H}_4(x) = \begin{pmatrix} x_1 \\ x_2 \\ -x_1 + x_3^5 x_1 \\ -x_2 - 5x_3^4 x_1^2 x_2 + x_3^5 x_2 \end{pmatrix}$$

is injective on $\mathcal{S}$ and admits the following left inverse, defined on $\{\xi \in \mathbb{R}^4 : \xi_1^2 + \xi_2^2 \neq 0\}$:

$$\mathbf{H}_4^{-1}(\xi) = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \left(\frac{(\xi_3+\xi_1)\xi_1 + \left[(\xi_4+\xi_2) + 3|(\xi_3+\xi_1)\lfloor \xi_1 \rceil^{\frac{3}{2}}|^{\frac{4}{5}} \xi_2 \right] \xi_2}{\xi_1^2+\xi_2^2} \right)^{\frac{1}{5}} \end{pmatrix}$$

However, $\mathbf{H}_4$ is not an immersion because of a singularity of its Jacobian at $x_3 = 0$. So the drift system is weakly differentially observable of order 4 on $\mathcal{S}$ but not strongly. The reader may check that it can be transformed into the continuous triangular normal form of dimension 4 given by (4.27). The trajectory given in Figure 4.1 on which the observers presented in Section 4.5 have been tested are in fact the image by $\mathbf{H}_4$ of the solution plotted in Figure 6.1.

6.5 Conclusion

Like for strongly differentially observable systems of order d_x, uniform instantaneous observability of systems whose drift system is weakly differentially observable systems of order $m > d_x$, may still imply the existence of an at least up-to-d_x+1-triangular normal form (6.3) of dimension m. But

- we have shown this under the additional assumption that the functions $\mathbf{H}_i(x) = (h(x), L_f h(x), ..., L_f^{i-1} h(x)$ are open maps. Actually it is sufficient that the properties $\mathcal{B}(2), \ldots, \mathcal{B}(d_x + 1)$ hold.

- the functions in the triangular form are possibly non Lipschitz, but only close to points where the rank of the Jacobian of $\mathbf{H}_i$ changes. Anyhow, uniform infinitesimal observability is necessary to have Lipschitz functions.

- for a non Lipschitz triangular normal form, convergence of the regular high gain observer may be lost, but, as we saw in Chapter 4, it is still possible to design asymptotic observers.

Although our result only gives a partial triangular form and with additional assumptions $\mathcal{B}(i)$, we have no counter example showing that uniform instantaneous observability is not sufficient to have a full continuous triangular form. The crucial point would be to prove Lemma 6.2.2 under this weaker condition, which unfortunately we have not managed to do.

Chapter 7

Transformation into a Hurwitz form: nonlinear Luenberger observers

Chapitre 7 – Transformation dans une forme Hurwitz : observateurs de Luenberger non linéaires. *Dans ce chapitre, nous montrons comment la méthodologie de Luenberger s'applique à des systèmes non linéaires commandés, i-e nous étendons ce qui a été fait dans [AP06] pour les systèmes autonomes. Cette méthode consiste à transformer le système en une forme Hurwitz par la résolution d'une EDP. Si cette transformation est injective, un observateur s'ensuit immédiatement. Le problème se résume donc à l'existence (et au calcul) d'une solution injective à une EDP. Nous montrons entre autres que cette EDP admet toujours des solutions dépendant du temps dont l' injectivité est assurée si le système est fortement différentiellement observable à un certain ordre et que les trajectoires sont bornées. Lorsque le système est seulement distingable en temps rétrograde, nous montrons qu'au moins une des solutions est injective pour presque tout choix de la matrice Hurwitz. Nous illustrons comment ces solutions peuvent être calculées en pratique sur des exemples physiques. Enfin, nous ajoutons un résultat concernant la possibilité d'utiliser une transformation stationnaire malgré la présence d'entrées dans le cas d'un système uniformément instantanément observable.*

Contents

Consider a general system of the form

$$\dot{x} = f(x,u) \qquad , \qquad y = h(x,u) \tag{7.1}$$

where x is the state in $\mathbb{R}^{d_x}$, y the measurement in $\mathbb{R}^{d_y}$, f and h sufficiently many times differentiable functions and $u : [0, +\infty) \to \mathbb{R}^{d_u}$ in $\mathcal{U}$, the set of considered inputs. Recall that we denote $X(x,t;s;u)$ the value at time s of the solution to System (7.1) with input u, initialized at x at time t, and $Y(x,t;s;u)$ the corresponding output function at time s.

In this chapter, we investigate the possibility of transforming System (7.1) into a Hurwitz form[1]

$$\dot{\xi} = A\,\xi + B\,y \tag{7.2}$$

with A Hurwitz in $\mathbb{R}^{d_\xi \times d_\xi}$, B a vector in $\mathbb{R}^{d_\xi \times dy}$, for some strictly positive integer d_ξ, i-e for each u in $\mathcal{U}$, find a transformation[2] $T : \mathbb{R}^{d_x} \times [0, +\infty) \to \mathbb{R}^{d_\xi}$ such that for any x in $\mathcal{X}$ and any time t in $[0, +\infty)$,

$$\frac{\partial T}{\partial x}(x,t) f(x, u(t)) + \frac{\partial T}{\partial t}(x,t) = A\,T(x,t) + B\,h(x, u(t)) \ . \tag{7.3}$$

Indeed, since the Hurwitz form (7.2) admits a trivial observer made of a copy of its dynamics, according to Theorem 2.2.1, it is sufficient that T becomes injective uniformly in time and in space at least after a certain time to obtain an observer for System (7.1).

We have seen in Chapter 5 that this problem has been solved in [AP06] for autonomous systems. Our goal is to extend those results to controlled/time-varying systems. Exactly as we saw for the high gain design in Section 5.2.1, two paths are possible : either we keep the stationary transformation obtained for some constant value of u (for instance the drift system at $u \equiv 0$) and hope the additional terms due to the presence of u do not prevent convergence, or we take a time-varying transformation taking into account (implicitly or explicitly) the input u.

As far as we know, no result concerning this problem exists in the literature apart from [Eng05, Eng07] which follows and extends [KE03]. The idea pursued in [Eng05] belongs to the first path : the transformation is stationary and the input is seen as a disturbance which must be small enough. Although the construction is extended in a cunning fashion to a larger class of inputs, namely those which can be considered as output of a linear generator model with small external input, this approach remains theoretic and restrictive. On the other hand, in [Eng07], the author rather tries to use a time-varying transformation but its injectivity is proved only under the so-called "finite-complexity" assumption, originally introduced in [KE03] for autonomous systems. Unfortunately, this property is very restrictive and hard to check. Besides, no indication about the dimension d_ξ is given and the transformation cannot be computed online because it depends on the whole past trajectory of the output.

That is why, in this chapter, we endeavor to give results of existence and injectivity of the transformation under more reasonable observability assumptions and keeping in mind the practical implementation of this method. We start by exploring the path of a time-varying transformation in Section 7.1. We show that the existence of the transformation itself is not a problem. On the other hand, its injectivity can be ensured by observability assumptions, similar to those presented in [AP06] for autonomous systems. Then, in Section 7.2, we show on practical examples how an explicit expression for such a transformation can be computed. Finally, in Section 7.3, we prove that, similarly to Theorem 5.2.2 for a high gain design, uniformly observable input-affine systems whose drift system is strongly differentially observable of order d_x, admit a Luenberger-type observer built with a stationary transformation.

Notations

1. Since h (resp Y) takes values in $\mathbb{R}^{d_y}$, we denote h_i (resp Y_i) its ith-component.

2. For some integer m, which will be chosen later in the chapter, we consider the dynamic extension of order m introduced in Definition 5.2.1 and use the corresponding notations :

[1] We could have considered a more general Hurwitz form $\dot{\xi} = A\,\xi + B(y)$ with B any nonlinear function, but taking B linear is sufficient to obtain satisfactory results.

[2] The function T depends on u in $\mathcal{U}$ and we should write T_u as in Theorem 2.2.1. But we drop this too heavy notation in this chapter to ease the comprehension. What is important is that the target Hurwitz form (7.2), namely d_ξ, A and B, be the same for all u in $\mathcal{U}$.

$\overline{u}_m = (u, \dot{u}, ..., u^{(m)})$, $\overline{\nu}_m = (\nu_0, ..., \nu_m)$, $\bar{f}$ the extended vector field

$$\bar{f}(x, \overline{\nu}_m, u^{(m+1)}) = \left(f(x, \nu_0), \nu_1, \ \dots \ , \nu_m, u^{(m+1)}\right)$$

and the extended measurement function

$$\bar{h}_i(x, \overline{\nu}_m) = h_i(x, \nu_0) \ .$$

We recall the reader that while $\overline{\nu}_m$ is an element $\mathbb{R}^{d_u(m+1)}$, $\overline{u}_m$ is a function defined on $[0, +\infty)$ such that $\overline{u}_m(s) = (u(s), \dot{u}(s), ..., u^{(m)}(s))$ is in $\overline{U}_m \subset \mathbb{R}^{d_u(m+1)}$ for all s in $[0, +\infty)$. For $1 \leq i \leq d_y$, the successive time derivatives of Y_i are related to the Lie derivatives of $\bar{h}_i$ along the vector fields $\bar{f}$, namely for $j \leq m$

$$\frac{\partial^j Y_i}{\partial s^j}(x, t; s; u) = L^j_{\bar{f}} \bar{h}_i(X(x, t; s; u), \overline{u}_m(s)) \ .$$

7.1 Time-varying transformation

The existence of a C^1 time-varying solution to PDE (7.3) is achieved thanks to the following lemma :

Lemma 7.1.1.

Consider d_ξ a strictly positive number, A a Hurwitz matrix in $\mathbb{R}^{d_\xi * d_\xi}$, B a matrix in $\mathbb{R}^{d_\xi \times d_y}$, and u an input function in $\mathcal{U}$. The function T^0 defined on $\mathcal{S} \times [0, +\infty)$ by

$$T^0(x, t) = \int_0^t e^{A(t-s)} B \, Y(x, t; s; u) \, ds \tag{7.4}$$

is a C^1 solution to PDE (7.3).

Proof : First, for any u in $\mathcal{U}$, and any s in $[0, +\infty)$, $(x, t) \mapsto Y(x, t; s; u) = h(X(x, t; s; u), s)$ is C^1, thus T^0 is C^1. Take x in $\mathcal{S}$ and t in $[0, +\infty)$. For any τ in $\mathbb{R}$,

$$X(X(x, t, t+\tau; u), t+\tau; s; u) = X(x, t; s; u).$$

Therefore,

$$\begin{aligned} T^0(X(x, t; t+\tau; u), t+\tau) &= \int_0^{t+\tau} e^{A(t+\tau-s)} B h(X(x, t, s; u), u(s)) ds \\ &= e^{A\tau} \, T^0(x, t) + e^{A\tau} \int_t^{t+\tau} e^{A(t-s)} B \, h(X(x, t; s; u), u(s)) ds \end{aligned}$$

and

$$\begin{aligned} \frac{T^0(X(x, t; t+\tau; u), t+\tau) - T^0(x, t)}{\tau} &= \frac{e^{A\tau} - I}{\tau} T^0(x, t) \\ &\quad + \frac{e^{A\tau}}{\tau} \int_t^{t+\tau} e^{A(t-s)} B \, h(X(x, t, s; u), u(s)) ds \ . \end{aligned}$$

Making τ tend to 0, we get PDE (7.3). ■

Note that extending directly what is done in [KE03, AP06] would rather lead us to the solution

$$T^\infty(x, t) = \int_{-\infty}^t e^{A(t-s)} B \, Y(x, t; s; u) \, ds \ .$$

The drawback is that some assumptions about the growth of Y have to be made to ensure its continuity, unless Y is bounded backward in time. As for the C^1 property, and even if the solutions are bounded backward in time, it is achieved only if the eigenvalues of A are

sufficiently negative. In fact, it is not absolutely needed that the solution be C^1, one could look for continuous solutions to

$$L_{(f,1)}T(x,t) = AT(x,t) + Bh(x,u(t))$$

as defined in Theorem 2.2.1 instead of PDE (7.3). The major disadvantage of this solution is rather that T^∞ is not easily computable since it depends on the values of u on $(-\infty, t]$. Nevertheless, it may still be useful. For example, that is the solution chosen in [PPO08] for the specific application of a permanent synchronous motor, where it is proved to be injective.

Unlike T^∞, T^0 depends only on the values of the input u on $[0,t]$. Therefore, it is theoretically computable online. However, for each couple (x,t), one would need to integrate backwards the dynamics (7.1) until time 0, which is quite heavy. If the input u is known in advance (for instance $u(t) = t$) it can also be computed offline. We will see in Section 7.2 on practical examples how we can find a solution to PDE (7.3) in practice, without relying on the expression T^0.

We conclude that a C^1 time-varying transformation into a Hurwitz form always exists, but the core of the problem is to ensure its injectivity.

7.1.1 Injectivity with strong differential observability

Assumptions

There exists a subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$ such that :

1. For any u in $\mathcal{U}$, any x in $\mathcal{S}$ and any time t in $[0,+\infty)$, $X(x,t;s;u)$ is in $\mathcal{S}$ for all s in $[0,+\infty)$.

2. The quantity
$$M_f = \sup_{\substack{x \in \mathcal{S} \\ \nu_0 \in U}} \left| \frac{\partial f}{\partial x}(x, \nu_0) \right|$$
is finite.

3. There exist d_y integers $(m_1, \dots, m_{d_y})$ such that the functions
$$H_i(x, \overline{\nu}_m) = \left(\bar{h}_i(x, \overline{\nu}_m), L_{\bar{f}} \bar{h}_i(x, \overline{\nu}_m), \dots, L_{\bar{f}}^{m_i - 1} \bar{h}_i(x, \overline{\nu}_m) \right) \tag{7.5}$$
defined on $\mathcal{S} \times \mathbb{R}^{d_u(m+1)}$ with $m = \max_i m_i$ and $1 \leq i \leq d_y$ verify :
   - for all u in $\mathcal{U}$, $H_i(\cdot, \overline{u}_m(0)))$ is Lipschitz on $\mathcal{S}$.
   - there exists L_H such that the function
   $$H(x, \overline{\nu}_m) = \left(H_1(x, \overline{\nu}_m), \dots, H_i(x, \overline{\nu}_m), \dots, H_{d_y}(x, \overline{\nu}_m) \right) \tag{7.6}$$
   verifies for any (x_1, x_2) in $\mathcal{S}^2$ and any $\overline{\nu}_m$ in $\overline{U}_m$
   $$|x_1 - x_2| \leq L_H |H(x_1, \overline{\nu}_m) - H(x_2, \overline{\nu}_m)|$$
   namely H is Lipschitz-injective on $\mathcal{S}$, uniformly with respect to $\overline{\nu}_m$ in $\overline{U}_m$.

4. For all $1 \leq i \leq d_y$, there exists L_i such that for all (x_1, x_2) in $\mathcal{S}^2$ and for all $\overline{\nu}_m$ in $\overline{U}_m$,
$$|L_{\bar{f}}^{m_i} \bar{h}_i(x_1, \overline{\nu}_m) - L_{\bar{f}}^{m_i} \bar{h}_i(x_2, \overline{\nu}_m)| \leq L_i |x_1 - x_2|$$
namely $L_{\bar{f}}^{m_i} \bar{h}_i(\cdot, \overline{\nu}_m)$ is Lipschitz on $\mathcal{S}$, uniformly with respect to $\overline{\nu}_m$ in $\overline{U}_m$.

We have the following result.

Theorem 7.1.1.

Suppose Assumptions 1-2-3-4 are satisfied. Consider Hurwitz matrices A_i in $\mathbb{R}^{m_i * m_i}$, with m_i defined in Assumption 3, and vectors B_i in $\mathbb{R}^{m_i}$ such that the pairs (A_i, B_i) are controllable. There exists a strictly positive real number $\bar{k}$ such that for all $k \geq \bar{k}$, for all input u in $\mathcal{U}$, there exists $\bar{t}_{k,u}$ such that any C^1 solution T to PDE (7.3) on $\mathcal{S} \times [0, +\infty)$ with

- $d_\xi = \sum_{i=1}^{d_y} m_i$
- A in $\mathbb{R}^{d_\xi \times d_\xi}$ and B in $\mathbb{R}^{d_\xi \times d_y}$ defined by

$$A = \begin{pmatrix} kA_1 & & & & \\ & \ddots & & & \\ & & kA_i & & \\ & & & \ddots & \\ & & & & kA_{d_y} \end{pmatrix} \qquad B = \begin{pmatrix} B_1 & & & & \\ & \ddots & & & \\ & & B_i & & \\ & & & \ddots & \\ & & & & B_{d_y} \end{pmatrix}$$

- $T(\cdot, 0)$ Lipschitz on $\mathcal{S}$

is such that $T(\cdot, t)$ is injective on $\mathcal{S}$ for all $t \geq \bar{t}_{k,u}$, uniformly in time and in space. More precisely, there exists a constant L_k such that for any (x_1, x_2) in $\mathcal{S}^2$, any u in $\mathcal{U}$ and any time $t \geq \bar{t}_{k,u}$

$$|x_1 - x_2| \leq L_k \, |T(x_1, t) - T(x_2, t)| \ .$$

Besides, for any $t \geq \bar{t}_{k,u}$, $T(\cdot, t)$ is an injective immersion on $\mathcal{S}$.

Note that the additional assumption "$T(\cdot, 0)$ Lipschitz on $\mathcal{S}$" is not very restrictive because the solution T can usually be chosen arbitrarily at initial time 0 (see examples in Section 7.2). In particular, the elementary solution T^0 found in Lemma 1 is zero at time 0 and thus clearly verifies this assumption.

Proof : Given the form of the matrices A and B, we have

$$T(x,t) = \big(T_1(x,t), \ldots, T_i(x,t), \ldots, T_{d_y}(x,t)\big) \tag{7.7}$$

with

$$\frac{\partial T_i}{\partial x}(x,t) f(x,u(t)) + \frac{\partial T_i}{\partial t}(x,t) = kA_i \, T_i(x,t) + B_i \, h_i(x,u(t)) \ . \tag{7.8}$$

Take u in $\mathcal{U}$, i in $\{1, \ldots, d_y\}$, x in $\mathcal{S}$ and t in $[0, +\infty)$. According to PDE (7.8), T_i satisfies for all s in $[0, +\infty)$,

$$\frac{d}{ds} T_i(X(x,t;s;u),s) = kA_i \, T_i(X(x,t;s;u),s) + B_i Y_i(x,t;s;u) \ .$$

Integrating between t and s, it follows that

$$T_i(X(x,t;s;u),s) = e^{kA_i(s-t)} \underbrace{T_i(X(x,t;t;u),t)}_{T_i(x,t)} + \int_t^s e^{kA_i(s-\tau)} B_i \, Y_i(x,t;\tau;u) d\tau$$

and thus,

$$T_i(x,t) = e^{kA_i(t-s)} \, T_i(X(x,t;s;u),s) + \int_s^t e^{kA_i(t-\tau)} B Y_i(x,t;\tau;u) d\tau \ .$$

applying this inequality at $s = 0$, we get

$$T_i(x,t) = e^{kA_i t} \, T_i(X(x,t;0;u),0) + T_i^0(x,t)$$

where T_i^0 is such that T^0 defined in (7.4) is

$$T^0(x,t) = \big(T_1^0(x,t), \ldots, T_i^0(x,t), \ldots, T_{d_y}^0(x,t)\big) \ .$$

But after m_i successive integration by parts in (7.4), we get,

$$T_i^0(x,t) = -A_i^{-m_i}\mathscr{C}_i\, K_i\, H_i(x,\overline{u}_m(t)) \\ +A_i^{-m_i}e^{kA_it}\mathscr{C}_iK_i\, H_i(X(x,t;0;u),\overline{u}_m(0)) + \frac{1}{k^{m_i}}A_i^{-m_i}R_i(x,t)$$

where $K_i = \mathrm{diag}\left(\frac{1}{k}\ ,\ \dots\ ,\ \frac{1}{k^{m_i}}\right)$, $\mathscr{C}_i$ is the invertible controllability matrix

$$\mathscr{C}_i = [A_i^{m_i-1}B_i\ ,\ \dots\ ,\ A_iB_i\ ,\ B_i]\ ,$$

$H_i(x,\overline{\nu}_m)$ is defined in (7.5), and R_i is the remainder :

$$R_i(x,t) = \int_0^t e^{kA_i(t-\tau)}B_i\, L_{\bar{f}}^{m_i}h_i\left(X(x,t;\tau;u),\overline{u}_m(\tau)\right)d\tau$$

We finally deduce that

$$T_i(x,t) = A_i^{-m_i}\,\mathscr{C}_i\,K_i\left(-H_i(x,\overline{u}_m(t)) + K_i^{-1}\mathscr{C}_i^{-1}\left(e^{kA_it}\,\Psi_i(X(x,t;0;u),0) + \frac{1}{k^{m_i}}R_i(x,t)\right)\right)$$

with $\Psi_i(x,t) = A_i^{m_i}T_i(x,t) + \mathscr{C}_iK_i\, H_i(x,\overline{u}_m(t))$.

Let us now consider x_1 and x_2 in $\mathcal{S}$, and t in $[0,+\infty)$. We are interested in the quantity $|T(x_1,t)-T(x_2,t)|$, and thus in $|T_i(x_1,t)-T_i(x_2,t)|$.

Thanks to Assumption 2, for any (x_1,x_2) in $\mathcal{S}$, and (t,τ) in $[0,+\infty)^2$, we have (see for instance [RM82])

$$|X(x_1,t;\tau;u) - X(x_2,t;\tau;u)| \le e^{M_f|\tau-t|}|x_1-x_2|\ . \tag{7.9}$$

By assumption $T_i(\cdot,0)$ and $H_i(\cdot,\overline{u}_m(0)))$ are Lipschitz on $\mathcal{S}$, thus there exists L_{Ψ_i} such that

$$|\Psi_i(X(x_1,t;0;u),0) - \Psi_i(X(x_2,t;0;u),0)| \le L_{\Psi_i}\,e^{M_ft}|x_1-x_2|\ .$$

Then, A_i being Hurwitz, there exists strictly positive numbers a_i and γ_i (see [RM82]) such that for all τ in $[0,t]$

$$\left|e^{kA_i(t-s)}\right| \le \gamma_i\, e^{-ka_i(t-s)}\ . \tag{7.10}$$

Using Assumption 4 and inequalities (7.9) and (7.10), we deduce that if $k > \dfrac{M_f}{a_i}$,

$$|R_i(x_1,t) - R_i(x_2,t)| \le L_i|B_i|\gamma_i\int_0^t e^{-(ka_i-M_f)(t-\tau)}d\tau\,|x_1-x_2| \le \frac{L_i|B_i|\gamma_i}{ka_i-M_f}\,|x_1-x_2|\,.$$

We finally deduce that

$$\begin{aligned}|T_i(x_1,t) - T_i(x_2,t)| \ge& |A_i^{-m_i}\,\mathscr{C}_i\,K_i|\,\left(|\Delta H_i| - |K_i^{-1}\mathscr{C}_i^{-1}|\left(\left|e^{kA_it}\right|\,|\Delta\Psi_i| + \frac{1}{k^{m_i}}|\Delta R_i|\right)\right)\\ \ge& \frac{|A_i^{-m_i}\,\mathscr{C}_i|}{k^{m_i}}\Big(\,|\Delta H_i| - k^{m_i}|\mathscr{C}_i^{-1}|\,\gamma_i\,L_{\Psi_i}\,e^{-(ka_i-M_f)t}\,|x_1-x_2|\\ &-|\mathscr{C}_i^{-1}|\,\gamma_i\frac{L_i|B_i|}{ka_i-M_f}\,|x_1-x_2|\Big)\end{aligned}$$

where ΔH_i, $\Delta\Psi_i$ and ΔR_i denote the difference of the functions $H_i(\cdot,\overline{u}_m(t))$, $\Psi_i(X(\cdot,t;0;u),0)$ and $R_i(\cdot,t)$ respectively, evaluated at x_1 and x_2. It follows (by norm equivalence), that there exists a constant c such that

$$\begin{aligned}|T(x_1,t) - T(x_2,t)| \ge& c\frac{\min_i(|A_i^{-m_i}\,\mathscr{C}_i|)}{k^m}\Bigg[\,|H(x_1,\overline{u}_m(t)) - H(x_2,\overline{u}_m(t))|\\ &-\left(\left(\sum_{i=1}^p k^{m_i}\gamma_i|\mathscr{C}_i^{-1}|\,L_{\Psi_i}\right)e^{-(ka-M_f)t} + \frac{\left(\sum_{i=1}^p L_i\gamma_i|\mathscr{C}_i^{-1}||B_i|\right)}{ka-M_f}\right)|x_1-x_2|\Bigg]\\ \ge& c\frac{\min_i(|A_i^{-m_i}\,\mathscr{C}_i|)}{k^m}\left(\frac{1}{L_H} - c_1\,k^m\,e^{-(ka-M_f)t} - c_2\frac{1}{ka-M_f}\right)|x_1-x_2|\end{aligned}$$

where m a, c_1, c_2 are constants independent from k and t defined by

$$m = \max_i m_i\quad,\quad a = \min_i a_i\quad,\quad c_1 = \sum_{i=1}^p\gamma_i|\mathscr{C}_i^{-1}|\,L_{\Psi_i}\quad,\quad c_2 = \sum_{i=1}^p L_i\gamma_i|\mathscr{C}_i^{-1}||B_i|\ .$$

We deduce that for

$$k \ge \frac{1}{a}\left(M_f + 4c_2L_H\right)\quad,\quad t \ge \frac{\ln(4k^mc_1L_H)}{ka-M_f}\ ,$$

we have

$$|T(x_1,t)-T(x_2,t)| \geq c\frac{\min_i(|A_i^{-m_i}\mathscr{C}_i|)}{k^m}\frac{1}{2L_H}|x_1-x_2|$$

i-e

$$|x_1-x_2| \leq 2L_H\frac{k^m}{c\min_i(|A_i^{-m_i}\mathscr{C}_i|)}|T(x_1,t)-T(x_2,t)| \tag{7.11}$$

and $T(\cdot,t)$ is injective on $\mathcal{S}$, uniformly in time. We conclude that the result holds with

$$\overline{k}=\frac{1}{a}(M_f+4c_2L_H) \quad , \quad L_k=2L_H\frac{k^m}{c\min_i(|A_i^{-m_i}\mathscr{C}_i|)} \quad , \quad \overline{t}_{k,u}=\max\left\{\frac{\ln(4k^mc_1L_H)}{ka-M_f},0\right\}$$

Since M_f, L_H and L_i (and thus c_2) are independent from u, $\overline{k}$ and L_k are the same for all u in $\mathcal{U}$, while $\overline{t}_{k,u}$ depends on u through L_{Ψ_i}.

Now, take any x in $\mathcal{S}$ and $t \geq \overline{t}_{k,u}$. For any v and any h such that $x+hv$ is in $\mathcal{S}$, we have

$$L_k|v| \leq \frac{|T(x+hv,t)-T(x,t)|}{|h|}$$

and by letting h go to 0, we get

$$L_k|v| \leq \left|\frac{\partial T}{\partial x}(x,t)v\right| .$$

Hence, $T(\cdot,t)$ is an immersion on $\mathcal{S}$. ■

Applying successively Lemma 7.1.1, Theorem 7.1.1 and Theorem 2.2.1, we conclude that under Assumptions 2, 3, 4, it is possible to write an observer for system (7.1) by choosing any (A_i,B_i) controllable and k sufficiently large.

Remark 6 It is important to note that $\overline{k}$ does not dependent on u, thanks to the fact that L_H, M_f and L_i given by Assumptions 2-3-4 are the same for all $\overline{\nu}_m$ in $\overline{U}_m$. However, the time $\overline{t}_{k,u}$ after which the solution becomes injective a priori depends on k and u. This is not a problem in practice since we only want to be sure that for k sufficiently large, any solution will become injective after a certain time. If we want this time $\overline{t}_{k,u}$ to be uniform in u, the Lipschitz constants of $H_i(\cdot,\overline{u}_m(0)))$ and of $T(\cdot,0)$ must be the same for all u in $\mathcal{U}$.

Remark 7 If we choose $m=\max_i m_i$ sufficiently large distinct strictly positive real numbers λ_j, and take $A_i=-\mathsf{diag}(\lambda_1,\ldots,\lambda_{m_i})$ and $B_i=(1,\ldots,1)^\top$, then, the PDEs to solve are simply

$$\frac{\partial T_{\lambda,i}}{\partial x}(x,t)f(x,u(t))+\frac{\partial T_{\lambda,i}}{\partial t}(x,t)=-\lambda T_{\lambda,i}(x,t)+h_i(x,u(t)) \tag{7.12}$$

for each $1 \leq i \leq d_y$ and λ in $\{\lambda_1,\ldots,\lambda_{m_i}\}$. Then, one take

$$T(x,t)=\left(T_{\lambda_1,1},\ldots,T_{\lambda_{m_1},1},\ldots,T_{\lambda_1,d_y},\ldots,T_{\lambda_{m_{d_y}},d_y}\right) .$$

Remark 8 Under Assumption 3-4, the system could also be transformed into a Lipschitz phase-variable form of dimension $d_y \times \max_i m_i \geq \sum_{i=1}^{d_y} m_i$ according to Theorem 5.2.1 and a high gain observer could be used. If we wanted to use only m_i derivatives for each input and obtain an observer of same dimension $\sum_{i=1}^{d_y} m_i$, each $L_f^{m_i}\overline{h}$ would have to satisfy an additional triangularity assumption. But in any case, the crucial difference with the Luenberger observer presented in this chapter is that the latter does not require the computation of the derivatives of the input (see examples in Section 7.2).

In order to check the assumptions of Theorem 7.1.1 more easily in practical cases, we have the following result :

Lemma 7.1.2.

Assume that $\mathcal{S}$ is compact and there exist d_y integers $(m_1, ..., m_{d_y})$ such that $\overline{U}_m$ with $m = \max_i m_i$ is compact and for any $\overline{\nu}_m$ in $\overline{U}_m$, $H(\cdot, \overline{\nu}_m)$ defined in (7.6) is an injective immersion[3] on $\mathcal{S}$. Then, Assumptions 2, 3, 4 are satisfied.

In other words, since the additional assumption "$T(\cdot, 0)$ Lipschitz on $\mathcal{S}$" made in Theorem 7.1.1 is automatically verified when $\mathcal{S}$ is compact, the result of Theorem 7.1.1 holds under the only assumptions of Lemma 7.1.2 if $\mathcal{S}$ satisfies Assumption 1.

Proof : First, $\mathcal{S}$ and $\overline{U}_m$ being compact, Assumptions 2 and 4 are satisfied. Besides, $H_i(\cdot, \overline{u}_m(0))$ is clearly Lipschitz on $\mathcal{S}$. The only thing to prove is the uniform Lipschitz-injectivity of H, which follows directly from Lemma A.3.5. ■

7.1.2 Injectivity with backward distinguishability ?

In the previous section, we have shown that finding an injective transformation into an Hurwitz form was possible under a strong differential observability property, namely that the function made of each output and a certain number of its derivatives was an injective immersion. We investigate in this section if injectivity is still ensured when we have only a weak differential observability or even only backward-distinguishability as in [AP06, Theorem 3] for autonomous systems (recalled in Section 5.1.2).

Theorem 7.1.2.

Take u in $\mathcal{U}$. Assume that for this input, System (7.1) is backward-distinguishable in time $\bar{t}_u$ on $\mathcal{S}$, i-e for any $t \geq \bar{t}_u$ and any (x_a, x_b) in $\mathcal{S}^2$,

$$Y(x_a, t; s; u) = Y(x_b, t; s; u) \quad \forall s \in [t - \bar{t}_u, t] \quad \Longrightarrow \quad x_a = x_b \ .$$

There exists a set $\mathcal{R}$ of zero-Lebesgue measure in $\mathbb{C}^{d_x+1}$ such that for any $(\lambda_1, \ldots, \lambda_{d_x+1})$ in $\Omega^{d_x+1} \setminus \mathcal{R}$ with $\Omega = \{\lambda \in \mathbb{C}\,,\, \Re(\lambda) < 0\}$, and any $t \geq \bar{t}_u$, the function T^0 defined in (7.4) with

- $d_\xi = d_y \times (d_x + 1)$
- A in $\mathbb{R}^{d_\xi \times d_\xi}$ and B in $\mathbb{R}^{d_\xi \times d_y}$ defined by

$$A = \begin{pmatrix} \tilde{A} & & & \\ & \ddots & & \\ & & \tilde{A} & \\ & & & \ddots \\ & & & & \tilde{A} \end{pmatrix} \quad , \quad B = \begin{pmatrix} \tilde{B} & & & \\ & \ddots & & \\ & & \tilde{B} & \\ & & & \ddots \\ & & & & \tilde{B} \end{pmatrix}$$

and

$$\tilde{A} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{d_x+1} \end{pmatrix} \quad , \quad \tilde{B} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} .$$

is such that $T^0(\cdot, t)$ is injective on $\mathcal{S}$ for $t > \bar{t}_u$.

Proof : Let us define for λ in $\mathbb{C}$, the function $T^0_\lambda : \mathcal{S} \times \mathbb{R}^+ \to \mathbb{C}^{d_y}$

$$T^0_\lambda(x, t) = \int_0^t e^{-\lambda(t-s)} Y(x, t; s; u)\, ds \ . \tag{7.13}$$

[3] $H(\cdot, \overline{\nu}_m)$ is injective on $\mathcal{S}$ and $\frac{\partial H}{\partial x}(x, \overline{\nu}_m)$ is full-rank for any x in $\mathcal{S}$

Given the structure of A and B, and with a permutations of the components,

$$T^0(x,t) = \left(T^0_{\lambda_1}(x,t), \ldots, T^0_{\lambda_{d_x+1}}(x,t)\right) \ .$$

We need to prove that T^0 is injective for almost all $(\lambda_1, \ldots, \lambda_{d_x+1})$ in Ω^{d_x+1} (in the sense of the Lebesgue measure). For that, we define the function

$$\Delta T(x_a, x_b, t, \lambda) = T^0_\lambda(x_a, t) - T^0_\lambda(x_b, t)$$

on $\Upsilon \times \Omega$ with

$$\Upsilon = \{(x_a, x_b, t) \in \mathcal{S}^2 \times (\bar{t}_u, +\infty) \ : \ x_a \neq x_b\} \ .$$

We are going to use the following lemma whose proof[4] can be found in [AP06]:

Lemma 7.1.3. Coron's lemma

Let Ω and Υ be open sets of $\mathbb{C}$ and $\mathbb{R}^{2d_x+1}$ respectively. Let $\Delta T : \Upsilon \times \Omega \to \mathbb{C}^{d_y}$ be a function which is holomorphic in λ for all $\underline{x}$ in Υ and C^1 in $\underline{x}$ for all λ in Ω. If for any $(\underline{x}, \lambda)$ in $\Upsilon \times \Omega$ such that $\Delta T(\underline{x}, \lambda) = 0$, there exists i in $\{1, \ldots, d_y\}$ and $k > 0$ such that $\dfrac{\partial^k \Delta T_i}{\partial \lambda^k}(\underline{x}, \lambda) \neq 0$, then the set

$$\mathcal{R} = \bigcup_{\underline{x} \in \Upsilon} \left\{(\lambda_1, \ldots, \lambda_{d_x+1}) \in \Omega^{d_x+1} \ : \ \Delta T(\underline{x}, \lambda_1) = \ldots = \Delta T(\underline{x}, \lambda_{d_x+1}) = 0\right\}$$

has zero Lebesgue measure in $\mathbb{C}^{d_x+1}$.

In our case, ΔT is clearly holomorphic in λ and C^1 in $\underline{x}$. Since for every $\underline{x}$ in Υ, $\lambda \mapsto \Delta T(\underline{x}, \lambda)$ is holomorphic on the connex set $\mathbb{C}$, its zeros are isolated and admit a finite multiplicity, unless it is identically zero on $\mathbb{C}$. In the latter case, we have in particular for any ω in $\mathbb{R}$

$$\int_{-\infty}^{+\infty} e^{-i\omega\tau} g(\tau)\, d\tau = 0$$

with g the function

$$g(\tau) = \begin{cases} Y(x_a, t; t-\tau; u) - Y(x_b, t; t-\tau; u) & , \quad \text{if } \tau \in [0, t] \\ 0 & , \quad \text{otherwise} \end{cases}$$

which is in $\mathcal{L}^2$. Thus, the Fourier transform of g is identically zero and we deduce that necessarily

$$Y(x_a, t; t-\tau; u) - Y(x_b, t; t-\tau; u) = 0$$

for almost all τ in $[0, t]$ and thus for all τ in $[0, t]$ by continuity. Since $t \geq \bar{t}_u$, it follows from the backward-distinguishability that $x_a = x_b$ but this is impossible because (x_a, x_b, t) is in Υ. We conclude that $\lambda \mapsto \Delta T(\underline{x}, \lambda)$ is not identically zero on $\mathbb{C}$ and the assumptions of the lemma are satisfied. Thus, $\mathcal{R}$ has zero measure and for all $(\lambda_1, \ldots, \lambda_{d_x+1})$ in $\mathbb{C}^{d_x+1} \setminus \mathcal{R}$, T^0 is injective on $\mathcal{S}$, by definition of $\mathcal{R}$. ■

Remark 9 The function T proposed by Theorem 7.1.2 takes complex values. To remain in the real frame, one should consider the transformation made of its real and imaginary parts, and instead of implementing for each i in $\{1, \ldots, d_y\}$ and each lambda

$$\dot{\hat{\xi}}_{\lambda,i} = -\lambda \hat{\xi}_\lambda + y_i$$

in $\mathbb{C}$, one should implement

$$\dot{\hat{\xi}}_{\lambda,i} = \begin{pmatrix} -\Re(\lambda) & -\Im(\lambda) \\ \Im(\lambda) & -\Re(\lambda) \end{pmatrix} \hat{\xi}_{\lambda,i} + \begin{pmatrix} y_i \\ 0 \end{pmatrix}$$

in $\mathbb{R}$. Thus, the dimension of the observer is $2 \times d_y \times (d_x + 1)$ in terms of real variables.

[4] More precisely, the result proved in [AP06] is for Υ open set of $\mathbb{R}^{2d_x}$ instead of $\mathbb{R}^{2d_x+1}$. But the proof turns out to be still valid with $\mathbb{R}^{2d_x+1}$ because the only constraint is that the dimension of Υ be strictly less than $2(d_\xi + 1)$.

Remark 10 It should be noted that Theorem 7.1.2 gives for each u in $\mathcal{U}$ a set $\mathcal{R}_u$ of zero measure in which not to choose the λ_i, but unfortunately, there is no guarantee that $\bigcup_{u\in\mathcal{U}} \mathcal{R}_u$ is also of zero-Lebesgue measure.

Remark 11 Unlike Theorem 7.1.1 which proved the injectivity of any solution T to PDE (7.3), Theorem 7.1.2 proves only the injectivity of T^0. Note though that as shown at the beginning of the proof of Theorem 7.1.1, any solution T writes

$$T(x,t) = e^{At}\, T(X(x,t;0;u),0) + T^0(x,t)$$

with A Hurwitz, and thus tends to the injective function T^0. We can thus expect T to become injective after a certain time. In fact, a way of ensuring the injectivity is to take, if possible, a solution T with the boundary condition

$$T(x,0) = 0 \quad \forall x \in \mathcal{S}\ ,$$

because in that case, necessarily, $T = T^0$.

We conclude from this section that there always exists a time-varying solution to PDE (7.3) which is injective under appropriate observability assumptions. It follows that the only remaining problem to address is the computation of such a solution without relying on the expression (7.4). This is done in the following section through practical examples.

7.2 Examples

7.2.1 Permanent Magnet Synchronous Motor (PMSM)

A first practical example which falls directly into the scope of this paper is the Luenberger observer presented in [HMP12] for a PMSM. We reproduce here the minimal information needed for comprehension, and we add the theoretical arguments which are not given in [HMP12]. The system can be modeled by

$$\dot{x} = u - Ri \quad , \quad y = |x - Li|^2 - \Phi^2 = 0 \tag{7.14}$$

where x is in $\mathbb{R}^2$, the voltages u and currents i are inputs in $\mathbb{R}^2$, the resistance R, impedance L and flux Φ are known scalar parameters and the measurement y is constantly zero. Here $d_y = 1$, so we can drop the subscript i. Since the dynamics are linear and the measurement quadratic in x, one can look for T_λ of the form :

$$T_\lambda(x,t) = |x|^2 + a_\lambda(t)^\top x + b_\lambda(t)$$

where the dynamics of a_λ and b_λ are to be chosen so that T_λ is solution of PDE (7.12). We can check that the dynamics

$$\begin{aligned} \dot{a}_\lambda &= -\lambda\, a_\lambda - 2(u - Ri) + 2Li \\ \dot{b}_\lambda &= -\lambda\, b_\lambda - a_\lambda^\top (u - Ri) + L^2|i|^2 - \Phi^2 \end{aligned} \tag{7.15}$$

make T_λ follow the dynamics

$$\dot{\xi}_\lambda = -\lambda \xi_\lambda + y = -\lambda \xi_\lambda$$

and a trivial solution is thus $\xi_\lambda = 0$. Let us now check whether the assumptions of Theorem 7.1.1 are verified. We suppose that i, $\dot{\widehat{i}}$, $\ddot{\widehat{i}}$ and u, $\dot{u}$ are bounded, so that the state x also remains bounded (since $y = 0$). Choosing $m = 3$, we have

$$H(x,u,i,\dot{u},\dot{\widehat{i}},\ddot{\widehat{i}}) = \begin{pmatrix} |x - Li|^2 - \Phi^2 \\ 2\eta^\top (x - Li) \\ 2\dot{\eta}^\top (x - Li) + 2\eta^\top \eta \end{pmatrix}$$

where we denote $\eta = u - Ri + L\dot{\widehat{i}}$. Thus, if we suppose besides that there exists $c > 0$ such that the inputs verify $|\det(\eta, \dot{\eta})| \geq c$, every assumption of Lemma 7.1.2 is satisfied. In fact, the inputs happen to be such that[5] $\det(\eta\,,\,\dot{\eta}) = w^3\Phi^2$, where ω is the rotor angular velocity. We conclude that all the conditions are verified when the inputs and their derivatives are bounded and the rotor angular velocity is away from zero.

Applying Theorem 7.1.1, it follows that for any three distinct and sufficiently large strictly positive λ_j, the function

$$T(x,t) = (T_{\lambda_1}(x,t), T_{\lambda_2}(x,t), T_{\lambda_3}(x,t))$$

becomes injective after a certain time (once the filters (7.15) have sufficiently converged). Implementing (7.15) for each λ_j, one can obtain after a certain time an estimate $\hat{x}$ of $x(t)$ for instance by :

$$\hat{x}(t) = -\begin{pmatrix} a_{\lambda_1}(t)^\top - a_{\lambda_3}(t)^\top \\ a_{\lambda_2}(t)^\top - a_{\lambda_3}(t)^\top \end{pmatrix}^{-1} \begin{pmatrix} b_{\lambda_1}(t) - b_{\lambda_3}(t) \\ b_{\lambda_2}(t) - b_{\lambda_3}(t) \end{pmatrix} .$$

Note that for this system, a classical gradient observer of smaller dimension exists ([LHN+10, MPH12]). The Luenberger observer proposed here offers the advantage of depending only on filtered versions of u and i, which can be useful in presence of significant noise. On the other hand, no high gain design would have been possible for this system without computing the derivatives of i, which is not desirable in practice.

7.2.2 Non-holomic vehicle

Another appropriate example is the celebrated non-holomic vehicle with dynamics

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& u_1 \cos(x_3) \\ \dot{x}_2 &=& u_1 \sin(x_3) \\ \dot{x}_3 &=& u_1 u_2 \end{array} \right. \quad , \quad y = (x_1, x_2) \tag{7.16}$$

where the inputs u_1 and u_2 correspond to the norm of vehicle velocity and the orientation of the front steering wheels respectively. A wide literature already exists on this system, and our goal here is only to show on another example how to solve PDE (7.12) for each component of the measurement. The dynamics and measurements being linear in x_1, x_2, $\cos(x_3)$, $\sin(x_3)$, it is quite natural to look for a function T linear in those quantities. Besides, x_1 and x_2 are independent so we look for $T_{\lambda,1}$ and $T_{\lambda,2}$, associated to measurement x_1 and x_2 respectively, of the form :

$$\begin{array}{rcl} T_{\lambda,1}(x,t) &=& a_\lambda(t)\, x_1 + b_\lambda(t) \cos(x_3) + c_\lambda(t) \sin(x_3) \\ T_{\lambda,2}(x,t) &=& \tilde{a}_\lambda(t)\, x_2 + \tilde{b}_\lambda(t) \cos(x_3) + \tilde{c}_\lambda(t) \sin(x_3) \; . \end{array}$$

By straightforward computations, we conclude that to satisfy PDE (7.12), we can take :

$$\begin{array}{c} \tilde{a}_\lambda = a_\lambda = \dfrac{1}{\lambda} \quad , \quad \tilde{b}_\lambda = -c_\lambda \quad , \quad \tilde{c}_\lambda = d_\lambda \\ \begin{array}{rcl} \dot{b}_\lambda &=& -\lambda\, b_\lambda - u_1 u_2\, c_\lambda - \dfrac{1}{\lambda} u_1 \\ \dot{c}_\lambda &=& -\lambda\, c_\lambda + u_1 u_2\, b_\lambda \; . \end{array} \end{array} \tag{7.17}$$

Then, $T_{\lambda,1}$ and $T_{\lambda,2}$ are solutions of

$$\begin{array}{rcl} \dot{\xi}_{\lambda,1} &=& -\lambda\, \xi_{\lambda,1} + x_1 \\ \dot{\xi}_{\lambda,2} &=& -\lambda\, \xi_{\lambda,2} + x_2 \end{array} \tag{7.18}$$

[5] $\eta = \Phi\omega \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$, with θ the motor angle, and $\omega = \dot{\theta}$. See Chapter 12 for more information on this system.

respectively. Besides, computing the successive derivatives of the measurements (x_1, x_2), we can see that H_1 and H_2 are injective immersions at the order m if at least u_1 or one of its first $m-2$ derivatives is nonzero. Therefore, if the state, the inputs u_1, u_2 and their derivatives remain in compact sets, and if there exist an integer $m \geq 2$ and a real number $c > 0$ such that for all t and all considered input u,

$$u_1(t)^2 + \dot{u}_1(t)^2 + ... + u_1^{(m-2)}(t)^2 \geq c \ ,$$

then all the assumptions in Lemma 7.1.2 are verified with $m_1 = m_2 = m$. Therefore, by choosing m strictly positive distinct real numbers λ_j, the function

$$T(x,t) = (T_{\lambda_1,1}(x,t), ..., T_{\lambda_m,1}(x,t), T_{\lambda_1,2}(x,t), ..., T_{\lambda_m,2}(x,t))$$

becomes injective after a certain time. Implementing (7.17) – (7.18) for each λ_j, we thus get an observer of dimension $4m$.

7.2.3 A time-varying transformation for an autonomous system ?

It was observed in [And05, Section 8.4] that it is sometimes useful to allow the transformation to be time-varying even for an autonomous system. Only results concerning stationary transformations were available at the time, so that the framework of dynamic extensions had to be used. This is no longer necessary thanks to Theorems 7.1.2 and 7.1.1. Indeed, consider for instance the system

$$\left\{ \begin{array}{rcl} \dot{x}_1 &=& x_2^3 \\ \dot{x}_2 &=& -x_1 \end{array} \right. \quad , \quad y = x_1 \tag{7.19}$$

which admits bounded trajectories, the quantity $x_1^2 + x_2^4$ being constant along the trajectories. This system is weakly differentially observable of order 2 on $\mathbb{R}^2$ since $x \mapsto \mathbf{H}_2(x) = (x_1, x_2^3)$ is injective on $\mathbb{R}^2$. It is thus a fortiori instantaneously backward-distinguishable and Theorem 5.1.3 holds. Applying Luenberger's methodology to this system would thus bring us to look for a stationary transformation T_λ into

$$\dot{\xi}_\lambda = -\lambda\, \xi_\lambda + x_1 \ , \tag{7.20}$$

for which a possible solution is

$$T_\lambda(x) = \int_{-\infty}^{0} e^{\lambda\, \tau} Y(x;\tau) d\tau \ .$$

Although the injectivity of $T = (T_{\lambda_1}, T_{\lambda_2}, T_{\lambda_3})$ is satisfied for a generic choice of $(\lambda_1, \lambda_2, \lambda_3)$ in $\{\lambda \in \mathbb{C} \ : \ \Re(\lambda) > 0\}^3$ according to Theorem 5.1.3, it is difficult to compute numerically and as far as we are concerned, we are not able to find an explicit expression.

Instead, it may be easier to look for a time-varying transformation and apply either Theorem 7.1.1 or 7.1.2. Given the structure of the dynamics, one can try to look for a transformation of the form

$$T_\lambda(x,t) = a_\lambda(t) x_2^3 + b_\lambda(t) x_2^2 + c_\lambda(t) x_2 + d_\lambda(t) x_1 + e_\lambda(t) \ . \tag{7.21}$$

It verifies the dynamics (7.20) if for instance

$$\begin{array}{rcl} \dot{a}_\lambda &=& -\lambda\, a_\lambda + d_\lambda \\ \dot{b}_\lambda &=& -\lambda\, b_\lambda + 3 a_\lambda y \\ \dot{c}_\lambda &=& -\lambda\, c_\lambda + 2 b_\lambda y \\ \dot{d}_\lambda &=& -\lambda\, d_\lambda + 1 \\ \dot{e}_\lambda &=& -\lambda\, e_\lambda + c_\lambda y \end{array}$$

Using Remark 11 and applying Theorem 7.1.2, we know that, by initializing the filters a_λ, b_λ, c_λ, d_λ and e_λ at 0 at time 0, $x \mapsto (T_{\lambda_1}(x,t), T_{\lambda_2}(x,t), T_{\lambda_3}(x,t))$ is injective on $\mathbb{R}^2$ for $t > 0$ and for a generic choice of $(\lambda_1, \lambda_2, \lambda_3)$ in $\{\lambda \in \mathbb{C} \ : \ \Re(\lambda) > 0\}^3$.

To reduce the dimension of the filters, we can take $d_\lambda(t) = \frac{1}{\lambda}$ and $a_\lambda(t) = \frac{1}{\lambda^2}$. In that case Theorem 7.1.2 cannot be properly applied because T_λ is not T^0_λ. However, we have found at least in simulations that injectivity is preserved after a certain time as shown in Figure 7.1.

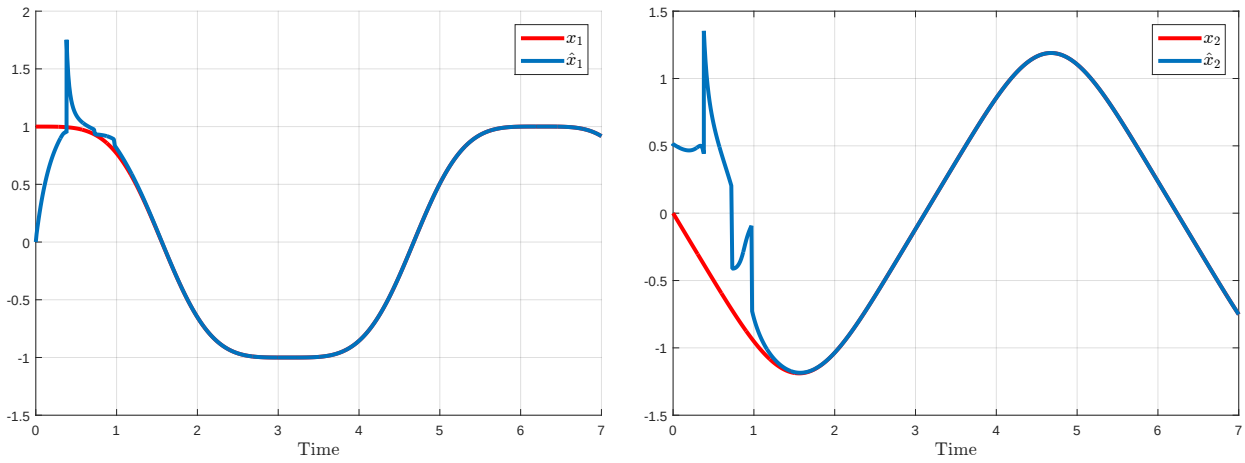


Figure 7.1: Nonlinear Luenberger observer for System (7.19) : dynamics (7.20) and transformations (7.21) (with $d_\lambda(t) = \frac{1}{\lambda}$ and $a_\lambda(t) = \frac{1}{\lambda^2}$) for $\lambda_1 = 5$, $\lambda_2 = 6$, $\lambda_3 = 7$. The transformation is inverted by searching numerically the common roots of two polynomials of order 3.

Note that since the system is strongly differentially observable of order 4 on $\mathcal{S} = \{(x_1, x_2) \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 \neq 0\}$, i-e $\mathbf{H}_4$ is an injective immersion on $\mathcal{S}$, Theorem 7.1.1 also says that, for any compact subset $\mathcal{C}$ of $\mathcal{S}$, by choosing 4 sufficiently large real strictly positive numbers λ_i, and for any initial conditions for the filters, $x \mapsto (T_{\lambda_1}(x,t), T_{\lambda_2}(x,t), T_{\lambda_3}(x,t), T_{\lambda_4}(x,t))$ becomes injective on $\mathcal{C}$ after some time.

7.3 Stationary transformation ?

We have just seen that a time-varying transformation could be used for an autonomous system. We investigate here the converse, i-e if a stationary transformation can be used for time-varying systems. Consider a control-affine single-output system

$$\dot{x} = f(x) + g(x)u \quad , \quad y = h(x) \in \mathbb{R} \tag{7.22}$$

In the high gain framework, we saw with Theorem (5.2.2) that if System 7.22 is uniformly instantaneously observable and its drift dynamics are differentially observable of order d_x, it is possible to keep the stationary transformation associated to the drift autonomous system, because the additional terms resulting from the presence of inputs are triangular and do not prevent the convergence of the observer. It turns out that, inspired from [AP06, Theorem 5], an equivalent result exists in the Luenberger framework.

Theorem 7.3.1.

Let $\lambda_1, \ldots, \lambda_{d_x}$ be any distinct strictly positive real numbers, A the Hurwitz matrix $\mathsf{diag}(-\lambda_1, \ldots, -\lambda_{d_x})$ in $\mathbb{R}^{d_x \times d_x}$, B the vector $(1, ..., 1)^\top$ in $\mathbb{R}^{d_x}$ and $\mathcal{S}$ an open subset of $\mathbb{R}^{d_x}$.

Assume that System (7.22) is uniformly instantaneously observable[6] on $\mathcal{S}$ and its drift system is strongly differentially observable[7] of order d_x on $\mathcal{S}$. Then, for any positive real number $\overline{u}$, any bounded open subsets $\mathcal{X}$, $\mathcal{X}'$ and $\mathcal{X}''$ of $\mathbb{R}^{d_x}$, and any C^∞ function $\chi : \mathbb{R}^{d_x} \to \mathbb{R}$ such that

- $\mathtt{cl}(\mathcal{X}) \subset \mathcal{X}' \subset \mathtt{cl}(\mathcal{X}') \subset \mathcal{X}'' \subset \mathtt{cl}(\mathcal{X}'') \subset \mathcal{S}$,
- for any u in $\mathcal{U}$, for all t in $[0, +\infty)$ and for all x_0 in $\mathcal{X}_0$, $|u(t)| \leq \overline{u}$ and $X(x_0; t; u)$ is in $\mathcal{X}$,
- $\chi(x) = \begin{cases} 1 & , \text{ if } x \in \mathtt{cl}(\mathcal{X}') \\ 0 & , \text{ if } x \notin \mathcal{X}'' \end{cases}$

there exists a strictly positive number $\overline{k}$ such that for any $k > \overline{k}$:

- the function $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ defined by

$$T(x) = \int_{-\infty}^{0} e^{-kA\tau} B\, h(\check{X}(x, \tau))\, d\tau$$

  where $\check{X}(x, \tau)$ denotes the value at time τ of the solution initialized at x at time 0 of the modified autonomous drift system

$$\dot{x} = \chi(x) f(x) \ ,$$

  is a diffeomorphism on $\mathcal{X}'$ and is solution to the PDE associated to the drift dynamics

$$\frac{\partial T}{\partial x}(x) f(x) = k\, A\, T(x) + B\, h(x) \quad \forall x \in \mathcal{X}' \ . \tag{7.23}$$

- there exists a Lipschitz function $\overline{\varphi}$ defined on $\mathbb{R}^{d_x}$ verifying

$$\overline{\varphi}(T(x)) = \frac{\partial T}{\partial x}(x) g(x) \qquad \forall x \in \mathcal{X}' \ , \tag{7.24}$$

  and such that, for any function $\mathcal{T} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ verifying

$$\mathcal{T}(T(x)) = x \qquad \forall x \in \mathcal{X}' \ ,$$

  the system

$$\dot{\hat{\xi}} = k\, A\, \hat{\xi} + B\, y + \overline{\varphi}(\hat{\xi})\, u \quad , \quad \hat{x} = \mathcal{T}(\hat{\xi}) \tag{7.25}$$

  is an observer for System 7.22 initialized in $\mathcal{X}_0$.

Remark 12 The function $\overline{\varphi}$ is defined on the open set $T(\mathcal{X}')$ by (7.24). If the trajectories of the observer state $\hat{\xi}$ remain in this set, there is no need to extend its domain of definition to the whole $\mathbb{R}^{d_x}$. Otherwise, the only constraint is that the global Lipschitz constant $\mathfrak{a}$ of the extension be such that $k \min |\lambda_i| > \mathfrak{a}\, \overline{u}$, to ensure the convergence of the observer. In the proof below, it is proved that such extensions exist for k sufficiently large (this is not trivial because $\mathfrak{a}$ could a priori depend on k).

Otherwise, instead of extending $\overline{\varphi}$ outside $T(\mathcal{X}')$, one could take

$$\overline{\varphi}(\xi) = \frac{\partial T}{\partial x}(\mathcal{T}(\xi)) g(\mathcal{T}(\xi))$$

[6] See Definition 2.2.1.
[7] See Definition 5.2.2.

but the way $\mathcal{T}$ is defined outside $T(\mathcal{X})$ must be such that :

$$\exists \alpha > 0 \ : \ \forall k \geq \overline{k}\,, \forall \hat{\xi} \in \mathbb{R}^{d_x}\,, \forall x \in \mathcal{X}\,, |T(x) - T(\mathcal{T}(\hat{\xi}))| \leq \alpha |T(x) - \hat{\xi}|\ .$$

The constraint here is that α must be independent from k. For instance, the function

$$\mathcal{T}(\xi) = \mathsf{Argmin}_{x \in \mathcal{X}'} |T(x) - \xi|$$

clearly works since

$$|T(x) - T(\mathcal{T}(\hat{\xi}))| \leq |T(x) - \hat{\xi}| + \underbrace{|\hat{\xi} - T(\mathcal{T}(\hat{\xi}))|}_{\leq |\hat{\xi} - T(x)|}\ .$$

Another more regular candidate is the McShane extension

$$\mathcal{T}(\xi) = \min_{x \in \mathcal{X}'} x + |T(x) - \xi|$$

which also verifies the requirement.

Proof : According to [And14, Proposition 3.3], there exists $\overline{k}_0$ such that for all $k \geq \overline{k}_0$, T is C^1 and verifies PDE (7.23). Now let us prove that it is injective on $\mathtt{cl}(\mathcal{X}')$ for k sufficiently large[8]. The drift system being strongly differentially observable of order d_x, the function

$$\mathbf{H}_{d_x}(x) = (h(x), L_f h(x), \ldots, L_f^{d_x}(x))$$

is an injective immersion on $\mathtt{cl}(\mathcal{X}')$ and by Lemma A.3.5, there exists $L_H > 0$ such that for all $(x_a, x_b)^2$ in $\mathtt{cl}(\mathcal{X}')^2$,

$$|\mathbf{H}_{d_x}(x_a) - \mathbf{H}_{d_x}(x_b)| \geq L_H |x_a - x_b|\ .$$

Besides, since $\chi f = f$ on $\mathtt{cl}(\mathcal{X}')$, after several integrations by parts, we obtain for all x in $\mathtt{cl}(\mathcal{X}')$

$$T(x) = A^{-d_x} \mathscr{C} \left(-K \mathbf{H}_{d_x}(x) + \frac{1}{k^{d_x}} R(x) \right) \tag{7.26}$$

where $K = \mathsf{diag}\left(\frac{1}{k}, ..., \frac{1}{k^{d_x}}\right)$, $\mathscr{C}$ is the invertible controllability matrix

$$\mathscr{C} = [A^{d_x - 1} B \ ... \ AB \ B]\ ,$$

and R the remainder

$$R(x) = \mathscr{C}^{-1} \int_{-\infty}^{0} e^{-kA\tau} B\, L_f^{d_x}(\check{X}(x, \tau))\, d\tau\ .$$

This latter integral makes sense on $\mathtt{cl}(\mathcal{X}')$ because :

-A being diagonal and denoting $a = \min_i |\lambda_i| > 0$, for all $\tau \in (-\infty, 0]$,

$$\left| e^{-kA\tau} \right| \leq e^{ka\tau}\ .$$

-By definition of the function χ, for all x in $\mathtt{cl}(\mathcal{X}')$, $\check{X}(x, \tau)$ is in $\mathtt{cl}(\mathcal{X}')$ for all τ, i-e $\tau \mapsto L_f^{d_x}(\check{X}(x, \tau))$ is bounded.

So now taking (x_a, x_b) in $\mathtt{cl}(\mathcal{X}')^2$, and considering the difference $|T(x_a) - T(x_b)|$, from (7.26), we obtain

$$|T(x_a) - T(x_b)| \geq \frac{|A^{-d_x}\mathscr{C}|}{k^{d_x}} (|\mathbf{H}_{d_x}(x_a) - \mathbf{H}_{d_x}(x_b)| - |R(x_a) - R(x_b)|)\ ,$$

and if R is Lipschitz with Lipschitz constant L_R, we get

$$|T(x_a) - T(x_b)| \geq \frac{|A^{-d_x}\mathscr{C}|}{k^{d_x}} (L_H - L_R) |x_a - x_b|\ .$$

In order to deduce the injectivity of T, we also need $L_R < L_H$ and we are going to prove that this is true for k sufficiently large. To compute L_R, let us find a bound of $\left|\frac{\partial R}{\partial x}(x)\right|$. By defining

$$c_0 = \max_{x \in \mathtt{cl}(\mathcal{X}')} \left| B \frac{\partial L_f^{d_x} h}{\partial x}(x) \right| \quad , \quad \rho_1 = \max_{x \in \mathtt{cl}(\mathcal{X}')} \left| \frac{\partial f}{\partial x}(x) \right|\ ,$$

[8]This proof is similar to that of [AP06, Theorem 4].

we have for all τ in $(-\infty, 0]$ and all x in $\mathtt{cl}(\mathcal{X}')$, $\left| B \frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(x,\tau)) \right| \leq c_0$ and[9]

$$\left| \frac{\partial \check{X}}{\partial x}(x,\tau) \right| \leq e^{-\rho_1 \tau} \ . \tag{7.27}$$

We conclude that for $k > \frac{\rho_1}{a}$, R is C^1 and there exists a positive constant c_1 such that for all x in $\mathtt{cl}(\mathcal{X}')$,

$$\left| \frac{\partial R}{\partial x}(x) \right| \leq |\mathscr{C}^{-1}| \int_{-\infty}^{0} \left| e^{-kA\tau} \right| \left| B \frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(x,\tau)) \right| \left| \frac{\partial \check{X}}{\partial x}(x,\tau) \right| d\tau \leq \frac{c_1}{ka - \rho_1}.$$

We finally obtain

$$|T(x_a) - T(x_b)| \geq L_T |x_a - x_b| \quad \forall (x_a, x_b) \in \mathtt{cl}(\mathcal{X}')^2 \tag{7.28}$$

where

$$L_T = \frac{|A^{-d_x}\mathscr{C}|}{k^{d_x}} \left(L_H - \frac{c_1}{ka - \rho_1} \right) \ ,$$

and T is injective on $\mathtt{cl}(\mathcal{X}')$ if $k \geq \overline{k}_1$ with

$$\overline{k}_1 = \max \left\{ \overline{k}_0, \frac{c_1 + \rho_1 L_H}{a L_H} \right\} \ .$$

Moreover, taking x in $\mathcal{X}'$, any v in $\mathbb{R}^m$ and h sufficiently small for $x + hv$ to be in $\mathcal{X}'$, it follows from (7.28) that

$$\left| \frac{T(x+hv) - T(x)}{h} \right| \geq L_T |v| \ ,$$

and making h tend to zero, we get

$$\left| \frac{\partial T}{\partial x}(x) v \right| \geq L_T |v|$$

and T is full-rank on $\mathcal{X}'$. So T is a diffeomorphism on $\mathcal{X}'$ for $k \geq \overline{k}_1$.

Now, let us show that System (7.25) is an observer for System (7.22). Suppose for the time being that we have shown that there exists a strictly positive number $\mathfrak{a}$ such that for any $k \geq \overline{k}_1$, there exists a function $\overline{\varphi}$ such that (7.24) holds and

$$|\overline{\varphi}(\hat{\xi}) - \overline{\varphi}(\xi)| \leq \mathfrak{a} \, |\hat{\xi} - \xi| \qquad \forall (\hat{\xi}, \xi) \in (\mathbb{R}^{d_x})^2 \ . \tag{7.29}$$

Take u in $\mathcal{U}$, x_0 in $\mathcal{X}_0$ $\hat{\xi}_0$ in $\mathbb{R}^{d_x}$, and consider the solution $X(x_0; t; u)$ of System (7.22) and any corresponding solution $\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0})$ of System (7.25). Since $X(x_0; t; u)$ remains in $\mathcal{X}$ by assumption, the error $e(t) = \hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0}) - T(X(x_0; t; u))$ verifies

$$\dot{e} = kA \ e + \left(\overline{\varphi}(\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0})) - \overline{\varphi}(T(X(x_0; t; u))) \right) u$$

and thus

$$\dot{\overbrace{e^\top e}} \leq -2(ka - \mathfrak{a}\,\overline{u}) \, e^\top e \ .$$

Defining $\overline{k}_2 = \max\{\overline{k}_1, \frac{\mathfrak{a}\overline{u}}{a}\}$, we conclude that e asymptotically converges to 0 if $k \geq \overline{k}_2$. Note that for this conclusion to hold, it is crucial to have $\mathfrak{a}$ independent from k. Now, consider an open set $\tilde{\mathcal{X}}$ such that $\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{X}} \subset \mathtt{cl}(\tilde{\mathcal{X}}) \subset \mathcal{X}'$. Since $T(X(x_0; t; u))$ remains in $T(\mathcal{X})$ and $\mathtt{cl}(T(\mathcal{X})) = T(\mathtt{cl}(\mathcal{X}))$ is contained in the open set $T(\tilde{\mathcal{X}})$, there exists a time $\bar{t}$ such that for all $t \geq \bar{t}$, $\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0})$ is in $T(\tilde{\mathcal{X}})$. $\mathcal{T} = T^{-1}$ is C^1 on the compact set $\mathtt{cl}(T(\tilde{\mathcal{X}}))$ and thus Lipschitz on that set. It follows that $\hat{X}((x_0, \hat{\xi}_0); t; u) = \mathcal{T}(\hat{\Xi}(\hat{\xi}_0; t; u, y_{x_0}))$ converges to $X(x_0; t; u)$.

It remains to show the existence of the functions $\overline{\varphi}$. Since System (7.22) is uniformly instantaneously observable and its drift system is strongly differentially observable of order d_x on $\mathcal{S}$, we know since [GB81] or with Theorem 6.3.1, that for all i in $\{1, \ldots, d_x\}$, there exists a Lipschitz function $\mathfrak{g}_i$ such that

$$L_g L_f^{i-1} h(x) \ = \ \mathfrak{g}_i(h(x), \ldots, L_f^{i-1}(x)) \quad \forall x \in \mathtt{cl}(\mathcal{X}) \ . \tag{7.30}$$

Consider the function

$$\begin{aligned} \varphi(x) &= \frac{\partial T}{\partial x}(x) g(x) \\ &= A^{-d_x} \mathscr{C} \left(\underbrace{- K \frac{\partial \mathbf{H}_{d_x}}{\partial x}(x) g(x)}_{\varphi_H(x)} + \underbrace{\frac{1}{k^{d_x}} \frac{\partial R}{\partial x}(x) g(x)}_{\varphi_R(x)} \right) \ . \end{aligned}$$

[9] Because $\psi(\tau) = \frac{\partial \check{X}}{\partial x}(x,\tau)$ follows the ODE $\frac{d\psi}{d\tau}(\tau) = \frac{\partial f}{\partial x}(\check{X}(x,\tau))\psi(\tau)$, and $\psi(0) = I$.

Let us first study φ_H. Notice that the ith-component of φ_H is $\varphi_{H,i} = \frac{1}{k^i} L_g L_f^{i-1} h(x)$ and according to (7.30), there exists L_i such that

$$|\varphi_{H,i}(\hat{x}) - \varphi_{H,i}(x)| \leq L_i \sum_{j=1}^{i} \left| \frac{1}{k^j} \left(L_f^{j-1}(\hat{x}) - L_f^{j-1}(x) \right) \right| \quad \forall (x,\hat{x}) \in \mathtt{cl}(\mathcal{X})^2$$

and thus L such that

$$|\varphi_H(\hat{x}) - \varphi_H(x)| \leq L |K\mathbf{H}_{d_x}(\hat{x}) - K\mathbf{H}_{d_x}(x)| \quad \forall (x,\hat{x}) \in \mathtt{cl}(\mathcal{X})^2 \ .$$

But using (7.26), we get

$$|K\mathbf{H}_{d_x}(\hat{x}) - K\mathbf{H}_{d_x}(x)| \leq |A^{d_x}\mathscr{C}^{-1}||T(\hat{x}) - T(x)| + \frac{1}{k^{d_x}} |R(\hat{x}) - R(x)| \quad \forall (x,\hat{x}) \in \mathtt{cl}(\mathcal{X})^2 \ .$$

We have seen that

$$|R(\hat{x}) - R(x)| \leq \frac{c_1}{ka - \rho_1} |\hat{x} - x| \quad \forall (x,\hat{x}) \in \mathtt{cl}(\mathcal{X})^2$$

and according to (7.28),

$$\frac{1}{k^{d_x}} |R(\hat{x}) - R(x)| \leq \frac{\frac{c_1}{ka-\rho_1}}{L_H - \frac{c_1}{ka-\rho_1}} |A^{d_x}\mathscr{C}^{-1}||T(\hat{x}) - T(x)| \quad \forall (x,\hat{x}) \in \mathtt{cl}(\mathcal{X})^2 \ .$$

We finally obtain, for any $(x,\hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$ and for any $k \geq \overline{k}_1$,

$$\begin{aligned} |\varphi_H(\hat{x}) - \varphi_H(x)| &\leq L |A^{d_x}\mathscr{C}^{-1}| \left(1 + \frac{\frac{c_1}{ka-\rho_1}}{L_H - \frac{c_1}{ka-\rho_1}} \right) |T(\hat{x}) - T(x)| \\ &\leq L |A^{d_x}\mathscr{C}^{-1}| \left(1 + \frac{c_1}{L_H(\overline{k}_1 a - \rho_1)} \right) |T(\hat{x}) - T(x)| \ . \end{aligned}$$

Let us now study the term $\varphi_R(x)$. For $(x,\hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$,

$$\varphi_R(\hat{x}) - \varphi_R(x) = \frac{1}{k^{d_x}} \mathscr{C}^{-1} \int_{-\infty}^{0} e^{-kA\tau} B(D_1(x,\hat{x},\tau) + D_2(x,\hat{x},\tau) + D_3(x,\hat{x},\tau)) d\tau$$

where

$$\begin{aligned} D_1(x,\hat{x},\tau) &= \left(\frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(x,\tau)) - \frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(\hat{x},\tau)) \right) \frac{\partial \check{X}}{\partial x}(\hat{x},\tau) g(\hat{x}) \\ D_2(x,\hat{x},\tau) &= \frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(x,\tau)) \left(\frac{\partial \check{X}}{\partial x}(\hat{x},\tau) - \frac{\partial \check{X}}{\partial x}(x,\tau) \right) g(\hat{x}) \\ D_3(x,\hat{x},\tau) &= \frac{\partial L_f^{d_x} h}{\partial x}(\check{X}(x,\tau)) \frac{\partial \check{X}}{\partial x}(x,\tau) \left(g(\hat{x}) - g(x) \right) \end{aligned}$$

Assuming that $L_f^{d_x} h$ is C^2 and g is C^1, it follows from (7.27) and the fact that $\check{X}(x,\tau)$ is in the compact set $\mathtt{cl}(\mathcal{X}')$ for all τ in $(-\infty, 0]$, that for all $(x,\hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$ and for all τ in $(-\infty, 0]$,

$$\begin{aligned} |D_1(x,\hat{x},\tau)| &\leq c_2 e^{-2\rho_1\tau} |x - \hat{x}| \\ |D_3(x,\hat{x},\tau)| &\leq c_3 e^{-\rho_1\tau} |x - \hat{x}| \ . \end{aligned}$$

As for D_2, posing $\varphi(\tau) = \frac{\partial \check{X}}{\partial x}(\hat{x},\tau) - \frac{\partial \check{X}}{\partial x}(x,\tau)$, and differentiating φ with respect to time, we get

$$\varphi(0) = 0 \quad , \quad \varphi'(\tau) = \frac{\partial f}{\partial x}(\check{X}(\hat{x},\tau))\varphi(\tau) + \left(\frac{\partial f}{\partial x}(\check{X}(\hat{x},\tau)) - \frac{\partial f}{\partial x}(\check{X}(x,\tau)) \right) \frac{\partial \check{X}}{\partial x}(x,\tau) \ . \tag{7.31}$$

Since for all τ in $(-\infty, 0]$ and for all $(x,\hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$,

$$\left| \frac{\partial f}{\partial x}(\check{X}(\hat{x},\tau)) \right| \leq \rho_1 \quad , \quad \left| \frac{\partial \check{X}}{\partial x}(x,\tau) \right| \leq e^{-\rho_1\tau}$$

and

$$\left| \frac{\partial f}{\partial x}(\check{X}(\hat{x},\tau)) - \frac{\partial f}{\partial x}(\check{X}(x,\tau)) \right| \leq c_4 e^{-\rho_1\tau} |x - \hat{x}| \ ,$$

we obtain by solving (7.31) in negative time and taking the norm

$$|D_2(\hat{x},x,\tau)| \leq \left(c_5 e^{-\rho_1\tau} + c_6 e^{-2\rho_1\tau} \right) |x - \hat{x}| \leq c_7 \, e^{-2\rho_1\tau} |x - \hat{x}|$$

for all τ in $(-\infty, 0]$ and all $(x, \hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$. Therefore, for all $k \geq \overline{k}_1$,

$$|\varphi_R(\hat{x}) - \varphi_R(x)| \leq \frac{1}{k^{d_x}} \frac{c_8}{ka - \rho_1} |x - \hat{x}| \leq \frac{\frac{c_9}{ka-\rho_1}}{L_H - \frac{c_1}{ka-\rho_1}} |T(x) - T(\hat{x})| \leq \frac{c_9}{L_H(\overline{k}_1 a - \rho_1)} |T(x) - T(\hat{x})| \ .$$

Finally, there exists a constant $\mathfrak{a}$ such that for all $k \geq \overline{k}_1$, and for all $(x, \hat{x})$ in $\mathtt{cl}(\mathcal{X})^2$,

$$|\varphi(\hat{x}) - \varphi(x)| \leq \mathfrak{a} \, |T(\hat{x}) - T(x)| \ . \tag{7.32}$$

Consider now the function

$$\overline{\varphi}(\xi) = \varphi(T^{-1}(\xi))$$

defined on $T(\mathcal{X}')$. According to (7.32), $\overline{\varphi}$ is Lipschitz on $T(\mathcal{X}')$, and with Kirszbraun-Valentine Theorem [Kir34, Val45], it admits a Lipschitz extension on $\mathbb{R}^{d_x}$ with same Lipschitz constant $\mathfrak{a}$, i-e such that (7.24) and (7.29) hold. This concludes the proof. ■

7.4 Conclusion

We have shown how a Luenberger methodology can be applied to nonlinear controlled systems. It is based on the resolution of a PDE, the solutions of which exist, transform the system into a linear asymptotically stable one, and become injective after a certain time. This injectivity is ensured if

- either the function made of the output and a certain number of its derivatives is Lipschitz-injective : this is verified when the system is strongly differentially observable and the trajectories are bounded.

- or the system is backward-distinguishable (uniformly in time), but in this case, injectivity is ensured for "almost all" choice of a diagonal complex matrix A (of sufficiently large dimension) in the sense of the Lebesgue measure in $\mathbb{C}$.

This methodology relies on finding a time-varying solution to a PDE, which always exists but may be difficult to compute. We have shown on practical examples how this can be done by a priori guessing its "structure".

Also, it is interesting to remember that as in the high gain paradigm, for uniformly instantaneously observable control-affine systems, we may use the stationary transformation associated to the autonomous drift system when it is strongly differentially observable of order d_x. The result does not stand for higher orders of differential observability, since it relies on the existence of Lipschitz functions $\mathfrak{g}_i$ such that $\mathfrak{g}_i(\mathbf{H}_i(x)) = L_g L_f^{i-1}(x)$, and we have seen in Chapter 6 that the Lipschitzness is lost when the drift system is differentially observable of higher order.

Part III

Expression of the dynamics of the observer in the system coordinates

Chapter 8

Motivation and problem statement

Chapitre 8 – Motivation et énoncé du problème. *Les Parties I-II montrent que l'on peut sous certaines conditions construire un observateur pour un système non linéaire en transformant sa dynamique en une forme favorable pour laquelle un observateur global est connu. Il s'ensuit que la dynamique du système et celle de l'observateur ne sont pas exprimées dans les mêmes coordonnées et évoluent même souvent dans des espaces de dimension différente. Afin d'obtenir une estimée de l'état du système, il est alors nécessaire d'inverser la transformation. Or, cette opération peut se révéler compliquée en pratique, notamment lorsqu'une expression explicite de l'inverse n'est pas connue, car elle repose alors sur la résolution d'un problème de minimisation couteux en calculs. C'est pour cette raison que nous avons développé une méthode permettant de ramener la dynamique de l'observateur dans les coordonnées initiales du système afin d'éviter l'inversion de la transformation. Dans ce chapitre, nous motivons cette démarche à l'aide d'exemples et donnons une première condition suffisante pour résoudre ce problème dans le cas où la transformation est stationnaire. Les chapitres suivants 9-10-11 seront consacrés à montrer comment remplir cette condition. De plus, la possible extension de ces résultats au cas où la transformation est non-stationnaire sera étudiée dans le chapitre 11, principalement à l'aide d'exemples tirés d'applications.*

Contents

Parts I-II have shown that it is possible, under certain conditions, to build an observer for a nonlinear system by transforming its dynamics into a favorable form for which a global observer is known. It follows that the dynamics of the system and of the observer are not expressed in the same coordinates and often even evolve in spaces of different dimensions. In order to obtain an estimate for the system state or even sometimes write the observer dynamics, it is necessary to invert the transformation. But this step can be difficult in practice, mostly when an explicit expression for the inverse is not available. Indeed, in this case, inversion usually relies on the resolution of a minimization problem with a heavy computation cost. That is why we have developed a methodology enabling to pull the dynamics of the observer back into the system

coordinates in order to avoid the inversion of the transformation, namely design an observer in the given coordinates[1]. In this chapter, we motivate and introduce this problem through examples and give a first sufficient condition to solve this problem in the case of a stationary transformation. The remaining chapters 9-10-11 will show how to satisfy this condition. Besides, the possible extension of those results to the case where the transformation is time-varying will be studied in Chapter 11 mainly through an example coming from an application. Note that we have submitted most of the results presented in this part in [BPA15] and [BPAew].

8.1 Example

To motivate the problem we shall tackle in this part of the thesis, we consider a harmonic oscillator with unknown frequency with dynamics

$$\left\{\begin{array}{rcll} \dot{x}_1 &=& x_2 & \\ \dot{x}_2 &=& -x_1x_3 & , \quad y = x_1 \\ \dot{x}_3 &=& 0 & \end{array}\right. \tag{8.1}$$

with state $x = (x_1, x_2, x_3)$ in $\mathbb{R}^2 \times \mathbb{R}_{>0}$ and measurement y. We are interested in estimating the state x of this system from the only knowledge of the function $t \mapsto y(t) = X_1(x,t)$. This problem has been widely studied in the literature ([HOD99, OPCTL02, Hou05, Hou12] among many others) and our goal is not to produce yet another observer for this system but rather to illustrate our methodology and the problems encountered throughout its implementation. This example is indeed sufficiently simple in terms of computations, but sufficiently rich in terms of underlying observability issues to be interesting throughout this part of the thesis.

For any solution with initial condition $x_1 = x_2 = 0$, y does not give any information on x_3. We thus restrict our attention to solutions evolving in $\mathcal{X}$ of the type

$$\mathcal{X} \;=\; \left\{ x \in \mathbb{R}^3 \;:\; x_1^2 + x_2^2 \in \left]\frac{1}{r}, r\right[\;,\; x_3 \in]0, r[\right\} \;, \tag{8.2}$$

where r is some arbitrary strictly positive real number. This set is forward-invariant by (8.1). Note also that System (8.1) is strongly differentially observable of order 4 on

$$\mathcal{S} = \left(\mathbb{R}^2 \setminus \{(0,0)\}\right) \times \mathbb{R}_+$$

containing $\mathcal{X}$, namely $\mathbf{H}_4$ defined by

$$\mathbf{H}_4(x) = \begin{pmatrix} h(x) \\ L_f h(x) \\ L_f^2 h(x) \\ L_f^3 h(x) \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ -x_1x_3 \\ -x_2x_3 \end{pmatrix}$$

is an injective immersion on $\mathcal{S}$.

8.1.1 High-gain design

According to Theorem 5.2.1 and Remark 4, we know that τ^* defined by

$$\tau^*(x) = \mathbf{H}_4(x) = (x_1, x_2, -x_1x_3, -x_2x_3) \tag{8.3}$$

transforms System 8.1 into a phase-variable form of dimension 4 for which a high-gain observer can be designed:

$$\dot{\hat{\xi}} \;=\; \mathcal{F}(\hat{\xi}, y) \;=\; \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\xi} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \Phi_4(\hat{\xi}) \end{pmatrix} + \begin{pmatrix} Lk_1 \\ L^2k_2 \\ L^3k_3 \\ L^4k_4 \end{pmatrix} [y - \hat{\xi}_1] \;, \tag{8.4}$$

[1] See Definition 2.1.1

where Φ_4 is defined by[2]

$$\Phi_4(\xi) = \mathsf{sat}_{r^3}(L_f^4 h(\tau(\xi)))$$

with τ any locally Lipschitz function defined on $\mathbb{R}^4$ verifying

$$\tau(\mathbf{H}_4(x)) = x \quad \forall x \in \mathcal{X} \ ,$$

r^3 may be replaced by any bound of $L_f^4 h$ on $\mathcal{X}$, and L is a sufficiently large strictly positive number depending on the Lipschitz constant of Φ_4, namely on the choice of τ and r. Wanting to highlight the role of the computation of the left-inverse τ, we get in fact a "raw" observer with dynamics

$$\dot{\hat{\xi}} \ = \ \varphi(\hat{\xi}, \hat{x}, y) \ = \ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\xi} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \mathsf{sat}_{r^3}(\hat{x}_1 \hat{x}_3^2) \end{pmatrix} + \begin{pmatrix} \ell k_1 \\ \ell^2 k_2 \\ \ell^3 k_3 \\ \ell^4 k_4 \end{pmatrix} [y - \hat{\xi}_1] \quad , \quad \hat{x} = \tau(\hat{\xi}) \ . \quad (8.5)$$

We deduce that the computation of the function τ (whose existence is guaranteed by the theorem) is crucial in the implementation of this observer, of course to deduce $\hat{x}$ from $\hat{\xi}$ but also to define the dynamics of the observer itself.

Although in this example an explicit and global expression[3] for τ can easily be found due to the simplicity of the transformation $\tau^* = \mathbf{H}_4$, it is not always the case in high gain designs for more complex applications. To overcome this problem, we may go with solving an optimization problem as

$$\hat{x} = \tau(\hat{\xi}) = \operatorname*{Argmin}_{\hat{x}} \left| \hat{\xi} - \tau^*(\hat{x}) \right|^2 \ .$$

8.1.2 Luenberger design

Instead of a high gain observer design as above, we may use a non linear Luenberger design. As explained in Section 5.1.2, the idea is to find a transformation into a Hurwitz form of the type :

$$\dot{\xi} \ = \ A\,\xi \ + \ B\,y$$

with ξ in $\mathbb{R}^{d_\xi}$, A a Hurwitz matrix and (A, B) a controllable pair. Indeed, this system admits as global observer

$$\dot{\hat{\xi}} \ = \ \varphi(\hat{\xi}, y) \ = \ A\,\hat{\xi} \ + \ B\,y \ . \quad (8.6)$$

Since the dynamics (8.1) are linear in (x_1, x_2), we can look for a transformation depending linearly in (x_1, x_2). Straightforward computations give :

$$\tau^*(x) \ = \ -(A^2 + x_3 I)^{-1}[ABx_1 + Bx_2] \ . \quad (8.7)$$

In particular, for a diagonal matrix $A = \mathsf{diag}(-\lambda_1, \ldots, -\lambda_{d_\xi})$ with $\lambda_i > 0$, and $B = (1, \ldots, 1)^\top$, this gives for i in $\{1, \ldots, d_\xi\}$:

$$\tau_i^*(x) = \frac{\lambda_i x_1 - x_2}{\lambda_i^2 + x_3} \ . \quad (8.8)$$

It is shown in [PMI06] that τ^* is injective on $\mathcal{S}$ if $d_\xi \geq 4$ for any distinct λ_i's in $(0, +\infty)$. More precisely, it is Lipschitz-injective on any compact subset of $\mathcal{S}$ and therefore, τ^* is an injective immersion[4] on $\mathcal{S}$. This is consistent with [AP06, Theorem 4] and the fact that the order of strong differentiability of this system is 4.

[2] The saturation function is defined by $\mathsf{sat}_M(s) \ = \ \min\{M, \max\{s, -M\}\}$.

[3] For instance, we can take $\tau(\xi) = \left(\xi_1 \, , \, \xi_2 \, , \, -\frac{\xi_1\xi_3 + \xi_4\xi_2}{\max\{\xi_1^2 + \xi_2^2, \frac{1}{r^2}\}}\right)$.

[4] Indeed, consider any x in $\mathcal{S}$ and $\mathcal{V}$ an open neighborhood of x such that $\mathtt{cl}(\mathcal{V})$ is contained in $\mathcal{S}$. According to the Lipschitz-injectivity of τ^* on $\mathtt{cl}(\mathcal{V})$, there exists $\mathfrak{a}$ such that for all v in $\mathbb{R}^3$ and for all h in $\mathbb{R}$ such that $x + hv$ is in $\mathcal{V}$, $|v| \leq \mathfrak{a}\frac{|\tau^*(x+hv) - \tau^*(x)|}{|h|}$ and thus by taking h to zero, $|v| \leq \mathfrak{a}\left|\frac{\partial \tau^*}{\partial x}(x)v\right|$ which means that $\dfrac{\partial \tau^*}{\partial x}(x)$ is full-rank.

Thus, since the trajectories of the system remain bounded, applying Corollary 2.2.1, there exists an observer for System (8.1) which is given by (8.6) and any continuous function τ satisfying

$$\tau(\tau^*(x)) = x \quad \forall x \in \mathcal{X} \ .$$

However, it is difficult to find an explicit expression of such a function, thus for this design, we would have to solve online :

$$\hat{x} = \tau(\hat{\xi}) = \underset{\hat{x}}{\text{Argmin}} \left| \hat{\xi} - \tau^*(\hat{x}) \right|^2 \ .$$

Note that a difference with the high gain observer above is that $\hat{x}$ is not involved in (8.6), i.e. the observer dynamics do not depend on τ.

8.1.3 General idea

In the following, we propose a methodology to write the dynamics of the given observers (8.5) and (8.6) directly in the x-coordinates[5] in order to eliminate the minimization step. This has been suggested by several researchers [DBGR92, MP03, AP13] in the case where the observer state $\hat{\xi}$ and the state estimate $\hat{x}$ are related by a diffeomorphism. We remove this restriction and complete the preliminary results presented in [AEP14].

In the example above, pulling the observer dynamics from the ξ-coordinates back to the x-coordinates appears impossible since x has dimension 3 whereas ξ has dimension 4. We overcome this difficulty by adding one component, say w, to x. Then, the dynamics of $(\hat{x}, \hat{w})$ can be obtained as an image of those of ξ if we have a diffeomorphism $(x, w) \mapsto \xi = \tau_e^*(x, w)$ "augmenting" the function $x \mapsto \tau^*(x)$ given in (8.3) or (8.7). We show in Chapter 9 that this can be done by complementing a full column rank Jacobian into an invertible matrix. Unfortunately, in doing so, the obtained diffeomorphism is rarely defined everywhere and we have no guarantee that the trajectory in $(\hat{x}, \hat{w})$ of the observer remains in the domain of definition of the diffeomorphism. We show in Chapter 10 how this new problem can be overcome via a diffeomorphism extension. The key point here is that the given observer dynamics (8.5) or (8.6) remain unchanged. This differs from other techniques as proposed in [MP03, AP13], which require extra assumptions such as convexity to preserve the convergence property.

8.2 Problem statement

8.2.1 Starting point

We consider a given system with dynamics :

$$\dot{x} = f(x, u) \quad , \qquad y = h(x, u) \ , \tag{8.9}$$

with x in $\mathbb{R}^{d_x}$, u a function in $\mathcal{U}$ with values in $U \subset \mathbb{R}^{d_u}$ and y in $\mathbb{R}^{d_y}$. The observation problem is to construct a dynamical system with input y and output $\hat{x}$, supposed to be an estimate of the system state x as long as the latter is in a specific set of interest denoted $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. As starting point here, we assume this problem is (formally) already solved but with maybe some implementation issues such as finding an expression of τ. More precisely,

Assumption $\mathcal{O}$: Converging observer in the ξ-coordinates

There exist an open subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$, a subset $\mathcal{X}$ of $\mathcal{S}$, a C^1 injective immersion $\tau^* : \mathcal{S} \to \mathbb{R}^{d_\xi}$,

[5] We will also refer to the x-coordinates as the "given coordinates" because they are chosen by the user to describe the model dynamics.

and a set[6] $\varphi\mathcal{T}$ of pairs (φ, τ) of functions such that :

- $\tau : \mathbb{R}^{d_\xi} \to \mathbb{R}^{d_x}$ is a left-inverse of τ^* on $\tau^*(\mathcal{X})$, i-e

$$\tau(\tau^*(x)) \;=\; x \qquad \forall x \in \mathcal{X} \tag{8.10}$$

- for any u in $\mathcal{U}$ and any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0, u) = +\infty$, the solution $X(x_0; t; u)$ of (8.9) remains in $\mathcal{X}$ for t in $[0, +\infty)$.

- for any u in $\mathcal{U}$, any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0, u) = +\infty$, and any $\hat{\xi}_0$ in $\mathbb{R}^{d_\xi}$, any solution $(X(x_0; t; u), \hat{\Xi}((x_0, \hat{\xi}_0); t; u))$ of the cascade system :

$$\dot{x} \;=\; f(x, u) \quad , \qquad y \;=\; h(x, u) \quad , \qquad \dot{\hat{\xi}} = \varphi(\hat{\xi}, \hat{x}, u, y) \quad , \qquad \hat{x} = \tau(\hat{\xi}) \; , \tag{8.11}$$

initialized at $(x_0, \hat{\xi}_0)$ and under the input u, is also defined on $[0, +\infty)$ and satisfies :

$$\lim_{t \to +\infty} \left| \Xi((x_0, \hat{\xi}_0); t; u) - \tau^*(X(x_0; t; u)) \right| \;=\; 0 \; . \tag{8.12}$$

Remark 13

1. The convergence property given by (8.12) is in the observer state space only. Property (8.10) is a necessary condition for this convergence to be transferred from the observer state space to the system state space. But as we saw earlier, we may need the injectivity of τ^* to be uniform in space, or equivalently τ to be uniformly continuous on $\mathbb{R}^{d_\xi}$, in order to conclude about a possible convergence in the x-coordinates. In that case, the couple $(\mathcal{F}, \mathcal{T})$ defined by
$$\mathcal{F}(\xi, u, y) = \varphi(\xi, \tau(\xi), u, y) \quad , \quad \mathcal{T}(\xi) = \tau(\xi)$$
is an observer for System (8.9) initialized in $\mathcal{X}_0$. Note that as in Corollary 2.2.1, this is achieved without further assumption in the case where $\mathcal{X}$ is bounded.

2. The reason why we make φ depend on $\hat{x}$, instead of simply taking $\mathcal{F}(\xi, u, y)$ as before, is that most of the time, and especially in a high gain design (see (8.5)), when expressing the dynamics of $\tau^*(x)$ as function of ξ to compute $\mathcal{F}$, we replace x by $\tau(\xi)$. Since we want here to avoid the computation of τ, we make this dependence explicit in φ.

3. The need for pairing φ and τ comes from this dependence because it may imply to change φ whenever we change τ. In the high-gain approach for instance, as in (8.5), when $\mathcal{X}$ is bounded, thanks to the gain L which can be chosen arbitrarily large, φ can be paired with any locally Lipschitz function τ provided its values are saturated whenever they are used as arguments of φ. On another hand, if, as in (8.6), φ does not depend on $\hat{x}$, then it can be paired with any τ.

Example 8.2.1 For System (8.1), $\mathcal{X}$ given in (8.2) being bounded, a set $\varphi\mathcal{T}$ satisfying Assumption $\mathcal{O}$ is made of pairs of

- a locally Lipschitz function τ satisfying

$$x = \tau(x_1, x_2, -x_1 x_3, -x_2 x_3) \qquad \forall x \in \mathcal{X} \tag{8.13}$$

and the function φ defined in (8.5), with L adapted to the properties of τ, if τ^* is defined by (8.3) ;

[6] The symbol $\varphi\mathcal{T}$ is pronounced *phitau*.

- or a continuous function τ satisfying

$$x = \tau\left(\frac{\lambda_1 x_1 - x_2}{\lambda_1^2 + x_3}, \frac{\lambda_2 x_1 - x_2}{\lambda_2^2 + x_3}, \frac{\lambda_3 x_1 - x_2}{\lambda_3^2 + x_3}, \frac{\lambda_4 x_1 - x_2}{\lambda_4^2 + x_3}\right) \qquad \forall x \in \mathcal{X} \tag{8.14}$$

and the function φ defined in (8.6) if τ^* is defined by (8.8). ▲

Although the problem of observer design seems already solved under Assumption $\mathcal{O}$, it can be difficult to find a left-inverse τ of τ^*. In the following, we consider that the function τ^* and the set $\mathcal{T}$ are given and we aim at avoiding the left-inversion of τ^* by expressing the observer for x in the, maybe augmented, x-coordinates.

8.2.2 A sufficient condition allowing the expression of the observer in the given x-coordinates

For the simpler case where the raw observer state $\hat{\xi}$ has the same dimension as the system state x, i.e. $d_x = d_\xi$, τ^*, in Assumption $\mathcal{O}$, is a diffeomorphism on $\mathcal{S}$ and we can express the observer in the given x-coordinates as :

$$\dot{\hat{x}} = \left(\frac{\partial \tau^*}{\partial x}(\hat{x})\right)^{-1} \varphi(\tau^*(\hat{x}), \hat{x}, u, y) \tag{8.15}$$

which requires a Jacobian inversion only. However, although, by assumption, the system trajectories remain in $\mathcal{S}$ where the Jacobian is invertible, we have no guarantee the ones of the observer do. Therefore, to obtain convergence and completeness of solutions, we must find means to ensure the estimate $\hat{x}$ does not leave the set $\mathcal{S}$, or equivalently that $\tau^*(\hat{x})$ remains in the image set $\tau^*(\mathcal{S})$. Observing that this problem obviously disappears if this set is the whole space $\mathbb{R}^{d_\xi}$, we address this point by modifying τ^* "marginally" in order to get $\tau^*(\mathcal{S}) = \mathbb{R}^{d_\xi}$.

In the more complex situation where $d_\xi > d_x$, τ^* is only an injective immersion. In [AEP14], it is proposed to augment the given x-coordinates in $\mathbb{R}^{d_x}$ with extra ones, say w, in $\mathbb{R}^{d_\xi - d_x}$ and correspondingly to augment the given injective immersion τ^* into a diffeomorphism $\tau_e^* : \mathcal{S}_a \to \mathbb{R}^{d_\xi}$, where $\mathcal{S}_a$ is an open subset of $\mathbb{R}^{d_\xi}$, which "augments" $\mathcal{S}$, i-e its Cartesian projection on $\mathbb{R}^{d_x}$ is contained in $\mathcal{S}$ and contains $\mathtt{cl}(\mathcal{X})$.

To help us find such an appropriate augmentation, we have the following sufficient condition.

Theorem 8.2.1.

Assume Assumption $\mathcal{O}$ holds and $\mathcal{X}$ is bounded. Assume also the existence of an open subset $\mathcal{S}_a$ of $\mathbb{R}^{d_\xi}$ containing $\mathtt{cl}(\mathcal{X} \times \{0\})$ and of a diffeomorphism $\tau_e^* : \mathcal{S}_a \to \mathbb{R}^{d_\xi}$ satisfying

$$\tau_e^*(x, 0) \;=\; \tau^*(x) \qquad \forall x \in \mathcal{X} \tag{8.16}$$

and

$$\tau_e^*(\mathcal{S}_a) = \mathbb{R}^{d_\xi} \ . \tag{8.17}$$

and such that, with let τ_{ex} denoting the x-component of the inverse of τ_e^*, there exists a function φ such that the pair (φ, τ_{ex}) is in the set $\mathcal{T}$ given by Assumption $\mathcal{O}$.

Under these conditions, for any u in $\mathcal{U}$ and any x_0 in $\mathcal{X}_0$ such that $\sigma^+(x_0, u) = +\infty$, any solution $(X(x_0; t; u), \hat{X}(x_0, \hat{x}_0, \hat{w}_0; t; u), \hat{W}(x_0, \hat{x}_0, \hat{w}_0; t, u))$, with initial condition $(\hat{x}_0, \hat{w}_0)$ in $\mathcal{S}_a$, of the cascade of System (8.9) with the observer :

$$\dot{\left[\begin{array}{c}\hat{x}\\ \hat{w}\end{array}\right]} = \left(\frac{\partial \tau_e^*}{\partial(\hat{x}, \hat{w})}(\hat{x}, \hat{w})\right)^{-1} \varphi(\tau_e^*(\hat{x}, \hat{w}), \hat{x}, u, y) \tag{8.18}$$

is also defined on $[0, +\infty)$ and satisfies :

$$\lim_{t\to+\infty} \left|\hat{W}(x_0, \hat{x}_0, \hat{w}_0; t; u)\right| + \left|X(x_0; t; u) - \hat{X}(x_0, \hat{x}_0, \hat{w}_0; t; u)\right| = 0 \ . \tag{8.19}$$

In other words, System (8.18) is an observer in the given coordinates[7] for System (8.9) initialized in $\mathcal{X}_0$.

The key point in the observer (8.18) is that, instead of left-inverting the function τ^* via τ as in (8.10), we invert only a matrix, exactly as in (8.15).

Proof : Take u in $\mathcal{U}$ and $(x_0, (\hat{x}_0, \hat{w}_0))$ in $\mathcal{X}_0 \times \mathcal{S}_a$ such that $\sigma^+(x_0, u) = +\infty$. $X(x_0; t; u)$ remains in $\mathcal{X}$ for t in $[0, +\infty)$ by assumption. Let $[0, \bar{t}[$ be the right maximal interval of definition of the solution $(X(x_0, t), \hat{X}(x_0, \hat{x}_0, \hat{w}_0; t; u), \hat{W}(x_0, \hat{x}_0, \hat{w}_0; t; u))$ when considered with values in $\mathcal{X} \times \mathcal{S}_a$. Assume for the time being $\bar{t}$ is finite. Then, when t goes to $\bar{t}$, either $(\hat{X}(x_0, \hat{x}_0, \hat{w}_0; t; u), \hat{W}(x_0, \hat{x}_0, \hat{w}_0; t; u))$ goes to infinity or to the boundary of $\mathcal{S}_a$. By construction $t \mapsto \Xi(t) := \tau_e^* \left(\hat{X}(\hat{x}_0, \hat{w}_0; t; u), \hat{W}(\hat{x}_0, \hat{w}_0; t; u)\right)$ is a solution of (8.11) on $[0, \bar{t}[$ with $\tau = \tau_{ex}$. From assumption $\mathcal{O}$ and since (φ, τ_{ex}) is in $\mathcal{F}$, it can be extended as a solution defined on $[0, +\infty[$ when considered with values in $\mathbb{R}^{d_\xi} = \tau_e^*(\mathcal{S}_a)$. This implies that $\Xi(\bar{t})$ is well defined in $\mathbb{R}^{d_\xi}$. Since, with (8.17), the inverse τ_e of τ_e^* is a diffeomorphism defined on $\mathbb{R}^{d_\xi}$, we obtain $\lim_{t\to\bar{t}} \left(\hat{X}(\hat{x}_0, \hat{w}_0; t; u), \hat{W}(\hat{x}_0, \hat{w}_0; t; u)\right) = \tau_e(\Xi(\bar{t}))$, which is an interior point of $\tau_e(\mathbb{R}^{d_\xi}) = \mathcal{S}_a$. This point being neither a boundary point nor at infinity, we have a contradiction. It follows that $\bar{t}$ is infinite.

Finally, with assumption $\mathcal{O}$, we have :

$$\lim_{t\to+\infty} \left|\tau_e^* \left(\hat{X}(\hat{x}_0, \hat{w}_0; t; u), \hat{W}(\hat{x}_0, \hat{w}_0; t; u)\right) - \tau^*(X(x_0; t; u))\right| = 0 \ .$$

Since $X(x_0; t; u)$ remains in $\mathcal{X}$, $\tau^*(X(x_0; t; u))$ equals $\tau_e^* \left(X(x_0; t; u), 0\right)$ and remains in the compact set $\tau^*(\mathtt{cl}(\mathcal{X}))$. So there exists a compact subset $\mathcal{C}$ of $\mathbb{R}^{d_\xi}$ and a time $t_\mathcal{C}$ such that $\tau_e^* \left(\hat{X}(\hat{x}_0, \hat{w}_0; t; u), \hat{W}(\hat{x}_0, \hat{w}_0; t; u)\right)$ is in $\mathcal{C}$ for all $t > t_\mathcal{C}$. Since τ_e^* is a diffeomorphism, its inverse τ_e is Lipschitz on the compact set $\mathcal{C}$. This implies (8.19). ■

With Theorem 8.2.1, we are left with finding a diffeomorphism τ_e^* satisfying the conditions listed in the statement :

- Equation (8.16) is about the fact that τ_e^* is an augmentation, with adding coordinates, of the given injective immersion τ^*. It motivates the following problem.

Problem 1. Immersion augmentation into a diffeomorphism

Given a set $\mathcal{X}$, an open subset $\mathcal{S}$ of $\mathbb{R}^{d_x}$ containing $\mathtt{cl}(\mathcal{X})$, and an injective immersion $\tau^* : \mathcal{S} \to \tau^*(\mathcal{S}) \subset \mathbb{R}^{d_\xi}$, the pair $(\tau_a^*, \mathcal{S}_a)$ is said to solve the problem of immersion augmentation into a diffeomorphism if $\mathcal{S}_a$ is an open subset of $\mathbb{R}^{d_\xi}$ containing $\mathtt{cl}(\mathcal{X} \times \{0\})$ and $\tau_a^* : \mathcal{S}_a \to \tau_a^*(\mathcal{S}_a) \subset \mathbb{R}^{d_\xi}$ is a diffeomorphism satisfying

$$\tau_a^*(x, 0) = \tau^*(x) \qquad \forall x \in \mathcal{X} \ .$$

We will present in Chapter 9 conditions under which Problem 1 can be solved via complementing a full column rank Jacobian of τ^* into an invertible matrix, i.e. via what we call Jacobian complementation.

- The condition expressed in (8.17), is about the fact that τ_e^* is surjective onto $\mathbb{R}^{d_\xi}$. This motivates us to introduce the surjective diffeomorphism extension problem

Problem 2. Surjective diffeomorphism extension

Given an open subset $\mathcal{S}_a$ of $\mathbb{R}^{d_\xi}$, a compact subset K of $\mathcal{S}_a$, and a diffeomorphism τ_a^*:

[7] See Definition 2.1.1.

extension problem if it satisfies

$$\tau_e^*(\mathcal{S}_a) = \mathbb{R}^{d_\xi} \quad , \qquad \tau_e^*(x,w) = \tau_a^*(x,w) \quad \forall (x,w) \in K.$$

This Problem 2 will be addressed in Chapter 10.

When Assumption $\mathcal{O}$ holds and $\mathcal{X}$ is bounded, by successively solving Problem 1 and Problem 2 with $\mathtt{cl}(\mathcal{X}\times\{0\}) \subset K \subset \mathcal{S}_a$, we get a diffeomorphism τ_e^* guaranteed to satisfy all the conditions of Theorem 8.2.1 except maybe the fact that the pair (φ, τ_{ex}) is in $\mathit{\varphi T}$. Fortunately, pairing a function φ with a function τ_{ex} obtained from a left inverse of τ_e^* is not as difficult as it seems, at least for general purpose observer designs such as high gain observers or nonlinear Luenberger observers. Indeed, we have already observed in point 3 of Remark 13 that if, as for Luenberger observers, there is a pair (φ, τ) in the set $\mathit{\varphi T}$ such that φ does not depend on τ, then we can associate this φ to any τ_{ex}. Also, for high gain observers, we need only that τ_{ex}, used as argument of φ, make it globally Lipschitz. This is obtained by modifying, if needed, this function outside a compact set, as the saturation function does in (8.5). We conclude from all this that our problem reduces to solving Problems 1 and 2.

Throughout Chapters 9-10, we will show how, step by step, we can express in the x-coordinates the high gain observer for the harmonic oscillator with unknown frequency introduced in Section 8.1.1. We will also show that our approach enables to ensure completeness of solutions of the observer presented in [GHO92] for a bioreactor. The various difficulties we shall encounter on this road will be discussed in Chapter 11. In particular, we shall see how they can be overcome thanks to a better choice of τ^* and of the pair (φ, τ) given by Assumption $\mathcal{O}$. We will also see that the same tools apply to the Luenberger observer presented in Section 8.1.2 for the oscillator. Finally, we will show in Chapter 11 that this methodology can be extended to the case where the transformation is time-varying through a very practical application related to aircraft landing.

Chapter 9

Around Problem 1 : augmenting an injective immersion into a diffeomorphism

Chapitre 9 – Autour du Problème 1 : augmenter une immersion injective en un difféomorphisme. *Une condition suffisante pour résoudre ce problème est de savoir compléter continûment le Jacobien (de rang plein) de la fonction en une matrice inversible. En effet, lorsque ceci est possible, une formule explicite de l'augmentation en un difféomorphisme est proposée. Ce chapitre est donc consacré au problème de complémentation continue d'une matrice rectangulaire de rang plein en une matrice carrée inversible. Plusieurs résultats sont donnés avec dans chaque cas des formules explicites ou des algorithmes constructifs, et sont illustrés grâce à l'exemple de l'oscillateur à fréquence inconnue.*

Contents

In [AEP14], we find the following sufficient condition for the augmentation of an immersion into a diffeomorphism.

Lemma 9.0.1. [AEP14]

Let $\mathcal{X}$ be a bounded set, $\mathcal{S}$ be an open subset of $\mathbb{R}^{d_x}$ containing $\mathtt{cl}(\mathcal{X})$, and $\tau^* : \mathcal{S} \to \tau^*(\mathcal{S}) \subset \mathbb{R}^{d_\xi}$ be an injective immersion. If there exists a bounded open set $\tilde{\mathcal{S}}$ satisfying

$$\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S}$$

and a C^1 function $\gamma : \mathcal{S} \to \mathbb{R}^{d_\xi \times (d_\xi - d_x)}$ the values of which are $d_\xi \times (d_\xi - d_x)$ matrices satisfying :

$$\det \left(\frac{\partial \tau^*}{\partial x}(x) \quad \gamma(x) \right) \neq 0 \qquad \forall x \in \mathtt{cl}(\tilde{\mathcal{S}}) \ , \tag{9.1}$$

then there exists a strictly positive real number ε such that the following pair[1] $(\tau_a^*, \mathcal{S}_a)$ solves Problem 1

$$\tau_a^*(x, w) \ = \ \tau^*(x) \ + \ \gamma(x)\, w \quad , \quad \mathcal{S}_a = \tilde{\mathcal{S}} \times B_\varepsilon(0) \ . \tag{9.2}$$

[1]For a positive real number ε and z_0 in $\mathbb{R}^p$, $B_\varepsilon(z_0)$ is the open ball centered at z_0 and with radius ε.

In other words, an injective immersion τ^* can be augmented into a diffeomorphism τ_a^* if we are able to find $d_\xi - d_x$ columns γ which are C^1 in x and which complement the full column rank Jacobian $\frac{\partial \tau^*}{\partial x}(x)$ into an invertible matrix.

Proof : The fact that τ_a^* is an immersion for ε small enough is established in [AEP14]. We now prove it is injective. Let ε_0 be a strictly positive real number such that the Jacobian of $\tau_a^*(x,w)$ in (9.2) is invertible for any (x,w) in $\mathtt{cl}(\tilde{\mathcal{S}} \times B_{\varepsilon_0}(0))$. Since $\mathtt{cl}(\tilde{\mathcal{S}} \times B_{\varepsilon_0}(0))$ is compact, not to contradict the Implicit function Theorem, there exists a strictly positive real number δ such that any two pairs (x_a, w_a) and (x_b, w_b) in $\mathtt{cl}(\tilde{\mathcal{S}} \times B_{\varepsilon_0}(0))$ which satisfy

$$\tau_a^*(x_a, w_a) \;=\; \tau_a^*(x_b, w_b) \quad , \quad (x_a, w_a) \;\neq\; (x_b, w_b) \tag{9.3}$$

satisfies also

$$|x_a - x_b| \;+\; |w_a - w_b| \;\geq\; \delta \; .$$

On another hand, since τ^* is continuous and injective on $\mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S}$, it has an inverse which is uniformly continuous on the compact set $\tau^*(\mathtt{cl}(\tilde{\mathcal{S}}))$ (see Lemma A.3.3). It follows that there exists a strictly positive real number η such that

$$\forall\, (x_a, x_b) \in \mathtt{cl}(\tilde{\mathcal{S}})^2 : |\tau^*(x_a) - \tau^*(x_b)| < \eta\,, \quad |x_a - x_b| < \frac{\delta}{2} \; .$$

But if (9.3) holds with w_a and w_b in $B_\varepsilon(0)$ with $\varepsilon \leq \varepsilon_0$, we have

$$\delta - 2\varepsilon \;\leq\; |x_a - x_b| \; , \quad |\tau^*(x_a) - \tau^*(x_b)| \;=\; |\gamma(x_a)w_a - \gamma(x_b)w_b| \;\leq\; 2\varepsilon \sup_{x \in \mathtt{cl}(\tilde{\mathcal{S}})} |\gamma(x)| \; .$$

We have a contradiction for all $\varepsilon \leq \min\left\{ \frac{3\delta}{4}, \frac{\eta}{2\varepsilon \sup_{x\in \mathtt{cl}(\tilde{\mathcal{S}})} |\gamma(x)|} \right\}$. So (9.3) cannot hold for such ε's, i.e. τ_a^* is injective on $\tilde{\mathcal{S}} \times B_\varepsilon(0)$. ■

Remark 14 Complementing a $d_\xi \times d_x$ full-rank matrix into an invertible one is equivalent to finding $d_\xi - d_x$ independent vectors orthogonal to that matrix. Precisely the existence of γ satisfying (9.1) is equivalent to the existence of a C^1 function $\tilde{\gamma} : \mathtt{cl}(\tilde{\mathcal{S}}) \to \mathbb{R}^{d_\xi \times (d_\xi - d_x)}$ the values of which are full rank matrices satisfying :

$$\tilde{\gamma}(x)^\top \frac{\partial \tau^*}{\partial x}(x) \;=\; 0 \qquad \forall x \in \mathtt{cl}(\tilde{\mathcal{S}}) \; . \tag{9.4}$$

Indeed, $\tilde{\gamma}$ satisfying (9.4) satisfies also (9.1) since the following matrices are invertible

$$\begin{pmatrix} \frac{\partial \tau^*}{\partial x}(x)^\top \\ \tilde{\gamma}(x)^\top \end{pmatrix} \left(\frac{\partial \tau^*}{\partial x}(x) \quad \tilde{\gamma}(x) \right) \;=\; \begin{pmatrix} \frac{\partial \tau^*}{\partial x}(x)^\top \frac{\partial \tau^*}{\partial x}(x) & 0 \\ 0 & \tilde{\gamma}(x)^\top \tilde{\gamma}(x) \end{pmatrix} \; .$$

Conversely, given γ satisfying (9.1), $\tilde{\gamma}$ defined by the identity below satisfies (9.4) and has full column rank

$$\tilde{\gamma}(x) = \left[I - \frac{\partial \tau^*}{\partial x}(x) \left[\frac{\partial \tau^*}{\partial x}(x)^\top \frac{\partial \tau^*}{\partial x}(x) \right]^{-1} \frac{\partial \tau^*}{\partial x}(x)^\top \right] \gamma(x) \; .$$

9.1 Submersion case

When $\tau^*(\mathtt{cl}(\tilde{\mathcal{S}}))$ is a level set of a submersion, we have the following complementation result :

Theorem 9.1.1.

Let $\mathcal{X}$ be a bounded set, $\tilde{\mathcal{S}}$ be a bounded open set and $\mathcal{S}$ be an open set satisfying

$$\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S} \; .$$

Let also $\tau^* : \mathcal{S} \to \tau^*(\mathcal{S}) \subset \mathbb{R}^{d_\xi}$ be an injective immersion. Assume there exists a C^2 function

$F : \mathbb{R}^{d_\xi} \to \mathbb{R}^{d_\xi - d_x}$ which is a submersion[2] at least on a neighborhood of $\tau^*(\tilde{\mathcal{S}})$ satisfying:

$$F(\tau^*(x)) = 0 \qquad \forall x \in \tilde{\mathcal{S}} \ , \tag{9.5}$$

then, with the C^1 function $x \mapsto \gamma(x) = \frac{\partial F}{\partial \xi}^T(\tau^*(x))$, the matrix in (9.1) is invertible for all x in $\tilde{\mathcal{S}}$ and the pair $(\tau_a^*, \mathcal{S}_a)$ defined in (9.2) solves Problem 1.

Proof : For all x in $\mathtt{cl}(\tilde{\mathcal{S}})$, $\frac{\partial \tau^*}{\partial x}(x)$ is right invertible and we have $\frac{\partial F}{\partial \xi}(\tau^*(x))\frac{\partial \tau^*}{\partial x}(x) = 0$. Thus, the rows of $\frac{\partial F}{\partial \xi}(\tau^*(x))$ are orthogonal to the column vectors of $\frac{\partial \tau^*}{\partial x}(x)$ and are independent since F is a submersion. The Jacobian of τ^* can therefore be completed with $\frac{\partial F}{\partial \xi}^T(\tau^*(x))$. The proof is completed with Lemma 9.0.1. ■

Remark 15 Since $\frac{\partial \tau^*}{\partial x}$ is of constant rank d_x on $\mathcal{S}$, the existence of such a function F is guaranteed at least locally by the constant rank Theorem.

Example 9.1.1 (Continuation of Example 8.2.1) Elimination of the $\hat{x}_i$ in the 4 equations given by the injective immersion τ^* defined in (8.3) leads to the function $F(\xi) = \xi_2\xi_3 - \xi_1\xi_4$ satisfying (9.5). It follows that a candidate for complementing:

$$\frac{\partial \tau^*}{\partial x}(x) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_3 & 0 & -x_1 \\ 0 & -x_3 & -x_2 \end{pmatrix} \tag{9.6}$$

is

$$\gamma(x) \ = \ \frac{\partial F}{\partial \xi}(\tau^*(x))^\top = (x_2x_3, -x_1x_3, x_2, -x_1)^\top \ .$$

This vector is nothing but the column of the minors of the matrix (9.6). It gives as determinant $(x_2x_3)^2 + (x_1x_3)^2 + x_2^2 + x_1^2$ which is never zero on $\mathcal{S}$.

Then, it follows from Lemma 9.0.1, that, for any bounded open set $\tilde{\mathcal{S}}$ such that $\mathcal{X} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S}$ the following function is a diffeomorphism on $\tilde{\mathcal{S}} \times B_\epsilon(0)$ for ε sufficiently small

$$\tau_a^*(x, w) = (x_1 + x_2x_3w, x_2 - x_1x_3w, -x_1x_3 + x_2w, -x_2x_3 - x_1w) \ .$$

With picking $\tau_e^* = \tau_a^*$, (8.18) gives us the following observer written in the given x-coordinates augmented with w :

$$\overbrace{\begin{pmatrix} \hat{x}_1 \\ \hat{x}_3 \\ \hat{x}_2 \\ \hat{w} \end{pmatrix}}^{\cdot} = \begin{pmatrix} 1 & \hat{x}_3\hat{w} & \hat{x}_2\hat{w} & \hat{x}_2\hat{x}_3 \\ -\hat{x}_3\hat{w} & 1 & -\hat{x}_1\hat{w} & -\hat{x}_1\hat{x}_3 \\ -\hat{x}_3 & \hat{w} & -\hat{x}_1 & \hat{x}_2 \\ -\hat{w} & -\hat{x}_3 & -\hat{x}_2 & -\hat{x}_1 \end{pmatrix}^{-1} \left[\begin{pmatrix} \hat{x}_2 - \hat{x}_1\hat{x}_3\hat{w} \\ -\hat{x}_1\hat{x}_3 + \hat{x}_2\hat{w} \\ -\hat{x}_2\hat{x}_3 - \hat{x}_1\hat{w} \\ \mathsf{sat}_{r^3}(\hat{x}_1\hat{x}_3^2) \end{pmatrix} + \begin{pmatrix} Lk_1 \\ L^2k_2 \\ L^3k_3 \\ L^4k_4 \end{pmatrix} [y - \hat{x}_1] \right] \tag{9.7}$$

Unfortunately the matrix to be inverted is non singular for $(\hat{x}, \hat{w})$ in $\tilde{\mathcal{S}} \times B_\varepsilon(0)$ only and we have no guarantee that the trajectories of this observer remain in this set. This shows that a further modification transforming τ_a^* into τ_e^* is needed to make sure that $\tau_e^{*-1}(\xi)$ belongs to this set whatever ξ in $\mathbb{R}^4$. This is Problem 2. ▲

The drawback of this Jacobian complementation method is that it asks for the knowledge of the function F. It would be better to simply have a universal formula relating the entries of the columns to be added to those of $\frac{\partial \tau^*}{\partial x}$.

[2] $F : \mathbb{R}^{d_\xi} \to \mathbb{R}^n$ with $d_\xi \geq n$ is a submersion on $\mathcal{V}$ if $\frac{\partial F}{\partial \xi}(\xi)$ is full-rank for all ξ in $\mathcal{V}$.

9.2 The $\tilde{P}[d_\xi, d_x]$ problem

Finding a universal formula for the Jacobian complementation problem amounts to solving the following problem.

Problem $\tilde{P}[d_\xi, d_x]$

For a pair of integers (d_ξ, d_x) such that $0 < d_x < d_\xi$, a C^1 matrix function $\tilde{\gamma} : \mathbb{R}^{m\times n} \to \mathbb{R}^{d_\xi \times (d_\xi - d_x)}$ solves the $\tilde{P}[d_\xi, d_x]$ problem if for any $d_\xi \times d_x$ matrix $\mathfrak{r} = (\mathfrak{r}_{ij})$ of rank d_x, the matrix $\begin{pmatrix} \mathfrak{r} & \tilde{\gamma}(\mathfrak{r}) \end{pmatrix}$ is invertible.

As a consequence of a theorem due to Eckmann [Eck06, §1.7 p. 126] and Lemma 9.0.1, we have

Theorem 9.2.1.

The $\tilde{P}[d_\xi, d_x]$ problem is solvable by a C^1 function $\tilde{\gamma}$ if and only if the pair (d_ξ, d_x) is in one of the following pairs

$$(\geq 2, d_\xi - 1) \qquad \text{or} \qquad (4, 1) \qquad \text{or} \qquad (8, 1)\ . \tag{9.8}$$

Moreover, for each of these pairs and for any bounded set $\mathcal{X}$, any bounded open set $\tilde{\mathcal{S}}$ and any open set $\mathcal{S}$ satisfying

$$\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S} \subset \mathbb{R}^{d_x}\ ,$$

and any injective immersion $\tau^* : \mathcal{S} \to \tau^*(\mathcal{S}) \subset \mathbb{R}^{d_\xi}$, the pair $(\tau_a^*, \mathcal{S}_a)$ defined in (9.2) with $\gamma(x) = \tilde{\gamma}\left(\frac{\partial \tau_a^*}{\partial x}(x)\right)$ solves Problem 1.

Proof : ["only if"] The following theorem is due to Eckmann.

Theorem 9.2.2. [Eck06]

For $d_\xi > d_x$, there exists a continuous function $\tilde{\gamma}_1 : \mathbb{R}^{d_\xi \times d_x} \to \mathbb{R}^{d_\xi}$ with non zero values and satisfying $\tilde{\gamma}_1(\mathfrak{r})^T \mathfrak{r} = 0$ for any $d_\xi \times d_x$ matrix $\mathfrak{r} = (\mathfrak{r}_{ij})$ of rank d_x if and only if (d_ξ, d_x) is in one of the following pairs

$$(\geq 2, d_\xi - 1) \qquad \text{or} \qquad (\text{even}, 1) \qquad \text{or} \qquad (7, 2) \qquad \text{or} \qquad (8, 3) \tag{9.9}$$

With Remark 14, any pair (d_ξ, d_x) for which $\tilde{P}[d_\xi, d_x]$ is solvable must be one in the list (9.9). The pair $(\geq 2, d_\xi - 1)$ is in the list (9.8). For the pair $(\text{even}, 1)$, we need to find $d_\xi - 1$ vectors to complement the given one into an invertible matrix. After normalizing the vector $\mathfrak{r}$ so that it belongs to the unit sphere $\mathbb{S}^{d_\xi - 1}$ and projecting each vector $\gamma_i(\mathfrak{r})$ of $\gamma(\mathfrak{r})$ onto the orthogonal complement of $\mathfrak{r}$, this complementation problem is equivalent to asking whether $\mathbb{S}^{d_\xi - 1}$ is parallelizable (since the $\gamma_i(\mathfrak{r})$ will be a basis for the tangent space at $\mathfrak{r}$ for each $\mathfrak{r} \in \mathbb{S}^{d_\xi - 1}$). It turns out that this problems admits solutions only for $d_\xi = 4$ or $d_\xi = 8$ (see [BM58]). So in the pairs $(\text{even}, 1)$ only $(4, 1)$ and $(8, 1)$ are in the list (9.8).

Finally, since $\tilde{P}[6, 1]$ has no solution, the pairs $(7, 2)$ and $(8, 3)$ cannot be in the list (9.8). Indeed, let $\mathfrak{r}$ be a full column rank $(d_\xi - 1) \times (d_x - 1)$ matrix. $\begin{pmatrix} \mathfrak{r} & 0 \\ 0 & 1 \end{pmatrix}$ is a full column rank $d_\xi \times d_x$ matrix. If if $\tilde{P}[d_\xi, d_x]$ has a solution, there exist a continuous $(d_\xi - 1) \times (d_\xi - d_x)$ matrix function $\tilde{\gamma}$ and a continuous row vector functions a^T such that $\begin{pmatrix} \tilde{\gamma}(\mathfrak{r}) & \mathfrak{r} & 0 \\ a(\mathfrak{r})^\top & 0 & 1 \end{pmatrix}$ is invertible. This implies that $\begin{pmatrix} \tilde{\gamma}(\mathfrak{r}) & \mathfrak{r} \end{pmatrix}$ is also invertible. So if $\tilde{P}[d_\xi, d_x]$ has a solution, $\tilde{P}[d_\xi - 1, d_x - 1]$ must have one. ■

[2] See Remark 14.

Proof : ["if"] For (d_ξ, d_x) equal to $(4,1)$ or $(8,1)$ respectively, possible solutions are

$$\tilde{\gamma}(\mathfrak{T}) = \begin{pmatrix} -\mathfrak{T}_2 & \mathfrak{T}_3 & \mathfrak{T}_4 \\ \mathfrak{T}_1 & -\mathfrak{T}_4 & \mathfrak{T}_3 \\ -\mathfrak{T}_4 & -\mathfrak{T}_1 & -\mathfrak{T}_2 \\ \mathfrak{T}_3 & \mathfrak{T}_2 & -\mathfrak{T}_1 \end{pmatrix} \quad , \quad \tilde{\gamma}(\mathfrak{T}) = \begin{pmatrix} \mathfrak{T}_2 & \mathfrak{T}_3 & \mathfrak{T}_4 & \mathfrak{T}_5 & \mathfrak{T}_6 & \mathfrak{T}_7 & \mathfrak{T}_8 \\ -\mathfrak{T}_1 & \mathfrak{T}_4 & -\mathfrak{T}_3 & \mathfrak{T}_6 & -\mathfrak{T}_5 & -\mathfrak{T}_8 & \mathfrak{T}_7 \\ -\mathfrak{T}_4 & -\mathfrak{T}_1 & \mathfrak{T}_2 & \mathfrak{T}_7 & \mathfrak{T}_8 & -\mathfrak{T}_5 & -\mathfrak{T}_6 \\ \mathfrak{T}_3 & -\mathfrak{T}_2 & -\mathfrak{T}_1 & \mathfrak{T}_8 & -\mathfrak{T}_7 & \mathfrak{T}_6 & -\mathfrak{T}_5 \\ -\mathfrak{T}_6 & -\mathfrak{T}_7 & -\mathfrak{T}_8 & -\mathfrak{T}_1 & \mathfrak{T}_2 & \mathfrak{T}_3 & \mathfrak{T}_4 \\ \mathfrak{T}_5 & -\mathfrak{T}_8 & \mathfrak{T}_7 & -\mathfrak{T}_2 & -\mathfrak{T}_1 & -\mathfrak{T}_4 & \mathfrak{T}_3 \\ \mathfrak{T}_8 & \mathfrak{T}_5 & -\mathfrak{T}_6 & -\mathfrak{T}_3 & \mathfrak{T}_4 & -\mathfrak{T}_1 & -\mathfrak{T}_2 \\ -\mathfrak{T}_7 & \mathfrak{T}_6 & \mathfrak{T}_5 & -\mathfrak{T}_4 & -\mathfrak{T}_3 & \mathfrak{T}_2 & -\mathfrak{T}_1 \end{pmatrix}$$

where $\mathfrak{T}_j$ is the jth component of the vector $\mathfrak{T}$. For $d_x = d_\xi - 1$, we have the identity

$$\det\begin{pmatrix} \mathfrak{T} & \tilde{\gamma}(\mathfrak{T}) \end{pmatrix} = \sum_{j=1}^{m} \tilde{\gamma}_j(\mathfrak{T}_{ij})\, M_{j,m}(\mathfrak{T}_{ij})$$

where $\tilde{\gamma}_j$ is the jth component of the vector-valued function $\tilde{\gamma}$ and the $M_{j,m}$, being the cofactors of $(\mathfrak{T}\ \tilde{\gamma}(\mathfrak{T}))$ computed along the last column, are polynomials in the given components $\mathfrak{T}_{ij}$. At least one of the $M_{j,m}$ is non-zero (because they are minors of dimension d_x of $\mathfrak{T}$ which is full-rank). So it is sufficient to take $\tilde{\gamma}_j(\mathfrak{T}_{ij}) = M_{j,m}(\mathfrak{T}_{ij})$. ■

In the following example we show how by exploiting some structure we can reduce the problem to one of these 3 pairs.

Example 9.2.1 (Continuation of Example 9.1.1) In Example 9.1.1, we have complemented the Jacobian (9.6) with the gradient of a submersion and observed that the components of this gradient are actually cofactors. We now know that this is consistent with the case $d_x = d_\xi - 1$. But we can also take advantage from the upper triangularity of the Jacobian (9.6) and complement only the vector $(-x_1, -x_2)$ by for instance $(x_2, -x_1)$. The corresponding vector γ is $\gamma(x) = (0, 0, x_2, -x_1)$. Here again, with Lemma 9.0.1, we know that, for any bounded open set $\tilde{\mathcal{S}}$ such that $\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S}$ the function

$$\tau_a^*(x, w) \ = \ (x_1\,,\ x_2\,,\ -x_1x_3 + x_2w\,,\ -x_2x_3 - x_1w)$$

is a diffeomorphism on $\tilde{\mathcal{S}} \times B_\epsilon(0)$. In fact, in this particular case ε can be arbitrary since the Jacobian of τ_a^* is full rank on $\tilde{\mathcal{S}} \times \mathbb{R}^{d_\xi - d_x}$. With picking $\tau_e^* = \tau_a^*$, (8.18) gives us the following observer :

$$\overbrace{\begin{pmatrix} \hat{x}_1 \\ \hat{x}_3 \\ \hat{x}_2 \\ \hat{w} \end{pmatrix}}^{\dot{}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\hat{x}_3 & \hat{w} & -\hat{x}_1 & \hat{x}_2 \\ -\hat{w} & -\hat{x}_3 & -\hat{x}_2 & -\hat{x}_1 \end{pmatrix}^{-1} \left[\begin{pmatrix} \hat{x}_2 \\ -\hat{x}_1\hat{x}_3 + \hat{x}_2\hat{w} \\ -\hat{x}_2\hat{x}_3 - \hat{x}_1\hat{w} \\ \mathsf{sat}_{r^3}(\hat{x}_1\hat{x}_3^2) \end{pmatrix} + \begin{pmatrix} Lk_1 \\ L^2k_2 \\ L^3k_3 \\ L^4k_4 \end{pmatrix} [y - \hat{x}_1] \right] \tag{9.10}$$

However, the singularity at $\hat{x}_1 = \hat{x}_2 = 0$ remains and equation (8.17) is still not satisfied. ▲

Given the very small number of cases where a universal formula exists, we now look for a more general solution to the Jacobian complementation problem.

9.3 Wazewski's theorem

Historically, the Jacobian complementation problem was first addressed by Wazewski in [Waz35]. His formulation was :

Wazewski's problem

> Given a continuous function $\mathfrak{T} : \mathcal{S} \subset \mathbb{R}^{d_x} \to \mathbb{R}^{d_\xi \times d_x}$, the values of which are full-rank $d_\xi \times d_x$ matrices, look for a continuous function $\gamma : \mathcal{S} \to \mathbb{R}^{d_\xi \times (d_\xi - d_x)}$ such that the matrix $\begin{pmatrix} \mathfrak{T}(x) & \gamma(x) \end{pmatrix}$ is invertible for all x in $\mathcal{S}$.

The difference with the previous section, is that here, we look for a continuous function γ of the argument x of $\mathfrak{x}(x)$ instead of continuous functions of $\mathfrak{x}$ itself.

Wazewski established that this other version of the problem admits a far more general solution :

Theorem 9.3.1. [Waz35, Theorems 1 and 3]

If $\mathcal{S}$, equipped with the subspace topology of $\mathbb{R}^{d_x}$, is a contractible space, then Wazewski's problem admits a solution. Besides, the function γ can be chosen $\mathcal{C}^\infty$ on $\mathcal{S}$.

Proof : The reader is referred to [Eck06, page 127] or [Dug66, pages 406-407] and to [Waz35, Theorems 1 and 3] for the complete proof of existence of a continuous function γ when $\mathcal{S}$ is contractible. We rather detail here the constructive main points of the proof originally given by Wazewski in the particular case where $\mathcal{S}$ is a parallelepiped, because it gives an insight on the explicit construction of γ. It is based on Remark 14, noting that, if we have the decomposition

$$\mathfrak{x}(x) = \begin{pmatrix} A(x) \\ B(x) \end{pmatrix}$$

with $A(x)$ invertible on some given subset $\mathcal{R}$ of $\mathcal{S}$, then

$$\gamma(x) = \begin{pmatrix} C(x) \\ D(x) \end{pmatrix}$$

makes $\begin{pmatrix} \mathfrak{x}(x) & \gamma(x) \end{pmatrix}$ invertible on $\mathcal{R}$ if and only if $D(x)$ is invertible on $\mathcal{R}$ and we have

$$C(x) = -(A^T(x))^{-1}B(x)^T D(x) \qquad \forall x \in \mathcal{R} \ . \tag{9.11}$$

Thus, C is imposed by the choice of D and choosing D invertible is enough to build γ on $\mathcal{R}$.

Also, if we already have a candidate

$$\begin{pmatrix} A(x) & C_0(x) \\ B(x) & D_0(x) \end{pmatrix}$$

on a boundary $\partial\mathcal{R}$ of $\mathcal{R}$ and $A(x)$ is invertible for all x in $\partial\mathcal{R}$, then, necessarily, $D_0(x)$ is invertible and $C_0(x) = -(A^T(x))^{-1}B(x)^T D_0(x)$ all x in $\partial\mathcal{R}$. Thus, to extend the construction of a continuous function γ inside $\mathcal{R}$ from its knowledge on the boundary $\partial\mathcal{R}$, it suffices to pick D as any invertible matrix satisfying $D = D_0$ on $\partial\mathcal{R}$. Because we can propagate continuously γ from one boundary to the other, Wazewski deduces from these two observations that, it is sufficient to partition the set $\mathcal{S}$ into adjacent sets $\mathcal{R}_i$ where a given $d_\xi \times d_\xi$ minor A_i is invertible. This is possible since $\mathfrak{x}$ is full-rank on $\mathcal{S}$. When $\mathcal{S}$ is a parallelepiped, he shows that there exists an ordering of the $\mathcal{R}_i$ such that the continuity of each D_i can be successively ensured. We illustrate this construction in Example 9.3.1 below.

Finally, it remains to show how this continuous function γ can be modified into a smoother one giving the same invertibility property. For this, we use a partition of unity. Let γ_i denote the ith column of γ. We start with modifying γ_1 into $\tilde{\gamma}_1$. Since $\mathfrak{x}$, γ and the determinant are continuous, for any x in $\mathcal{S}$, there exists a strictly positive real number r_x, such that, may be after changing γ_1 into $-\gamma_1$,

$$\det\left(\mathfrak{x}(y)\ \gamma_1(x)\ \gamma_{2:d_\xi - d_x}(y)\right) > 0 \ , \qquad \forall y \in B_{r_x}(x) \ , \tag{9.12}$$

where $\gamma_{i:j}$ denotes the matrix composed of the i^{th} to j^{th} columns of γ. The family of sets $(B_{r_x}(x))_{x\in\mathcal{S}}$ is an open cover of $\mathcal{S}$. Therefore, by [Hir76, Theorem 2.1], there exists a subordinate C^∞ partition of unity, i.e. there exist a family of C^∞ functions $\psi_x : \mathcal{S} \to \mathbb{R}_{\geq 0}$ such that

$$\texttt{Supp}\,(\psi_x) \subset B_{r_x}(x) \quad \forall x \in \mathcal{S} \ , \tag{9.13}$$

$$\{\texttt{Supp}\,(\psi_x)\}_{x\in\mathcal{S}} \text{ is locally finite} \ , \tag{9.14}$$

$$\sum_{x\in\mathcal{S}} \psi_x(y) = 1 \quad \forall y \in \mathcal{S} \quad . \tag{9.15}$$

With this, we define the function $\tilde{\gamma}_1$ on $\mathcal{S}$ by

$$\tilde{\gamma}_1(y) = \sum_{x\in\mathcal{S}} \psi_x(y)\gamma_1(x) \ .$$

This function is well-defined and C^∞ on $\mathcal{S}$ because the sum is finite at each point according to (9.14). Using multi-linearity of the determinant, we have, for all y in $\mathcal{S}$,

$$\det\left(\mathfrak{r}(y)\ \tilde{\gamma}_1(y)\ \gamma_{2:d_\xi-d_x}(y)\right) = \sum_{x\in\mathcal{S}} \psi_x(y) \det\left(\mathfrak{r}(y)\ \gamma_1(x)\ \gamma_{2:d_\xi-d_x}(y)\right) \ .$$

Thanks to (9.14), at each point y in $\mathcal{S}$, there is a finite number of $\psi_x(y)$ which are not zero. Also, the right hand side is the sum of non negative terms because of (9.12) and the non negativeness of the ψ_x, and one of these terms is strictly positive because of (9.12) and (9.15). Therefore, we can replace the continuous function γ_1 by the C^∞ function $\tilde{\gamma}_1$ as a first column of γ. Then we follow exactly the same procedure for γ_2 with this modified γ. By proceeding this way, one column after the other, we get our result. ■

The following corollary is a consequence of Lemma 9.0.1 and provides another answer to Problem 1.

Corollary 9.3.1.

Let $\mathcal{X}$ be a bounded set, $\mathcal{S}$ be an open subset of $\mathbb{R}^{d_x}$ containing $\mathtt{cl}(\mathcal{X})$ and which, equipped with the subspace topology of $\mathbb{R}^{d_x}$, is a contractible space. Let also $\tau^* : \mathcal{S} \to \tau^*(\mathcal{S}) \subset \mathbb{R}^{d_\xi}$ be an injective immersion. There exists a C^∞ function γ such that, for any bounded open set $\tilde{\mathcal{S}}$ satisfying

$$\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S}$$

we can find a strictly positive real number ε such that the pair $(\tau_a^*, \mathcal{S}_a)$ defined in (9.2) solves Problem 1.

Example 9.3.1 Consider the function

$$\mathfrak{r}(x) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_3 & 0 & -x_1 \\ 0 & -x_3 & -x_2 \\ \frac{\partial\wp}{\partial x_1}x_3 & \frac{\partial\wp}{\partial x_2}x_3 & \wp \end{pmatrix} \quad , \qquad \wp(x_1,x_2) = \max\left\{0, \frac{1}{r^2} - (x_1^2+x_2^2)\right\}^4 .$$

$\mathfrak{r}(x)$ has full rank 3 for any x in $\mathbb{R}^3$, since $\wp(x_1,x_2) \neq 0$ when $x_1 = x_2 = 0$. To follow Wazewski's construction, let δ be a strictly positive real number and consider the following 5 regions of $\mathbb{R}^3$ (see Figure 9.1)

$$\mathcal{R}_1 =]-\infty, -\delta] \times \mathbb{R}^2 \quad , \qquad \mathcal{R}_2 = [-\delta,\delta] \times [\delta, +\infty] \times \mathbb{R},$$
$$\mathcal{R}_3 = [-\delta,\delta]^2 \times \mathbb{R} \quad , \qquad \mathcal{R}_4 = [-\delta,\delta] \times [-\infty, -\delta] \times \mathbb{R} \quad , \qquad \mathcal{R}_5 = [\delta, +\infty[\times \mathbb{R}^2.$$

We select δ sufficiently small in such a way that $\wp$ is not 0 in $\mathcal{R}_3$.

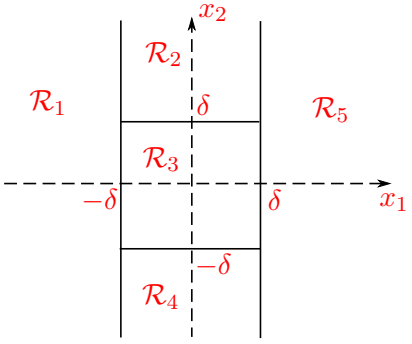


Figure 9.1: Projections of the regions $\mathcal{R}_i$ on $\mathbb{R}^2$.

We start Wazewski's algorithm in $\mathcal{R}_3$. Here, the invertible minor A is given by rows 1, 2 and 5 of $\mathfrak{T}$ (full-rank lines of $\mathfrak{T}$) and B by rows 3 and 4. With picking D as the identity, C is $(A^T)^{-1}B$ according to (9.11). D gives rows 3 and 4 of γ and C gives rows 1, 2 and 5 of γ.
Then we move to the region $\mathcal{R}_2$. There the matrix A is given by rows 1, 2 and 4 of $\mathfrak{T}$, B by rows 3 and 5. Also D, along the boundary between $\mathcal{R}_3$ and $\mathcal{R}_2$, is given by rows 3 and 5 of γ obtained in the previous step. We extrapolate this inside $\mathcal{R}_2$ by keeping D constant in planes $x_1 =$constant. An expression for C and therefore for γ follows.
We do exactly the same thing for $\mathcal{R}_4$.
Then we move to the region $\mathcal{R}_1$. There the matrix A is given by rows 1, 2 and 3 of $\mathfrak{T}$, B by rows 4 and 5. Also D, along the boundary between $\mathcal{R}_1$ and $\mathcal{R}_2$, between $\mathcal{R}_1$ and $\mathcal{R}_3$ and between $\mathcal{R}_1$ and $\mathcal{R}_4$, is given by rows 4 and 5 of γ obtained in the previous steps. We extrapolate this inside $\mathcal{R}_1$ by kipping D constant in planes $x_2 =$constant. An expression for C and therefore for γ follows.
We do exactly the same thing for $\mathcal{R}_5$.

Note that this construction produces a continuous γ, but we could have extrapolated D in a smoother way to obtain γ as smooth as necessary. ▲

Although Wazewski's method provides a more general answer to the problem of Jacobian complementation than the few solvable $\tilde{P}[d_\xi, d_x]$ problems, the explicit expressions of γ given in Section 9.2 are preferred in practice (when the couple (d_ξ, d_x) is in the list (9.8)) to Wazewski's costly computations.

We have given several methods to solve Problem 1, but to apply Theorem 8.2.1, we also need to solve Problem 2.

Chapter 10

Around Problem 2 : image extension of a diffeomorphism

Chapitre 10 – Autour du Problème 2 : extension d'image d'un difféomorphisme. *Dans ce chapitre, nous étudions comment un difféomorphisme peut être étendu pour que son image couvre l'espace $\mathbb{R}^{d_\xi}$ entier, c'est-à-dire pour qu'il devienne surjectif. Dans certains cas, la construction de l'extension est explicite et est illustrée à partir d'exemples. En particulier, nous montrons que la résolution du Problème 2 garantie la complétude des solutions de l'observateur présenté dans [GHO92] pour un bioréacteur.*

Contents

We study now how a diffeomorphism can be augmented to make its image be the whole set $\mathbb{R}^{d_\xi}$, i.e. to make it surjective. In certain cases, the construction of the extension is explicit and is illustrated on examples. In particular, we show that solving Problem 2 guarantees completeness of solutions of the observer presented in [GHO92] for a bioreactor.

10.1 A sufficient condition

There is a rich literature reporting very advanced results on the diffeomorphism extension problem. In the following some of the techniques are inspired from [Hir76, Chapter 8] and [Mil65, pages 2, 7 to 14 and 16 to 18](among others). Here we are interested in the particular aspect of this topic which is the diffeomorphism image extension as described by Problem 2. A very first necessary condition about this problem is in the following remark.

Remark 16 Since τ_e^*, obtained solving Problem 2, makes the set $\mathcal{S}$ diffeomorphic to $\mathbb{R}^{d_\xi}$, $\mathcal{S}$ must be contractible.

One of the key technical property which will allow us to solve Problem 2 can be phrased as follows.

Property $\mathfrak{C}$

An open subset E of $\mathbb{R}^{d_\xi}$ is said to verify property $\mathfrak{C}$ if there exist a C^1 function $\kappa : \mathbb{R}^{d_\xi} \to \mathbb{R}$, a bounded[1] C^1 vector field χ, and a closed set K_0 contained in E such that:

1. $E = \left\{ z \in \mathbb{R}^{d_\xi} \ : \ \kappa(z) < 0 \right\}$
2. K_0 is globally attractive for χ
3. we have the following transversality property:

$$\frac{\partial \kappa}{\partial z}(z)\chi(z) < 0 \qquad \forall z \in \mathbb{R}^{d_\xi} : \ \kappa(z) = 0.$$

The two main ingredients of this condition are the function κ and the vector field χ which, both, have to satisfy the transversality property $\mathfrak{C}.3$. In the case where only the function κ is given satisfying $\mathfrak{C}.1$ and with no critical point on the boundary of E, its gradient could play the role of χ. But then for K_0 to be globally attractive we need at least to remove all the possible critical points that κ could have outside K_0. This task is performed for example on Morse functions in the proof of the h-Cobordism Theorem [Mil65]. We are in a much simpler situation when χ is given and makes E forward invariant.

Lemma 10.1.1.

Let E be a bounded open subset of $\mathbb{R}^{d_\xi}$, χ be a bounded C^1 vector field , and K_0 be a compact set contained in E such that:

1. K_0 is globally asymptotically stable for χ
2. E is forward invariant for χ.

For any strictly positive real number $\overline{d}$, there exists a bounded set $\mathcal{E}$ such that

$$\mathtt{cl}(E) \subset \mathcal{E} \subset \{z \in \mathbb{R}^{d_\xi}, \ \inf_{z_E \in E} |z - z_E| \leq \overline{d}\}$$

and $\mathcal{E}$ verifies Property $\mathfrak{C}$.

This Lemma roughly says that if E does not satisfy conditions $\mathfrak{C}.1$ or $\mathfrak{C}.3$ but is forward invariant for χ, then Condition $\mathfrak{C}$ is satisfied by an arbitrarily close superset of E. Its proof is given in Appendix B.1.

Our main result on the diffeomorphism image extension problem is:

Theorem 10.1.1.

Let $\mathcal{S}_a$ be an open subset of $\mathbb{R}^{d_\xi}$ and τ_a^*: $\mathcal{S}_a \to \mathbb{R}^{d_\xi}$ be a diffeomorphism. If
a) either $\tau_a^*(\mathcal{S}_a)$ verifies property $\mathfrak{C}$,
b) or $\mathcal{S}_a$ is C^2-diffeomorphic to $\mathbb{R}^{d_\xi}$ and τ_a^* is C^2,
then for any compact set K in $\mathcal{S}_a$, there exists a diffeomorphism $\tau_e^* : \mathcal{S}_a \to \mathbb{R}^{d_\xi}$ solving Problem 2.

The proof of case a) of this theorem is given in Section 10.2. It provides an explicit construction of τ_e^*. The proof of case b) can be found in Appendix B.3. For the time being, we observe that a direct consequence is :

[1] If not replace χ by $\frac{\chi}{\sqrt{1+|\chi|^2}}$.

Corollary 10.1.1.

Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^{d_x}$, $\mathcal{S}_a$ be an open subset of $\mathbb{R}^{d_\xi}$ containing $K = \mathtt{cl}(\mathcal{X} \times \{0\})$ and $\tau_a^* : \mathcal{S}_a \to \tau_a^*(\mathcal{S}_a)$ be a diffeomorphism such that
a) either $\tau_a^*(\mathcal{S}_a)$ verifies property $\mathfrak{C}$,
b) or $\mathcal{S}_a$ is C^2-diffeomorphic to $\mathbb{R}^{d_\xi}$ and τ_a^* is C^2.
Then, there exists a diffeomorphism $\tau_e^* : \mathcal{S}_a \to \mathbb{R}^{d_\xi}$, such that

$$\tau_e^*(\mathcal{S}_a) = \mathbb{R}^{d_\xi} \quad , \quad \tau_e^*(x, 0) \; = \; \tau_a^*(x, 0) \qquad \forall x \in \mathcal{X} \; .$$

Thus, if besides the pair $(\tau_a^*, \mathcal{S}_a)$ solves Problem 1, then $(\tau_e^*, \mathcal{S}_a)$ solves Problems 1 and 2.

10.2 Proof of part a) of Theorem 10.1.1

We have the following technical lemma :

Lemma 10.2.1.

Let E be an open strict subset of $\mathbb{R}^{d_\xi}$ verifying Condition $\mathfrak{C}$. For any closed subset K of E, lying at a strictly positive distance of the boundary of E, there exists a diffeomorphism ϕ: $\mathbb{R}^{d_\xi} \to E$, such that ϕ is the identity function on K.

A constructive proof of this lemma is given in Appendix B.2 and provides an explicit expression for ϕ which will be used in Example 10.2.1 and Section 10.3. Its construction is illustrated on Figure 10.1.

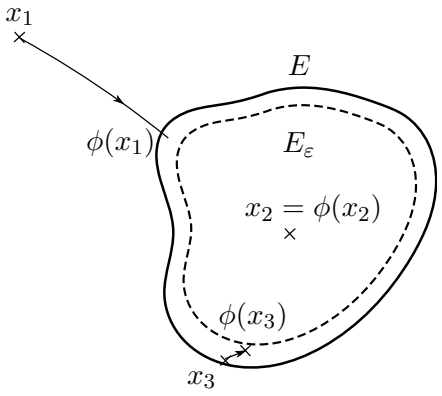


Figure 10.1: Sketch of the construction of the diffeomorphism ϕ in Lemma 10.2.1 : one follow the flow κ given by Condition $\mathfrak{C}$ for a more or less long time depending on the initial point. E_ε denotes the set where ϕ is the identity. ε measuring the "width" of $E \setminus E_\varepsilon$ can be chosen sufficiently small for K to be included in E_ε.

In the case a) of Theorem 10.1.1, we suppose that $\tau_a^*(\mathcal{S}_a)$ satisfies $\mathfrak{C}$. Now, τ_a^* being a diffeomorphism on an open set $\mathcal{S}_a$, the image of any compact subset K of $\mathcal{S}_a$ is a compact subset of $\tau_a^*(\mathcal{S}_a)$. According to Lemma 10.2.1, there exists a diffeomorphism ϕ from $\mathbb{R}^{d_\xi}$ to $\tau_a^*(\mathcal{S}_a)$ which is the identity on $\tau_a^*(K)$. Thus, the function $\tau_e^* \; = \; \phi^{-1} \circ \tau_a^*$ solves Problem 2 and the theorem is proved.

Example 10.2.1 (Continuation of Example 9.1.1) In Example 9.1.1, we have introduced the function

$$F(\xi) = \xi_2\xi_3 - \xi_1\xi_4 \; \triangleq \; \frac{1}{2}\xi^\top M \xi$$

as a submersion on $\mathbb{R}^4\backslash\{0\}$ satisfying

$$F(\tau^*(x)) = 0, \tag{10.1}$$

where τ^* is the injective immersion given in (8.3). With it we have augmented τ^* as

$$\tau_a^*(x,w) = \tau^*(x) + \frac{\partial F}{\partial \xi}^T(\tau^*(x))\, w = \tau^*(x) + M\tau^*(x)\, w$$

which is a diffeomorphism on $\mathcal{S}_a = \tilde{\mathcal{S}}\times]-\varepsilon,\varepsilon[$ for some strictly positive real number ε.

To modify τ_a^* in τ_e^* satisfying $\tau_e^*(\mathcal{S}_a) = \mathbb{R}^4$, we let K be the compact set

$$K = \mathtt{cl}(\tau_a^*(\mathcal{X}\times\{0\})) \ \subset \tau_a^*(\mathcal{S}_a) \subset \mathbb{R}^4\ .$$

With Lemma 10.2.1, we know that, if $\tau_a^*(\mathcal{S}_a)$ verifies property $\mathfrak{C}$, there exists a diffeomorphism ϕ defined on $\mathbb{R}^4$ such that ϕ is the identity function on the compact set K and $\phi(\mathbb{R}^4) = \tau_e^*(\mathcal{S}_a)$. In that case, as seen above, the diffeomorphism $\tau_e^* = \phi^{-1}\circ\tau_a^*$ defined on $\mathcal{S}_a$ is such that $\tau_e^* = \tau_a^*$ on $\mathcal{X}\times\{0\}$ and $\tau_e^*(\mathcal{S}_a) = \mathbb{R}^4$, i-e would be a solution to Problems 1 and 2. Unfortunately this is impossible. Indeed, due to the observability singularity at $x_1 = x_2 = 0$, $\tilde{\mathcal{S}}$ (and thus $\mathcal{S}_a$) is not contractible. Therefore, there is no diffeomorphism τ_e^* such that $\tau_e^*(\mathcal{S}_a) = \mathbb{R}^4$. We will see in Section 11.1 how this problem can be overcome. For the time being, we show that it is still possible to find τ_e^* such that $\tau_e^*(\mathcal{S}_a)$ covers "almost all" $\mathbb{R}^4$. The idea is to find an approximation E of $\tau_a^*(\mathcal{S}_a)$ verifying property $\mathfrak{C}$ and apply the same method on E. Indeed, if E is close enough to $\tau_a^*(\mathcal{S}_a)$, one can expect to have $\tau_e^*(\mathcal{S}_a)$ "almost equal to" $\mathbb{R}^4$.

With (10.1) and since $M^2 = I$, we have,

$$F(\tau_a^*(x,w)) = |\tau^*(x)|^2\, w\ .$$

Since $\mathcal{S}_a$ is bounded, there exists $\delta > 0$ such that the set

$$E = \left\{\xi\in\mathbb{R}^4 :\ F(\xi)^2 < \delta\right\}$$

contains $\tau_a^*(\mathcal{S}_a)$ and thus the compact set K. Let us show that E verifies property $\mathfrak{C}$. We pick

$$\kappa(\xi) \ = \ F(\xi)^2 - \delta \ = \ \left(\frac{1}{2}\xi^T M\xi\right)^2 - \delta\ .$$

and consider the vector field χ

$$\chi(\xi) = -2\frac{\partial\kappa}{\partial\xi}(\xi) = -[\xi^T M\xi]\, M\xi \qquad \text{or more simply} \qquad \chi(\xi) = -\xi\ .$$

The latter implies the transversality property $\mathfrak{C}.3$ is verified. Besides, the closed set $K_0 = \{0\}$ is contained in E and is globally attractive for the vector field χ.

Then Lemma 10.2.1 gives the existence of a diffeomorphism $\phi : \mathbb{R}^4 \to E$ which is the identity on K and verifies $\phi(\mathbb{R}^4) = E$. We obtain an expression of ϕ by following the constructive proof of this Lemma (see Appendix B.2). Let E_ε be the set

$$E_\varepsilon = \left\{\xi\in\mathbb{R}^4 :\ \left(\frac{1}{2}\xi^T M\xi\right)^2 < e^{-4\varepsilon}\,\delta\right\}\ .$$

It contains K. Let also $\nu : [-\varepsilon,+\infty[\to\mathbb{R}$ and $\mathfrak{t} : \mathbb{R}^4\setminus E_\varepsilon \to \mathbb{R}$ be the functions defined as

$$\nu(t) \ = \ \frac{(t+\varepsilon)^2}{2\varepsilon + t} \quad , \qquad \mathfrak{t}(\xi) = \frac{1}{4}\ln\frac{\left(\frac{1}{2}\xi^T M\xi\right)^2}{\delta}\ . \tag{10.2}$$

$\mathfrak{t}(\xi)$ is the time that a solution of $\dot{\xi} = \chi(\xi) = -\xi$ with initial condition ξ needs to reach the boundary of E i.e. $e^{-\mathfrak{t}(\xi)}\xi$ belongs to the boundary of E. From the proof Lemma 10.2.1, we know the function $\phi : \mathbb{R}^4 \to E$ defined as :

$$\phi(\xi) = \begin{cases} \xi & , \quad \text{if } \left(\frac{1}{2}\xi^T M \xi\right)^2 \leq e^{-4\varepsilon}\delta, \\ e^{-\nu(\mathfrak{t}(\xi))}\xi & , \quad \text{otherwise,} \end{cases} \tag{10.3}$$

is a diffeomorphism $\phi : \mathbb{R}^4 \to E$ which is the identity on K and verifies $\phi(\mathbb{R}^4) = E$.

As explained above, we use ϕ to replace τ_a^* by the diffeomorphism $\tau_e^* = \phi^{-1} \circ \tau_a^*$ also defined on $\mathcal{S}_a$. But, because $\tau_a^*(\mathcal{S}_a)$ is a strict subset of E, $\tau_e^*(\mathcal{S}_a)$ is a strict subset of $\mathbb{R}^4$, i.e. equation (8.17) is not satisfied. Nevertheless, for any trajectory of the observer $t \mapsto \hat{\xi}(t)$ in $\mathbb{R}^4$, our estimate defined by $(\hat{x}, \hat{w}) = \tau_e^{*-1}(\hat{\xi})$ will be such that $\tau_a^*(\hat{x}, \hat{w})$ remains in E, along this trajectory i-e $|\tau^*(\hat{x})|^2\, \hat{w} < \delta$. This ensures that, far from the observability singularity where $|\tau^*(\hat{x})| = 0$, $\hat{w}$ remains sufficiently small to keep the invertibility of the Jacobian of τ_e^*. But we still have a problem close to the observability singularity, i.e. when $(\hat{x}_1, \hat{x}_2)$ is close to the origin. We shall see in Section 11.1 how to avoid this difficulty via a better choice of the initial injective immersion τ^*. ▲

10.3 Application : bioreactor

As a less academic illustration we consider the model of bioreactor presented in [GHO92] :

$$\dot{x}_1 = \frac{a_1 x_1 x_2}{a_2 x_1 + x_2} - u x_1 \ , \ \dot{x}_2 = -\frac{a_3 a_1 x_1 x_2}{a_2 x_1 + x_2} - u x_2 + u a_4 \ , \ y = x_1$$

where the a_i's are strictly positive real numbers and the control u verifies : $0 < u_{min} < u(t) < u_{max} < a_1$. This system evolves in the set $\mathcal{S} = \left\{x \in \mathbb{R}^2 : x_1 > \varepsilon_1 \ , \ x_2 > -a_2 x_1\right\}$ which is forward invariant. A high gain observer design leads us to consider the function $\tau^* : \mathcal{S} \to \mathbb{R}^2$ defined as :

$$\tau^*(x_1, x_2) \ = \ (x_1, \dot{x}_1|_{u=0}) \ = \ \left(x_1, \frac{a_1 x_1 x_2}{a_2 x_1 + x_2}\right) \ .$$

It is a diffeomorphism onto

$$\tau^*(\mathcal{S}) \ = \ \left\{\xi \in \mathbb{R}^2 \ : \ \xi_1 > 0 \ , \ a_1 \xi_1 > \xi_2\right\} \ .$$

The image by τ^* of the bioreactor dynamics is of the form

$$\dot{\xi}_1 \ = \ \xi_2 + g_1(\xi_1) u \quad , \qquad \dot{\xi}_2 \ = \ \varphi_2(\xi_1, \xi_2) + g_2(\xi_1, \xi_2) u$$

for which the following high gain observer can be built:

$$\dot{\hat{\xi}}_1 = \hat{\xi}_2 + g_1(\xi_1) u - k_1 \ell (\xi_1 - y) \quad , \quad \dot{\hat{\xi}}_2 = \varphi_2(\xi_1, \xi_2) + g_2(\xi_1, \xi_2) u - k_2 \ell (\xi_1 - y) \, , \tag{10.4}$$

where k_1 and k_2 are strictly positive real numbers and ℓ sufficiently large. As in [GHO92], τ^* being a diffeomorphism the dynamics of this observer in the x-coordinates are

$$\dot{\hat{x}} \ = \ \begin{pmatrix} \frac{a_1 \hat{x}_1 \hat{x}_2}{a_2 \hat{x}_1 + \hat{x}_2} - u \hat{x}_1 \\ -\frac{a_3 a_1 \hat{x}_1 \hat{x}_2}{a_2 \hat{x}_1 + \hat{x}_2} - u \hat{x}_2 + u a_4 \end{pmatrix} + \ell \begin{pmatrix} 1 & 0 \\ -1 & \frac{(a_2 \hat{x}_1 + \hat{x}_2)^2}{a_1 a_2 \hat{x}_1^2} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} (\xi_1 - y) \ . \tag{10.5}$$

Unfortunately the right hand side is singular at $\hat{x}_1 = 0$ or $\hat{x}_2 = -a_1 \hat{x}_1$. $\mathcal{S}$ being forward invariant, the system trajectories stay away from the singularity. But nothing guarantees the same property holds for the observer trajectories given by (10.5). In other words, since τ^* is

already a diffeomorphism, Problem 1 is solved with $d_\xi = d_x$, $\tau_a^* = \tau^*$ and $\mathcal{S}_a = \mathcal{S}$. But (8.17) is not satisfied, i.e. Problem 2 must be solved.

To construct the extension τ_e^* of τ_a^*, we view the image $\tau_a^*(\mathcal{S}_a)$ as the intersection $\tau_a^*(\mathcal{S}_a) = E_1 \cap E_2$ with :

$$E_1 \;=\; \Big\{(\xi_1,\xi_2)\in\mathbb{R}^2,\ \xi_1 > \varepsilon_1\Big\} \quad , \qquad E_2 \;=\; \Big\{(\xi_1,\xi_2)\in\mathbb{R}^2,\ a_1\xi_1 > \xi_2\Big\}\ .$$

This exhibits the fact that $\tau_a^*(\mathcal{S}_a)$ does not satisfy the property $\mathfrak{C}$ since its boundary is not C^1. We could smoothen this boundary to remove its "corner". But we prefer to exploit its particular "shape" and proceed as follows :

1. We build a diffeomorphism $\phi_1 : \mathbb{R}^2 \to E_1$ which acts on ξ_1 without changing ξ_2.
2. We build a diffeomorphism $\phi_2 : \mathbb{R}^2 \to E_2$ which acts on ξ_2 without changing ξ_1.
3. Denoting $\phi = \phi_2 \circ \phi_1 : \mathbb{R}^2 \to E_1 \cap E_2$, we take $\tau_e^* = \phi^{-1} \circ \tau_a^* : \mathcal{S}_a \to \mathbb{R}^2$.

To build ϕ_1 and ϕ_2, we follow the procedure given in the proof of Lemma 10.2.1 since E_1 and E_2 satisfy property $\mathfrak{C}$ with :

$$\kappa_1(\xi) = \varepsilon_1 - \xi_1\ ,\ \kappa_2(\xi) = \xi_2 - a_1\xi_1\ ,\ \chi_1(\xi) = \begin{pmatrix} -(\xi_1 - 1) \\ 0 \end{pmatrix}\ ,\ \chi_2(\xi) = \begin{pmatrix} 0 \\ -(\xi_2 + 1) \end{pmatrix}.$$

By following the same steps as in Example 10.2.1, with ε an arbitrary small strictly positive real number and ν defined in (10.2), we obtain :

$$\left|\begin{array}{l} \mathfrak{t}_1(\xi) = \ln\frac{1-\xi_1}{1-\varepsilon} \\ E_{\varepsilon,1} = \Big\{(\xi_1,\xi_2)\in\mathbb{R}^2,\ \xi_1 > 1 - \frac{1-\varepsilon}{e^\varepsilon}\Big\} \\ \phi_1(\xi) = \begin{cases} \xi & , \quad \text{if } \xi \in E_{\varepsilon,1} \\ \frac{\xi_1 - 1}{e^{\nu(\mathfrak{t}_1(\xi))}} + 1\ , & \text{otherwise} \end{cases} \end{array}\right| \quad \left|\begin{array}{l} \mathfrak{t}_2(\xi) = \ln\frac{\xi_2+1}{a_1\xi_1+1}\ , \\ E_{\varepsilon,2} = \Big\{(\xi_1,\xi_2)\in\mathbb{R}^2,\ \xi_2 \le \frac{a_1\xi_1+1}{e^\varepsilon} - 1\Big\} \\ \phi_2(\xi) = \begin{cases} \xi & , \quad \text{if } \xi \in E_{\varepsilon,2} \\ \frac{\xi_2+1}{e^{\nu(\mathfrak{t}_2(\xi))}} - 1\ , & \text{otherwise} \end{cases} \end{array}\right| \tag{10.6}$$

We remind the reader that, in the ξ-coordinates, the observer dynamics are not modified. The difference between using τ^* or τ_e^* is seen in the $\hat{x}$-coordinates only. And, by construction it has no effect on the system trajectories since we have

$$\tau^*(x) = \tau_e^*(x) \quad \forall x \in \mathcal{S}\ \text{“}-\varepsilon\text{”}\ .$$

As a consequence the difference between τ^* and τ_e^* is significant only during the transient, making sure, for the latter, that $\hat{x}$ never reaches a singularity of the Jacobian of τ_e^*.

We present in Figure 10.2 the results in the ξ coordinates (to allow us to see the effects of both τ^* and τ_e^*) of a simulation with (similar to [GHO92]) :

$$a_1 = a_2 = a_3 = 1\ ,\ a_4 = 0.1$$
$$u(t) = 0.08 \text{ for } t \le 10\ , \quad = 0.02 \text{ for } 10 \le t \le 20\ , \quad = 0.08 \text{ for } t \ge 20$$
$$x(0) = (0.04, 0.07), \quad \hat{x}(0) = (0.03, 0.09), \quad \ell = 5.$$

The solid black curves are the singularity locus. The red curve represents the bioreactor solution. The magenta curve represents the solution of the observer built with τ_e^*. It evolves freely in $\mathbb{R}^2$ according to the dynamics (10.4), not worried by any constraints. The blue curve represents its image by ϕ which brings it back inside the constrained domain where τ^{*-1} can then be used. This means these two curves represent the same object but viewed in different coordinates.

The solution of the observer built with τ^* would coincide with the magenta curve up to the point it reaches one solid black curve of a singularity locus. At that point it leaves $\tau^*(\mathcal{S})$ and consequently stops existing in the x-coordinates.

As proposed in [MP03, AP13], instead of keeping the raw dynamics (10.4) untouched as above, another solution would be to modify them to force ξ to remain in the set $\tau^*(\mathcal{S})$. For

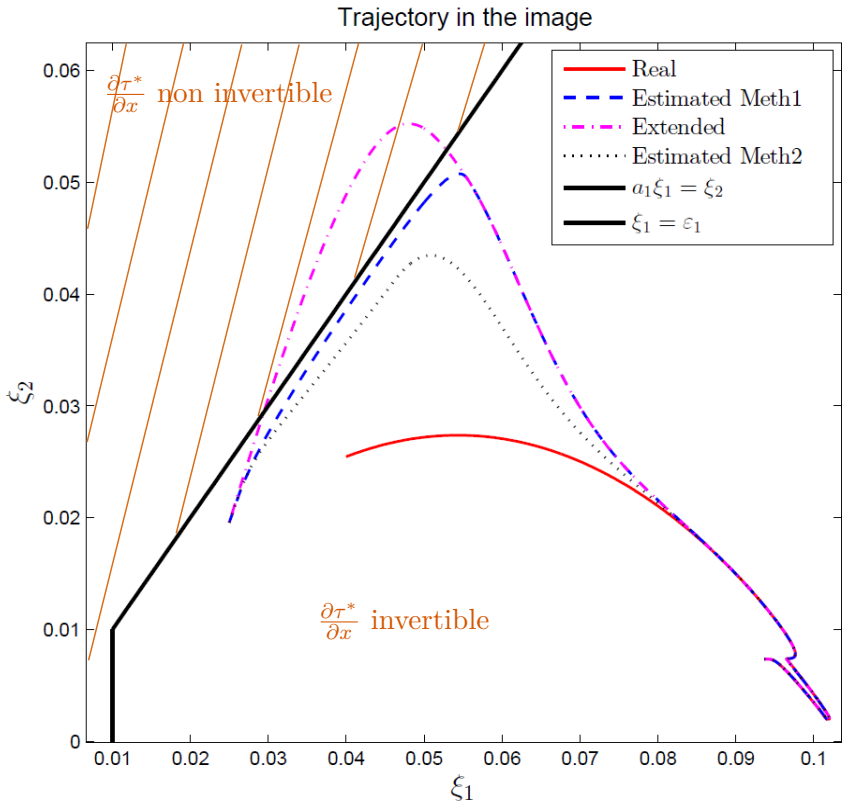


Figure 10.2: Bioreactor and observers solutions in the ξ-coordinates

instance, taking advantage of the convexity of this set, the modification proposed in [AP13] consists in adding to (10.4) the term

$$\mathcal{M}(\xi) = -g\, S_\infty \frac{\partial \mathfrak{h}}{\partial \xi}(\xi)^T\, \mathfrak{h}(\xi) \quad , \qquad \mathfrak{h}(\xi) \;=\; \begin{pmatrix} \max\{\kappa_1(\xi)+\varepsilon, 0\}^2 \\ \max\{\kappa_2(\xi)+\varepsilon, 0\}^2 \end{pmatrix} \tag{10.7}$$

with S_∞ a symetric positive definite matrix depending on (k_1, k_2, ℓ), ε an arbitrary small real number and g a sufficiently large real number. The solution corresponding to this modified observer dynamics is shown in Figure 10.2 with the dotted black curve. As expected it stays away from the the singularities locus in a very efficient way. But, for this method to apply, we have the restriction that $\tau^*(\mathcal{S})$ should be convex, instead of satisfying the less restrictive property $\mathfrak{C}$. Moreover, to guarantee that ξ is in $\tau^*(\mathcal{S})$, g has to be large enough and even larger when the measurement noise is larger. On the contrary, when the observer is built with τ_e^*, there is no need to tune properly any parameter to obtain convergence, at least theoretically. Nevertheless there maybe some numerical problems when ξ becomes too large or equivalently $\phi(\xi)$ is too close to the boundary of $\tau^*(\mathcal{S})$. To overcome this difficulty we can select the "thickness" of the layer (parameter ε in (10.6)) sufficiently large. Actually instead of “opposing” the two methods, we suggest to combine them when possible. The modification (10.7) makes sure ξ does not go too far outside the domain, and τ_e^* makes sure that $\hat{x}$ does not cross the singularity locus.

10.4 Conclusion

Joining Corollaries 9.3.1 and 10.1.1, we obtain the following answer to our problem :

Corollary 10.4.1.

Let $\mathcal{X}$ be a bounded subset of $\mathbb{R}^{d_x}$, $\mathcal{S}$ be an open subset of $\mathbb{R}^{d_x}$ and $\tau^* : \mathcal{S} \to \mathbb{R}^{d_\xi}$ be an injective immersion. Assume there exists an open bounded contractible set $\tilde{\mathcal{S}}$ which is C^2-diffeomorphic to $\mathbb{R}^{d_x}$ and such that

$$\mathtt{cl}(\mathcal{X}) \subset \tilde{\mathcal{S}} \subset \mathtt{cl}(\tilde{\mathcal{S}}) \subset \mathcal{S} \ .$$

There exists a strictly positive number ε and a diffeomorphism $\tau_e^* : \mathcal{S}_a \to \mathbb{R}^{d_\xi}$ with $\mathcal{S}_a = \tilde{\mathcal{S}} \times B_\varepsilon(0)$, such that

$$\tau_e^*(x,0) = \tau^*(x) \quad \forall x \in \mathcal{X} \quad , \qquad \tau_e^*(\mathcal{S}_a) = \mathbb{R}^{d_\xi} \ ,$$

namely $(\tau_e^*, \mathcal{S}_a)$ solves Problems 1-2.

We conclude that if $\mathcal{X}$, $\mathcal{S}$ and τ^* given by Assumption $\mathcal{O}$ verify the conditions of Corollary 10.4.1, then Problems 1-2 can be solved and Theorem 8.2.1 holds, i-e an observer can be expressed in the given x-coordinates.

Chapter 11

Generalizations and applications

Chapitre 11 – Généralisations et applications. *Dans les chapitres 9 et 10, nous avons donné (en particuliier à travers Corollaire 10.4.1) des conditions permettant de résoudre les Problèmes 1 et 2 lorsque l'hypothèse $\mathcal{O}$ est vérifiée et $\mathcal{X}$ est borné. Cependant, il arrive que ces conditions ne soient pas satisfaites et nous montrons dans ce chapitre comment résoudre les Problèmes 1 et 2 grâce à un meilleur choix de τ^* et $\varphi_{\mathcal{T}}$, donnés par l'hypothèse $\mathcal{O}$. En particulier, ceci permet d'écrire un observateur dans les coordonnées x pour l'oscillateur à fréquence inconnue* (8.1)*, à la fois par la voie du grand gain* (8.4) *et de Luenberger* (8.6)*. Enfin, nous montrons à travers un exemple tiré d'une application, comment la méthodologie présentée dans cette Partie III peut être étendue au cas où la transformation τ^* dépend du temps.*

Contents

Throughout Chapters 9 and 10, we have given (in particular in Corollary 10.4.1) conditions allowing to solve Problem 1 and Problem 2 when Assumption $\mathcal{O}$ holds and $\mathcal{X}$ is bounded.

However, it can happen that those conditions are not satisfied and we show in this chapter how to solve both Problems 1 and 2 via a better choice of the data given by Assumption $\mathcal{O}$, namely τ^* and $\varphi_{\mathcal{T}}$. In particular, this enables to write an observer in the x-coordinates for the oscillator with unknown frequency (8.1) both via the high gain (8.4) and Luenberger (8.6) designs.

Finally, we show through an application in aircraft landing how the methodology presented in this Part III can be extended to the case where the transformation τ^* is time-varying.

11.1 Modifying τ^* and $\varphi_{\mathcal{T}}$ given by Assumption $\mathcal{O}$

The sufficient conditions given in Chapters 9 and 10, to solve Problem 1 and Problem 2 in order to fulfill the requirements of Theorem 8.2.1, impose conditions on the dimensions or on the domain

of injectivity $\mathcal{S}$ which are not always satisfied : contractibility for Jacobian complementation and diffeomorphism extension, limited number of pairs (d_ξ, d_x) for the $\tilde{P}[d_\xi, d_x]$ problem, etc. Expressed in terms of our initial problem, these conditions are limitations on the data f, h and τ^* that we have considered. In the following, we show by means of examples that, in some cases, these data can be modified in such a way that our various tools apply and give a satisfactory solution. Such modifications are possible since we restrict our attention to system solutions which remain in $\mathcal{X}$. Therefore the data f, h and τ^* can be arbitrarily modified outside this set. For example we can add "fictitious" components to the measured output y as long as their value is known on $\mathcal{X}$.

11.1.1 For contractibility

It may happen that the set $\mathcal{S}$ attached to τ^* is not contractible, for example due to an observability singularity. We have seen that Jacobian complementation and image extension may be prevented by this (see Theorem 9.3.1 and Remark 16). A possible approach to overcome this difficulty when we know the system trajectories stay away from the singularities is to add a fictitious output traducing this information :

Example 11.1.1 (Continuation of Example 9.2.1) The observer (9.10) we have obtained at the end of Example 9.2.1 for the harmonic oscillator with unknown frequency is not satisfactory because of the singularity at $\hat{x}_1 = \hat{x}_2 = 0$. To overcome this difficulty we add, to the given measurement $y = x_1$, the following one

$$y_2 = h_2(x) \;=\; \wp(x_1, x_2)\, x_3$$

with

$$\wp(x_1, x_2) \;=\; \max\left\{0, \frac{1}{r^2} - (x_1^2 + x_2^2)\right\} \;. \tag{11.1}$$

By construction this function is zero on $\mathcal{X}$ and y_2 can thus be considered as an extra measurement with zero as constant value. The interest of y_2 is to give access to x_3 even at the singularity $x_1 = x_2 = 0$. Indeed, consider the new function τ^* defined as

$$\tau^*(x) \;=\; (x_1\,,\; x_2\,,\; -x_1x_3\,,\; -x_2x_3\,,\; \wp(x_1, x_2)\, x_3) \;. \tag{11.2}$$

τ^* is C^1 on $\mathbb{R}^3$ and its Jacobian is :

$$\frac{\partial \tau^*}{\partial x}(x) \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -x_3 & 0 & -x_1 \\ 0 & -x_3 & -x_2 \\ \frac{\partial \wp}{\partial x_1} x_3 & \frac{\partial \wp}{\partial x_2} x_3 & \wp \end{pmatrix} \;, \tag{11.3}$$

which has full rank 3 on $\mathbb{R}^3$, since $\wp(x_1, x_2) \neq 0$ when $x_1 = x_2 = 0$. It follows that the singularity has disappeared and this new τ^* is an injective immersion on the entire $\mathbb{R}^3$ which is contractible.

We have shown in Example 9.3.1 how Wazewski's algorithm allows us to get in this case a C^2 function $\gamma : \mathbb{R}^3 \to \mathbb{R}^4$ satisfying :

$$\det\left(\frac{\partial \tau^*}{\partial x}(x)\; \gamma(x)\right) \;\neq\; 0 \qquad \forall x \in \mathbb{R}^3 \;.$$

This gives us $\tau_a^*(x, w) \;=\; \tau^*(x) + \gamma(x)w$ which is a C^2-diffeomorphism on $\mathbb{R}^3 \times B_\varepsilon(0)$, with ε sufficiently small. Furthermore, $\mathcal{S}_a = \mathbb{R}^3 \times B_\varepsilon(0)$ being now diffeomorphic to $\mathbb{R}^5$, Corollary 10.1.1 applies and provides an extension τ_e^* of τ_a^* satisfying Problems 1 and 2. ▲

11.1.2 For a solvable $\tilde{P}[d_\xi, d_x]$ problem

If we are in a case that cannot be reduced to a solvable $\tilde{P}[d_\xi, d_x]$ problem, we may try to modify d_ξ by adding arbitrary rows to $\frac{\partial \tau^*}{\partial x}$. We illustrate this technique with the following example.

Example 11.1.2 (Continuation of Example 11.1.1) In Example 11.1.1, by adding the fictitious measured output $y_2 = h_2(x)$, we have obtained another function τ^* for the harmonic oscillator with unknown frequency which is an injective immersion on $\mathbb{R}^3$. In this case, we have $d_x = 3$ and $d_\xi = 5$ which gives a pair not in (9.8). But, as already exploited in Example 9.2.1, the first 2 rows of the Jacobian $\frac{\partial \tau^*}{\partial x}$ in (11.3) are independent for all x in $\mathbb{R}^3$. It follows that our Jacobian complementation problem reduces to complement the vector $(-x_1, -x_2, \wp(x_1, x_2))^\top$. This is a problem with pair $(3, 1)$ which is still not in the list (9.8). Instead, the pair $(4, 1)$ is, so that the vector $(-x_1, -x_2, \wp(x_1, x_2), 0)^\top$ can be complemented via a universal formula. We have added a zero component, without changing the full rank property. Actually this vector is extracted from the Jacobian of

$$\tau^*(x) = (x_1\,,\; x_2\,,\; -x_1x_3\,,\; -x_2x_3\,,\; \wp(x_1, x_2)\, x_3\,,\; 0)\;. \tag{11.4}$$

In the high gain observer paradigm, this zero we have added can come from another (fictitious) measured output $y_3 = 0$. As we saw in the proof of Theorem (9.2.1), a complement of $(-x_1, -x_2, \wp(x_1, x_2), 0)^\top$ is

$$\begin{pmatrix} x_2 & -\wp(x_1,x_2) & 0 \\ -x_1 & 0 & -\wp(x_1,x_2) \\ 0 & -x_1 & -x_2 \\ \wp(x_1,x_2) & x_2 & -x_1 \end{pmatrix}$$

and thus a complement of $\frac{\partial \tau^*}{\partial x}(x)$ is

$$\gamma(x) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ x_2 & -\wp(x_1,x_2) & 0 \\ -x_1 & 0 & -\wp(x_1,x_2) \\ 0 & -x_1 & -x_2 \\ \wp(x_1,x_2) & x_2 & -x_1 \end{pmatrix}$$

which gives with the formula (9.2)

$$\tau_a^*(x, w) = \Big(\; x_1\,,\; x_2\,,\; [-x_1x_3 + x_2w_1 - \wp(x_1,x_2)w_2]\,,\; [-x_2x_3 - x_1w_1 - \wp(x_1,x_2)w_3]\,,$$
$$[\wp(x_1,x_2)x_3 - x_1w_2 - x_2w_3]\,,\; [\wp(x_1,x_2)w_1 + x_2w_2 - x_1w_3)]\Big)\,.$$

The determinant of the Jacobian of τ_a^* thus defined is $(x_1^2 + x_2^2 + \wp(x_1, x_2)^2)^2$ which is nowhere 0 on $\mathbb{R}^6$. Hence τ_a^* is locally invertible. Actually it is diffeomorphism from $\mathbb{R}^6$ onto $\mathbb{R}^6$ since we can express $\xi = \tau_a^*(x, w)$ as

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\;,\quad \begin{pmatrix} -\xi_1 & \xi_2 & -\wp(\xi_1,\xi_2) & 0 \\ -\xi_2 & -\xi_1 & 0 & -\wp(\xi_1,\xi_2) \\ \wp(\xi_1,\xi_2) & 0 & -\xi_1 & -\xi_2 \\ 0 & \wp(\xi_1,\xi_2) & \xi_2 & -\xi_1 \end{pmatrix} \begin{pmatrix} x_3 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} \xi_3 \\ \xi_4 \\ \xi_5 \\ \xi_6 \end{pmatrix}\;,$$

where the matrix on the left is invertible by construction. Since $\tau_a^*(\mathbb{R}^6) = \mathbb{R}^6$, there is no need for an image extension and we simply take $\tau_e^* = \tau_a^*$. To have all the assumptions of Theorem 8.2.1 satisfied, it remains to find a function φ such that (τ_{ex}, φ) is in the set $\mathcal{F}$, the function τ_{ex} being the x-component of the inverse of τ_e^*. Since the first four components of τ^* are the same

as in (8.3), the first four components of φ are given in (8.5). It remains to define the dynamics of $\hat{\xi}_5$ and $\hat{\xi}_6$. Exploiting the fact that, for x in $\mathcal{X}$,

$$y_2 = 0 \quad , \quad \dot{y}_2 = \dot{\overbrace{\wp(x_1,x_2)x_3}} \; = \; 0 \quad , \qquad y_3 = 0 \quad , \quad \dot{y}_3 \; = \; 0 \; ,$$

one can simply choose

$$\dot{\hat{\xi}}_5 = 0 - a(\hat{\xi}_5 - y_2) = -a\,\hat{\xi}_5 \quad , \quad \dot{\hat{\xi}}_6 = 0 - b(\hat{\xi}_6 - y_3) = -b\,\hat{\xi}_6$$

for some strictly positive numbers a and b, which finally leads to the function

$$\varphi(\xi,\hat{x},y) \; = \; \begin{pmatrix} \xi_2 + Lk_1(y-\hat{x}_1) \\ \xi_3 + L^2k_2(y-\hat{x}_1) \\ \xi_4 + L^3k_3(y-\hat{x}_1) \\ \mathsf{sat}_{r^3}(\hat{x}_1\hat{x}_3^2) + L^4k_4(y-\hat{x}_1) \\ -a\,\xi_5 \\ -b\,\xi_6 \end{pmatrix}$$

With picking L large enough, φ can be paired with any function $\tau : \mathbb{R}^6 \to \mathbb{R}^6$ which is locally Lipschitz, and thus in particular with τ_{ex}. Therefore, Theorem 8.2.1 applies and gives the following observer for the harmonic oscillator with unknown frequency

$$\begin{pmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{w}}_1 \\ \dot{\hat{w}}_2 \\ \dot{\hat{w}}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -\hat{x}_3 - \frac{\partial\wp}{\partial\hat{x}_1}\hat{w}_2 & \hat{w}_1 - \frac{\partial\wp}{\partial x_2}\hat{w}_2 & -\hat{x}_1 & \hat{x}_2 & -\wp & 0 \\ -\hat{w}_1 - \frac{\partial\wp}{\partial\hat{x}_1}\hat{w}_3 & -\hat{x}_3 - \frac{\partial\wp}{\partial x_2}\hat{w}_3 & -\hat{x}_2 & -\hat{x}_1 & 0 & -\wp \\ \frac{\partial\wp}{\partial x_1}\hat{x}_3 - \hat{w}_2 & \frac{\partial\wp}{\partial x_2}\hat{x}_3 - \hat{w}_3 & \wp & 0 & -\hat{x}_1 & -\hat{x}_2 \\ \frac{\partial\wp}{\partial x_1}\hat{w}_1 - \hat{w}_3 & \frac{\partial\wp}{\partial x_2}\hat{w}_1 + \hat{w}_2 & 0 & \wp & \hat{x}_2 & -\hat{x}_1 \end{pmatrix}^{-1} \times \tag{11.5}$$

$$\times \begin{pmatrix} \hat{x}_2 + Lk_1(y-\hat{x}_1) \\ [-\hat{x}_1\hat{x}_3 + \hat{x}_2\hat{w}_1 - \wp(\hat{x}_1,\hat{x}_2)\hat{w}_2] + L^2k_2(y-\hat{x}_1) \\ [-\hat{x}_2\hat{x}_3 - \hat{x}_1\hat{w}_1 - \wp(\hat{x}_1,\hat{x}_2)\hat{w}_3] + L^3k_3(y-\hat{x}_1) \\ \mathsf{sat}_{r^3}(\hat{x}_1\hat{x}_3^2) + L^4k_4(y-\hat{x}_1) \\ -a\,[\wp(\hat{x}_1,\hat{x}_2)\hat{x}_3 - \hat{x}_1\hat{w}_2 - \hat{x}_2\hat{w}_3] \\ -b\,[\wp(\hat{x}_1,\hat{x}_2)\hat{w}_1 + \hat{x}_2\hat{w}_2 - \hat{x}_1\hat{w}_3)] \end{pmatrix}.$$

It is globally defined and globally convergent for any solution of the oscillator initialized in the set $\mathcal{X}$ given in (8.2). Results of a simulation are given in Figure 11.1. Notice that the observer converges despite the fact that $\hat{x}_1$ and $\hat{x}_2$ are initialized at the singularity. This would not have been possible with observer (9.7), i-e without adding the fictitious output. By the way, observe that w_2 and w_3 present a violent peak at $t = 0$. This is due to the fact that $\hat{x}_1$ and $\hat{x}_2$ are around the singularity, where only the fictitious output (which has a very small but non zero value) preserves the invertibility of the Jacobian. We used a step-variable integration scheme to take this into account. ▲

Remark 17 It is interesting to notice that the manifold $\hat{\xi}_5 = \hat{\xi}_6 = 0$ is invariant. This implies the existence of an observer with order reduced to 4. One could thus wonder if it could be expressed with coordinates $(x, \bar{w})$ in $\mathbb{R}^4$, instead of (x, w) in $\mathbb{R}^6$, i-e if maybe there existed a diffeomorphism $\bar{\tau}_e = (\bar{\tau}_{ex}, \bar{\tau}_{ew})$ such that

$$\begin{array}{rcccl} x & = & \bar{\tau}_{ex}(\xi_1,\xi_2,\xi_3,\xi_4) & = & \tau_{ex}(\xi_1,\xi_2,\xi_3,\xi_4,0,0) \\ \bar{w} & = & \bar{\tau}_{ew}(\xi_1,\xi_2,\xi_3,\xi_4) & & \end{array}$$

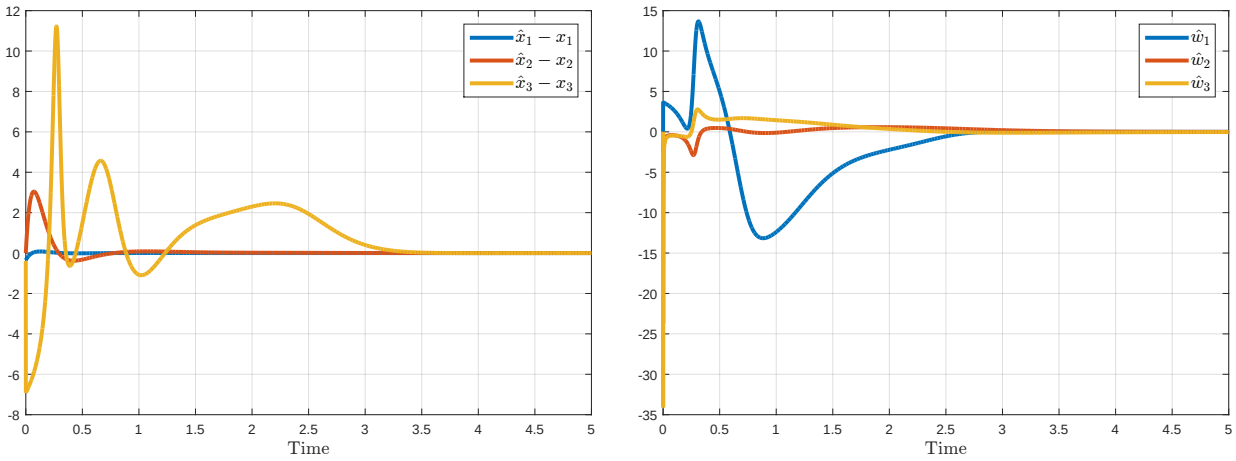


Figure 11.1: High gain observer (11.5) with $\hat{x}_1 = \hat{x}_2 = \hat{x}_3 = 0$ at the singularity, $L = 3$, $k_1 = 10$, $k_2 = 35$, $k_3 = 50$, $k_4 = 24$. The simulation was done with a step-variable Euler algorithm.

But then its Jacobian would necessarily be of the form :

$$\frac{\partial \bar{\tau}_e}{\partial \xi}(\xi_1, \xi_2, \xi_3, \xi_4) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ * & * & -\frac{\xi_1}{\xi_1^2+\xi_2^2+\wp(\xi_1,\xi_2)^2} & -\frac{\xi_2}{\xi_1^2+\xi_2^2+\wp(\xi_1,\xi_2)^2} \\ * & * & * & * \end{pmatrix}$$

which is singular for $\xi_1 = \xi_2 = 0$.

11.1.3 A universal complementation method

In the previous example, we have made the Jacobian complementation possible by increasing d_ξ with augmenting the number of coordinates of τ^*. Actually if we augment τ^* with d_x zeros the possibility of a Jacobian complementation is guaranteed. Indeed pick any C^1 function B the values of which are $d_\xi \times d_\xi$ matrices with positive definite symmetric part, we can complement $\begin{pmatrix} \frac{\partial \tau^*}{\partial x}(x) \\ 0 \end{pmatrix}$ which is full column rank with $\gamma = \begin{pmatrix} -B(x) \\ \frac{\partial \tau^*}{\partial x}(x)^\top \end{pmatrix}$. This follows from the identity (Schur complement) involving invertible matrices

$$\begin{pmatrix} \frac{\partial \tau^*}{\partial x}(x) & -B(x) \\ 0 & \frac{\partial \tau^*}{\partial x}(x)^\top \end{pmatrix} \begin{pmatrix} 0 & I \\ I & B(x)^{-1}\frac{\partial \tau^*}{\partial x}(x) \end{pmatrix} = \begin{pmatrix} -B(x) & 0 \\ \frac{\partial \tau^*}{\partial x}^\top(x) & \frac{\partial \tau^*}{\partial x}(x)^\top B(x)^{-1}\frac{\partial \tau^*}{\partial x}(x) \end{pmatrix}.$$

So we have here a universal method to solve Problem 1. Its drawback is that the dimension of the state increases by d_ξ, instead of $d_\xi - d_x$.

11.2 A global example : Luenberger design for the oscillator

Let us now come back to the Luenberger observer presented in Section 8.1.2 for the oscillator with unknown frequency. Although an inversion of the transformation was proposed in [PMI06] based on the resolution of a minimization problem, we want to show here how this step can be avoided.

Recall that the transformation is given by

$$\tau^*(x) = \left(\frac{\lambda_1 x_1 - x_2}{\lambda_1^2 + x_3}, \frac{\lambda_2 x_1 - x_2}{\lambda_2^2 + x_3}, \frac{\lambda_3 x_1 - x_2}{\lambda_3^2 + x_3}, \frac{\lambda_4 x_1 - x_2}{\lambda_4^2 + x_3} \right)$$

and its Jacobian

$$\frac{\partial \tau^*}{\partial x}(x) = \begin{pmatrix} \frac{\lambda_1}{\lambda_1^2+x_3} & -\frac{1}{\lambda_1^2+x_3} & -\frac{\tau_1^*(x)}{\lambda_1^2+x_3} \\ \frac{\lambda_2}{\lambda_2^2+x_3} & -\frac{1}{\lambda_2^2+x_3} & -\frac{\tau_2^*(x)}{\lambda_2^2+x_3} \\ \frac{\lambda_3}{\lambda_3^2+x_3} & -\frac{1}{\lambda_3^2+x_3} & -\frac{\tau_3^*(x)}{\lambda_3^2+x_3} \\ \frac{\lambda_4}{\lambda_4^2+x_3} & -\frac{1}{\lambda_4^2+x_3} & -\frac{\tau_4^*(x)}{\lambda_4^2+x_3} \end{pmatrix} .$$

The complementation is quite easy because there is only one dimension to add : we could just add a column $\gamma(x)$ consisting of the corresponding minors as suggested in Section 9.2. However, this would produce a diffeomorphism on $\mathcal{X} \times B_\varepsilon(0)$ for some ε, where $\mathcal{X}$ defined in (8.2) is not contractible due to the observability singularity at $x_1 = x_2 = 0$. Therefore, no image extension is possible and it would be necessary to ensure that $\hat{w}$ remains small and $(\hat{x}_1, \hat{x}_2)$ far from $(0, 0)$ by some other means. Like for the high gain observer, we thus try to remove this singularity.

Again, we assume the system solutions remain in $\mathcal{X}$ and add the same fictitious output y_2 as before, which vanishes in $\mathcal{X}$ and which is non zero when (x_1, x_2) is close to the origin namely :

$$y_2 = \wp(x_1, x_2)x_3$$

where $\wp$ is defined in (11.1). Once again, it is possible to show[1] that by adding y_2 to τ^*, the observability singularity disappears, namely the function

$$\tau^*(x) = \left(\frac{\lambda_1 x_1 - x_2}{\lambda_1^2 + x_3}, \frac{\lambda_2 x_1 - x_2}{\lambda_2^2 + x_3}, \frac{\lambda_3 x_1 - x_2}{\lambda_3^2 + x_3}, \frac{\lambda_4 x_1 - x_2}{\lambda_4^2 + x_3}, \wp(x_1, x_2)x_3 \right)$$

is an injective immersion on

$$\tilde{\mathcal{S}} = \mathbb{R}^2 \times \mathbb{R}_+ .$$

Although the Jacobian complementation problem is solvable for this τ^* according to Wazewski's theorem 9.3.1 because $\tilde{\mathcal{S}}$ is contractible, we want to avoid the lengthy computations entailed by this method. We are going to see in the following that it is possible if one rather take (as before)

$$\tau^*(x) = \left(\frac{\lambda_1 x_1 - x_2}{\lambda_1^2 + x_3}, \frac{\lambda_2 x_1 - x_2}{\lambda_2^2 + x_3}, \frac{\lambda_3 x_1 - x_2}{\lambda_3^2 + x_3}, \frac{\lambda_4 x_1 - x_2}{\lambda_4^2 + x_3}, \wp(x_1, x_2)x_3, 0 \right) \tag{11.6}$$

which is still an injective immersion on $\tilde{\mathcal{S}}$. Although the utility of this zero seems questionable at this point, we will point out its interest in the subsequent computations. The new Jacobian takes the form

$$\frac{\partial \tau^*}{\partial x}(x) = \begin{pmatrix} \frac{\lambda_1}{\lambda_1^2+x_3} & -\frac{1}{\lambda_1^2+x_3} & -\frac{\tau_1^*(x)}{\lambda_1^2+x_3} \\ \frac{\lambda_2}{\lambda_2^2+x_3} & -\frac{1}{\lambda_2^2+x_3} & -\frac{\tau_2^*(x)}{\lambda_2^2+x_3} \\ \frac{\lambda_3}{\lambda_3^2+x_3} & -\frac{1}{\lambda_3^2+x_3} & -\frac{\tau_3^*(x)}{\lambda_3^2+x_3} \\ \frac{\lambda_4}{\lambda_4^2+x_3} & -\frac{1}{\lambda_4^2+x_3} & -\frac{\tau_4^*(x)}{\lambda_4^2+x_3} \\ A(x) & B(x) & C(x) \\ 0 & 0 & 0 \end{pmatrix} .$$

[1] In [PMI06], it is shown that for any $r > 0$, there exists $L_r > 0$ such that for all (x_a, x_b) in $\mathbb{R}^2 \times (0, r)$, $|x_{1,a} - x_{1,b}| + |x_{2,a} - x_{2,b}| + \frac{x_{1,a}+x_{1,b}+x_{2,a}+x_{2,b}}{2}|x_{3,a} - x_{3,b}| \leq L_r|\tau^*{}_{14}(x_a) - \tau^*{}_{14}(x_b)|$ where $\tau^*{}_{14}$ denotes the first four components of τ^*. Therefore, $\tau^*{}_{14}(x_a) = \tau^*{}_{14}(x_b)$ implies that $x_{1,a} = x_{1,b}$ and $x_{2,a} = x_{2,b}$: either one of them is non zero and in that case, the inequality says that we have also $x_{3,a} = x_{3,b}$, or they are all zero but then $\tau^*{}_5(x_a) = \tau^*{}_5(x_b)$ implies that $x_{3,a} = x_{3,b}$. We conclude that τ^* is injective on $\tilde{\mathcal{S}}$. Now, applying the inequality between x and $x + hv$ and making h go to zero, we get that $\frac{\partial \tau^*{}_{14}}{\partial x}(x)v = 0$ implies that $v_1 = v_2 = 0$ and $v_3 = 0$ if either x_1 or x_2 is nonzero. If they are both zero, $\frac{\partial \tau^*{}_5}{\partial x}(x)v = 0$ with $v_1 = v_2 = 0$ gives $v_3 = 0$. Thus, $\frac{\partial \tau^*}{\partial x}(x)$ is full-rank.

Let us first simplify the matrix to be complemented by noticing that

$$M(x_3,\lambda_i)\,\frac{\partial \tau^*}{\partial x}(x) = \begin{pmatrix} 1 & 0 & m_1(x) \\ 0 & 1 & m_2(x) \\ 0 & 0 & m_3(x) \\ 0 & 0 & m_4(x) \\ A(x) & B(x) & C(x) \\ 0 & 0 & 0 \end{pmatrix} \tag{11.7}$$

where $M(x_3,\lambda_i)$ is the invertible matrix

$$M(x_3,\lambda_i) = \begin{pmatrix} \mathfrak{D}^{-1}(\lambda_i) & 0_{4\times 2} \\ 0_{2\times 4} & I_{2\times 2} \end{pmatrix} \begin{pmatrix} \lambda_1^2+x_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^2+x_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3^2+x_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4^2+x_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and $\mathfrak{D}(\lambda_i)$ is an appropriate Vandermonde matrix associated to the λ_i. So now we are left with complementing the matrix given by (11.7). Observing that right-multiplying (11.7) by the invertible matrix $N(x) = \begin{pmatrix} 1 & 0 & -m_1(x) \\ 0 & 1 & -m_2(x) \\ 0 & 0 & 1 \end{pmatrix}$ gives $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & m_3(x) \\ 0 & 0 & m_4(x) \\ A(x) & B(x) & m_5(x) \\ 0 & 0 & 0 \end{pmatrix}$, with

$$m_5(x) = C(x) - m_1(x)A(x) - m_2(x)B(x)\ ,$$

we conclude first that the vector $(m_3(x), m_4(x), m_5(x), 0)$ in $\mathbb{R}^4$ is non-zero on $\tilde{\mathcal{S}}$ and then that (11.7) can be simply complemented by complementing the vector $(m_3(x), m_4(x), m_5(x), 0)$ into an invertible 4×4 matrix. Note that this is the solvable problem $\tilde{P}[4,1]$ from (9.8), and without adding the 0 output y_3, we would have obtained $\tilde{P}[3,1]$ which is not solvable. An explicit solution to $\tilde{P}[4,1]$ is given in Section 9.2, but we can here also exploit the very particular structure of the vector and use the remark made in Section 11.1.3 that the matrix

$$P(x) = \begin{pmatrix} m_3(x) & -1 & 0 & 0 \\ m_4(x) & 0 & -1 & 0 \\ m_5(x) & 0 & 0 & -1 \\ 0 & m_3(x) & m_4(x) & m_5(x) \end{pmatrix}$$

is invertible as soon as $(m_3(x), m_4(x), m_5(x))$ is non-zero.

Reversing the transformations, we thus manage to extend the Jacobian of τ^* into a matrix of dimension 6 whose determinant is non-zero on $\tilde{\mathcal{S}}$. Adding three state components to the system state, we obtain a diffeomorphism τ_a^* on $\mathcal{S}_a = \tilde{\mathcal{S}} \times B_\varepsilon$, with ε sufficiently small. All this leads to the observer :

$$\begin{pmatrix} \dot{\hat{x}} \\ \dot{\hat{w}} \end{pmatrix} = \left(\frac{\partial \tau_a^*}{\partial x}(\hat{x},\hat{w})\right)^{-1} (A\tau_a^*(\hat{x},\hat{w}) + By_1) \tag{11.8}$$

where $B = [1,1,1,1,0,0]^\top$, $A = \texttt{diag}(-\lambda_1,-\lambda_2,-\lambda_3,-\lambda_4,-\mu,-\gamma)$ and μ and γ are two strictly positive real numbers. The expression of the Jacobian of the extended function is omitted here due to its complexity, but it can be obtained by straightforward symbolic computations.

The singularity at $(\hat{x}_1,\hat{x}_2) = 0$ has disappeared, but we still need to ensure that $\hat{x}_3$ remains positive, or at least greater than $-\min\{\lambda_i^2\}$. Besides, unlike the high gain observer (11.5), the

invertibility of the extended Jacobian is only guaranteed for w in B_ε. To make sure the solutions remain in $\mathcal{S}_a = \tilde{\mathcal{S}} \times B_\varepsilon$, we should solve Problem 2 namely extend τ_a^* into a diffeomorphism τ_e^* whose image of $\mathcal{S}_a$ covers $\mathbb{R}^6$. Since $\mathcal{S}_a$ is diffeomorphic to $\mathbb{R}^6$, we know it is theoretically possible by Theorem 10.1.1 and replacing τ_a^* by the new surjective diffeomorphism τ_e^* in (11.8) would give an observer whose solutions are ensured to exist for all t.

Unfortunately, due to the complexity of the expression of τ_a^*, we are not yet able to achieve such an extension. The consequence is that there may exist a set of initial conditions and parameters such that the corresponding trajectory of observer (11.8) encounters a singularity of the jacobian of τ_a^* and thus diverges. A way of reducing this set is to approximate the image of $\mathcal{S}_a$ by τ_a^*, as proposed in Example 10.2.1. In the present case, we have (denoting τ_{14}^* the first four components of τ^* defined in (11.6)),

$$(\xi_1,\xi_2,\xi_3,\xi_4) = \tau_{14}^*(x) \quad \Longleftrightarrow \quad \begin{pmatrix} \lambda_1^2\xi_1 & -\lambda_1 & 1 & \xi_1 \\ \lambda_2^2\xi_2 & -\lambda_2 & 1 & \xi_2 \\ \lambda_3^2\xi_3 & -\lambda_3 & 1 & \xi_3 \\ \lambda_4^2\xi_4 & -\lambda_4 & 1 & \xi_4 \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0$$

and thus

$$F(\tau^*(x)) = F(\tau_a^*(x,0)) = 0$$

where F is the quadratic function defined by

$$F(\xi) = \det \begin{pmatrix} \lambda_1^2\xi_1 & -\lambda_1 & 1 & \xi_1 \\ \lambda_2^2\xi_2 & -\lambda_2 & 1 & \xi_2 \\ \lambda_3^2\xi_3 & -\lambda_3 & 1 & \xi_3 \\ \lambda_4^2\xi_4 & -\lambda_4 & 1 & \xi_4 \end{pmatrix} .$$

Therefore, replacing $\xi^\top M \xi$ by $F(\xi)$ in (10.2)-(10.3) gives a diffeomorphism (ϕ, Id) from $\mathbb{R}^6$ to

$$E = \left\{ \xi \in \mathbb{R}^6 : F(\xi)^2 < \delta \right\}$$

and taking $\tau_e^* = \phi^{-1} \circ \tau_a^*$ instead of τ_a^* ensures that for any observer solution $t \mapsto \hat{\xi}(t)$, our estimate defined by $(\hat{x},\hat{w}) = \tau_e^{*-1}(\hat{\xi})$ will be such that $\tau_a^*(\hat{x},\hat{w})$ remains in E. When δ goes to zero, E gets closer to $\tau_a^*(\tilde{\mathcal{S}} \times \{0\})$ and thus we can hope that $\hat{w}$ will remain sufficiently small to keep the invertibility of the Jacobian of τ_e^*. We indeed observe in simulations that taking τ_e^* instead of τ_a^* enables to ensure completeness of some of the solutions which otherwise diverge with τ_a^*. An example is given in Figure 11.2 : before $t = 0.05$, the observer trajectory is close to a singularity, $\hat{w}$ tends to become very large (see Figure 11.2(b)), so does $F(\hat{\xi})$, but ϕ enables to reduce $F(\hat{\xi})$ (see Figure 11.2(d)) and thus prevent $\hat{w}$ from becoming too large and encounter the singularity. Unfortunately, although the set of initial conditions leading to uncomplete solutions is reduced by this method, it does not completely disappears.

11.3 Generalization to a time-varying τ^*

In Assumption $\mathcal{O}$, it is supposed that the transformation τ^* from the given x-coordinates to the ξ-coordinates is stationary. But we have seen in Part II that it is sometimes easier/necessary to consider a time-varying transformation which depends on the input, and apply Theorem 2.2.1. It is thus legitimate to wonder if the methodology presented in this part is still useful. In fact, the same tools can be applied in the sense that :

- Assumption $\mathcal{O}$ should now provide for each u in $\mathcal{U}$ a C^1 function $\tau^* : \mathbb{R}^{d_x} \times \mathbb{R} \to \mathbb{R}^{d_\xi}$, subsets $\mathcal{S}_t$ and $\mathcal{X}_t$ of $\mathbb{R}^{d_x}$ and a set $\varphi\mathcal{T}$ made of couples (φ,τ) such that for all t in $[0,+\infty)$, $x \mapsto \tau^*(x,t)$ is an injective immersion on $\mathcal{S}_t$, for all x_0 in $\mathcal{X}_0$ and all all t in $[0,+\infty)$, $X(x_0;t;u)$ is in $\mathcal{X}_t$, for all x in $\mathcal{X}_t$, $\tau(\tau^*(x,t),t) = x$ and φ is such that the appropriate convergence in the ξ-coordinates is achieved.

- Problems 1 and 2 can then be solved applying the tools of Chapters 9 and 10 on $x \mapsto \tau^*(x,t)$ for each t. This leads to a function $\tau_e^* : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\xi - d_x} \times \mathbb{R} \to \mathbb{R}^{d_\xi}$ and open subsets $\mathcal{S}_{a,t}$ of $\mathbb{R}^{d_\xi}$ containing $\mathcal{X} \times \{0\}$ such that for all t in $[0,+\infty)$, $(x,w) \mapsto \tau_e^*(x,w,t)$ is a diffeomorphism on $\mathcal{S}_{a,t}$ verifying :
$$\tau_e^*(x,0,t) \;=\; \tau^*(x,t) \qquad \forall x \in \mathcal{X}_t \tag{11.9}$$
and
$$\tau_e^*(\mathcal{S}_{a,t},t) = \mathbb{R}^{d_\xi} \;. \tag{11.10}$$

- In order to ensure
$$\dot{\overbrace{\tau_e^*(\hat{x},\hat{w},t)}} = \varphi(\tau_e^*(\hat{x},\hat{w},t),\hat{x},u,y) \;,$$
and conclude as before that
$$\lim_{t\to+\infty} \left| \tau_e^*\left(\hat{X}(\hat{x}_0,\hat{w}_0;t;u),\hat{W}(\hat{x}_0,\hat{w}_0;t;u),t\right) - \tau^*(X(x_0;t;u),t) \right| \;=\; 0 \;, \tag{11.11}$$
we must take into account the dependence of τ_e^* on t and take :
$$\dot{\overbrace{\begin{bmatrix} \hat{x} \\ \hat{w} \end{bmatrix}}} = \left(\frac{\partial \tau_e^*}{\partial(\hat{x},\hat{w})}(\hat{x},\hat{w},t)\right)^{-1} \left(\varphi(\tau_e^*(\hat{x},\hat{w},t),\hat{x},u,y) - \frac{\partial \tau_e^*}{\partial t}(\hat{x},\hat{w},t)\right) \;. \tag{11.12}$$

- Finally, to conclude from (11.11), that $\hat{x}$ converges to x and $\hat{w}$ to 0, we further need that the injectivity of $(x,w) \mapsto \tau_e^*(x,w,t)$ be uniform in t. When the dependence on t of τ_e^* comes from the input (and its derivatives), this property is often satisfied, in particular when those signals are bounded in time (see Lemma A.3.5). Note that a special attention should also be given to the set $\mathcal{S}_{a,t}$ which could be of the form $\mathcal{S}_t \times B_{\varepsilon(t)}$ with ε going to 0 with t. Thus, it should be checked that ε is lower bounded. A justification as to why this should be true in practice appears in the next section.

We give in the following section some elements of justification and then we illustrate this on an example about aircraft landing.

11.3.1 Partial theoretical justification

Suppose that for all t in $[0,+\infty)$, $x \mapsto \tau^*(x,t)$ is an injective immersion on some open set $\mathcal{S}_t$. Consider the extended system
$$\left\{ \begin{array}{rcl} \dot{x} &=& f(x,u(t)) \\ \dot{t} &=& 1 \end{array} \right. \quad , \quad \underline{y} = \begin{pmatrix} h(x,u(t)) \\ t \end{pmatrix}$$
with state $\underline{x} = (x,t)$. Then, the function
$$\underline{\tau^*}(\underline{x}) = (\tau^*(x,t),t)$$
is an injective immersion on
$$\underline{\mathcal{S}} = \{(x,t) \in \mathbb{R}^{d_x} \times [0,+\infty) \,:\, x \in \mathcal{S}_t\}$$
and complementing its Jacobian
$$\frac{\partial \underline{\tau^*}}{\partial \underline{x}}(\underline{x}) = \begin{pmatrix} \frac{\partial \tau^*}{\partial x}(x,t) & \frac{\partial \tau^*}{\partial t}(x,t) \\ 0 & 1 \end{pmatrix}$$
on $\underline{\mathcal{S}}$ is equivalent to complementing that of $x \mapsto \tau^*(x,t)$ on $\mathcal{S}_t$ for each t. Indeed, if $\gamma(x,t)$ is a complementation of $\frac{\partial \tau^*}{\partial x}(x,t)$ on $\mathcal{S}_t$ for each t, $\underline{\gamma}(\underline{x}) = \begin{pmatrix} \gamma(x,t) \\ 0 \end{pmatrix}$ is a complementation for

$\frac{\partial \tau^*}{\partial \underline{x}}(\underline{x})$ on $\underline{\mathcal{S}}$. And conversely, if $\underline{\gamma}(\underline{x}) = \begin{pmatrix} \gamma(x,t) \\ \alpha \end{pmatrix}$ complements $\frac{\partial \tau^*}{\partial \underline{x}}(\underline{x})$, then $\gamma(x,t) - \frac{\partial \tau^*}{\partial t}(x,t)$ complements $\frac{\partial \tau^*}{\partial x}(x,t)$.

We conclude that it is not restrictive to look for a complementation of the Jacobian of $x \mapsto \tau^*(x,t)$ at each time t. Assume it has been done and take

$$\underline{\gamma}(\underline{x}) = \begin{pmatrix} \gamma(x,t) \\ 0 \end{pmatrix} .$$

Following the methodology, we consider

$$\underline{\tau_a^*}(\underline{x}, w) = \underline{\tau^*}(\underline{x}) + \underline{\gamma}(\underline{x})w = \begin{pmatrix} \tau^*(x,t) + \gamma(x,t)w \\ t \end{pmatrix} = \begin{pmatrix} \tau_a^*(x,w,t) \\ t \end{pmatrix} .$$

Beware that Lemma 9.0.1 does not apply directly because $\underline{\mathcal{S}}$ is not bounded, thus we cannot directly conclude that there exists $\varepsilon > 0$ such that $\underline{\tau_a^*}$ is a diffeomorphism on $\underline{\mathcal{S}_a} = \underline{\mathcal{S}} \times B_\varepsilon$. However, the reader may check in the proof of [AEP14, Proposition 2] that if $\frac{\partial \tau^*}{\partial \underline{x}}(x,t)$, $\gamma(x,t)$ and $\frac{\partial \gamma}{\partial \underline{x}}(x,t)$ are bounded on $\underline{\mathcal{S}}$, the Jacobian of $\underline{\tau_a^*}$ is full-rank on $\underline{\mathcal{S}_a}$ for some ε sufficiently small. This condition is often verified in practice when the inputs are bounded. It follows that we can reasonably assume Problem 1 solved, and leaving aside Problem 2, this leads to an observer of the type (denoting $\underline{\tau_a^*}(x,w,t)$ rather than $\underline{\tau_a^*}(x,t,w)$)

$$\overbrace{\begin{bmatrix} \hat{x} \\ \hat{w} \\ \hat{t} \end{bmatrix}}^{\cdot} = \left(\frac{\partial \underline{\tau_a^*}}{\partial(x,w,t)}(\hat{x},\hat{w},\hat{t})\right)^{-1} \begin{pmatrix} \varphi(\tau_a^*(\hat{x},\hat{w},\hat{t}),\hat{x},u,\underline{y}) \\ \varphi_1(\hat{t},\underline{y}) \end{pmatrix}$$

where φ_1 should be an observer for t and we have

$$\left(\frac{\partial \underline{\tau_a^*}}{\partial(x,w,t)}\right)^{-1} = \begin{pmatrix} \frac{\partial \underline{\tau_a^*}}{\partial(x,w)} & \frac{\partial \underline{\tau_a^*}}{\partial t} \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \left(\frac{\partial \underline{\tau_a^*}}{\partial(x,w)}\right)^{-1} & -\left(\frac{\partial \underline{\tau_a^*}}{\partial(x,w)}\right)^{-1} \frac{\partial \underline{\tau_a^*}}{\partial t} \\ 0 & 1 \end{pmatrix} .$$

Of course, t being well known without any noise, we can replace $\hat{t}$ by t and φ_1 by the constant function 1. This finally gives the "reduced order" observer (11.12).

11.3.2 Application to image-based aircraft landing

In [GBC+15a, GBC+15b], the authors use image-processing to estimate the deviations of an aircraft with respect to the run-away during a landing operation thanks to vision sensors such as cameras and inertial sensors embarked on the aircraft. The objective is to make landing possible without relying on external technologies or any knowledge about the run-away. In order to estimate the position of the plane, the idea is to follow the change of position of particular points and/or particular lines on the images provided by the cameras. A strategic choice of those points/lines must be made in order to guarantee observability during the whole duration of the landing operation : for instance, a point may disappear from the image, and a line can stop moving on the image in some particular alignment conditions, thus providing no (or only partial) information about the movement of the aircraft. A full study of those methods can be found in [Gib16]. A possible choice ensuring observability is to follow on the image the position of the two lateral lines of the run-away and the reference point at the end of the run-away. It

gives the following model :

$$\left\{\begin{array}{rcl} \dot\theta_1 &=& \sigma_1(\rho_1,\theta_1,t)+\pi_1(\rho_1,\theta_1,t)\eta \\ \dot\rho_1 &=& \sigma_2(\rho_1,\theta_1,t)+\pi_2(\rho_1,\theta_1,t)\eta \\ \dot\theta_2 &=& \sigma_1(\rho_2,\theta_2,t)+\pi_1(\rho_2,\theta_2,t)\eta \\ \dot\rho_2 &=& \sigma_2(\rho_2,\theta_2,t)+\pi_2(\rho_2,\theta_2,t)\eta \\ \dot\nu_1 &=& (V_H\nu_1-V_X)\eta \\ \dot\nu_2 &=& (V_H\nu_2-V_Y)\eta \\ \dot\eta &=& V_H\eta^2 \end{array}\right. \quad , \quad y=(\theta_1,\rho_1,\theta_2,\rho_2,\nu_1,\nu_2)$$

where (θ_i,ρ_i) and (ν_1,ν_2) are the measured position on the image of the two lines and the point respectively, the functions σ and π are defined by

$$\begin{array}{rcl} \sigma_1(\rho,\theta,t) &=& -\omega_1\rho\cos\theta-\omega_2\rho\sin\theta-\omega_3 \\ \sigma_2(\rho,\theta,t) &=& (1+\rho^2)(\omega_1\sin\theta-\omega_2\cos\theta) \\ \pi_1(\rho,\theta,t) &=& (a\sin\theta-b\cos\theta)(v_1\cos\theta+v_2\sin\theta-v_3\rho) \\ \pi_2(\rho,\theta,t) &=& (a\rho\cos\theta+b\rho\sin\theta+c)(v_1\cos\theta+v_2\sin\theta-v_3\rho) \end{array}$$

where the aircraft velocities v and ω expressed in the camera frame, the aircraft velocities V_X, V_Y and V_H expressed in the runway frame, and camera orientations (a,b,c) are known input signals.

Denoting $x_m=(\theta_1,\rho_1,\theta_2,\rho_2,\nu_1,\nu_2)$ the measured part of the state, we obtain a model with state

$$x=(x_m,\eta)\in\mathbb{R}^7$$

and dynamics of the form[2]

$$\left\{\begin{array}{rcl} \dot x_m &=& \Sigma(x_m,t)+\Pi(x_m,t)\,\eta \\ \dot\eta &=& V_H(t)\,\eta^2 \end{array}\right. \quad , \quad y=x_m\ , \tag{11.13}$$

where the action of the input $u=(a,b,c,v,w,V_X,V_Y,V_H)$ is represented by a time-dependence[3] to simplify the notations in the rest of this section. This system is observable if and only if the unmeasured state η can be uniquely determined from the knowledge of the measured state x_m. From the structure of the dynamics, we notice that this is possible if the quantity

$$\delta(x_m,t)=\Pi(x_m,t)^\top\Pi(x_m,t) \tag{11.14}$$

never vanishes. It is the case in practice, thanks to a sensible choice of lines and point (see [Gib16] for a thorough observability analysis during several landing operations).

A high-gain observer

Assumptions

- The input signal $u=(v,w,V_X,V_Y,V_H)$ and its first derivative are bounded in time.

- There exists a strictly positive number ε and a compact subset $\mathcal{C}$ of $\mathbb{R}^7$ such that for

[2]Note that whatever the number of chosen lines and points in the image, the model can always be written in this form, only the dimensions of x_m and the input change.

[3]This comes back to choosing one particular input law, but the reader may check that the same design works for any input such that the observability assumption and the saturation by $\overline{\Phi}$ in (11.18) are valid.

where for each time t, we define

$$\mathcal{S}_t = \{x \in \mathbb{R}^7 \,,\ \delta(x_m, t) \geq \varepsilon\} \ .$$

In other words, δ remains greater than ε along any solution of the system, making it observable. Under this assumption, we know that the state x can be reconstructed from the measurement x_m and its first derivative. We thus consider the transformation $\tau_0^* : \mathbb{R}^7 \times \mathbb{R} \to \mathbb{R}^{12}$ made of y and its first derivative i-e :

$$\tau_0^*(x,t) = \overline{\mathbf{H}}_2(x, u(t)) = \begin{pmatrix} x_m \\ \Sigma(x_m,t) + \Pi(x_m,t)\,\eta \end{pmatrix} \ . \tag{11.16}$$

For any t in $\mathbb{R}$, $\tau_0^*(\cdot\,,\,t)$ is an injective immersion on $\mathcal{S}_t$. Since u, $\dot{u}$ and the trajectories are bounded, we deduce from Theorem 5.2.1 and Remark 4 that τ_0^* transforms the system into a phase variable form

$$\left\{ \begin{array}{rcl} \dot{\xi}_m &=& \xi_d \\ \dot{\xi}_d &=& \Phi_2(\xi, u(t), \dot{u}(t)) \end{array} \right. \quad , \quad y = \xi_m \tag{11.17}$$

where ξ_m denotes the first six components of ξ and ξ_d the six others, and Φ_2 can be defined by

$$\Phi_2(\xi, \nu_0, \nu_1) = \mathsf{sat}(L_{\overline{f}}^2 \overline{h}(\tau_0(\xi,t), \nu_0, \nu_1), \overline{\Phi}) \tag{11.18}$$

with $\overline{f}$ and $\overline{h}$ as defined in Definition 5.2.1, $\overline{\Phi}$ a bound of $L_{\overline{f}}^2 \overline{h}(x, \nu_0, \nu_1)$ for x in $\mathcal{C}$ and (ν_0, ν_1) bounded by the bound for $(u, \dot{u})$, and $\xi \mapsto \tau_0(\xi, \cdot)$ any locally Lipschitz function defined on $\mathbb{R}^{12}$ such that it is a left-inverse[4] of $x \mapsto \tau_0^*(x, \cdot)$ for x in $\mathcal{X}_t$.

We have the following observer for System (11.17):

$$\left\{ \begin{array}{rcl} \dot{\hat{\xi}}_m &=& \hat{\xi}_d + Lk_1(y - \hat{\xi}_m) \\ \dot{\hat{\xi}}_d &=& \Phi_2(\hat{\xi}, u, \dot{u}) + L^2 k_2 (y - \hat{\xi}_m) \end{array} \right. \quad , \quad y = \xi_m \tag{11.19}$$

with $k_1, k_2 > 0$ and L sufficiently large. Although a left inverse τ_0 of τ_0^* can be found in that case, and an estimate $\hat{x}$ of x could be computed by $\hat{x} = \tau_0(\hat{\xi}, t)$ as proposed by Theorem 2.2.1, we would like to express the dynamics of this observer directly in the x-coordinates.

Observer in the given coordinates

Fictitious output Following the same idea as for the oscillator with unknown frequency, we start by removing the injectivity singularity of τ_0^* outside of $\mathcal{S}_t$, i-e we look for an alternative function τ^* which is an injective immersion on $\mathbb{R}^7$. Notice that the function

$$\wp(x_m, t) = \max\left\{\varepsilon - \delta(x_m, t), 0\right\}^4 \tag{11.20}$$

is zero in $\mathcal{S}_t$ and nonzero outside of $\mathcal{S}_t$. According to (**??**), this function remains equal to 0 along the solutions and therefore so does the fictitious output

$$y_7 = \wp(x_m, t)\eta \ .$$

It follows that y_7 can be considered as an extra measurement traducing the information of observability. Consider now the function

$$\tau^*(x,t) = (\tau_0^*(x,t)\,,\ \wp(x_m,t)\eta) \ .$$

[4]Take for instance $\tau_0(\xi,t) = \left(\xi_m, \frac{\Pi(\xi_m)^T(\xi_d - \Sigma(\xi_m,t))}{\max\{\delta(\xi_m,t),\varepsilon\}}\right)$.

Unlike $\tau_0^*(\cdot,t)$, $\tau^*(\cdot,t)$ is an injective immersion on the whole space $\mathbb{R}^7$ for all t. Indeed, $\Pi(x_m,t)$ and $\wp(x_m,t)$ cannot be zero at the same time so that the new coordinate $\wp(x,t)\eta$ enables to have the information on η when Π is zero. Besides, its Jacobian

$$\frac{\partial \tau^*}{\partial x}(x,t) = \begin{pmatrix} I_{6\times 6} & 0_{6\times 1} \\ * & \Pi(x_m,t) \\ * & \wp(x_m,t) \end{pmatrix} \tag{11.21}$$

is full-rank everywhere.

Immersion augmentation into diffeomorphism by Jacobian complementation. Following the methodology presented in this Part III, we extend the injective immersion $\tau^*(\cdot,t)$ into a diffeomorphism. The first step consists in finding a C^1 matrix $\gamma(x,t)$ in $\mathbb{R}^{13\times 6}$ such that the matrix

$$\left(\frac{\partial \tau^*}{\partial x}(x,t) \quad , \quad \gamma(x,t)\right)$$

is invertible for any x and any t. In others words, we want to complement the full-rank rectangular matrix $\frac{\partial \tau^*}{\partial x}(x,t)$ with 6 vectors in $\mathbb{R}^{13}$ which make it square and invertible. Thanks to the identity block, it is in fact sufficient to find 6 independent vectors in $\mathbb{R}^7$ which complement the vector $\begin{pmatrix}\Pi(x_m,t)\\ \wp(x_m,t)\end{pmatrix}$. A first solution would be to implement Wazewski's algorithm on $\mathbb{R}^6$ which is contractible, like in Example 9.3.1, but this leads to rather tedious computations. Since Problem $\tilde{P}[7,1]$ is not in the list (9.8) of cases admitting universal formulas, we could had another fictitious output $y_8 = 0$ like we did for the oscillator to recover a solvable problem $\tilde{P}[8,1]$. We present here another path which does not necessitate lengthy computations nor an additional output. The idea comes from the remark made in Section 11.1.3 that when $\begin{pmatrix}\frac{\partial \tau^*}{\partial x}\\ 0\end{pmatrix}$ is full rank, it can always be complemented by $\begin{pmatrix}-I\\ \frac{\partial \tau^*}{\partial x}^\top\end{pmatrix}$ because the resulting matrix has a determinant equal to $\det\left(\frac{\partial \tau^*}{\partial x}^\top \frac{\partial \tau^*}{\partial x}\right) \neq 0$. In our case, we remark that the determinant of the matrix $\begin{pmatrix}\Pi(x_m,t) & -I_{6\times 6}\\ \wp(x_m,t) & \Pi(x_m,t)^\top\end{pmatrix}$ is equal to $\wp(x_m,t) + \Pi(x_m,t)^T\Pi(x_m,t)$ which never vanishes by definition. Thus, a possible candidate for complementation is :

$$\gamma(x_m,t) = \begin{pmatrix} 0_{6\times 6} \\ -I_{6\times 6} \\ \Pi(x_m,t)^\top \end{pmatrix} .$$

As recommended by Lemma 9.0.1, we now introduce the extension of τ^* defined on $\mathbb{R}^7\times\mathbb{R}^6\times\mathbb{R}$ by

$$\tau_e^*(x,w,t) = \tau^*(x,t) + \gamma(x_m,t)w \ . \tag{11.22}$$

Besides, thanks to the fact that γ does not depend on η, we have :

$$\frac{\partial \tau_e^*}{\partial (x,w)}(x,w,t) = \begin{pmatrix} \mathsf{Id}_{6\times 6} & 0_{6\times 1} & 0_{6\times 6} \\ * & \Pi(x_m,t) & -\mathsf{Id}_{6\times 6} \\ * & \wp(x_m,t) & \Pi(x_m,t)^\top \end{pmatrix}$$

which is invertible for any (x,w) in $\mathbb{R}^{13}$ and any time t. In fact, as for the high gain observer for the oscillator, $\tau_e^*(\cdot,t)$ is a diffeomorphism on $\mathbb{R}^{13}$ such that $\tau_e^*(\mathbb{R}^{13},t) = \mathbb{R}^{13}$ for any t. Thus, we have managed to transform an injective immersion $\tau^*(\cdot,t) : \mathbb{R}^7 \to \mathbb{R}^{13}$ into a surjective diffeomorphism $\tau_e^*(\cdot,\cdot,t) : \mathbb{R}^{13} \to \mathbb{R}^{13}$.

Observer in the given coordinates As suggested at the beginning of this section, we consider the observer :

$$\dot{\overbrace{\begin{bmatrix} \hat{x} \\ \hat{w} \end{bmatrix}}} = \left(\frac{\partial \tau_e^*}{\partial(x,w)}(\hat{x},\hat{w},t)\right)^{-1} \left(\varphi(\tau_e^*(\hat{x},\hat{w},t),\hat{x},t,y) - \frac{\partial \tau_e^*}{\partial t}(\hat{x},\hat{w},t)\right) \tag{11.23}$$

where φ is defined on $\mathbb{R}^{13} \times \mathbb{R}^7 \times \mathbb{R} \times \mathbb{R}^6$ by

$$\varphi(\hat{\xi},\hat{x},t,y) = \begin{pmatrix} \hat{\xi}_d + Lk_1(y - \hat{\xi}_m) \\ \mathsf{sat}(L_f^2 \overline{h}(\hat{x},u(t),\dot{u}(t)),\overline{\Phi}) + L^2 k_2(y - \hat{\xi}_m) \\ -a\hat{\xi}_a \end{pmatrix}$$

with $\hat{\xi} = (\hat{\xi}_m, \hat{\xi}_d, \hat{\xi}_a) \in \mathbb{R}^6 \times \mathbb{R}^6 \times \mathbb{R}$, a any strictly positive number. A result of a simulation is given in Figure 11.3.

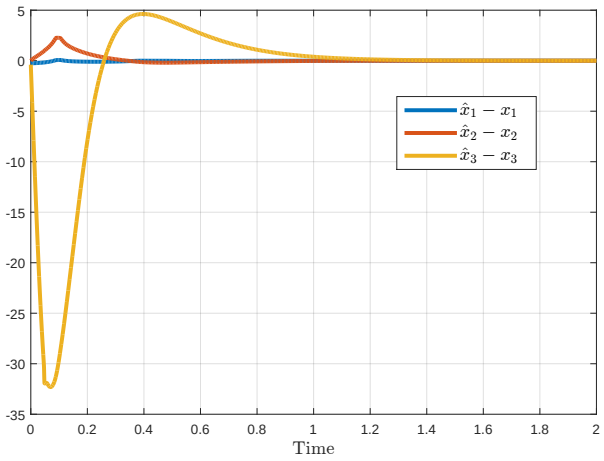
11.4 Conclusion

We have presented a method to express the dynamics of an observer in the given system coordinates, thereby avoiding the difficult left-inversion of an injective immersion. It assumes the knowledge of an injective immersion and a converging observer for the immersed system through Assumption $\mathcal{O}$.

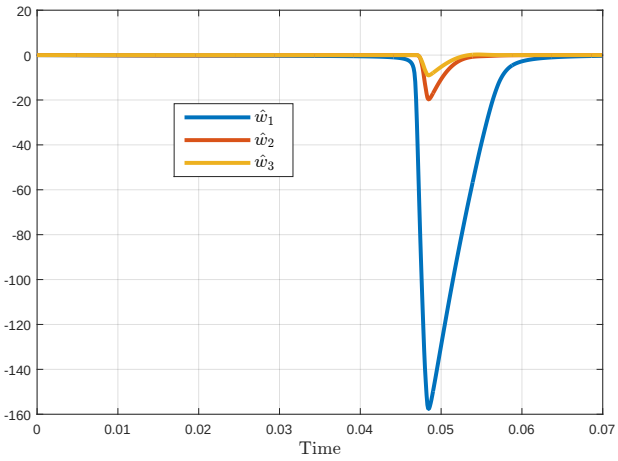
The idea is not to modify this observer dynamics but to map it back to the given coordinates in a different way. Our construction involves two tools : the augmentation of an injective immersion into a diffeomorphism through a Jacobian complementation (Chapter 9) and the extension of the image of the obtained diffeomorphism to enlarge the domain where the observer solutions can go without encountering singularities (Chapter 10).

For those tools to be usable, some assumptions on the domain of injectivity must be verified, but we have seen how they can be fulfilled in practice through a wise choice of the transformation τ^*, and how those tools also extend to the case where the transformation is time-varying.

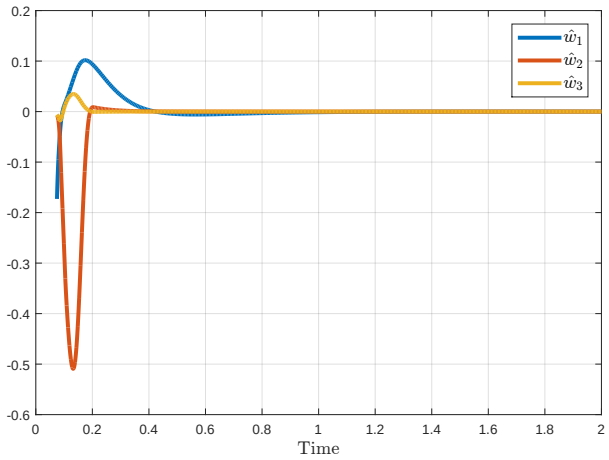
To conclude from an implementation point of view, the tools presented in Chapter 9 to augment an injective immersion into a diffeomorphism are sufficiently constructive and general to be applicable in practice. We have even given a universal complementation method in this chapter. The main limitation of this method rather appears when wanting to extend the image of this diffeomorphism. Indeed, the only constructive result presented in Chapter 10 requires this set to be precisely known and also to satisfy some extra conditions. Although we have shown that it is sometimes possible to use an approximation, it lacks in generality and this step constitutes a significant difficulty in practice. Other solutions may exist and need to be developed, in particular to keep the amplitude of the extra coordinates $\hat{w}$ small to preserve the invertibility of the Jacobian.



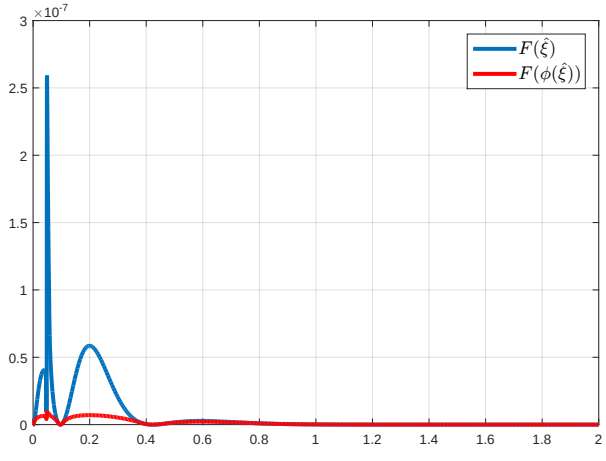
(a) $\hat{x} - x$



(b) $\hat{w}$ for $t \in [0, 0.07]$

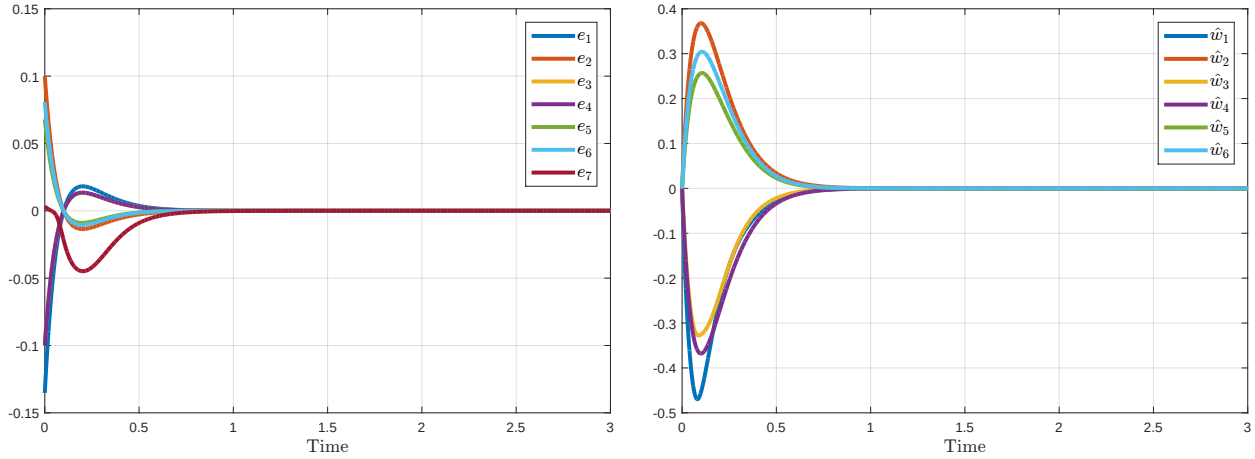


(c) $\hat{w}$ for $t \in [0.07, 2]$



(d) Value of F before and after applying ϕ

Figure 11.2: Luenberger observer (11.8) with $\hat{x}_1 = 0.08$, $\hat{x}_2 = \hat{x}_3 = 0$, $\lambda_1 = 6$, $\lambda_2 = 9$, $\lambda_3 = 14$, $\lambda_4 = 15$, and $\tau_e^* = \phi^{-1} \circ \tau_a^*$ instead of τ_a^*. The simulation was done with a variable step Euler algorithm.



(a) $e_i = \hat{x}_i - x_i$

(b) $\hat{w}_i$

Figure 11.3: Observer (11.23) with $L = 10$ and $k_1 = k_2 = 1$. The simulation was run on Simulink.

Part IV

Observers for permanent magnet synchronous motors with unkown parameters

Chapter 12

Short introduction to permanent magnet synchronous motors

Chapitre 12 – Courte introduction aux moteurs synchrones à aimant permanent. *Dans ce chapitre, nous présentons rapidement le fonctionnement et le modèle d'un moteur synchrone à aimant permanent (MSAP). En pratique, il est crucial de savoir estimer en ligne la position du rotor et sa vitesse de rotation, ceci avec un minimum de capteurs pour des raisons de coût et de contraintes mécaniques. En particulier, des chercheurs ont développé le contrôle "sensorless", c'est-à-dire basé seulement sur les mesures des variables électriques (tensions et intensités) et non mécaniques (angle du moteur, vitesse). En particulier, des observateurs de type gradient ont été proposés et sont rappelés ici. Cependant, ces observateurs dépendent le plus souvent de paramètres tels que la résistance et le flux de l'aimant, qui peuvent varier significativement avec la température. Il est donc important de trouver des observateurs de la position du rotor qui sont indépendants de ces paramètres, voire qui en donnent une estimation : c'est le problème considéré dans cette partie.*

A Permanent Magnet Synchronous Motor is composed of a permanent magnet rotor placed inside a stator made of windings whose repartition and currents are chosen in order to create a rotating magnetic field in the airgap of the machine. A torque is then produced on the permanent rotor magnet due to magnetic attraction, thus inducing the rotor to rotate. Compared to other commonly used induction machines (see Figure 12.1), the absence of rotor windings and external rotor excitation reduces the maintenance costs as well as losses in the rotor, and makes PMSMs highly efficient.

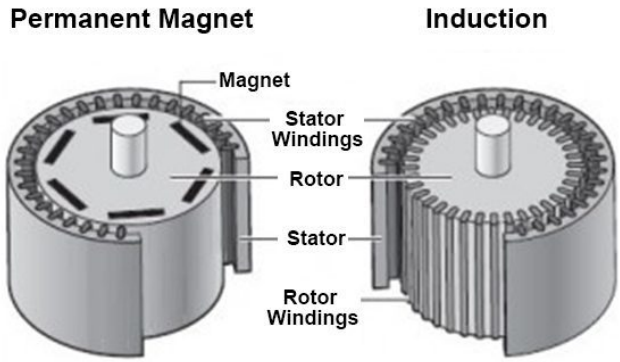


Figure 12.1: Permanent Magnet vs Induction Motor

Using Joule's and Faraday's laws, a PMSM model expressed in a fixed stator frame reads

$$\dot{\Psi} = u - R\,i \tag{12.1}$$

where Ψ is the total flux generated by the stator windings and the permanent magnet, (u, i) are the voltage and intensity of the current in the fixed stator frame and R the stator winding resistance. The quantities u, i and Ψ are two dimensional vectors. The way the total flux Ψ is related to the rotor angle θ_r differs depending on the geometry of the rotor and stator. When the repartition of the windings and the profile of the magnet are perfectly symmetric, the motor is said to be *non-salient* and the total flux may be expressed simply as

$$\Psi = L\,i + \Phi \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{12.2}$$

where L is the stator inductance, Φ the magnet's flux, and $\theta = n_p\theta_r$ the electrical phase, with n_p the number of poles (winding pairs) of the stator. This relation implies

$$|\Psi - L\,i|^2 - \Phi^2 = 0 \tag{12.3}$$

$$\theta = \arg(\Psi - L\,i)\ . \tag{12.4}$$

This model may appear unorthodox to those who are rather used to models of the type

$$\begin{aligned} L\,\dot{\widehat{i}} &= u - R\,i - \Phi\,n_p\,\omega_r \begin{pmatrix} -\sin(n_p\theta_r) \\ \cos(n_p\theta_r) \end{pmatrix} \\ \dot{\theta}_r &= \omega_r \\ J\,\dot{\omega}_r &= \Phi\,n_p\,i^\top \begin{pmatrix} -\sin(n_p\theta_r) \\ \cos(n_p\theta_r) \end{pmatrix} - \tau_L \end{aligned} \tag{12.5}$$

where J is the inertia of the rotor and τ_L the load torque. However, they should observe that the electrical part of this model (first line) is actually obtained by plugging (12.2) into (12.1). But this operation makes ω_r appear and they are then forced to integrate it in the model with the mechanical part (third line). The drawback is that it depends on two new parameters J and τ_L which must be either known or estimated. That is why we rather keep the model made of (12.1)-(12.2).

To minimize the cost and increase the reliability of PMSMs, it is strategic to make progress on estimating online the rotor position θ_r and speed $\omega_r = \dot{\theta}_r$, with a minimum of sensors and fast algorithms. To this end, researchers have developed the so-called "sensorless" control which uses no measurement of mechanical variables, only of electrical ones, namely (u, i). Indeed, cost as well as mechanical constraints often render the integration of position sensors troublesome, or even impossible.

According to (12.4), in the case where L and i are known, an estimate of θ can be simply recovered from an estimate of the total flux Ψ. Thus, it is enough to design an observer for the system

$$\dot{\Psi} = u - R\,i \quad , \quad y = |\Psi - Li|^2 - \Phi^2 = 0\ , \tag{12.6}$$

with known inputs (u, i) and where the information given by (12.3) is used as a measurement. A review of the first steps in that direction was given in [AW06]) and a Luenberger observer was proposed in [PPO08]. More recently in [LHN$^+$10], was proposed the very simple gradient observer

$$\dot{\hat{\Psi}} = u - Ri - 2q\,(\hat{\Psi} - Li)\,\left(\left|\hat{\Psi} - Li\right|^2 - \Phi^2\right)\ , \tag{12.7}$$

which turned out to be extremely effective in practice as rotor position estimator. However, from a theoretical view point, it was proved in [OPA$^+$11] to be only conditionally convergent

: it may admit several equilibrium points depending on the rotation speed ω. In fact, later in [MPH12], the author showed that the following minor modification

$$\dot{\hat{\Psi}} = u - Ri - 2q\left(\hat{\Psi} - Li\right) \max\left(\left|\hat{\Psi} - Li\right|^2 - \Phi^2, 0\right) \tag{12.8}$$

enables to achieve global asymptotic stability thanks to convexity arguments.

All these observers typically require the knowledge of the resistance R, magnet flux Φ and inductance L. Unfortunately while L may be considered known and constant (as long as there is no magnetic saturation), R and Φ do vary significantly with the temperature and these variations should be taken into account in the observer. For example, for a given injected current, when the magnet's temperature increases, its magnetic flux decreases, and the produced torque becomes smaller. Therefore, an online estimation of the magnet's flux enables to :

- adapt the control law in real time and thus ensure a torque control which is robust to the machine's temperature ;
- have an estimation of the rotor's temperature
- have an estimation of the magnet's magnetization degradation with time.

That is why efforts have been made to look for position observers which do not rely on the knowledge of those parameters, or even better, which also estimates them. For instance, [HMP12, BPO15a] have proposed observers which are independent from the magnet flux. We complete this line of research in Chapter 13 by extending the gradient observer (12.7) with the estimation of Φ : global convergence is established when the rotation speed stays away from zero and its performances are compared to that of other existing observers. As for the resistance, in [ROH$^+$16], the authors propose and study via simulations an adaptive observer to make the gradient observer previously mentioned independent from the resistance. However, the convergence is not ensured and actually we show in Chapter 14 that the system is not observable when R is unknown unless other informations are added. When those informations are available, we propose a novel Luenberger observer.

Chapter 13

Rotor position estimation with unknown magnet flux

Chapitre 13 – Estimation de la position d'un rotor lorsque le flux des aimants est inconnu. *Dans ce chapitre, nous proposons un nouvel observateur "sensorless" qui estime la position du rotor sans avoir à connaître le flux des aimants : seules les mesures des intensités et courants, et les valeurs de l'inductance et de la résistance sont nécessaires. Cet observateur étend l'observateur gradient introduit dans [LHN+10] en ajoutant l'estimation du flux des aimants, et le rend globalement convergent si la vitesse de rotation ne s'approche pas de zéro. Nous étudions sa sensibilité aux incertitudes de résistance et inductance, ainsi qu'à la présence de saillance. Ses performances en boucle ouverte sont illustrées par des simulations sur des données réelles et comparées à d'autres observateurs indépendants du flux existant dans la littérature, à la fois en terme de coût en calcul et de robustesse.*

Contents

In this chapter, we address the problem of estimating the rotor position of a PMSM without relying on the knowledge of the magnet's flux, i-e when only electrical measurements and (approximate) knowledge of the resistance and inductance are available.

First steps in this direction are reported in [HMP12] with the design of a Luenberger observer (see [Hen14] for a much more detailed analysis and Section 7.2.1), and in [BPO15a, BPO+15b, BBP+16], with the design of an observer based on tools from parameter linear identification. In fact, we will show that those two observers rely on the same regression equation but the former solves it at each time whereas the latter solves it as time goes on with a gradient-like scheme. Convergence comes under an assumption of invertibility of the regressor matrix for the former, and on a persistent excitation condition for the latter.

In the same line of research, we propose here a new observer which extends the gradient observer from [LHN+10] with the estimation of the magnet's flux, and makes it globally convergent provided the rotation speed remains away from zero. We study its sensitivity to uncertainties

on the resistance and inductance and to the presence of saliency. Its performances in open-loop are illustrated via simulations on real data and compared to the other previously mentioned magnet flux independent observers in terms of computational cost and robustness.

The content of this chapter was presented in [BP17].

Notations The rotation matrix is denoted

$$\mathcal{R}(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} .$$

13.1 Gradient observer

Since Φ is unknown in this chapter, we consider the system

$$\begin{cases} \dot{\Psi} &= u - Ri \\ \dot{\Phi} &= 0 \\ y &= |\Psi - Li|^2 - \Phi^2 \end{cases} \tag{13.1}$$

with inputs (u, i), known parameters (R, L), state (Ψ, Φ) and measurement y which is constantly zero. We introduce the corresponding gradient observer

$$\begin{cases} \dot{\hat{\Psi}} &= u - Ri - 2q\,(\hat{\Psi} - Li)\,\left(\left|\hat{\Psi} - Li\right|^2 - \hat{\Phi}^2\right) \\ \dot{\hat{\Phi}} &= q\,\hat{\Phi}\left(\left|\hat{\Psi} - Li\right|^2 - \hat{\Phi}^2\right) \end{cases} \tag{13.2}$$

where q is an arbitrary strictly positive real number. It is a straightforward extension of observer (12.7) with the estimation of Φ.

Theorem 13.1.1.

Consider (ψ, Φ) in $\mathbb{R}^2 \times (0, +\infty)$ and inputs $u, i : \mathbb{R} \to \mathbb{R}^2$ such that there exists strictly positive numbers $\underline{\omega}_0$, $\overline{\omega}_0$ and $\overline{\omega}_1$ such that the solution $(\Psi(\psi; t; u, i), \Phi)$ of (13.1) verifies

$$0 < \underline{\omega}_0 \leq \dot{\theta}(t) \leq \overline{\omega}_0 \quad , \quad \ddot{\theta}(t) \leq \overline{\omega}_1$$

with

$$\theta(t) = \arg(\Psi(\psi; t; u, i) - Li(t)) .$$

For any strictly positive real number q, for any $(\hat{\psi}, \hat{\phi})$ in $\mathbb{R}^2 \times (0, +\infty)$, the solution $(\hat{\Psi}(\hat{\psi}, \hat{\phi}; t; u, i), \hat{\Phi}(\hat{\psi}, \hat{\phi}; t; u, i))$ of (13.2) satisfies

$$\lim_{t\to\infty} |\hat{\Psi}(t) - \Psi(t)| + |\hat{\Phi}(t) - \Phi| = 0 ,$$

where we have used the abbreviation $\hat{\Psi}(t) = \hat{\Psi}(\hat{\psi}, \hat{\phi}; t; u, i)$, $\hat{\Phi}(t) = \hat{\Phi}(\hat{\psi}, \hat{\phi}; t; u, i)$ and $\Psi(t) = \Psi(\psi; t; u, i)$.

In other words, System (13.2) is an observer for System (13.1) for the solutions with a bounded rotation speed which remains away from zero. Of course, taking $\hat{\theta}$ as the argument of $\hat{\Psi} - Li$, we also obtain

$$\lim_{t\to\infty} \hat{\theta}(t) - \theta(t) = 0 .$$

Proof : The proof of Theorem 13.1.1 is lengthy and technical, so we only give here the most important steps. The whole proof is available in Appendix C.

Consider a solution (Ψ, Φ) of (13.1), with Φ in $(0, \infty)$ and define

$$\theta(t) = \arg(\Psi(t) - Li(t)) ,$$

so that

$$\Psi(t) = Li(t) + \Phi \begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix} . \tag{13.3}$$

Pick $(\hat{\psi}, \hat{\phi})$ in $\mathbb{R}^2 \times (0, \infty)$, and $q > 0$. To ease the notations, we denote the corresponding solution of (13.2) $(\hat{\Psi}(t), \hat{\Phi}(t))$. According to (13.3), it is enough to prove that

$$\lim_{t\to\infty} \left[\hat{\Psi}(t) - Li(t)\right] - \Phi \begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix} = 0$$

and

$$\lim_{t\to\infty} \hat{\Phi}(t) = \Phi .$$

To simplify our task, we transform the solution $\left(Li + \Phi \begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix}, \Phi\right)$ into an equilibrium. Thus, we carry out the analysis in the coordinates

$$\begin{pmatrix} X_d \\ X_q \end{pmatrix} = \mathcal{R}(-\theta)\ (\Psi - Li) \quad , \quad \begin{pmatrix} \hat{X}_d \\ \hat{X}_q \end{pmatrix} = \mathcal{R}(-\theta)\ \left(\hat{\Psi} - Li\right) .$$

With (13.3), we obtain

$$\begin{pmatrix} X_d \\ X_q \end{pmatrix} = \begin{pmatrix} \Phi \\ 0 \end{pmatrix}$$

and it is enough to show that

$$\lim_{t\to\infty} \hat{X}_d(t) = \Phi \quad , \quad \lim_{t\to\infty} \hat{X}_q(t) = 0 \quad , \quad \lim_{t\to\infty} \hat{\Phi}(t) = \Phi .$$

In those coordinates, the dynamics of the observer reads :

$$\left\{ \begin{array}{rcl} \dot{\hat{X}}_d &=& \omega \hat{X}_q - 2q\hat{X}_d \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2\right) \\ \dot{\hat{X}}_q &=& -\omega \hat{X}_d + \omega\Phi - 2q\hat{X}_q \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2\right) \\ \dot{\hat{\Phi}} &=& q\,\hat{\Phi} \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2\right) \end{array} \right. \tag{13.4}$$

where $\omega(t) = \dot{\theta}(t)$. The set $\Omega = \mathbb{R}^2 \times (0, +\infty)$ being forward invariant for these dynamics, we study the behavior of its solutions when they are in Ω. The proof consists in finding a Lyapunov function decreasing along the solutions of (13.4), and proving convergence to $(\Phi, 0, \Phi)$ which is the only equilibrium point in Ω. More precisely :

1.The function

$$V(\hat{X}_d, \hat{X}_q, \hat{\Phi}) = \frac{\hat{\Phi}^4}{4} + \frac{1}{2}\hat{\Phi}^2(\hat{X}_d^2 + \hat{X}_q^2) - \Phi\hat{\Phi}^2\hat{X}_d + \frac{\Phi^4}{4}$$

is a Lyapunov function. It satisfies :

$$\dot{V} = -q\,\hat{\Phi}^2(\hat{\Phi}^2 - (\hat{X}_d^2 + \hat{X}_q^2))^2 \leq 0 .$$

2.Any solution of (13.4) starting in Ω is bounded and is defined in Ω for all t in $[0, +\infty)$. Then, thanks to Barbalat's lemma,

$$\lim_{t\to+\infty} \hat{\Phi}(t)(\hat{\Phi}(t)^2 - (\hat{X}_d(t)^2 + \hat{X}_q(t)^2)) = 0 \quad , \quad \lim_{t\to+\infty} \hat{\Phi}(t)\hat{X}_q(t) = 0 \quad , \quad \lim_{t\to+\infty} \hat{\Phi}(t)(\hat{X}_d(t) - \Phi) = 0 .$$

3.It is not possible to have $\liminf_{t\to+\infty} \hat{\Phi}(t) = 0$.

■

Theorem 13.1.1 tells us that unlike for observer (12.7), no convexification is needed to achieve global convergence of the gradient observer (13.2). Hence, even when the parameter Φ is known, we may prefer to use observer (13.2) instead of observer (12.8). In this way, although the observer state is augmented with $\hat{\Phi}$, we get global convergence without knowing Φ.

13.2 Alternative path

The observer presented in the previous section is based on System (13.1) which is nonlinear because of its output function. Fortunately, this function is quadratic in (Ψ, Φ), and $(\dot{\Psi}, \dot{\Phi})$ does not depend on (Ψ, Φ). Hence linearity can be obtained by time derivation. Namely, we have

$$\dot{y} = 2\,(\Psi - Li)^T\,(u - Ri - \widehat{\dot{Li}})$$

which is linear in Ψ and independent from Φ. The new problem we face now is the presence of the time derivative $\widehat{\dot{Li}}$. A well known fix to this, is to use a strictly causal filter. Namely, let

$$\dot{\eta} \;=\; -\lambda(\eta + y) \quad , \quad y_f = \eta + y \tag{13.5}$$

with λ any complex number with strictly positive real part. It is easy to check that the evaluation of $y_f + (c + 2Li)^T\Psi - (z + L^2|i|^2)$, along any solution, decreases as $\exp(-\lambda t)$ when c and z are solutions of

$$\left\{ \begin{array}{rcl} \dot{c} &=& -\lambda c - 2\lambda Li - 2(u - Ri) \\ \dot{z} &=& -\lambda z + c^T(u - Ri) - \lambda L^2|i|^2\ . \end{array} \right. \tag{13.6}$$

So, instead of the design model (13.1), we can use

$$\dot{\Psi} = u - Ri \quad , \quad y_f = -(c + 2Li)^T\Psi + (z + L^2|i|^2) \tag{13.7}$$

with inputs (u, i, c, z), state Ψ and measurement y_f. Also because of (13.5), we pick y_f constantly zero as we did above with y. System (13.7) can be seen as a linear time varying system and therefore any observer design for such systems apply. It can be a Kalman filter or more simply the following gradient observer

$$\left\{ \begin{array}{l} \dot{c} = -\lambda c - 2\lambda Li - 2(u - Ri) \\ \dot{z} = -\lambda z + c^T(u - Ri) - \lambda L^2|i|^2 \\ \dot{\hat{\Psi}} = u - Ri + \gamma\,(c + 2Li)\Big(-(c + 2Li)^T\hat{\Psi} + z + L^2|i|^2\Big). \end{array} \right. \tag{13.8}$$

where γ is an arbitrary strictly positive real number. In [BPO15a], the authors propose the following non minimal version of this observer :

$$\left\{ \begin{array}{rcl} \dot{\xi}_{14} &=& u - Ri \\ \dot{\xi}_5 &=& -\lambda(\xi_5 - |\xi_{14} - Li|^2) \\ \dot{\xi}_{89} &=& \gamma\Omega\left(y - \Omega^T\xi_{89}\right) \end{array} \right. \quad , \quad y = -\lambda|\xi_{14} - Li|^2 - \lambda\xi_5 \tag{13.9}$$

with

$$\begin{array}{rcl} \hat{\Psi} &=& \xi_{14} + \xi_{89} \\ \Omega &=& -\lambda(c + 2Li) \end{array}$$

where c verifies the dynamics (13.8) and we have the relation

$$z = \xi_{14}^T(c + \xi_{14}) + \xi_5\ .$$

with z satisfying (13.8).

Convergence of these observers (13.8) or (13.9) is guaranteed as long as Ω satisfies a persistent excitation condition which, as proved in [BPO15a], holds when the rotation speed is sufficiently rich.

Inspired from nonlinear Luenberger observers, another observer is proposed in [HMP12]. It consists in using m filters of the type (13.6), with poles λ_k, with k in $\{1, \ldots, m\}$, to obtain m equations in $\hat{\Psi}$

$$(c_k + 2Li)^T\hat{\Psi} - (z_k + L^2|i|^2) \;=\; 0 \tag{13.10}$$

which are solved in a least square sense. It is proved in [Hen14] that the matrix of the $c_k + Li$ is full column rank when ω stays away from 0, $m \geq 3$ and the λ_k are chosen in a generic way.

Actually, observer (13.8), observer (13.9) of [BPO15a], or the one in [HMP12], are identical except in their way of solving in $\hat{\Psi}$ equations (13.10). The former two solve (13.10) with only one λ ($m = 1$) but dynamically along time. The later solves them at each time, with at least two λ ($m \geq 2$).

In the remainder of the chapter, we intend to compare the performances of observer (13.2) introduced in the previous section with those of this other family of observers, in particular observer (13.8).

13.3 Performances

13.3.1 Computational cost

We already see that the smaller dimension of observer (13.2) and its great simplicity of implementation provides a significant advantage. Indeed, in our matlab simulations, CPU time was found to be twice smaller than for the other observers presented in Section 13.2. This numerical efficiency constitutes an important feature since those observers are intended to run online where processing power is often limited.

13.3.2 Sensitivity to the presence of saliency when i_d is constant

According to [BC98], the simplest way to take saliency into account in the model of a PMSM is to keep (12.1) but to replace the expression (12.2) of the total flux by

$$\Psi = L_0 i + L_1 \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix} i + \Phi \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{13.11}$$

where L_1 is a second order inductance. Thanks to the identity

$$\begin{pmatrix} \cos 2\theta + 1 & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta + 1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = 2 \begin{pmatrix} \cos\theta & 0 \\ \sin\theta & 0 \end{pmatrix}$$

the above expression of Ψ can be rewritten as

$$\Psi - (L_0 - L_1) i = (\Phi + 2L_1 i_d) \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \tag{13.12}$$

with the notation

$$i_{dq} \; = \; \begin{pmatrix} i_d \\ i_q \end{pmatrix} = \mathcal{R}(-\theta)\, i \; . \tag{13.13}$$

This shows that, when i_d is constant, we recover exactly the design model (13.1) provided we replace L and Φ by

$$L_s = L_0 - L_1 \quad , \quad \Phi_s = |\Phi + 2L_1 i_d| \; .$$

Hence Theorem 1 holds in the case with saliency at least when the signals obtained from the motor are such that i_d is constant. Specifically, by implementing observer (13.2) with L_s instead of L, we directly obtain :

$$\lim_{t\to\infty} |\hat{\Psi}(t) - \Psi(t)| + |\hat{\Phi}(t) - \Phi_s| = 0 \; .$$

This means that $\hat{\Psi}$ converges to Ψ and $\hat{\Phi}$ to the "equivalent flux" Φ_s. But this time, it is not sufficient to compute the argument of $\hat{\Psi} - L_s i$ to obtain an estimate of θ, since according to

(13.12), it converges either to θ or $\theta+\pi$ depending on the sign of $\Phi + 2L_1 i_d$. In fact, defining θ_0 as

$$\theta_0 = \arg(\Psi - L_s i)$$

and $i_{dq,0}$ as

$$i_{dq,0} \;=\; \begin{pmatrix} i_{d,0} \\ i_{q,0} \end{pmatrix} = \mathcal{R}(-\theta_0)\, i$$

we have :

- if $\Phi + 2L_1 i_d > 0$, then $\Phi_s = \Phi + 2L_1 i_d$, $\theta_0 = \theta$, $i_{d,0} = i_d$ and $\Phi_s - 2L_1 i_{d,0} = \Phi > 0$
- if $\Phi + 2L_1 i_d < 0$, then $\Phi_s = -\Phi - 2L_1 i_d$, $\theta_0 = \theta + \pi$, $i_{d,0} = -i_d$ and $\Phi_s - 2L_1 i_{d,0} = -\Phi < 0$.

Therefore, computing

$$\hat\theta_0 = \arg(\hat\Psi - L_s i)$$

and $\hat i_{dq,0}$ defined by

$$\hat i_{dq,0} \;=\; \begin{pmatrix} \hat i_{d,0} \\ \hat i_{q,0} \end{pmatrix} = \mathcal{R}(-\hat\theta_0)\, i \;,$$

and taking

$$\begin{array}{rcll} \hat\theta & = & \hat\theta_0 & \text{if } \hat\Phi - 2L_1 \hat i_{d,0} \geq 0 \\ \hat\theta & = & \hat\theta_0 + \pi & \text{otherwise} \;, \end{array}$$

we obtain

$$\lim_{t\to\infty} \hat\theta(t) - \theta(t) = 0 \;.$$

This convergence is a clear argument in favor of observer (13.2) with respect to observer (12.8). Indeed, the flexibility provided by the estimation of Φ enables to apply the same observer to salient motors without losing convergence of θ. The same conclusions hold for the observers presented in Section 13.2. Not to be forgotten, all this holds when i_d is constant.

13.3.3 Sensitivity to errors on R and L when (i_d, i_q, ω) is constant

In Theorem 1, we claimed convergence for observer (13.2) assuming perfect knowledge of the resistance and the inductance and the absence of saliency. Then, in the latter subsection, we extended this result to salient models as long as the current in the dq frame i_d is constant. Given the fact that the non salient models can easily be obtained from the salient ones by taking $L_1 = 0$, we keep here the more general model with saliency made of (12.1) and (13.11).

In this section, we study the possible consequences of using in the observers approximations $\hat R$ and $\hat L$ of R and L_s. For this we restrict our attention to the case where $\mathcal{R}(-\theta)\, i = i_{dq}$ and ω are constant. This configuration is often considered in practice, since it corresponds to a constant rotation speed with a constant load torque. In this case, the model has an asymptotic behavior given by

$$u = \mathcal{R}(\theta) u_{dq} \;, \quad i = \mathcal{R}(\theta) i_{dq} \;, \quad \Psi = \mathcal{R}(\theta)\Psi_{dq}$$

where u_{dq}, i_{dq} and Ψ_{dq} are constants satisfying

$$\omega J \Psi_{dq} \;=\; u_{dq} - R i_{dq} \quad , \qquad \Psi_{dq} - L_s i_{dq} \;=\; \begin{pmatrix} \Phi_s \\ 0 \end{pmatrix}$$

where

$$J \;=\; \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \;.$$

Let Ψ_{eq} be defined as

$$\Psi_{eq} = \frac{1}{\omega} J^{-1} \mathcal{R}(\theta)(u_{dq} - \hat R i_{dq})$$

It satisfies

a) $$\dot{\Psi}_{eq} = u - \hat{R}i$$

i-e the same dynamics as Ψ but with $\hat{R}$ instead of R.

b) $$\Psi_{eq} - \hat{L}i = \mathcal{R}(\theta) \underbrace{\left(\frac{1}{\omega} J^{-1} \left(u_{dq} - \hat{R} i_{dq}\right) - \hat{L} i_{dq}\right)}_{\text{constant}} . \tag{13.14}$$

Thus, with Φ_{eq} the constant real number defined as

$$\begin{aligned} \Phi_{eq} &= \left| \tfrac{1}{\omega} J^{-1} \left(u_{dq} - \hat{R} i_{dq}\right) - \hat{L} i_{dq} \right| \\ &= \left| \begin{pmatrix} \Phi_s \\ 0 \end{pmatrix} + \left([R - \hat{R}] \tfrac{J^{-1}}{\omega} + L_s - \hat{L} \right) i_{dq} \right| \end{aligned}$$

we have

$$|\Psi_{eq} - \hat{L}i|^2 - \Phi_{eq}^2 = 0 .$$

It follows that (Ψ_{eq}, Φ_{eq}) is solution of the model (13.1) if we replace (R, L) by $(\hat{R}, \hat{L})$. So, according to Theorem 13.1.1, the observer (13.2), implemented with $\hat{R}$ and $\hat{L}$, gives

$$\lim_{t \to \infty} |\hat{\Psi}(t) - \Psi_{eq}(t)| + |\hat{\Phi}(t) - \Phi_{eq}| = 0 .$$

Hence $\hat{\Phi}$ converges to $\left| \begin{pmatrix} \Phi_s \\ 0 \end{pmatrix} + \left([R - \hat{R}] \tfrac{J^{-1}}{\omega} + L_s - \hat{L} \right) i_{dq} \right|$. And with $\hat{\theta}$ computed as the argument of $\hat{\Psi} - \hat{L}i$, we have asymptotically

$$\begin{aligned} \left| \hat{\Psi} - \hat{L}i \right| \begin{pmatrix} \cos(\hat{\theta} - \theta) \\ \sin(\hat{\theta} - \theta) \end{pmatrix} &= \mathcal{R}(-\theta) \left(\Psi_{eq} - \hat{L}i\right) \\ &= \frac{1}{\omega} J^{-1} \left(u_{dq} - \hat{R} i_{dq}\right) - \hat{L} i_{dq} \\ &= \begin{pmatrix} \Phi_s \\ 0 \end{pmatrix} + \left([R - \hat{R}] \frac{J^{-1}}{\omega} + L_s - \hat{L} \right) i_{dq} , \end{aligned} \tag{13.15}$$

where we have used (13.14). In other words the error $\hat{\theta} - \theta$ converges to the argument of $\begin{pmatrix} \Phi_s \\ 0 \end{pmatrix} + \left([R - \hat{R}] \tfrac{J^{-1}}{\omega} + L_s - \hat{L} \right) i_{dq}$. Up to the first order, this is exactly the same result as the one obtained in [Hen14] for the Luenberger observer presented in [HMP12]. Of course we recover the fact that, without any errors on R and L, the asymptotic value of $\hat{\Phi}$ is Φ_s and $\hat{\theta}$ converges to θ.

We illustrate formula (13.15) via simulations with ideal data obtained for $L = 0.65$ mH, $R = 0.167\,\Omega$, $\Phi = 7.3$ mWb, $i_d = -3.46$ A, $i_q = 6$ A, for two different regimes. The results are given in Table 13.1 for observers (13.2) and (13.8). Both observers were implemented with an Euler scheme with $dt = 1.2\,10^{-4}$ s and give similar results. The reader may check that the absolute error on θ and the relative error on Φ correspond exactly to the expected theoretical errors.

13.4 Tests with real data

To illustrate the results above about the sensitivity with respect to the parameters, to saliency, but also to noise, we apply in open-loop (and offline) the observers (13.2) and (13.8) to real data obtained from two PMSM used in test beds at IFPEN : Motor 1 and Motor 2. The available data are the measurements of voltages u_m and currents i_m in the $\alpha\beta$ fixed frame, the measurement of the rotor position θ_m, the physical parameters given in Table 13.2.

		$R+1\%R$		$L+1\%L$	
ω	Obs	$\tilde{\theta}$ (rad)	$\tilde{\Phi}/\Phi$	$\tilde{\theta}$ (rad)	$\tilde{\Phi}/\Phi$
500 rpm	(13.2)	0.015	2.6 %	$5.4\,10^{-3}$	0.3 %
	(13.8)	0.015	2.6 %	$5.2\,10^{-3}$	0.3 %
2000 rpm	(13.2)	$3.8\,10^{-3}$	0.7 %	$5.4\,10^{-3}$	0.3 %
	(13.8)	$3.3\,10^{-3}$	0.6 %	$4.9\,10^{-3}$	0.3 %

Table 13.1: Sensitivity of observers (13.2) and (13.8) with respect to R and L at two different electrical rotation speeds with the notation $\tilde{\theta} = |\hat{\theta} - \theta|$ and $\tilde{\Phi}/\Phi = \frac{|\hat{\Phi}-\Phi|}{\Phi}$.

Parameter	Motor 1	Motor 2
Regime	variable : Figure 13.1	constant : 2000 rpm
L_d	0.72 mH	0.142 mH
L_q	0.78 mH	0.62 mH
Φ	8.94 mWb	18.5 mWb
R	0.151 Ω	0.023 Ω
Pairs of poles (n_p)	10	2

Table 13.2: Parameters for Motor 1 and 2.

The norms of u_m and i_m for each motor are given in Figure 13.2. Note that unlike Motor 2, Motor 1 is submitted consecutively to four regimes : around 150 rpm, 450 rpm, 1000 rpm and finally 1500 rpm (see Figure 13.1).

The motors differ in terms of saliency. According to [BC98], L_0 and L_1 in (13.11) are given by

$$L_0 = \frac{L_d + L_q}{2} \quad , \quad L_1 = \frac{L_d - L_q}{2} \ .$$

and therefore

$$L_s = L_0 - L_1 = L_q \ .$$

We conclude that saliency is weak for Motor 1 ($\frac{L_1}{L_0} \approx 4\%$), but dominant for Motor 2 ($\frac{L_1}{L_0} \approx 80\%$).

We have implemented the observers using the measured values u_m and i_m as u and i, and an explicit Euler scheme with the sample time ($dt_1 = 10^{-4}$ s, $dt_2 = 2\,10^{-5}$ s). We chose the parameters of the observers to ensure the responses have all approximately the same time constant ($\gamma_{(13.2)} = 20000$, $\gamma_{(13.8)} = 50000$, $\lambda = 50$) and so that convergence is obtained in less than two rotations of the motor. The results are presented in Figures 13.3-13.4. The performances are globally better for Motor 1 than Motor 2, but it is mainly due to the fact that the data were noisier for the latter.

For θ (Figure 13.3), both observers provide similar results, with a final oscillatory error of amplitude smaller than 0.05 rad for Motor 1 (0.09 rad for the last regime) and 0.12 rad for Motor 2. But (the mean value of) the estimation $\hat{\theta}$ does not converge to the measurement θ_m. There

are static errors. They are likely due, in part at least, to an offset in the sensor for θ_m. But there is more since, according to Figure 13.3(a), these biases depend on the regime. One explanation comes from (13.15) where the regime ω appears explicitly. Another possible explanation has been proposed and studied in [Hen14]. It is the effects of the dynamics of the sensors providing the measurements u_m and i_m. When they are modelled simply by

$$\dot{i}_m \;=\; -\tau_i(i_m - i) \quad , \qquad \dot{u}_m \;=\; -\tau_u(u_m - u)$$

the phase shift of these first order systems (depending on the regime) is directly translated in a static error on $\hat{\theta}$ and consequently on $\hat{\Phi}$. We refer the reader to [Hen14] for more details.

Concerning Φ (Figures 13.4), although both observers provide again the same mean for the final errors, the transient of observer (13.8) seems to be more oscillatory. This difference could be explained by the fact that $\hat{\Phi}$ is directly estimated by observer (13.2) while it is reconstructed from the norm of $\hat{\Psi} - L_q i$ for observer (13.8). Here again (the mean value of) $\hat{\Phi}$ does not tend to Φ. Let us concentrate on the data from Motor 2 and from the first regime of Motor 1, where the norm of the current is constant. Assuming that the offset $\hat{\theta} - \theta_m$ mentioned above is only due to the position of the sensor and therefore that $\hat{\theta}$ is actually the correct rotor position, we compute i_d as the first component of $\mathcal{R}(-\hat{\theta})i$ and find

$$\begin{array}{ll} \text{Motor 1:} & i_{d,1} = -4.2\,\text{A} \\ \text{Motor 2:} & i_{d,2} = -201\,\text{A}\ . \end{array}$$

If the values of R, L_d, L_q and Φ in Table 13.2 are correct, we can expect $\hat{\Phi}$ to tend to $\Phi_s = |\Phi + 2L_1 i_d|$, i-e

$$\begin{array}{ll} \text{Motor 1:} & \Phi_{s,1} = 9.2\,\text{mWb} \\ \text{Motor 2:} & \Phi_{s,2} = 115\,\text{mWb}\ . \end{array}$$

This is verified for both motors on Figures 13.4(a) (first regime) and 13.4(b). We could conclude that the values of R and L used in the observers are correct. Unfortunately we cannot go further in the analysis since, for the other regimes in Figure 13.4(a), the steady state is not reached.

13.5 Conclusion

We have introduced a new rotor position observer for sensorless permanent magnet synchronous motors (PMSM). It is designed from a non salient model and uses measurements of voltages and current, and estimations of resistance and inductance. But it does not require the knowledge of the magnet flux. We have claimed its convergence in an ideal context and for a rotating motor.

We have compared it with the equivalent observers proposed in [HMP12, Hen14] and [BPO$^+$15b]. The main difference is that this new observer is less demanding in terms of computations. On the other hand it gives qualitatively the same kind of performance, in terms of speed of convergence, sensitivity to errors in the resistance or the inductance and also in presence of saliency.

At least three important issues remain to be addressed:

a) Sensitivity to measurement noise or more interestingly the definition of a tuning policy in presence of such disturbances. This kind of study has been made in [Hen14] for the Luenberger observer proposed in [HMP12] . The same kind of tools should be useful in our context.

b) Use of the observer in closed loop. Tests via simulations or test beds for the observers in [HMP12] and [BPO$^+$15b] are reported in those papers. But as far as we know no theoretical results are yet available.

c) Extension to non salient models. We are unaware of any observer for this case.

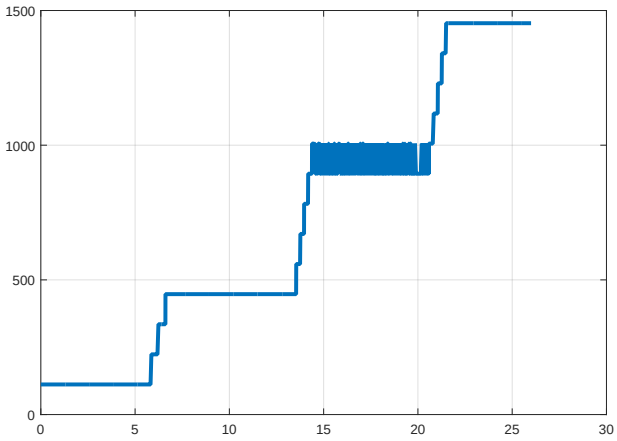
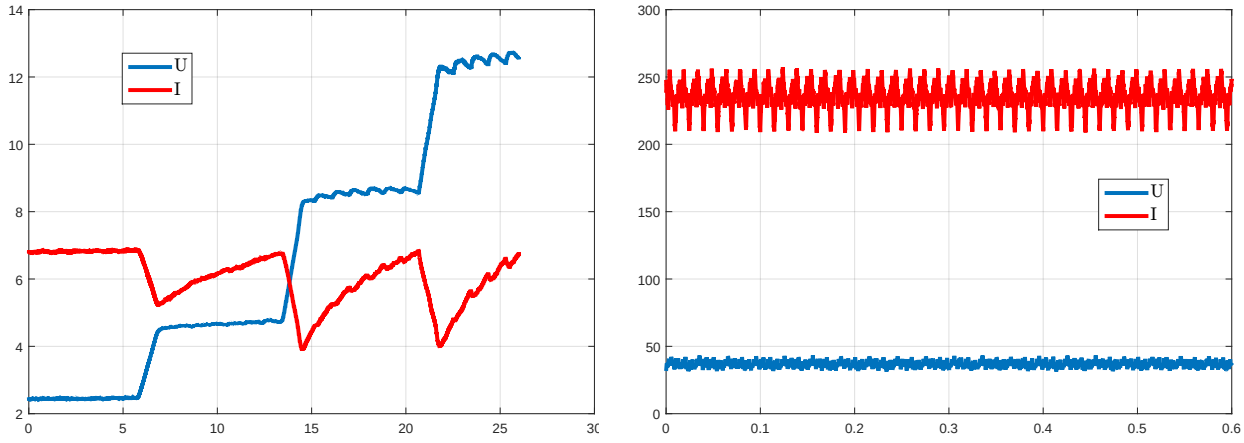


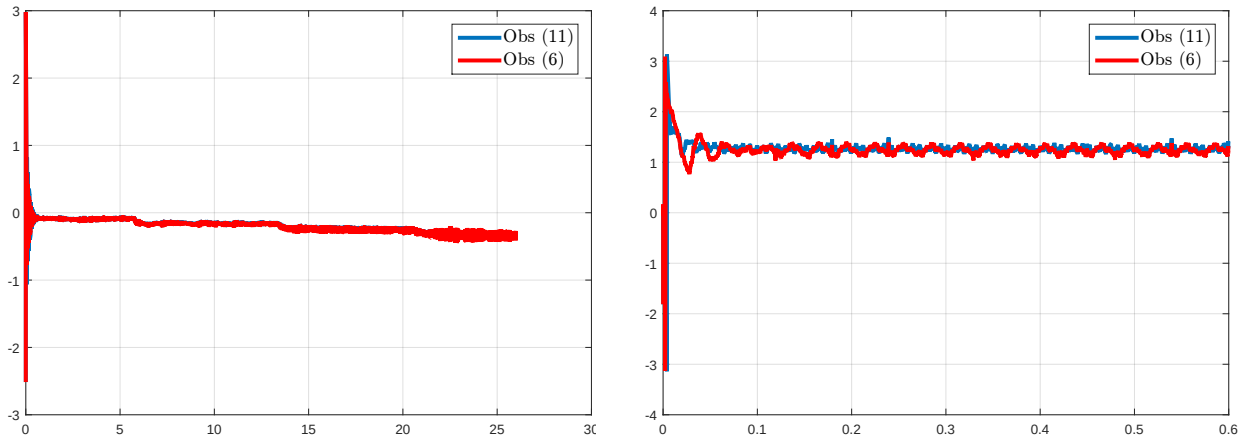
Figure 13.1: Regime of Motor 1 (rpm).



(a) Motor 1

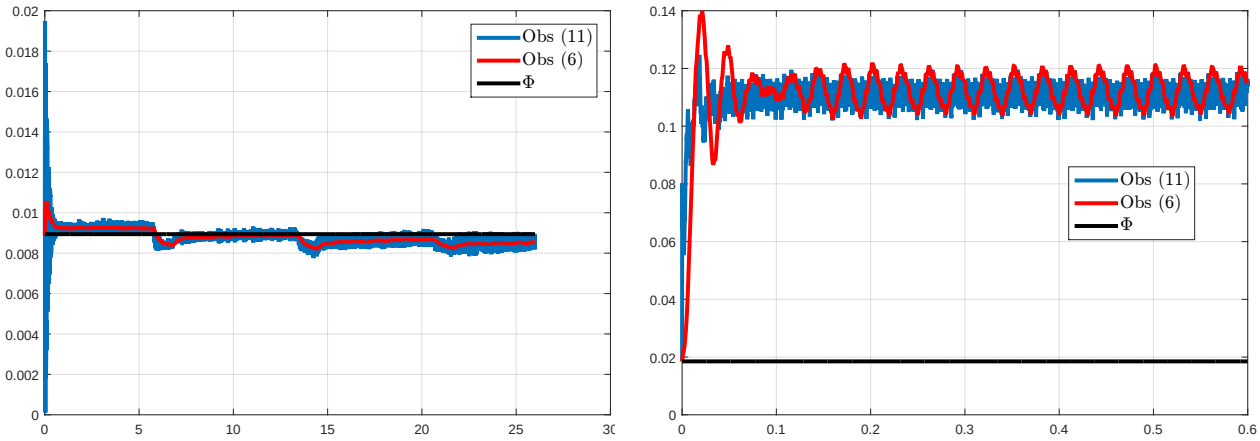
(b) Motor 2

Figure 13.2: Norm of the voltage u_m (V) and current i_m (A).



(a) Motor 1 (b) Motor 2

Figure 13.3: Error $\hat{\theta} - \theta_m$ (rad) given by observers (13.2) and (13.8) , where θ_m is a measurement of θ.



(a) Motor 1 (b) Motor 2

Figure 13.4: $\hat{\Phi}$ given by observers (13.2) and (13.8) compared to Φ.

Chapter 14

Rotor position estimation with unknown resistance

Chapitre 14 – Estimation de la position du rotor lorsque la résistance est inconnue. *Nous montrons dans ce chapitre que contrairement à (Ψ, Φ), le couple (Ψ, R) n'est pas observable avec la seule information que $y(t) = 0$ pour tout t. Cependant, lorsque ω et i_d sont non nuls, il ne peut exister que six solutions indistingables maximum, la resistance étant l'une des racines d'un polynôme de degré 6. De plus, dans le cas particulier où ω, i_d et i_q sont constants, nous prouvons que le nombre de solutions possibles est réduit à deux, avec deux valeurs bien identifiées pour la résistance, qui sont distinctes sauf si i_q est nul. Il apparaît alors que ces deux solutions peuvent en fait être dissociées si le signe de i_q (c'est-à-dire le mode d'utilisation du moteur) est connu. Cette propriété nous permet de proposer une stratégie d'observation, basée sur une synthèse de Luenberger. Ses performances sont testées et illustrées en simulation.*

Contents

We have seen in the previous chapter that it is possible to estimate both Ψ and Φ at the same time. In this chapter, we suppose the magnet flux Φ known, but the resistance R unknown and we wonder if it is possible to estimate both Ψ and R. So we consider the system

$$\left\{ \begin{array}{rcl} \dot{\Psi} & = & u - R\, i \\ \dot{R} & = & 0 \\ y & = & |\Psi - L\, i|^2 - \Phi^2 \end{array} \right. \tag{14.1}$$

with inputs (u, i), known parameters (Φ, L), state (Ψ, R) and the knowledge that y is constantly zero.

To ease the reading of this chapter, some of the proofs are summarized with only their most important steps, or even omitted when they are of no particular interest. Their detailed version is available in Appendix D.

14.1 Observability

Before looking for an observer, we need to check the observability of the system. To do that, we consider the time-varying system

$$\left\{ \begin{array}{rcl} \dot{x} & = & u - x_3\, i \\ \dot{x}_3 & = & 0 \\ y & = & |x - L\, i|^2 - \Phi^2 \end{array} \right. \tag{14.2}$$

with L and Φ given, and where u and i are time signals such that there exists a particular solution $(x = \Psi, x_3 = R)$ of (14.2) verifying

$$y(t) = 0 \quad \forall t \ .$$

This means that there exists a (unique) time-signal θ such that for all t,

$$\Psi(t) = L\, i(t) + \Phi \left(\begin{array}{c} \cos(\theta(t)) \\ \sin(\theta(t)) \end{array} \right) \ . \tag{14.3}$$

In the following, we denote

$$z = \left(\begin{array}{c} \cos\theta \\ \sin\theta \end{array} \right) \quad , \quad i_{dq} = \left(\begin{array}{c} i_d \\ i_q \end{array} \right) = \mathcal{R}(-\theta)\, i \quad , \quad \omega = \dot{\theta} \ .$$

We want to know whether, given the time signals (u, i) and the parameters (L, Φ), the particular solution (Ψ, R) is the unique solution to System (14.2) verifying $y(t) = 0$ for all t. Note that this is somehow a weak notion of observability since it is for a particular trajectory of y.

14.1.1 A first observability result

We start from the following result :

Theorem 14.1.1.

If

a) for all t, $\omega(t) = 0$

or

b) for all t such that $\omega \neq 0$, $i_d(t) = 0$, $i_q(t) \neq 0$ and $\frac{\omega}{i_q}$ is constant

there exists an infinite number of solutions to System (14.2) verifying $y(t) = 0$ for all t. Otherwise, if besides $|i(t)| \neq 0$ for all t, there exist at most 6 solutions.

Proof : Consider a solution (x, x_3) to System (14.2) verifying for all t

$$0 = y(t) = |x(t) - Li(t)|^2 - \Phi^2 \ .$$

x is necessarily of the form

$$x(t) = x_0 + \int_0^t u(\tau)d\tau - x_3 \int_0^t i(\tau)d\tau$$

with

$$\dot{x}_0 = 0 \quad , \quad \dot{x}_3 = 0 \ ,$$

and finding (x, x_3) is equivalent to finding (x_0, x_3). It follows that for all t

$$\begin{aligned} 0 &= |x(t) - Li(t)|^2 - |x_0 - Li(0)|^2 \\ &= [x(t) - x_0 - L(i(t) - i(0))]^\top [x(t) + x_0 - L(i(t) + i(0))] \\ &= \tilde{\eta}(x_3, t)^\top [2(x_0 - Li(0)) + \tilde{\eta}(x_3, t)] \end{aligned}$$

where we have defined

$$\tilde{\eta}(x_3, t) = \int_0^t u(\tau)d\tau - x_3 \int_0^t i(\tau)d\tau - L(i(t) - i(0)) \ . \tag{14.4}$$

We deduce that for any time t,

$$2\tilde{\eta}(x_3, t)^\top (x_0 - Li(0)) = -\tilde{\eta}(x_3, t)^\top \tilde{\eta}(x_3, t) = -|\tilde{\eta}(x_3, t)|^2 \ .$$

Therefore, unless x_3 makes $\tilde{\eta}(x_3, t_1)$ and $\tilde{\eta}(x_3, t_2)$ colinear for any (t_1, t_2), there exits at most one possible value of x_0 for each x_3.

The rest of the proof then consists in showing that[1] :

1. for x_3 such that $\tilde{\eta}(x_3, .)$ is not constant, there exist couples (t_1, t_2) such that $\tilde{\eta}(x_3, t_1)$ and $\tilde{\eta}(x_3, t_2)$ are not colinear. x_0 is then uniquely determined by the value of x_3, which must be the root of a polynomial of degree 6. Therefore, there are at most 6 solutions (x, x_3) such that $\tilde{\eta}(x_3, .)$ is not constant.
2. to the values of x_3 such that $\tilde{\eta}(x_3, .)$ is constant, is associated an infinite number of solutions (x, x_3).
3. if x_3 makes $\tilde{\eta}(x_3, .)$ constant, it must satisfy for all t

$$\begin{aligned} (R - x_3) i_d(t) &= 0 \\ (R - x_3) i_q(t) &= -\omega(t)\Phi \ . \end{aligned} \tag{14.5}$$

We can thus distinguish the following cases :

- if $\omega(t) = 0$ for all t, there exists at least one constant value of x_3 solution to System (14.5) for all t. Thus, $\tilde{\eta}(x_3, \cdot)$ is constant and there is an infinite number of solutions (x, x_3).
- if for all t such that $\omega(t) \neq 0$, $i_d(t) = 0$, $i_q(t) \neq 0$, and $\frac{\omega}{i_q}$ is constant, there exists a constant value of x_3 solution to System (14.5) for all t and thus an infinity of solutions (x, x_3).
- otherwise, there exist no solutions to System (14.5). Therefore, $\tilde{\eta}(x_3, \cdot)$ cannot be constant and there are at most 6 solutions (x, x_3) to our observability problem.

■

We recover the fact that the system is not observable when the rotating speed is zero (this is the case even when R is known). In the usual case where there exists at least a time t for which $\omega(t) i_d(t) \neq 0$, this result says that there exist maximum 6 possible solutions (x, x_3), with x_3 given by the roots of a polynomial of order 6. In order to conclude that the system is not observable, we need to know more about those roots. In particular, the polynomial may not have 6 distinct real roots and even if it does, they may not be constant with time.

To get more information, one could study in detail this polynomial of order 6 obtained in the proof. But its expression is too complex and the next section shows how a stronger notion of differential observability enables to get a more precise idea of this polynomial.

14.1.2 Differential observability of order 3

Let us consider the stronger observability question : is $(\Psi(t), x_3)$ the only solution at time t of $y(t) = \dot{y}(t) = \ddot{y}(t) = 0$? Of course, in the cases of non observability identified in Theorem 14.1.1, the answer is no. But we want to study in more details what happens in the other cases, in particular when there exists a time t such that

$$|i(t)| \neq 0 \quad \text{and} \quad \omega(t) \neq 0 \quad \text{and} \quad i_d(t) \neq 0 \text{ or } i_q(t) = 0 \ ,$$

[1] See Appendix D.1.1.

which is equivalent to

$$\omega(t) \neq 0 \quad \text{and} \quad i_d(t) \neq 0 \ .$$

Consider the function $\overline{\mathbf{H}}_3$ made of $\Big(y(t), \dot{y}(t), \ddot{y}(t)\Big)$:

$$\overline{\mathbf{H}}_3(x, x_3, t) \ = \ \begin{pmatrix} |x - L\, i(t)|^2 - \Phi^2 \\ 2(x - L\, i(t))^\top (u(t) - x_3\, i - L\dot{\widehat{i}}(t)) \\ 2(x - L\, i(t))^\top (\dot{u}(t) - x_3 \dot{\widehat{i}}(t) - L\ddot{\widehat{i}}(t) + 2|u(t) - x_3\, i(t) - L\dot{\widehat{i}}(t)|^2 \end{pmatrix} \ .$$

Our problem consists in looking for the solutions in (x, x_3) of

$$\overline{\mathbf{H}}_3(x, x_3, t) = 0 \ .$$

We have the following result :

Theorem 14.1.2.

Consider a time t such that $\omega(t) \neq 0$ and $i_d(t) \neq 0$. There are as many solutions (x, x_3) to the equation

$$\overline{\mathbf{H}}_3(x, x_3, t) = 0 \ ,$$

as the number of real roots of the following polynomial of order 6 :

$$P(x_3, t) = \omega(t)^6 \Phi^6 \left[\left(1 - \frac{(R - x_3)}{\omega(t)\Phi} \left(\dot{\widehat{\left(\frac{i_d}{\omega} \right)}}(t) - 2 i_q(t) \right) + \frac{(R - x_3)^2}{\omega(t)^2 \Phi^2} \mu(t) |i(t)|^2 \right)^2 \right.$$
$$\left. - \left(1 + \frac{(R - x_3)}{\omega(t)\Phi} 2 i_q + \frac{(R - x_3)^2}{\omega(t)^2 \Phi^2} |i(t)|^2 \right)^3 \right] \quad (14.6)$$

where[2]

$$\mu(t) \ = \ \frac{1}{\omega(t)} \frac{\left[i(t)^\top J \dot{\widehat{i(t)}} \right]}{|i(t)|^2} \quad , \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \ .$$

Proof : It is appropriate to introduce the following degree one polynomial of x_3

$$\eta(x_3, t) \ = \ u(t) - x_3 i(t) - L\dot{\widehat{i}}(t) \ , \quad (14.7)$$

so that $\overline{\mathbf{H}}_3$ actually reads

$$\overline{\mathbf{H}}_3(x, x_3, t) \ = \ \begin{pmatrix} |x - L\, i(t)|^2 - \Phi^2 \\ 2\eta(x_3, t)^\top (x - L\, i(t)) \\ 2\dot{\eta}(x_3, t)^\top (x - L\, i(t)) + 2|\eta(x_3, t)|^2 \end{pmatrix} \ .$$

It is interesting to note that $\eta(x_3, t) = \dot{\tilde{\eta}}(x_3, t)$, where $\tilde{\eta}$ is defined in (14.4), and what is done in this proof is somehow the differential version of the proof of Theorem 14.1.1 To study how many solutions in (x, x_3) the equation $\mathbf{H}_3(x, x_3, t) = 0$ has, we note that the second and third component is a linear system in $x - L\, i$. So our approach is to solve this system and replace in the first component. This gives a function of x_3 only. Hence the first question is invertibility of the linear system, i.e. colinearity of $\eta(x_3, t)$ and $\dot{\eta}(x_3, t)$.

Assume that $\eta(x_3, t)$ is non zero and is colinear with $\dot{\eta}(x_3, t)$, namely $\dot{\eta}(x_3, t) = \lambda \eta(x_3, t)$. Then, $\overline{\mathbf{H}}_3(x, x_3, t) = 0$ gives

$$\eta(x_3, t)^\top (x - Li(t)) = 0 \quad , \quad \lambda \eta(x_3, t)^\top (x - Li(t)) = -|\eta(x_3, t)|^2$$

[2] μ is the ratio between ω and the rotation speed of i.

and necessarily $\eta(x_3,t)=0$ which is impossible. Therefore, colinearity can only happen if $\eta(x_3,t)=0$. But differentiating (14.3) with respect to time and combining this expression with (14.7), we get

$$\eta(x_3,t) \;=\; -\omega(t)\Phi Jz(t) + (R-x_3)i(t)\;. \tag{14.8}$$

By imposing $\eta(x_3,t)$ to be zero and multiplying by $\mathcal{R}(-\theta)$, we recover System (14.5) which does not admit any solution if $\omega(t)$ and $i_d(t)$ are nonzero. We conclude that for all x_3, $\eta(x_3,t)$ and $\dot\eta(x_3,t)$ are not colinear.

It follows that we can get x from the second and third components of $\overline{\mathbf{H}}_3$, namely

$$\begin{array}{rcl}(x-L\,i(t))^\top \eta(x_3,t) &=& 0\\ (x-L\,i(t))^\top \dot\eta(x_3,t) &=& -|\eta(x_3,t)|^2\end{array} \tag{14.9}$$

i-e

$$x - L\,i(t) \;=\; \frac{|\eta(x_3,t)|^2}{\eta(x_3,t)^\top J\dot\eta(x_3,t)} J\eta(x_3,t)\;.$$

Inserting this expression in the first component of $\overline{\mathbf{H}}_3$ gives

$$\Phi^2 \;=\; \left|\frac{|\eta(x_3,t)|^2}{\eta(x_3,t)^\top J\dot\eta(x_3,t)} J\eta(x_3,t)\right|^2 \;=\; \frac{|\eta(x_3,t)|^6}{[\eta(x_3,t)^\top J\dot\eta(x_3,t)]^2}$$

and x_3 is a root of the following polynomial

$$P(x_3,t) = \Phi^2[\eta(x_3,t)^\top J\dot\eta(x_3,t)]^2 \;-\; |\eta(x_3,t)|^6\;.$$

Differentiating (14.8), we get

$$\dot\eta(x_3,t) = -\dot\omega(t)\Phi Jz(t) - \omega(t)^2\Phi z(t) + (R-x_3)\dot{\widehat{i}}(t)\;, \tag{14.10}$$

which yields

$$\begin{array}{rcl}\det\Big(\eta(x_3,t)\,,\,\dot\eta(x_3,t)\Big) &=& \eta(x_3,t)^\top J\dot\eta(x_3,t)\\ &=& \omega^3\Phi^2 - (R-x_3)\Phi\left[\omega^2 i^\top Jz - \dot\omega\, i^\top z + \omega\dot{\widehat{i}}^{\,\top} z\right] + (R-x_3)^2\, i^\top J\dot{\widehat{i}}\\ &=& \omega^3\Phi^2 - (R-x_3)\Phi\omega^2\left[-2i_q + \frac{\dot{\widehat{i_d}}}{\omega}\right] + (R-x_3)^2\, i^\top J\dot{\widehat{i}}\end{array} \tag{14.11}$$

where we have used the fact that $i^\top z = i_d$ and $i^\top Jz = -i_q$. Inserting those expressions in the expression of P, we get the polynomial (14.6). The coefficient of degree 6 is $|i|^6$ which is non zero by assumption. We conclude that there are at most 6 possible values for x_3, and since the value of x is imposed by that of x_3, we get the result. ■

With this result, we are not much further advanced than with Theorem 14.1.1, but at least we have a more precise expression of the polynomial. The reader may check in particular that $x_3 = R$ is a possible root. But since the degree is even, there is at least another real root (which can be equal to R too). In order to have a better idea of those roots, we study the usual case where ω, i_d and i_q are constant.

14.1.3 Particular case where ω, i_d and i_q are constant

We have the following corollary :

Corollary 14.1.1.

Assume ω, i_d and i_q are constants such that $\omega \neq 0$ and $i_d \neq 0$. P has only two roots given by

$$x_3 = R \quad,\quad x_3 = R + \frac{2\Phi\omega\, i_q}{|i|^2}\;.$$

Therefore, the equation $\overline{\mathbf{H}}_3(x,x_3,t) = 0$ admits one solution if $i_q = 0$ and two distinct

| solutions if $i_q \neq 0$.

Proof : In this particular case, $i^\top J \dot{\widehat{i}} = \omega |i|^2$ so that $\mu(t) = 1$ and

$$P(x_3) = -\omega^6 \Phi^6 \left(1 + \frac{(R-x_3)}{\omega\Phi} 2i_q + \frac{(R-x_3)^2}{\omega^2\Phi^2}|i|^2\right)^2 \frac{(R-x_3)}{\omega\Phi}\left(2i_q + \frac{(R-x_3)}{\omega\Phi}|i|^2\right) .$$

The polynomial $|i|^2 X^2 + 2i_q X + 1$ has a discriminant equal to $-4i_d^2 < 0$ and does not admit any real root. The conclusion follows. Note that in this case, according to (14.11), P also writes

$$P(x_3) = -\Phi^2 \det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)^2 \frac{(R-x_3)}{\omega\Phi}\left(2i_q + \frac{(R-x_3)}{\omega\Phi}|i|^2\right) . \tag{14.12}$$

■

The conclusion from this theorem is that the system is not differentially observable of order 3 unless $i_q = 0$. This does not mean that the system is not observable because the solution corresponding to $x_3 = R + \frac{2\Phi\omega\, i_q}{|i|^2}$ may not be admissible for System (14.2). Actually, it turns out that both solutions are truly indistinguishable :

Theorem 14.1.3.

Assume ω, i_d and i_q are constants such that $\omega \neq 0$ and $i_d \neq 0$. There exist exactly two indistinguishable solutions (x, x_3) to System (14.2) verifying $y(t) = 0$ for all t. They are of the form (Ψ, R) and (Ψ_δ, R_δ) with

$$R_\delta = R + \frac{2\Phi\omega\, i_q}{|i|^2} .$$

Proof : See Appendix D.1.2. ■

We conclude that the system is not observable if $i_q \neq 0$. However, the problem is well-identified with only two possible solutions and the following result shows how they can be dissociated by adding an extra information, namely the sign of i_q.

Theorem 14.1.4.

Assume ω, i_d and i_q are constants such that $\omega \neq 0$ and $i_d \neq 0$. Consider both solutions (Ψ, R) and (Ψ_δ, R_δ) given by Theorem 14.1.3, and their associated[3] (θ, i_{dq}), $(\theta_\delta, i_{dq,\delta})$. We have

$$\begin{aligned} i_{d,\delta} &= i_d \\ i_{q,\delta} &= -i_q \end{aligned}$$

so that both solutions can be distinguished by the sign of their corresponding i_q.

Besides, if $(\hat R, \hat\theta)$ is is one of the solutions $\Big\{(R,\theta),(R_\delta,\theta_\delta)\Big\}$, then the other solution is[4]

$$\left(\hat R + \frac{2\Phi\hat\omega\,\hat{\widehat{i_q}}}{|i|^2}\,,\ \hat\theta + \arctan_2\left(2\hat{\widehat{i_q}}\hat{\widehat{i_d}}\,,\, 1 - 2\hat{\widehat{i_q}}^2\right)\right) .$$

Proof : See Appendix D.1.3. ■

[3] $i_{dq,\delta} = \begin{pmatrix} i_{d,\delta} \\ i_{q,\delta} \end{pmatrix} = \mathcal{R}(-\theta_\delta)\, i$.

[4] $\hat{\widehat{i_{dq}}} = \begin{pmatrix} \hat{\widehat{i_d}} \\ \hat{\widehat{i_q}} \end{pmatrix} = \mathcal{R}(-\hat\theta) i$

We conclude that the additional information of the sign of i_q makes the system observable ! If fact, the sign of i_q determines the mode of use of the machine : if $i_q > 0$, the torque is positive and the machine acts as a motor, whereas if $i_q < 0$, the torque is negative and the machine acts as a generator. In other words, both solutions can be distinguished if we know the mode of use of the motor.

This result also says that if an estimation $\hat{R}$ among $\{R, R_\delta\}$ is available (for instance thanks to an observer), it is possible to find the other candidate, at least when the rotation speed ω is known or estimated. Therefore, if the sign of i_q is known or if an imprecise sensor gives an idea of θ, the right solution can be picked online. Of course, the smaller i_q the more difficult to know its sign or to choose between θ and θ_δ, but also the smaller the error if we choose the wrong one...

Remark 18 In fact, from a physical point of view, those two values of R correspond to two systems with same total energy but with different energy repartition. Indeed, the dynamics of a PMSM in the dq-coordinates can be modeled by

$$\left\{\begin{array}{rcl} L\dot{\widehat{i_d}} & = & -Ri_d + \omega L i_q + u_d \\ L\dot{\widehat{i_q}} & = & -Ri_q - \omega L i_d - \omega\Phi + u_q \\ \dot{\omega} & = & \Phi i_q - \tau \end{array}\right.$$

where τ is the external torque. The total energy of the system varies along

$$\frac{L}{2}\dot{\widehat{i_d^2 + i_q^2 + \omega^2}} = -R(i_d^2 + i_q^2) + i^\top u - \tau\omega = -R|i|^2 + i^\top u - \tau\omega \ .$$

Thus, an equilibrium with i_d, i_q and ω constant is such that

$$-R|i|^2 + i^\top u - \tau\omega = 0 \quad , \quad \Phi i_q = \tau \ .$$

Now, either $\tau = \tau_0 > 0$, in which case $R = \frac{u^\top i - \tau_0\omega}{|i|^2}$, either $\tau = -\tau_0$, and $R = \frac{u^\top i + \tau_0\omega}{|i|^2}$. The two values of R differ by $\frac{2\omega\tau_0}{|i|^2}$, i-e $\frac{2\omega\Phi|i_q|}{|i|^2}$, which is exactly what we found in our observability analysis. We conclude that both solutions have the same total energy, but in the first one energy is produced by the motor and lost in friction, and in the second one external energy is given to the motor and is dissipated in the motor by a larger resistance.

We conclude from this observability analysis that System (14.2) is not observable when ω or i_d remains at 0. However, when ω and i_d are nonzero, the number of indistinguishable trajectories is reduced to maximum 6 : the possible values of R are the roots of a polynomial P of order 6 given by (14.6). Unfortunately, we have not been able to say more about those roots unless ω, i_d and i_q are constant. In that case, there are exactly two indistinguishable trajectories and they can be distinguished with additional information on the resistance or simply the sign of i_q. In the next section, we propose an algorithm to estimate those solutions based on a Luenberger observer.

14.2 Observer design

For λ in $\mathbb{R}_+^*$, we define the function

$$T_\lambda(x, x_3, t) = \lambda^2\, x^\top x + \lambda\, c_\lambda(t)^\top x + \lambda\, x_3\, b_\lambda(t)^\top x + a_\lambda(t) x_3 + d_\lambda(t) x_3^2 \tag{14.13}$$

on $\mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}$, with a_λ, b_λ, c_λ, and d_λ the outputs of the following filters :

$$\dot{a_\lambda} = \lambda(-a_\lambda + c_\lambda^\top i - b_\lambda^\top u) \tag{14.14}$$
$$\dot{b_\lambda} = \lambda(-b_\lambda + 2i) \tag{14.15}$$
$$\dot{c_\lambda} = \lambda(-c_\lambda - 2u - 2\lambda L i) \tag{14.16}$$
$$\dot{d_\lambda} = \lambda(-d_\lambda + b_\lambda^\top i) \ . \tag{14.17}$$

We have the following result :

Lemma 14.2.1.

For any λ in $\mathbb{R}_+^*$, for any initial conditions in the filters (14.14)-(14.17), any solution (Ψ, R) to System (14.1) such that $y(t) = 0$ for all t, and any solution Z_λ to the dynamics

$$\dot{z_\lambda} = \lambda(-z_\lambda + c_\lambda^\top u - \lambda^2 L^2 |i|^2 + \lambda^2 \Phi^2) \tag{14.18}$$

verify

$$\lim_{t\to\infty} Z_\lambda(t) - T_\lambda(\Psi(t), R, t) = 0 \ .$$

Proof : Straightforward computations show that $t \to T_\lambda(\Psi(t), R, t)$ follows the dynamics (14.18), hence the result. ■

This means that by implementing filters (14.14)-(14.17) and (14.18) with any initial conditions, one can obtain an estimate of $T_\lambda(\Psi(t), R, t)$. Since our goal is to estimate (Ψ, R), we are interested in the injectivity of the function T_λ. Theorem 7.1.2 tells us that by choosing a sufficiently large number m of eigenvalues λ_i, the function $T = (T_{\lambda_1}, ..., T_{\lambda_m})$ is injective if the system is backward-distinguishable. We have seen that when ω, i_d and i_q are constant, two states (Ψ, R) and (Ψ_δ, R_δ) are not distinguishable by the dynamics, and thus necessarily $T(\Psi(t), R, t) = T(\Psi_\delta(t), R_\delta, t)$ for all t. This means that it is hopeless to prove the injectivity of T, but it may still be possible to recover the (at most 6 !) possible values of (Ψ, R).

14.2.1 An algorithm for the inversion of T

Consider three strictly positive real numbers λ_1, λ_2, λ_3. We deduce from Lemma 14.2.1 that by defining the function

$$T(x, x_3, t) = \begin{pmatrix} T_{\lambda_1}(x, x_3, t) \\ T_{\lambda_2}(x, x_3, t) \\ T_{\lambda_3}(x, x_3, t) \end{pmatrix} = m_\lambda \, x^\top x + \Lambda \left(c(t) + x_3 \, b(t)\right) x + a(t) \, x_3 + d(t) \, x_3^2 \tag{14.19}$$

on $\mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}$, we have

$$\lim_{t\to\infty} Z(t) - T(\Psi(t), R, t) = 0 \ ,$$

where we have denoted

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \quad , \quad Z = \begin{pmatrix} Z_{\lambda_1} \\ Z_{\lambda_2} \\ Z_{\lambda_3} \end{pmatrix} \quad , \quad m_\lambda = \begin{pmatrix} \lambda_1^2 \\ \lambda_2^2 \\ \lambda_3^2 \end{pmatrix}$$

$$a = \begin{pmatrix} a_{\lambda_1} \\ a_{\lambda_2} \\ a_{\lambda_3} \end{pmatrix} \quad , \quad b = \begin{pmatrix} b_{\lambda_1}^\top \\ b_{\lambda_2}^\top \\ b_{\lambda_3}^\top \end{pmatrix} \quad , \quad c = \begin{pmatrix} c_{\lambda_1}^\top \\ c_{\lambda_2}^\top \\ c_{\lambda_3}^\top \end{pmatrix} \quad , \quad d = \begin{pmatrix} d_{\lambda_1} \\ d_{\lambda_2} \\ d_{\lambda_3} \end{pmatrix} \ .$$

Thus, implementing the filters (14.14)-(14.17) and (14.18) for three values of λ gives an estimate of $T(\Psi(t), R, t)$, and we would like to invert T, i-e find the possible candidates (x, x_3) for a given $T(x, x_3, t)$.

To do that, we consider the matrix

$$M_\lambda = \begin{pmatrix} \lambda_2^2 & -\lambda_1^2 & 0 \\ 0 & \lambda_3^2 & -\lambda_2^2 \end{pmatrix}$$

which is such that

$$M_\lambda m_\lambda = 0 \ . \tag{14.20}$$

We have the following result :

Theorem 14.2.1.

Consider any $(\lambda_1, \lambda_2, \lambda_3)$ in $(\mathbb{R}_+^*)^3$, any initial conditions of the filters (14.14)-(14.17) and (14.18), and define

$$\mathcal{M}(x_3, t) = M_\lambda \Lambda \Big(c(t) + x_3\, b(t) \Big) \ . \tag{14.21}$$

Assume the input (u, i) is bounded. Then, any solution (Ψ, R) to System (14.1) such that there exists $\underline{\delta}$ such that for all t,

$$y(t) = 0 \quad , \quad \Big| \det \Big(\mathcal{M}(R,t) \Big) \Big| \geq \underline{\delta} > 0 \ ,$$

verifies

$$\lim_{t \to +\infty} \Psi(t) - \chi(R,t) = 0 \quad , \quad \lim_{t \to +\infty} J(R,t) = 0$$

where

$$\chi(x_3, t) = \mathcal{M}(x_3,t)^{-1} \Big(M_\lambda\, Z(t) - M_\lambda a(t)\, x_3 - M_\lambda d(t)\, x_3^2 \Big) \tag{14.22}$$

and

$$J(x_3, t) = m_\lambda^\top \Big(Z(t) - T(\chi(x_3,t), x_3, t) \Big) \ . \tag{14.23}$$

Proof : Observe that

$$M_\lambda\, T(x, x_3, t) = M_\lambda \Lambda \Big(c(t) + x_3\, b(t) \Big)\, x + M_\lambda a(t)\, x_3 + M_\lambda d(t)\, x_3^2$$

is linear in x. This means that for any x_3 and any t such that the matrix $\mathcal{M}(x_3, t)$ is invertible, x is solution of :

$$x = \mathcal{M}(x_3,t)^{-1} \Big(M_\lambda\, T(x, x_3, t) - M_\lambda a(t)\, x_3 - M_\lambda d(t)\, x_3^2 \Big) \ .$$

Thus, $(x, x_3) = (\Psi(t), R)$ satisfies this equation for all t and we have

$$|\Psi(t) - \chi(R,t)| \leq \Big| \mathcal{M}(R,t)^{-1} \Big|\, |M_\lambda| |Z(t) - T(\Psi(t), R, t)| \, .$$

Lemma 14.2.1 gives the result if

$$\Big| \mathcal{M}(R,t)^{-1} \Big| = \frac{1}{\Big| \det \Big(\mathcal{M}(R,t) \Big) \Big|} |\mathcal{M}^*(R,t)|$$

is upper-bounded in time, where $\mathcal{M}^*(R,t)$ is the comatrix of $\mathcal{M}(R,t)$. $t \mapsto \mathcal{M}^*(R,t)$ is a continuous function of the coefficients of c and b which are filtered versions of the bounded input (u, i) and which are thus bounded. Since $\Big| \det \Big(\mathcal{M}(R,t) \Big) \Big|$ is lower-bounded away from 0, the conclusion follows. ■

This leads us to introduce the following algorithm :

Algorithm

Implement filters (14.14)-(14.17) and (14.18) for three strictly positive real numbers λ_1, λ_2, λ_3, and at any time t, find an estimate $\hat{R}$ of R with

$$\hat{R}(t) = \mathsf{Argmin}_{x_3 \in \mathbb{R}^+} |J(x_3, t)| \ , \tag{14.24}$$

and an estimate $\hat{\Psi}$ of Ψ with

$$\hat{\Psi}(t) = \chi(\hat{R}(t), t) \ .$$

In fact, χ captures the information given by T in the direction of M_λ and the remaining information along the orthogonal direction, i-e along m_λ, is used in J to determine x_3.

Theorem 14.2.1 says that $(\hat{\Psi}, \hat{R}) = (\Psi, R)$ should be (asymptotically) a possible solution of this algorithm whenever $\mathcal{M}(R,t)$ is invertible for all t. Its implementation thus raises two questions :

- Is the matrix $\mathcal{M}(R,t)$ invertible for any t, or, more precisely, is $\left|\det\Big(\mathcal{M}(R,t)\Big)\right|$ lower-bounded ?
- Is R the only solution to the minimization problem at least after a certain time ? If no, which are the other solutions ?

Note that at each time t, the determinant of $\mathcal{M}(x_3,t)$ is a polynomial of order 2 in x_3, so that $\mathcal{M}(x_3,t)$ is invertible for all x_3 except maybe for two values $\{z_1(t),z_2(t)\}$. Then, $\chi(x_3,t)$ is a two-dimensional matrix made of rational fractions in x_3 with numerator of degree 3 and denominator of degree 2, defined everywhere except at $\{z_1(t),z_2(t)\}$. We conclude that $J(x_3,t)$ is a rational fraction with numerator of degree 6 and denominator of degree 4 defined everywhere except maybe at the two roots of the determinant of $\mathcal{M}(x_3,t)$.

Remark 19 Since x_3 is one-dimensional and we often have a fairly good idea of the interval in which lies the true value R, the resolution of the minimization problem can easily be managed with a one-dimensional grid, which can either be fixed around the initial guess $\hat{R}(0)$ or placed at each iteration around the previously found value $\hat{R}(t)$. This latter option enables to follow the slow variations of R with the temperature. Also, since R is fairly constant, it may not be necessary to update $\hat{R}$ at each iteration.

Remark 20 This algorithm necessitates the implementation of 7 filters (b_λ and c_λ are of dimension 2, and a_λ, d_λ and z_λ of dimension 1) for three values of λ, namely 21 filters. An alternative solution with only 14 filters will be given in Section 14.2.3.

14.2.2 Link with observability

The following technical lemma shows that there is a tight link between the quantities of interest for the observer, and those encountered during the observability study above.

Lemma 14.2.2.

We have the following relations :

$$\det\Big(\mathcal{M}(x_3,t)\Big) = 4\underbrace{\lambda_2^2(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)(\lambda_3-\lambda_1)}_{O(\lambda^5)}\det\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)+O(\lambda^4)\ , \quad (14.25)$$

and if (x_3,t) is such that $\mathcal{M}(x_3,t)$ and $(\eta(x_3,t),\dot{\eta}(x_3,t))$ are invertible

$$\begin{pmatrix}\eta(x_3,t)^\top\\ \dot{\eta}(x_3,t)^\top\end{pmatrix}(\chi(x_3,t)-Li) = \begin{pmatrix}0\\ -|\eta(x_3,t|^2\end{pmatrix}+O\left(\frac{1}{\lambda}\right) \quad (14.26)$$

$$J(x_3,t) = \underbrace{(\lambda_1^4+\lambda_2^4+\lambda_3^4)}_{O(\lambda^4)}\frac{P(x_3,t)}{\det\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)^2}+O(\lambda^3) \quad (14.27)$$

with η defined in (14.7), P in (14.6), and the notation $O(\lambda^k)$ indicates a term $f(\lambda_1,\lambda_2,\lambda_3,x_3,t)$ such that $\left|\dfrac{f(\alpha\lambda_1,\alpha\lambda_2,\alpha\lambda_3,x_3,t)}{\alpha^k}\right|$ is bounded when α goes to $+\infty$.

Proof : This is done by developing the solutions of the filters with respect to λ. See Appendix D.2.1. ■

It follows that when the λ_i are sufficiently large, $\mathcal{M}$ and J are closely related to $\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)$ and P respectively. We can thus hope to transfer the known properties of those functions to $\mathcal{M}$ and J.

About Equation (14.25)

From (14.25), we get the impression that the invertibility of $\mathcal{M}(x_3,t)$ is related to that of $\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)$, at least for λ_i sufficiently large. Actually, we have a more precise result :

Theorem 14.2.2.

Consider $(\tilde{\lambda}_1,\tilde{\lambda}_2,\tilde{\lambda}_3)$ three distinct strictly positive real numbers and assume that the inputs (u,i) and their derivatives are bounded.

Then, for any x_3 and any $\underline{d}$ such that for all t,

$$\Big|\det\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)\Big| \geq \underline{d} > 0\ ,$$

there exists $\underline{\alpha} > 0$ and $\underline{\delta} > 0$ such that for any $\alpha \geq \underline{\alpha}$,

$$\Big|\det\Big(\mathcal{M}(x_3,t)\Big)\Big| \geq \underline{\delta}$$

for all t when choosing

$$(\lambda_1,\lambda_2,\lambda_3) = (\alpha\tilde{\lambda}_1,\alpha\tilde{\lambda}_2,\alpha\tilde{\lambda}_3)\ .$$

In particular, if there exists $\underline{\omega} > 0$ such that $|\omega(t)| \geq \underline{\omega}$ for all t, there exists $\underline{\alpha} > 0$ and $\underline{\delta} > 0$ such that for all $\alpha \geq \underline{\alpha}$, $\Big|\det\Big(\mathcal{M}(R,t)\Big)\Big| \geq \underline{\delta}$ for all t.

Proof : See Appendix D.2.2. ■

We conclude that, if ω is lower-bounded away from zero, it is possible to guarantee the invertibility of $\mathcal{M}(R,t)$ for all t by taking the λ_i sufficiently large. In that case, any x_3 making $\mathcal{M}(x_3,t)$ non invertible at some time t cannot be R and can be put aside in the algorithm.

About Equation (14.26)

(14.26) implies that $\chi(x_3,t)$ is solution to the same system (14.9) (at the first order of $\frac{1}{\lambda}$) as x in the observability analysis. Therefore, whenever $(\eta(x_3,t),\dot{\eta}(x_3,t))$ is invertible, $\chi(x_3,t)$ corresponds to x in the observability analysis, and further $\big||\chi(x_3,t)-Li|^2-\Phi^2\big|$ corresponds to $P(x_3,t)$, still at the first order in $\frac{1}{\lambda}$. Thus, in order to find x_3, one could minimize $J(x_3,t) = \big||\chi(x_3,t)-Li|^2-\Phi^2\big|$ instead of (14.23). But the injection of the input i in the criteria increases its sensitivity to noise. Note that this option is exploited in the next section 14.2.3.

About Equation (14.27)

(14.27) implies that, for large values of λ_i, the criteria $J(x_3,t)$ roughly behaves like $\dfrac{P(x_3,t)}{\det\Big(\eta(x_3,t)\,,\,\dot{\eta}(x_3,t)\Big)^2}$ which is also a rational fraction with numerator of degree 6 and denominator of degree 4. Therefore, we can hope that, by choosing λ_i sufficiently large, one can ensure that J does not have more roots than P, and minimizing J is closely linked to finding the roots of P. Since P is perfectly known with Corollary 14.1.1 when ω, i_d and i_q are constant, it is possible to state the following result :

Theorem 14.2.3.

Let $(\tilde{\lambda}_1,\tilde{\lambda}_2,\tilde{\lambda}_3)$ be any three distinct strictly positive real numbers.

Assume the inputs (u,i) are bounded, and ω, i_d and i_q are constant such that $\omega \neq 0$ and $i_d \neq 0$. Then, for any initial conditions in the filters and for any $0 < \varepsilon < 1$, there exists $\underline{\alpha} > 0$

such that for all $\alpha \geq \underline{\alpha}$, by choosing

$$(\lambda_1, \lambda_2, \lambda_3) = (\alpha\lambda_1, \alpha\lambda_2, \alpha\lambda_3) \ ,$$

we have :

- there exists $\underline{\delta} > 0$ such that $\left|\det\left(\mathcal{M}(R,t)\right)\right| \geq \underline{\delta}$ for all t.
- for all t, the only two roots of $\det\left(\mathcal{M}(x_3,t)\right)$ are complex and situated in the annulus[5] $C(R, \underline{r_\varepsilon}, \overline{r_\varepsilon})$ with

$$\underline{r_\varepsilon} = \frac{\omega\Phi}{|i|}(1-\varepsilon) \quad , \quad \overline{r_\varepsilon} = \frac{\omega\Phi}{|i|}(1+\varepsilon)$$

  In other words, $\mathcal{M}(x_3,t)$ is invertible and $J(x_3,t)$ is defined for all x_3 in $\mathbb{R}$ and all t.
- for all t, $J(\cdot,t)$ admits in $[R-r_\varepsilon, R+r_\varepsilon]$
  - only one zero $\hat{R}_1(t)$ if $i_q > \frac{1-\varepsilon}{2}|i|$;
  - two zeros $(\hat{R}_1(t), \hat{R}_2(t))$ if $i_q < \frac{1-\varepsilon}{2}|i|$.

Proof : The proof of this result relies on Rouché's theorem. See Appendix D.2.3 ■

Remark 21 Unfortunately, we cannot say anything about the number of zeros of $J(\cdot,t)$ outside of $[R-r_\varepsilon, R+r_\varepsilon]$. Indeed, $J(\cdot,t)$ admits (complex) poles outside of $B_{\underline{r_\varepsilon}}(R)$ (the roots of $\det(\mathcal{M}(\cdot,t))$), and Rouché's theorem would only tell us that it admits at most 6 zeros, which we already know.

We conclude from this study, that when $|\omega|$ is lower-bounded away from zero, the invertibility of $\mathcal{M}(R,t)$ (and lower-boundedness of $\left|\det\left(\mathcal{M}(R,t)\right)\right|$) can be ensured for all t by taking the λ_i sufficiently large. According to Theorem 14.2.1, this means that $\lim_{t\to+\infty} J(R,t) = 0$ and R should appear among the minimizers of $|J(\cdot,t)|$ at least after a certain time.

In particular, when ω, i_d and i_q are constant with $\omega \neq 0$ and $i_d \neq 0$, J has only one or two zeros in the vicinity of R. Note that (Ψ, R) and (Ψ_δ, R_δ) identified in Corollary (14.1.1) are both solution to the dynamics and are both such that $y(t) = 0$ for all t. Therefore, Theorem 14.2.1 apply to both and we have in fact :

$$\lim_{t\to+\infty} J(R,t) = \lim_{t\to+\infty} J(R_\delta,t) = 0 \ .$$

This means that the two zeros of J expected with Theorem 14.2.3 are likely to be R and $R_\delta = R + \frac{2\omega\Phi i_q}{|i|^2}$ asymptotically.

In fact, although we are not able to prove it theoretically at this point, simulations seem to indicate that $P(\cdot,t)$ has always only two roots, as soon as $i_d(t) \neq 0$ and $\omega(t) \neq 0$. Therefore, $J(\cdot,t)$ has, at least after a certain time, also two roots, with one converging to R. The problem of course is that a numerical minimization of $|J(\cdot,t)|$ might return the "wrong" root. So, how to detect this situation, and how to deduce the "right" root ? Here are some elements of solution :

- most of the time, an interval for the value of R is known and the minimization can be carried out on this interval. When both roots are far apart, there might be only one in the interval of interest.
- in the case where ω and i_{dq} are constant, Theorem 14.1.4 shows how to detect whether the solution is the "right" one, if the sign of i_q is known. It also provides the exact expression of the other candidate, which can be computed by estimating the rotation speed, i-e $\hat{\dot{\theta}}$.

[5]The annulus $C(a, r_0, r_1)$ is the set of points such that $r_0 < |x-a| < r_1$.

- even in the general case, when ω and i_{dq} are not moving too fast, the two solutions may still be associated to two values of i_q of opposite sign. Therefore, the detection may still be possible if this sign is known. As for computing the other candidate, although the value given by Theorem 14.1.4 is not exact, it can enable to switch the basin of attraction and obtain the right estimate at the following iteration.

An account on the efficiency of this strategy in simulations is provided in Section 14.3.

14.2.3 Alternate observer with a reduced number of filters

Before commenting some simulations, we want to signal to the reader the existence of an observer involving a smaller number of filters, and thus a reduced computational cost.

Indeed, the dynamics (14.18) can be rewritten as

$$\dot{\overline{z_\lambda - \lambda^2\Phi^2}} = \lambda(-(z_\lambda - \lambda^2\Phi^2) + c_\lambda^\top u - \lambda^2 L^2|i|^2) \ .$$

Therefore, we can take

$$\begin{aligned}\tilde{T_\lambda}(x, x_3, t) &= T_\lambda(x, x_3, t) - \lambda^2\Phi^2 \\ &= \lambda^2(|x - Li|^2 - \Phi^2) + \lambda\,(c_\lambda(t) + 2\lambda\, Li)^\top x + \lambda\, x_3\, b_\lambda(t)^\top x + a_\lambda(t)x_3 + d_\lambda(t)x_3^2 - \lambda^2 L^2|i|^2\end{aligned}$$

which is such that $\tilde{T_\lambda}(\Psi, R, t)$ is solution of

$$\dot{\tilde{z_\lambda}} = \lambda(-\tilde{z_\lambda} + c_\lambda^\top u - \lambda^2 L^2|i|^2) \ . \tag{14.28}$$

Besides, since $(|\Psi - Li|^2 - \Phi^2) = 0$ along the solutions of interest, we can even take

$$\begin{aligned}\tilde{T_\lambda}(x, x_3, t) &= \lambda\,(c_\lambda(t) + 2\lambda\, Li + \lambda\, x_3\, b_\lambda(t))^\top x + a_\lambda(t)x_3 + d_\lambda(t)x_3^2 - \lambda^2 L^2|i|^2 \\ &= -\lambda\,\mu_\lambda(x_3, t)^\top x + a_\lambda(t)x_3 + d_\lambda(t)x_3^2 - \lambda^2 L^2|i|^2\end{aligned}$$

which is linear in x and we have like before :

$$\lim_{t\to\infty} \tilde{Z_\lambda}(t) - \tilde{T_\lambda}(\Psi(t), R, t) = 0 \ .$$

The drawback of this solution is that we use the measurement i directly in $\tilde{T_\lambda}$ and thus the estimation may be biased by noise. However, the fact that it is already linear in x suggests that it is sufficient to implement the filters (14.14)-(14.17) and (14.28) for only two values of λ to obtain x as a function of x_3. Then, the value of x_3 can be obtained by minimizing $(|x - Li|^2 - \Phi^2)$.

So consider two strictly positive real numbers λ_1 and λ_2. By defining the function

$$\tilde{T}(x, x_3, t) = \begin{pmatrix} \tilde{T}_{\lambda_1}(x, x_3, t) \\ \tilde{T}_{\lambda_2}(x, x_3, t) \end{pmatrix} = \tilde{\mathcal{M}}(x_3, t)\, x + \tilde{a}(t)\, x_3 + \tilde{d}(t)\, x_3^2 - L^2|i|^2\tilde{m_\lambda}$$

on $\mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}$, we have

$$\lim_{t\to\infty} \tilde{Z}(t) - \tilde{T}(\Psi(t), R, t) = 0 \ ,$$

where we have denoted

$$\tilde{\Lambda} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad , \quad \tilde{Z} = \begin{pmatrix} \tilde{Z}_{\lambda_1} \\ \tilde{Z}_{\lambda_2} \end{pmatrix} \quad , \quad \tilde{m_\lambda} = \begin{pmatrix} \lambda_1^2 \\ \lambda_2^2 \end{pmatrix}$$

$$\tilde{a} = \begin{pmatrix} \tilde{a}_{\lambda_1} \\ \tilde{a}_{\lambda_2} \end{pmatrix} \quad , \quad \tilde{\mathcal{M}}(x_3, t) = -\tilde{\Lambda} \begin{pmatrix} \mu_{\lambda_1}(x_3, t)^\top \\ \mu_{\lambda_2}(x_3, t)^\top \end{pmatrix} \quad , \quad \tilde{d} = \begin{pmatrix} \tilde{d}_{\lambda_1} \\ \tilde{d}_{\lambda_2} \end{pmatrix} \ .$$

Since x can be simply deduced from $\tilde{T}$ by inversion of $\tilde{\mathcal{M}}$, it is natural to try the following simple algorithm :

Alternate algorithm

Implement filters (14.14)-(14.17) and (14.28) for two strictly positive real numbers λ_1 and λ_2, and at any time t, find an estimate $\hat{R}(t)$ of R by

$$\hat{R}(t) = \mathsf{Argmin}_{x_3 \in \mathbb{R}^+} |\tilde{J}(x_3, t)| \ ,$$

where

$$\tilde{\chi}(x_3, t) = \tilde{\mathcal{M}}(x_3, t)^{-1} \left(\tilde{Z}(t) - \tilde{a}(t)\, x_3 - \tilde{d}(t)\, x_3^2 + L^2 |i|^2 \tilde{m}_\lambda \right) \ , \tag{14.29}$$

$$\tilde{J}(x_3, t) = |\tilde{\chi}(x_3, t) - Li|^2 - \Phi^2 \ , \tag{14.30}$$

and an estimate $\hat{\Psi}(t)$ of $\Psi(t)$ by

$$\hat{\Psi}(t) = \tilde{\chi}(\hat{R}(t), t) \ .$$

Once again, it leads to the questions of invertibility of the matrix $\tilde{\mathcal{M}}$ and uniqueness of solutions to the minimization problem. But in the same spirit as Lemma 14.2.2, it is possible to show that

$$\det(\tilde{\mathcal{M}}(x_3, t)) = 4 \underbrace{(\lambda_2 - \lambda_1)}_{O(\lambda)} \det \left(\eta(x_3, t), \dot{\eta}(x_3, t) \right) + O(1)$$

$$\begin{pmatrix} \eta(x_3, t)^\top \\ \dot{\eta}(x_3, t)^\top \end{pmatrix} (\tilde{\chi}(x_3, t) - Li(t)) = \begin{pmatrix} 0 \\ -|\eta(x_3, t)|^2 \end{pmatrix} + O\left(\frac{1}{\lambda}\right)$$

$$\tilde{J}(x_3, t) = \frac{P(x_3, t)}{\det \left(\eta(x_3, t)\ ,\ \dot{\eta}(x_3, t) \right)^2} + O\left(\frac{1}{\lambda}\right) \ ,$$

so that the same conclusions hold.

The main advantage of this algorithm is that the filters are implemented for only two λ (instead of three), thus reducing the dimension of the state from 21 to 14. However, the measurement i is used directly in the computation of $\tilde{\chi}$ and $\tilde{J}$, which, in presence of noise, can significantly deteriorate the invertibility of $\tilde{\chi}$ and the estimation of $\hat{R}$ and $\hat{\Psi}$.

14.3 Simulations

Model and scenario. The simulations presented in this chapter are based on ideal data produced by a general PMSM model of the type (12.5), where the input u is chosen to follow a desired rotation speed ω_R. The details of this model and of the controller is of no interest here, as long as the produced signals are solution to our model (12.6). The speed scenario chosen to test our observer is shown on Figure 14.1. The corresponding signals (u, i) are given in Figure 14.2. Note that at $t = 3$, although the speed setpoint is constant, an external torque is added, resulting in a transient behavior in the signals. This torque then remains constant throughout the simulation.

Observer algorithm. Choose strictly positive real numbers G and dt_R, a one-dimensional grid $\mathcal{G}$ of the interval $[-G, G]$, and three distinct strictly positive real numbers λ_1, λ_2, λ_3. We assume the machine is used as a motor, i-e that i_q is positive. The observer consists of the following modules :

- Implementation of filters (14.14)-(14.17) and (14.18).
- Computation of

$$\hat{\Psi}(t) = \chi(\hat{R}, t) \ ,$$

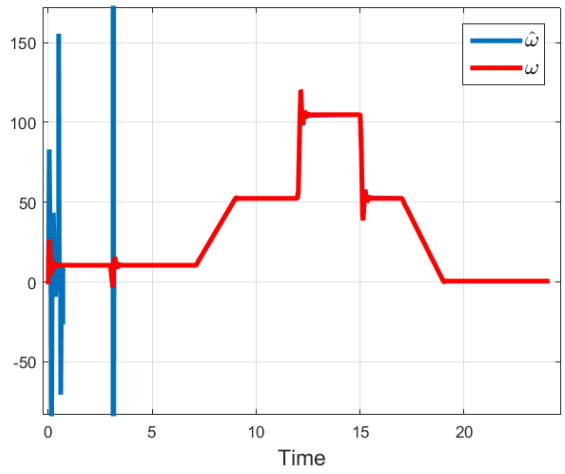
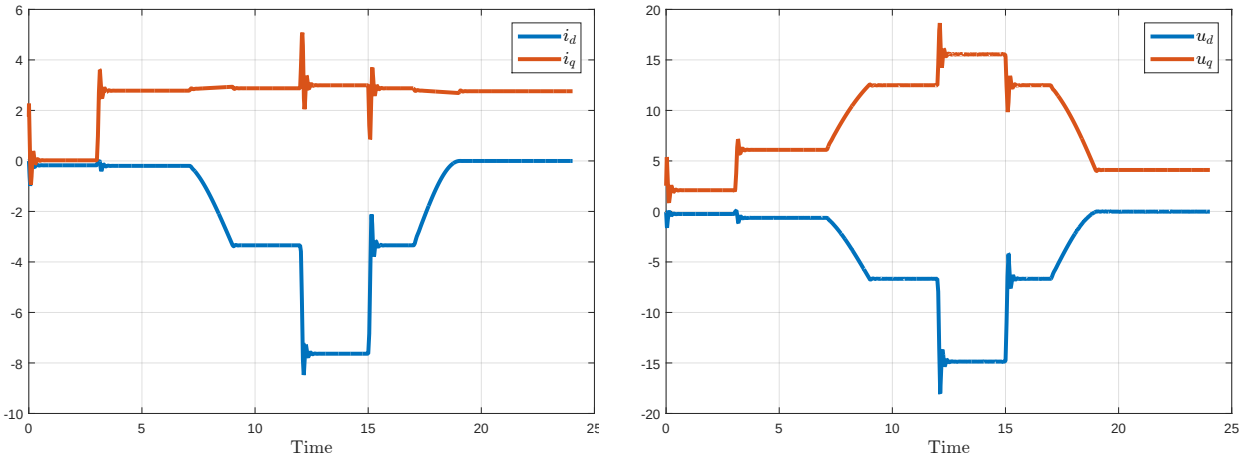


Figure 14.1: Rotation speed $\omega = \dot{\theta}$ and estimated rotation speed $\hat{\omega} = \dot{\hat{\theta}}$. The estimation algorithm starts at $t = 0.5$.



(a) $i_{dq} = \mathcal{R}(-\theta)i$ (b) $u_{dq} = \mathcal{R}(-\theta)u$

Figure 14.2: Currents and voltages in the rotating frame.

$$\hat{\theta}(t) = \arg(\hat{\Psi}(t) - Li(t)) \ ,$$

at each time t, given the current value of $\hat{R}$.

- Estimation of $\hat{\omega}(t) = \dot{\hat{\theta}}(t)$ at each time t (see below).
- Update of the value of $\hat{R}$ every $dt_R > 0$ with the following algorithm :

$\hat{R}_1 = \mathsf{Argmin}_{x_3 \in \hat{R}+\mathcal{G}} |J(x_3, t)|$
$\hat{\Psi}_1 = \chi(\hat{R}_1, t)$
$\hat{\theta}_1 = \arg(\hat{\Psi}_1 - Li(t))$
$i_{q,1} = \Big[-\sin(\theta_1) \, , \cos(\theta_1) \Big] i(t)$
if $i_{q,1} \geq 0$ **then**
  $\hat{R} = \hat{R}_1$
else
  $\hat{R}_2 = \hat{R}_1 + \frac{2\Phi\hat{\omega}(t) i_{q,1}}{|i(t)|^2}$

$\delta = \hat{R}_2 - \hat{R}_1$
if $|\delta| > G$ **then**
  $\hat{R} = \hat{R}_2$
else if $\delta > 0$ **then**
  $\hat{R} = \mathsf{Argmin}_{x_3 \in \hat{R} + (\mathcal{G} \cap [\frac{\delta}{2}, G])} |J(x_3, t)|$
else
  $\hat{R} = \mathsf{Argmin}_{x_3 \in \hat{R} + (\mathcal{G} \cap [-G, \frac{\delta}{2}])} |J(x_3, t)|$
end if
end if

In other words, the minimum of J is computed on the grid $\hat{R} + \mathcal{G}$ centered at the current value $\hat{R}$ and, if the corresponding i_q is positive, this value is kept. Otherwise, we take the other candidate given by Theorem 14.1.3, or rather, if this other value is in the grid where J has already been computed, the true minimum of J around this value is computed. This latter step can be removed and postponed to the next iteration, but it offers the possibility of correcting the estimate given by Theorem 14.1.3 when ω, i_d and i_q are not constant and/or when $\hat{w}$ is not exact.

Note that instead of starting the estimation process right away, one can wait for the filters to reach their steady-state, i-e "forget" their initial conditions. In the simulations presented here, we waited for 0.5s.

Estimation of $\hat{\omega} = \dot{\hat{\theta}}$. In order to implement the previous algorithm, ω and thus $\hat{\omega} = \dot{\hat{\theta}}$ needs to be estimated. This can be done in numerous ways including dirty derivatives, exact differentiators etc. A raw dirty derivative approach would be to take :

$$\left\{ \begin{array}{rcl} \dot{\hat{\theta}}_e & = & \hat{\omega}_e - \ell(\hat{\theta}_e - y) \\ \dot{\hat{\omega}}_e & = & -\ell^2(\hat{\theta}_e - y) \end{array} \right. , \quad y = \hat{\theta}$$

with ℓ sufficiently large to compensate for the neglected $\dot{\hat{\omega}}$. But the correction $(\hat{\theta}_e - y)$ is not a good idea because it does not vanish at $2k\pi$. A possible solution is to take the correction $\arctan_2(\sin(\hat{\theta}_e - y), \cos(\hat{\theta}_e - y))$ instead, but convergence is not guaranteed.

Another idea is to consider as measurements $x_1 = \cos(\hat{\theta})$ and $\tilde{x}_1 = \sin(\hat{\theta})$ and build a high gain observer for $(x_1, \dot{x}_1, \tilde{x}_1, \dot{\tilde{x}}_1)$, from which $\hat{\omega}$ can easily be deduced. This method offers the advantage of making no approximation on $\dot{\hat{\omega}}$, but it leads to an observer of dimension 4.

An intermediate solution is to use a reduced order observer of dimension 3 of the form :

$$\left\{ \begin{array}{rcl} \dot{\hat{\chi}} & = & -(\hat{\gamma} - k + \ell^2)\, y - \ell \hat{\chi} \\ \dot{\hat{\tilde{\chi}}} & = & -(\hat{\gamma} - k + \ell^2)\, \tilde{y} - \ell \hat{\tilde{\chi}} \\ \dot{\hat{\gamma}} & = & 2k\,(\hat{\chi} + \ell\, y)\, y + 2k\,(\hat{\tilde{\chi}} + \ell\, \tilde{y})\, \tilde{y} \end{array} \right. , \quad y = \cos\hat{\theta}\ , \quad \tilde{y} = \sin\hat{\theta}$$

and

$$\hat{\omega}_e = \hat{\chi}^2 + \hat{\tilde{\chi}}^2 - \ell^2 \quad \text{or} \quad \hat{\omega}_e = y\tilde{\chi} - \tilde{y}\,\chi\ .$$

It is possible to prove[6] that $\hat{\omega}_e$ converges to $\hat{\omega}$, at least when $\hat{\omega}$ is constant.

This latter observer is used for the estimation of $\hat{\omega}$ (and thus ω) shown in Figure 14.1 with $\ell = 1000$ and $k = 500$.

[6] Take $x_2 = \dot{x}_1 = -\hat{w}\sin\hat{\theta}$, $\tilde{x}_2 = \dot{\tilde{x}}_1 = \hat{\omega}\cos\hat{\theta}$, $\chi = x_2 - \ell x_1$, $\tilde{\chi} = \tilde{x}_2 - \ell\tilde{x}_1$ and $\gamma = \hat{\omega}^2 + kx_1^2 + k\tilde{x}_1^2 = \hat{\omega}^2 + k$. Denoting $e_\chi = \hat{\chi} - \chi$, $\tilde{e}_\chi = \hat{\tilde{\chi}} - \tilde{\chi}$ and $e_\gamma = \hat{\gamma} - \gamma$, we get $\dot{\overbrace{ke_\chi^2 + k\tilde{e}_\chi^2 + \frac{1}{2}e_\gamma^2}} = -2k\ell(e_\chi^2 + \tilde{e}_\chi^2)$, and thus $\lim e_\chi = 0$ and $\lim \tilde{e}_\chi = 0$. Hence the convergence of $\hat{\omega}_e$.

Results The results of the simulations are presented in Figure 14.3, for two grids with amplitude $G = 1$ and $G = 0.1$ respectively.

Observe that with $G = 1$, the algorithm finds the right value of R in two iterations only, whereas with $G = 0.1$, it takes a longer time before R can appear in the grid. In fact, for a same precision, the broader the grid, the higher the chances of R appearing in it, but the larger the number of points and computation time, and also the higher the chances of having several minima in the grid. In practice, one know roughly well the initial value of the resistance, so that a grid with small amplitude can be chosen, which is then going to follow R throughout the experiment, in the case where it evolves due to temperature.

The evolution of the criteria J during the simulation with $G = 1$ is shown in Figure 14.4. One can see that the minimum is well marked around $R = 1.45$.

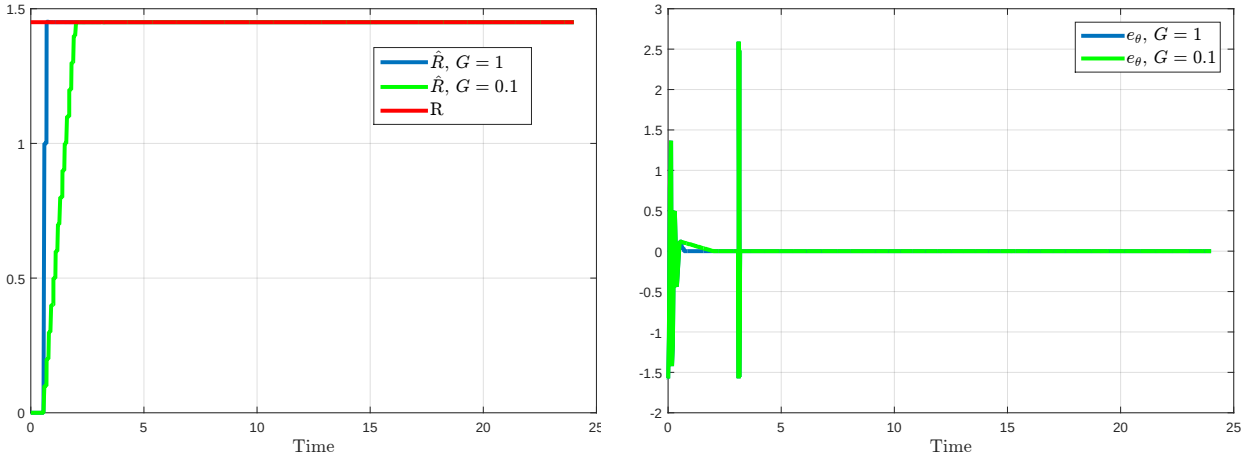
As for the estimation of θ, it naturally converges once $\hat{R}$ has converged. It is interesting to observe the peak in the error around $t = 3$ (which in turn appears on $\hat{\omega}$). This is due to the sudden addition of a torque which destabilizes i_d and ω and makes them go through 0. We have seen that in this case, observability is lost and $\mathcal{M}(R, t)$ is likely to be non invertible (the assumption of Theorem 14.2.2 is no longer verified). This event is not visible on $\hat{R}$ because it is not updated at those precise moments.

14.4 Conclusion

Unlike (Ψ, Φ), the couple (Ψ, R) is not observable from the only knowledge that $y(t) = 0$ for all t. However, when ω and i_d are non zero, there are at most six indistinguishable solutions, the resistance being one of the roots of a polynomial of degree 6. Besides, in the particular case where ω, i_d and i_q are constant, the number of possible solutions is reduced to two, with two well-identified values for the resistance. But those solutions turn out to be distinguishable if the sign of i_q (i-e the mode of use of the machine) is known. This information has enabled us to propose an observation strategy based on a Luenberger design. It remains now to test this observer on real data, and to understand the affect of saliency on this algorithm.

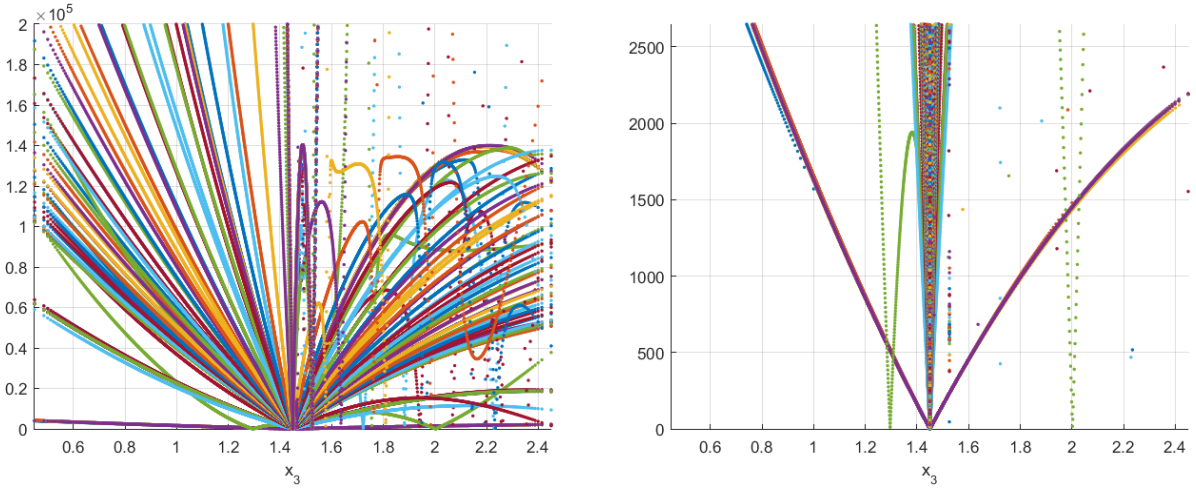
Note that in this context of non observability, it would be impossible to write the dynamics of the observer in the original coordinates (x, x_3) as recommended in Part III (the transformation is not even injective). Interestingly, the step of inversion of the transformation via minimization is crucial to the design because it allows to incorporate the additional information about the sign of i_q and to use a discontinuous strategy, which has no influence on the dynamics of the observer.

As a final remark, this example strongly advocates in favor of the Luenberger methodology. Indeed, we are not aware of any other observer construction which could work in this case. In particular, a high gain design is out of the question for two reasons : first, it would necessarily involve the derivatives of (u, i) which is undesirable in practice, and also, the dynamics of the observer depend on the inverse of the transformation.



(a) $\hat{R}$ (b) $e_\theta = \hat{\theta} - \theta$

Figure 14.3: Results of the observer algorithm with $\lambda_1 = 20$, $\lambda_2 = 30$, $\lambda_3 = 40$, $dt_R = 0.1$, and two grids with amplitude $G = 1$ and $G = 0.1$ respectively. The estimation starts at $t = 0.5$.



(b) Zoom around $R = 1.45$

Figure 14.4: Plot of the criteria $J(\cdot, t)$ on the grid with $G = 1$ at each iteration where $\hat{R}$ is updated, i-e every $dt_R = 0.1$.

Bibliography

[ABS13] V. Andrieu, G. Besançon, and U. Serres. Observability necessary conditions for the existence of observers. *Conference on Decision and Control*, 2013.

[AEP14] V. Andrieu, J.-B. Eytard, and L. Praly. Dynamic extension without inversion for observers. *IEEE Conference on Decision and Control*, pages 878–883, 2014.

[AK01] M. Arcak and P. Kokotovic. Observer-based control systems with slope-restricted nonlinearities. *IEEE Transactions on Automatic Control*, 46, 2001.

[Ala07] M. Alamir. *Nonlinear Observers and Applications*, volume 363, chapter Nonlinear Moving Horizon Observers : Theorey and Real-Time Implementation, pages 139–179. Springer, 2007.

[And05] V. Andrieu. *Bouclage de sortie et observateur*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2005.

[And14] V. Andrieu. Convergence speed of nonlinear Luenberger observers. *SIAM Journal on Control and Optimization*, 52(5):2831–2856, 2014.

[AP05] A. Astolfi and L. Praly. Global complete observability and output-to-state stability imply the existence of a globally convergent observer. *Mathematical Control Signals Systems*, 18(1):1–34, 2005.

[AP06] V. Andrieu and L. Praly. On the existence of a Kazantzis–Kravaris / Luenberger observer. *SIAM Journal on Control and Optimization*, 45(2):432–456, 2006.

[AP13] D. Astolfi and L. Praly. Output feedback stabilization for siso nonlinear systems with an observer in the original coordinate. *IEEE Conference on Decision and Control*, pages 5927 – 5932, 2013.

[APA06] V. Andrieu, L. Praly, and A. Astolfi. Nonlinear output feedback design via domination and generalized weighted homogeneity. *IEEE Conference on Decision and Control*, 2006.

[APA08] V. Andrieu, L. Praly, and A. Astolfi. Homogeneous approximation, recursive observer design, and output feedback. *SIAM Journal on Control and Optimization*, 47(4):1814–1850, 2008.

[APA09] V. Andrieu, L. Praly, and A. Astolfi. High gain observers with updated gain and homogeneous correction terms. *Automatica*, 45(2):422–428, 2009.

[AW06] P.P Acarnley and J.F Watson. Review of position-sensorless operation of brushless permanent-magnet machines. *IEEE Transactions on Industrial Electronics*, 53(2):352–362, 2006.

[BB97] G. Besançon and G. Bornard. On characterizing a class of observer forms for nonlinear systems. *European Control Conference*, 1997.

[BBD96] J.P. Barbot, T. Boukhobza, and M. Djemai. Sliding mode observer for triangular input form. *IEEE Conference on Decision and Control*, 2:1489 – 1490, 1996.

[BBH96] G. Besançon, G. Bornard, and H. Hammouri. Observer synthesis for a class of nonlinear control systems. *European Journal of Control*, 3(1):176–193, 1996.

[BBP+16] A. Bobtsov, D. Bazylev, A. Pyrkin, S. Aranovsky, and R. Ortega. A robust nonlinear position observer for synchronous motors with relaxed excitation conditions. *International Journal of Control*, 2016.

[BC98] M. Bodson and J. Chiasson. Differential-geometric methods for control of electric motors. *Int. J. Robust. Nonlinear Control*, 8:923–254, 1998.

[Bes99] G. Besançon. Further results on high gain observers for nonlinear systems. *IEEE Conference on Decision and Control*, 3:2904–2909, 1999.

[BH91] G. Bornard and H. Hammouri. A high gain observer for a class of uniformly observable systems. *Conference on Decision and Control*, 1991.

[BM58] R. Bott and J. Milnor. On the parallelizability of the spheres. *Bullettin of American Mathematical Society*, 64(3):87–89, 1958.

[BP17] P. Bernard and L. Praly. Robustness of rotor position observer for permanent magnet synchronous motors with unknown magnet flux. *IFAC World Congress*, 2017.

[BPA15] P. Bernard, L. Praly, and V. Andrieu. Tools for observers based on coordinate augmentation. *IEEE Conference on Decision and Control*, 2015.

[BPA17a] P. Bernard, L. Praly, and V. Andrieu. Observers for a non-lipschitz triangular form. *Automatica*, 82:301–313, 2017.

[BPA17b] P. Bernard, L. Praly, and V. Andrieu. On the triangular canonical form for uniformly observable controlled systems. *Automatica*, 85:293–300, 2017.

[BPAew] P. Bernard, L. Praly, and V. Andrieu. Expressing an observer in given coordinates by augmenting and extending an injective immersion to a surjective diffeomorphism. *SIAM Journal on Control and Optimization*, 2015, Under review.

[BPO15a] A. Bobtsov, A. Pyrkin, and R. Ortega. A new approach for estimation of electrical parameters and flux observation of permanent magnet synchronous motors. *Int. J. Adapt. Control Signal Process.*, 30:1434–1448, 2015.

[BPO+15b] A. Bobtsov, A. Pyrkin, R. Ortega, S. Vukosavic, A. Stankovic, and E. Panteley. A robust globally convergent position observer for the permanent magnet synchronous motor. *Automatica*, 61:47–54, 2015.

[BRG89] D. Bossane, D. Rakotopara, and J. P. Gauthier. Local and global immersion into linear systems up to output injection. *Conference on Decision and Control*, pages 2000–2004, 1989.

[BS04] J. Back and J.H. Seo. Immersion of non-linear systems into linear systems up to output injection : Characteristic equation approach. *Internation Journal of Control*, 77(8):723–734, 2004.

[BS15] S. Bonnabel and J-J. Slotine. A contraction theory-based analysis of the stability of the deterministic Extended Kalman Filter. *IEEE Transactions on Automatic Control*, 60(2):565–569, 2015.

[BT07] G. Besançon and A. Ticlea. An immersion-based observer design for rank-observable nonlinear systems. *IEEE Transactions on Automatic Control*, 52(1):83–88, 2007.

[BZ83] D. Bestle and M. Zeitz. Canonical form observer design for nonlinear time variable systems. *International Journal of Control*, 38:419–431, 1983.

[Che84] C-T Chen. *Linear system theory and design.* CBS College Publishing, 1984.

[CZM16] E. Cruz-Zavala and J. A. Moreno. Lyapunov functions for continuous and discontinuous differentiators. *IFAC Symposium on Nonlinear Control Systems*, 2016.

[CZMF11] E. Cruz-Zavala, J.A. Moreno, and L. Fridman. Uniform robust exact differentiator. *EEE Transactions on Automatic Control*, 56(11):2727–2733, 2011.

[DBGR92] F. Deza, E. Busvelle, J.P. Gauthier, and D. Rakotopara. High gain estimation for nonlinear systems. *Systems & Control Letters*, 18:295–299, 1992.

[Die60] J. Dieudonné. *Foundations of Modern Analysis.* Academic Press, 1960.

[Dug66] J. Dugundgi. *Topology.* Allyn and Bacon, 1966.

[Eck06] B. Eckmann. *Mathematical survey lectures* 1943 − 2004. Springer, 2006.

[EKNN89] S.V. Emelyanov, S.K. Korovin, S.V. Nikitin, and M.G. Nikitina. Observers and output differentiators for nonlinear systems. *Doklady Akademii Nauk*, 306:556–560, 1989.

[Eng05] R. Engel. Exponential observers for nonlinear systems with inputs. Technical report, Universität of Kassel, Department of Electrical Engineering, 2005.

[Eng07] R. Engel. Nonlinear observers for Lipschitz continuous systems with inputs. *International Journal of Control*, 80(4):495–508, 2007.

[Fil88] A. Filippov. *Differential equations with discontinuous right-hand sides.* Mathematics and its Applications Kluwer Academic Publishers Group, 1988.

[FK83] M. Fliess and I. Kupka. A finiteness criterion for nonlinear input-output differential systems. *SIAM Journal of Control and Optimization*, 21(5):721–728, 1983.

[Fli82] M. Fliess. Finite-dimensional observation-spaces for non-linear systems. In Heidelberg Springer, Berlin, editor, *Hinrichsen D., Isidori A. (eds) Feedback Control of Linear and Nonlinear Systems. Lecture Notes in Control and Information Sciences*, volume 39, pages 73–77, 1982.

[GB81] J-P. Gauthier and G. Bornard. Observability for any u(t) of a class of nonlinear systems. *IEEE Transactions on Automatic Control*, 26:922 – 926, 1981.

[GBC+15a] V. Gibert, L. Burlion, A. Chriette, J. Boada, and F. Plestan. New pose estimation scheme in perspective vision system during civil aircraft landing. *IFAC Symposium on Robot Control*, 2015.

[GBC+15b] V. Gibert, L. Burlion, A. Chriette, J. Boada, and F. Plestan. Nonlinear observers in vision system : Application to civil aircraft landing. *European Control Conference*, 2015.

[Gel74] A. Geld. *Applied Optimal Estimation.* MIT Press : Cambridge, 1974.

[GHO92] J.-P. Gauthier, H. Hammouri, and S. Othman. A simple observer for nonlinear systems application to bioreactors. *IEEE Transactions on Automatic Control*, 37(6):875–880, 1992.

[Gib16] V. Gibert. *Analyse d'observabilité et synthèse d'observateurs robustes pour l'atterrissage basé vision d'avions de ligne sur des pistes inconnues.* PhD thesis, Ecole Centrale de Nantes, 2016.

[GK01] J-P. Gauthier and I. Kupka. *Deterministic observation theory and applications.* Cambridge University Press, 2001.

[GMP96] A. Glumineau, C. H. Moog, and F. Plestan. New algebro-geometric conditions for the linearization by input-output injection. *IEEE Transactions on Automatic Control*, 41(4):598–603, 1996.

[GRHS00] J.L. Gouzé, A. Rapaport, and M.Z. Hady-Sadok. Interval observers for uncertain biological systems. *Ecological Modelling*, 133:45–56, 2000.

[Gua02] M. Guay. Observer linearization by output-dependent time-scale transformations. *IEEE Transactions on Automatic Control*, 47(10), 2002.

[Hah67] W. Hahn. *Stability of Motion.* Springer-Verlag, 1967.

[Ham08] Y. Hamami. Observateur de Kazantzis-Kravaris/Luenberger dans le cas d'un système instationnaire. Technical report, MINES ParisTech, 2008.

[HBB10] H. Hammouri, G. Bornard, and K. Busawon. High gain observer for structured multi-output nonlinear systems. *IEEE Transactions on Automatic Control*, 55(4):987–992, 2010.

[HC91] H. Hammouri and F. Celle. Some results about nonlinear systems equivalence for the observer synthesis. In *New Trends in Systems Theory*, volume 7 of *New Trends in Systems Theory*, pages 332–339. Birkhäuser, 1991.

[Hen14] N. Henwood. *Estimation en ligne de paramètres de machines électriques pour véhicule en vue d'un suivi de la température de ses composants.* PhD thesis, Control and System Center, MINES ParisTech, https://pastel.archives-ouvertes.fr/pastel-00958055, 2014.

[Hir76] M. Hirsch. *Differential topology.* Springer, 1976.

[HK77] R. Hermann and A.J. Krener. Nonlinear controllability and observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.

[HK96] H. Hammouri and M. Kinnaert. A new procedure for time-varying linearization up to output injection. *Systems & Control Letters*, 28:151–157, 1996.

[HM90] H. Hammouri and J. De Leon Morales. Observer synthesis for state-affine systems. *IEEE Conference on Decision and Control*, pages 784–785, 1990.

[HMP12] N. Henwood, J. Malaizé, and L. Praly. A robust nonlinear luenberger observer for the sensorless control of SM-PMSM : Rotor position and magnets flux estimation. *IECON Conference on IEEE Industrial Electronics Society*, 2012.

[HOD99] L. Hsu, R. Ortega, and G.R. Damm. A globally convergent frequency estimator. *IEEE Transactions on Automatic Control*, 44(4):698–713, 1999.

[Hou05] M. Hou. Amplitude and frequency estimator of a sinusoid. *IEEE Transactions on Automatic Control*, 50(6):855–858, 2005.

[Hou12] M. Hou. Parameter identification of sinusoids. *IEEE Transactions on Automatic Control*, 57(2), 2012.

[Jaz70] A. H. Jazwinski. *Stochastic processes and filtering theory.* Academic Press, 1970.

[JG96] P. Jouan and J.P. Gauthier. Finite singularities of nonlinear systems. Output stabilization, observability, and observers. *Journal of Dynamical and Control Systems*, 2(2):255–288, 1996.

[Jou03] P. Jouan. Immersion of nonlinear systems into linear systems modulo output injection. *SIAM Journal on Control and Optimization*, 41(6):1756–1778, 2003.

[Kal60] R.E Kalman. Contributions to the theory of optimal control. *Conference on Ordinary Differential Equations*, 1960.

[KB61] R.E Kalman and R.S Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 108:83–95, 1961.

[KE03] G. Kreisselmeier and R. Engel. Nonlinear observers for autonomous lipshitz continuous systems. *IEEE Transactions on Automatic Control*, 48(3):451–464, 2003.

[Kel87] H. Keller. Nonlinear observer by transformation into a generalized observer canonical form. *Internation Journal of Control*, 46(6):1915–1930, 1987.

[Kha02] H. Khalil. *Nonlinear Systems, 3rd Edition.* Prentice Hall, 2002.

[KI83] A.J. Krener and A. Isidori. Linearization by output injection and nonlinear observers. *Systems & Control Letters*, 3:47–52, 1983.

[Kir34] M. D. Kirszbraun. über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22:77–108, 1934.

[KK98] N. Kazantzis and C. Kravaris. Nonlinear observer design using Lyapunov's auxiliary theorem. *Systems and Control Letters*, 34:241–247, 1998.

[KP13] H. K. Khalil and L. Praly. High-gain observers in nonlinear feedback control. *Int. J. Robust. Nonlinear Control*, 24, April 2013.

[KR85] A. J. Krener and W. Respondek. Nonlinear observers with linearizable dynamics. *SIAM Journal of Control and Optimization*, 23(2):197–216, 1985.

[KX03] A.J. Krener and M. Xiao. Nonlinear observer design in the Siegel domain. *SIAM Journal on Control and Optimization*, 41(3):932–953, 2003.

[KX06] A.J. Krener and M. Xiao. Nonlinear observer design for smooth systems. *Chaos in Automatic Control, W. Perruguetti and J.-P. Barbot, Eds., Taylor and Francis*, pages 411–422, 2006.

[Leb82] D. Leborgne. *Calcul Différentiel et Géometrie.* Presse Universitaire de France, 1982.

[Lee13] J. M. Lee. *Introduction to Smooth Manifolds.* Springer, 2013.

[Lev b] A. Levant. Higher-order sliding modes and arbitrary-order exact robust differentiation. *Proceedings of the European Control Conference*, pages 996–1001, 2001 b.

[Lev03] A. Levant. Higher-order sliding modes, differentiation and output-feedback control. *International Journal of Control*, 76(9-10):924–941, 2003.

[Lev05] A. Levant. Homogeneity approach to high-order sliding mode design. *Automatica*, 41(5):823–830, 2005.

[LHN+10] J. Lee, J. Hong, K. Nam, R. Ortega, L. Praly, and A. Astolfi. Sensorless control of surface-mount permanent-magnet synchronous motors based on a nonlinear observer. *IEEE Transactions on Power Electronics*, 25(2):290–297, 2010.

[Lue64] D. Luenberger. Observing the state of a linear system. *IEEE Transactions on Military Electronics*, 8:74–80, 1964.

[LZA03] H. Lin, G. Zhai, and P. J. Antsaklis. Set-valued observer design for a class of uncertain linear systems with persistent disturbance. *American Control Conference*, 2003.

[McS34] E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934.

[Mil65] J. Milnor. *Lectures on the h-cobordism Theorem. Notes by L. Siebenmann and J. Sondow*. Princeton University Press, 1965.

[MP03] M. Maggiore and K.M. Passino. A separation principle for a class of non uniformly completely observable systems. *IEEE Transactions on Automatic Control*, 48, July 2003.

[MPH12] J. Malaizé, L. Praly, and N. Henwood. Globally convergent nonlinear observer for the sensorless control of surface-mount permanent magnet synchronous machines. *IEEE Conference on Decision and Control*, 2012.

[MV00] J. Moreno and A. Vargas. Approximate high-gain observers for uniformly observable nonlinear systems. In *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, volume 1, pages 784–789. IEEE, 2000.

[NYN04] H. Nakamura, Y. Yamashita, and H. Nishitani. Lyapunov functions for homogeneous differential incolusions. *IFAC Symposium on Nonlinear Control Systems*, 2004.

[OPA+11] R. Ortega, L. Praly, A. Astolfi, J. Lee, and K. Nam. Estimation of rotor position and speed of permanent magnet synchronous motors with guaranteed stability. *IEEE Transactions on Control Systems Technology*, 19(3):601–614, 2011.

[OPCTL02] G. Obregón-Pulido, B. Castillo-Toledo, and A. Loukianov. A globally convergent estimator for n frequencies. *IEEE Transactions on Automatic Control*, 47(5):857–863, 2002.

[ORSM15] F.-A. Ortiz-Ricardez, T. Sanchez, and J.-A. Moreno. Smooth lyapunov function and gain design for a second order differentiator. *IEEE Conference on Decision and Control*, pages 5402–5407, 2015.

[PdNC91] L. Praly, B. d'Andréa Novel, and J.-M. Coron. Lyapunov design of stabilizing controllers for cascaded systems. *IEEE Transactions on Automatic Control*, 36(10):1177–1181, 1991.

[PG97] F. Plestan and A. Glumineau. Linearization by generalized input-output injection. *Systems & Control Letters*, 31:115–128, 1997.

[PJ04] L. Praly and Z.P. Jiang. Linear output feedback with dynamic high gain for nonlinear systems. *Systems and Control Letters*, 53:107–116, 2004.

[PMI06] L. Praly, L. Marconi, and A. Isidori. A new observer for an unknown harmonic oscillator. *Symposium on Mathematical Theory of Networks and Systems*, 2006.

[PPO08] F. Poulain, L. Praly, and R. Ortega. An observer for permanent magnet synchronous motors with currents and voltages as only measurements. *IEEE Conference on Decision and Control*, 2008.

[Qia05] C. Qian. A homogeneous domination approach for global output feedback stabilization of a class of nonlinear systems. *Proceedings of the American Control Conference*, 2005.

[QL06] C. Qian and W. Lin. Recursive observer design, homogeneous approximation, and nonsmooth output feedback stabilization of nonlinear systems. *IEEE Transactions on Automatic Control*, 51(9), 2006.

[RM82] A. Michel R. Miller. *Ordinary Differential Equations*. Academic Press, 1982.

[RM04] A. Rapaport and A. Maloum. Design of exponential observers for nonlinear systems by embedding. *International Journal of Robust and Nonlinear Control*, 14:273–288, 2004.

[ROH+16] J-G. Romero, R. Ortega., Z. Han, T. Devos, and F. Malrait. An adaptive flux observer for the permanent magnet synchronous motor. *Int. J. Adapt. Control Signal Process.*, 30:473–487, 2016.

[RPN04] W. Respondek, A. Pogromski, and H. Nijmeijer. Time scaling for observer design with linearizable error dynamics. *Automatica*, 40:277–285, 2004.

[RZ13] F. Rotella and I. Zambettakis. On functional observers for linear time-varying systems. *IEEE Transactions on Automatic Control*, 58(5), 2013.

[Sho92] A. Shoshitaishvili. On control branching systems with degenerate linearization. *IFAC Symposium on Nonlinear Control Systems*, pages 495–500, 1992.

[SL16] H. Shim and D. Liberzon. Nonlinear observers robust to measurement disturbances in an ISS sense. *IEEE Transactions on Automatic Control*, 61(1), 2016.

[Smi01] G. Smirnov. *Introduction to the theory of differential inclusions*, volume 41. Graduate studies in Mathematics, 2001.

[Son89] E. Sontag. Smooth stabilization implies coprime factorization. *IEEE Transactions on Automatic Control*, 34(4), 1989.

[SP11] R. Sanfelice and L. Praly. On the performance of high-gain observers with gain adaptation under measurement noise. *Automatica*, 47:2165–2176, 2011.

[SW95] E. Sontag and Y. Wang. On characterizations of the input-to-state stability property. *Systems & Control Letter*, 24:351–359, 1995.

[Tee96] A. R. Teel. Nonlinear small gain theorem for the analysis of control systems with saturations. *IEEE Transactions on Automatic Control*, 41(9):1256–1270, 1996.

[Tor89] A. Tornambe. Use of asymptotic observers having high gains in the state and parameter estimation. *IEEE Conference on Decision and Control*, 2:1791–1794, 1989.

[Tru07] J. Trumpf. Observers for linear time-varying systems. *Linear Algebra and its Applications*, 425:303–312, 2007.

[Val45] F. A. Valentine. A lipschitz condition preserving extension for a vector function. *American Journal of Mathematics*, 67(1):83–93, 1945.

[Waz35] T. Wazewski. *Sur les matrices dont les éléments sont des fonctions continues*, volume 2. Composito Mathematica, 1935. p. 63-68.

[Wil69] F. W. Wilson. Smoothing derivatives of functions and applications. *Transactions American Mathematical Society*, 139:413–428, 1969.

[YL04] B. Yang and W. Lin. Homogeneous observers, iterative design, and global stabilization of high-order nonlinear systems by smooth output feedback. *IEEE Transactions on Automatic Control*, 49(7):1069–1080, 2004.

[Zei84] M. Zeitz. Observability canonical (phase-variable) form for nonlinear time-variable systems. *International Journal of Systems Science*, 15(9):949–958, 1984.

[Zim94] G. Zimmer. State observation by on-line minimization. *International Journal of Control*, 60(4):595–606, 1994.

Appendix A

Technical lemmas

In this appendix, we give the proof to some general technical lemmas used throughout this thesis.

A.1 About homogeneity

Lemma A.1.1.

Let η be a continuous functions defined on $\mathbb{R}^{n+1}$ and f a continuous function defined on $\mathbb{R}^n$. Let $\mathcal{C}$ be a compact subset of $\mathbb{R}^n$. Assume that, for all x in $\mathcal{C}$ and s in $\mathtt{S}(f(x))$,

$$\eta(x,s) < 0 \ .$$

Then, there exists $\alpha > 0$ such that for all x in $\mathcal{C}$ and s in $\mathtt{S}(f(x))$

$$\eta(x,s) < -\alpha \ .$$

Proof : Assume that for all $k > 0$, there exists x_k in $\mathcal{C}$ and s_k in $\mathtt{S}(f(x_k)) \subset [-1,1]$ such that

$$0 > \eta(x_k, s_k) \geq -\frac{1}{k} \ .$$

Then, $\eta(x_k, s_k)$ tends to 0 when k tends to infinity. Besides, there exists a subsequence (k_m) such that x_{k_m} tends to x^* in $\mathcal{C}$ and s_{k_m} tends to s^* in $[-1,1]$. Since η is continuous, it follows that $\eta(x^*, s^*) = 0$ and we will have a contradiction if $s^* \in \mathtt{S}(f(x^*))$. If $f(x^*)$ is not zero, then by continuity of f, s^* is equal to the sign of $f(x^*)$, and otherwise, $s^* \in [-1,1] = \mathtt{S}(f(x^*))$. Thus, $s^* \in \mathtt{S}(f(x^*))$ in all cases. ■

Lemma A.1.2.

Let η be a function defined on $\mathbb{R}^n$ homogeneous with degree d and weight vector $r = (r_1, ..., r_n)$, and V a positive definite proper function defined on $\mathbb{R}^n$ homogeneous of degree d_V with same weight vector r. Define $\mathcal{C} = V^{-1}(\{1\})$. If there exists α such that for all x in $\mathcal{C}$

$$\eta(x) < \alpha \ ,$$

then for all x in $\mathbb{R}^n \setminus \{0\}$,

$$\eta(x) < \alpha V(x)^{\frac{d}{d_V}} \ .$$

Proof : Let x in $\mathbb{R}^n \setminus \{0\}$. We have $\bar{x} = \dfrac{x_i}{V(x)^{\frac{r_i}{d_V}}}$ in $\mathcal{C}$. Thus $\eta(\bar{x}) < \alpha$ and by homogeneity

$$\frac{1}{V(x)^{\frac{d}{d_V}}}\eta(x) < \alpha$$

which gives the required inequality. ■

Lemma A.1.3.

Let η be a homogeneous function of degree d and weight vector r defined on $\mathbb{R}^n$ by

$$\eta(x) = \max_{s \in \mathtt{S}(f(x))} \tilde{\eta}(x, s)$$

where $\tilde{\eta}$ is a continuous function defined on $\mathbb{R}^{n+1}$ and f a continuous function defined on $\mathbb{R}^n$. Consider a continuous function γ homogeneous with same degree and weight vector such that, for all x in $\mathbb{R}^n \setminus \{0\}$ and s in $\mathtt{S}(f(x))$

$$\begin{array}{l} \gamma(x) \geq 0 \ , \\ \gamma(x) = 0 \quad \Rightarrow \quad \tilde{\eta}(x, s) < 0 \ . \end{array}$$

Then, there exists $k_0 > 0$ such that, for all x in $\mathbb{R}^n \setminus \{0\}$,

$$\eta(x) - k_0 \ \gamma(x) < 0 \ .$$

Proof : Define the homogeneous definite positive function $V(x) = \sum_{i=1}^{n} |x_i|^{\frac{d}{r_i}}$ and consider the compact set $\mathcal{C} = V^{-1}(\{1\})$. Assume that for all $k > 0$, there exists x_k in $\mathcal{C}$ and s_k in $\mathtt{S}(f(x_k))$ such that

$$\tilde{\eta}(x_k, s_k) \geq k \ \gamma(x_k) \geq 0$$

$\tilde{\eta}$ is continuous, and thus bounded on the compact set $\mathcal{C} \times [-1, 1]$. Therefore, $\gamma(x_k)$ tends to 0 when k tends to infinity. Besides, there exists a subsequence (k_m) such that x_{k_m} tends to x^* in $\mathcal{C}$ and s_{k_m} tends to s^* in $[-1, 1]$. It follows that $\gamma(x^*) = 0$ since γ is continuous. But with the same argument as in the proof of Lemma A.1.1, we have $s^* \in \mathtt{S}(f(x^*))$. It yields that $\tilde{\eta}(x^*, s^*) < 0$ by assumption and we have a contradiction.

Therefore, there exists k_0 such that

$$\tilde{\eta}(x, s) - k_0 \ \gamma(x) < 0$$

for all x in $\mathcal{C}$ and all s in $\mathtt{S}(f(x))$. Thus, with Lemma A.1.1 there exists $\alpha > 0$ such that

$$\tilde{\eta}(x, s) - k_0 \ \gamma(x) \leq -\alpha$$

so that

$$\eta(x) - k_0 \ \gamma(x) < 0$$

for any x in $\mathcal{C}$. The result follows applying Lemma A.1.2. ■

Lemma A.1.4.

Consider a positive bounded continuous function $t \mapsto c(t)$ and an absolutely continuous function $t \mapsto \nu(t)$ both defined on $[0, \bar{t})$ and such that

$$\text{for almost all } t \text{ in } [0, \bar{t}) \text{ such that } \nu(t) \geq c(t) \text{ then } \dot{\nu}(t) \leq -\nu(t)^d$$

with d in $]0, 1[$. Then, for all t in $[0, \bar{t})$,

$$\nu(t) \leq \max\left\{0, \max\{\nu(0) - c(0), 0\}^{1-d} - t\right\}^{1/(1-d)} + \sup_{s \in [0,t]} c(s) \ .$$

Proof : Let t be in $[0, \bar{t})$ and $c_t = \sup_{s \in [0,t]} c(s)$. For almost all $s \leq t$ such that $\nu(s) \geq \nu_t$, $\dot{\nu}(s) \leq -\nu(s)^d$, and thus

$$\begin{array}{rl} \dot{\overline{\max\{\nu(s) - c_t, 0\}}} & \leq -\nu(s)^d \\ & \leq -\max\{\nu(s) - c_t, 0\}^d \ . \end{array}$$

This inequality is also true when $\nu(s) < c_t$, therefore it is true for almost all $s \leq t$. It follows that for all $s \leq t$

$$\begin{array}{rl} \max\{\nu(s) - c_t, 0\}^{1-d} & \leq \max\{\nu(0) - c_t, 0\}^{1-d} - s \\ & \leq \max\{\nu(0) - c(0), 0\}^{1-d} - s \ , \end{array}$$

i-e

$$\max\{\nu(s) - c_t, 0\} \leq \max\left\{0, \left\{\max\{\nu(0) - c(0), 0\}^{1-d} - s\right\}\right\}^{\frac{1}{1-d}}$$

and finally, for all $s \leq t$

$$\nu(s) \leq \max\left\{0, \left\{\max\{\nu(0) - c(0), 0\}^{1-d} - s\right\}\right\}^{\frac{1}{1-d}} + c_t.$$

Taking this inequality at $s = t$ gives the required result. ■

Lemma A.1.5.

For any (x_a, x_b) in $\mathbb{R}^2$, for any $p \geq 1$, we have

- $\left|\lfloor x_a \rceil^{\frac{1}{p}} - \lfloor x_b \rceil^{\frac{1}{p}}\right| \leq 2^{1-\frac{1}{p}} |x_a - x_b|^{\frac{1}{p}}$
- $(|x_a| + |x_b|)^{\frac{1}{p}} \leq |x_a|^{\frac{1}{p}} + |x_b|^{\frac{1}{p}}$.

Proof : The second inequality is just the definition of the concavity of $x \mapsto x^{\frac{1}{p}}$ on $\mathbb{R}^+$. As for the first one, it is enough to prove it for $|x_a| \geq |x_b|$ (otherwise exchange them) and x_a non negative (otherwise take $(-x_a, -x_b)$). Besides, since it clearly holds for $x_b = 0$, we only have to prove (for $x = \frac{x_a}{|x_b|}$),

$$x^{\frac{1}{p}} \pm 1 \leq 2^{1-\frac{1}{p}}(x \pm 1)^{\frac{1}{p}} \quad \forall x \geq 1 \ .$$

First, by concavity of $x \mapsto x^{\frac{1}{p}}$, $\frac{1}{2}x^{\frac{1}{p}} + \frac{1}{2}1^{\frac{1}{p}} \leq \left(\frac{x+1}{2}\right)^{\frac{1}{p}}$ which gives the required inequality for the case "+". Besides, still by concavity of $x \mapsto x^{\frac{1}{p}}$, we have for $x \geq 1$, $\frac{x-1}{x}x^{\frac{1}{p}} + \frac{1}{x}0^{\frac{1}{p}} \leq \left(\frac{x-1}{x}x + \frac{1}{x}0\right)^{\frac{1}{p}}$ and $\frac{1}{x}x^{\frac{1}{p}} + \frac{x-1}{x}0^{\frac{1}{p}} \leq \left(\frac{1}{x}x + \frac{x-1}{x}0\right)^{\frac{1}{p}}$. Adding those two inequalities gives the case "−". ■

A.2 About continuity

Lemma A.2.1.

Let $\psi : \mathbb{R}^n \to \mathbb{R}^q$ be a continuous function on a compact subset $\mathcal{C}$ of $\mathbb{R}^n$. There exists a concave class $\mathcal{K}$ function ρ such that for all (x_a, x_b) in $\mathcal{C}^2$

$$|\psi(x_a) - \psi(x_b)| \leq \rho(|x_a - x_b|) \ .$$

Proof : Define the function

$$\rho_0(s) = \max_{x \in \mathcal{C}, |e| \leq s} |\psi(x + e) - \psi(x)|$$

which is increasing and such that $\rho_0(0) = 0$. Let us show that it is continuous at 0. Let (s_n) a sequence converging to 0. For all n, there exists x_n in $\mathcal{C}$ and e_n such that $|e_n| \leq s_n$ and $\rho_0(s_n) = |\psi(x_n + e_n) - \psi(x_n)|$. Since $\mathcal{C}$ is compact, there exist x^* in $\mathcal{C}$, e^* and subsequences of (x_n) and (e_n) converging to x^* and e^* respectively. But e^* is necessarily 0 and by continuity of ψ, $\rho_0(s_n)$ tends to 0. Now, the function, defined by the Riemann integral

$$\rho_1(s) = \begin{cases} \frac{1}{s}\displaystyle\int_s^{2s} \rho_0(s)ds + s & , \ s > 0 \\ 0 & , \ s = 0 \end{cases}$$

is continuous, strictly increasing and such that $\rho_0(s) \leq \rho_1(s)$. Besides, taking $\bar{s} = \max_{(x_a, x_b) \in \mathcal{C}^2} |x_a - x_b|$, there exists a concave class $\mathcal{K}$ function ρ such that for all s in $[0, \bar{s}]$, $\rho_1(s) \leq \rho(s)$ (see [McS34] for instance). Finally, we have :

$$|\psi(x_a) - \psi(x_b)| \ \leq \ \rho(|x_a - x_b|) \qquad \forall (x_a, x_b) \in \mathcal{C}^2 \ .$$

■

Lemma A.2.2.

Consider a function $\psi : \mathbb{R}^n \to \mathbb{R}$. Assume that there exist a compact set $\mathcal{C}$ of $\mathbb{R}^n$ and a function ρ of class $\mathcal{K}$ such that for all (x_a, x_b) in $\mathcal{C}^2$

$$|\psi(x_a) - \psi(x_b)| \leq \rho(|x_a - x_b|) \ .$$

Define the function $\hat{\psi} : \mathbb{R}^n \to [-\overline{\psi}, \overline{\psi}]$ by[1]

$$\hat{\psi}(z) = \mathsf{sat}_{\overline{\psi}}(\psi(z))$$

with $\overline{\psi} = \max_{z \in \mathcal{C}} \psi(z)$. Then, for any compact subset $\tilde{\mathcal{C}}$ strictly contained[2] in $\mathcal{C}$, there exists a positive real number c such that for all (x_a, x_b) in $\mathbb{R}^n \times \tilde{\mathcal{C}}$,

$$|\hat{\psi}(x_a) - \hat{\psi}(x_b)| \leq c\rho(|x_a - x_b|) \tag{A.1}$$

Proof : Since $\mathcal{C}$ strictly contains $\tilde{\mathcal{C}}$, we have :

$$\delta = \inf_{(x_a, x_b) \in (\mathbb{R}^n \setminus \mathcal{C}) \times \tilde{\mathcal{C}}} |x_a - x_b| > 0 \ .$$

First, for x_b in $\tilde{\mathcal{C}}$, $\hat{\psi}(x_b) = \psi(x_b)$. Now, if x_a is in $\mathcal{C}$, then we have $\hat{\psi}(x_a) = \psi(x_a)$ and consequently (A.1) holds for $c \geq 1$. If $x_a \notin \mathcal{C}$, we have, for all x_b in $\tilde{\mathcal{C}}$,

$$\begin{array}{rcl} |x_a - x_b| & \geq & \delta \ , \\ |\hat{\psi}(x_a) - \hat{\psi}(x_b)| & \leq & 2\overline{\psi} \leq 2\overline{\psi}\frac{\rho(|x_a - x_b|)}{\rho(\delta)} \ , \end{array}$$

and (A.1) holds for $c \geq \frac{2\overline{\psi}}{\rho(\delta)}$. ■

A.3 About injectivity

In this appendix, we consider two continuous functions $\Psi : \mathbb{R}^n \to \mathbb{R}^r$ and $\gamma : \mathbb{R}^n \to \mathbb{R}^q$ and a subset $\mathcal{S}$ of $\mathbb{R}^n$ such that

$$\Psi(x_a) = \Psi(x_b) \quad \forall (x_a, x_b) \in \mathcal{S}^2 \ : \ \gamma(x_a) = \gamma(x_b) \ . \tag{A.2}$$

In the particular case where Ψ is the identity function, (A.2) characterizes the injectivity of γ.

Lemma A.3.1.

There exists a function ψ defined on $\gamma(\mathcal{S})$ such that

$$\Psi(x) = \psi(\gamma(x)) \quad \forall x \in \mathcal{S} \ . \tag{A.3}$$

Proof : Define the map ψ on $\gamma(\mathcal{S})$ as

$$\psi(z) = \bigcup_{\substack{x \in \mathcal{S} \\ \gamma(x) = z}} \{\Psi(x)\} \ .$$

For any z in $\gamma(\mathcal{S})$, the set $\psi(z)$ is non-empty and single-valued because according to (A.2), if $z = \gamma(x_a) = \gamma(x_b)$, then $\Psi(x_a) = \Psi(x_b)$. Therefore, we can consider ψ as a function defined on $\gamma(\mathcal{S})$ and it verifies (A.3). ■

[1] The saturation function $\mathsf{sat}_M(\cdot)$ is defined by $\mathsf{sat}_M(x) = \max\{\min\{x, M\}, -M\}$

[2] By strictly contained, we mean that $\tilde{C} \subset \mathcal{C}$ and the distance between $\tilde{\mathcal{C}}$ and the complement of $\mathcal{C}$, namely $\mathbb{R}^n \setminus \mathcal{C}$, is strictly positive.

Lemma A.3.2.

Consider any compact subset $\mathcal{C}$ of $\mathcal{S}$. There exists a concave class $\mathcal{K}$ function ρ such that for all (x_a, x_b) in $\mathcal{C}^2$

$$|\Psi(x_a) - \Psi(x_b)| \leq \rho\left(|\gamma(x_a) - \gamma(x_b)|\right) \ . \tag{A.4}$$

Proof : We denote $D(x_a, x_b) = |\gamma(x_a) - \gamma(x_b)|$. Let

$$\rho_0(s) = \max_{\substack{(x_a, x_b) \in \mathcal{C}^2 \\ D(x_a, x_b) \leq s}} |\Psi(x_a) - \Psi(x_b)|$$

This defines properly a non decreasing function with non negative values which satisfies :

$$|\Psi(x_a) - \Psi(x_b)| \ \leq \ \rho_0(D(x_a, x_b)) \qquad \forall (x_a, x_b) \in \mathcal{C}^2 \ .$$

Also $\rho_0(0) = 0$. Indeed if not there would exist (x_a, x_b) in $\mathcal{C}^2$ satisfying :

$$D(x_a, x_b) \ = \ 0 \quad , \qquad |\Psi(x_a) - \Psi(x_b)| \ > \ 0 \ .$$

But this contradicts Equation (A.2).
Moreover, it can be shown that this function is also continuous at $s = 0$. Indeed, let $(s_k)_{k\in\mathbb{N}}$ be a sequence converging to 0. For each k, there exist $(x_{a,k}, x_{b,k})$ in $\mathcal{C}^2$ which satisfies $D(x_{a,k}, x_{b,k}) \leq s_k$ and $\rho_0(s_k) = |\Psi(x_{a,k}) - \Psi(x_{b,k})|$. The sequence $(x_{a,k}, x_{b,k})_{k\in\mathbb{N}}$ being in a compact set, it admits an accumulation point (x_a^*, x_b^*) which, because of the continuity of D must satisfy $D(x_a^*, x_b^*) = 0$ and therefore with (A.2) also $\Psi(x_a^*) - \Psi(x_b^*) \ = \ 0$. It follows that $\rho_0(s_k)$ tends to 0 and ρ_0 is continuous at 0. Proceeding with the same regularization of ρ_0 as in the proof of Lemma A.2.1, the conclusion follows. ■

Lemma A.3.3.

Consider any compact subset $\mathcal{C}$ of $\mathcal{S}$. There exists a uniformly continuous function ψ defined on $\mathbb{R}^q$ such that

$$\Psi(x) = \psi(\gamma(x)) \quad \forall x \in \mathcal{C} \ .$$

Proof : Consider ψ and ρ given by Lemmas A.3.1 and A.3.2 respectively. For any (z_a, z_b) in $\gamma(\mathcal{C})^2$, there exists (x_a, x_b) in $\mathcal{C}^2$ such that $z_a = \gamma(x_a)$ and $z_b = \gamma(x_b)$. Applying (A.4) to (x_a, x_b) and using (A.3), we have

$$|\psi(z_a) - \psi(z_b)| \ \leq \ \rho(|z_a - z_b|) \ .$$

ρ being concave, we deduce from [McS34, Theorem 2] (applied to each of the r real-valued components of ψ) that ψ admits a uniformly continuous extension defined on $\mathbb{R}^q$. Note that the extension of each component preserves the modulus of continuity ρ, so that the global extension has a modulus of continuity equal to $c\rho$ for some $c > 0$ depending only on the choice of the norm on $\mathbb{R}^r$. ■

When $q \leq n$ and γ is full-rank on $\mathcal{C}$, the function ψ is even C^1:

Lemma A.3.4.

Assume that $q \leq n$ and $\dfrac{\partial \gamma}{\partial x}$ is full-rank on $\mathcal{S}$, namely γ is a submersion on $\mathcal{S}$. Then, $\gamma(\mathcal{S})$ is open and there exists a C^1 function ψ defined on $\gamma(\mathcal{S})$ such that

$$\Psi(x) = \psi(\gamma(x)) \quad \forall x \in \mathcal{S} \ .$$

Proof : γ is an open map according to [Lee13, Proposition 4.28], thus $\gamma(\mathcal{S})$ is open. Consider the function ψ given by Lemma A.3.1 and take any z^* in $\gamma(\mathcal{S})$. There exists x^* in $\mathcal{S}$ such that $z^* = \gamma(x^*)$. γ being full-rank at x^*, according to the constant rank theorem, there exists an open neighborhood $\mathcal{V}$ of x^* and C^1 diffeomorphisms $\psi_1 : \mathbb{R}^n \to \mathcal{V}$ and $\psi_2 : \mathbb{R}^q \to \gamma(\mathcal{V})$ such that for all $\tilde{x}$ in $\mathbb{R}^n$:

$$\gamma(\psi_1(\tilde{x})) = \psi_2(\tilde{x}_1, \ldots, \tilde{x}_q) \ .$$

It follows that for all z in $\gamma(\mathcal{V})$

$$\gamma(\psi_1(\psi_2^{-1}(z), 0)) = z$$

namely γ admits a C^1 right-inverse γ^{ri} defined on $\gamma(\mathcal{V})$ which is an open neighborhood of z^*. Therefore, $\psi = \Psi \circ \gamma^{ri}$ and ψ is C^1 at z^*. ■

A direct consequence from those results is that any continuous function $\gamma : \mathbb{R}^n \to \mathbb{R}^q$ injective on a compact set $\mathcal{C}$ admits a uniformly continuous left-inverse ψ defined on $\mathbb{R}^q$ (take $\Psi = \mathsf{Id}$). The previous lemma does not apply because γ cannot be a submersion. However, we will show now that when γ is full-rank (i-e an immersion), this left-inverse can be taken Lipschitz on $\mathbb{R}^q$.

Due to needs in Chapters 5 or 7, we generalize those results to the case where the function γ depends on another parameter w evolving in a compact set:

Lemma A.3.5.

Let $\gamma : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^q$ be a continuous function and compact sets $\mathcal{C}_x$ and $\mathcal{C}_w$ of $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively such that for all w in $\mathcal{C}_w$, $x \mapsto \gamma(x, w)$ is injective on $\mathcal{C}_x$.

Then, there exist a concave class $\mathcal{K}$ function ρ, such that for all (x_a, x_b) in $\mathcal{C}_x^2$ and all w in $\mathcal{C}_w$,

$$|x_a - x_b| \leq \rho(|\gamma(x_a, w) - \gamma(x_b, w)|) \ ,$$

and a function ψ defined on $\mathbb{R}^q \times \mathbb{R}^p$ and a strictly positive number c such that

$$x = \psi(\gamma(x, w), w) \quad \forall (x, w) \in \mathcal{C}_x \times \mathcal{C}_w$$

and

$$|\psi(z_a, w) - \psi(z_b, w)| \leq c\rho(|z_a - z_b|)$$

i-e $z \mapsto \psi(z, w)$ is uniformly continuous on $\mathbb{R}^q$, uniformly in w.

If besides for all w in $\mathcal{C}_w$, $x \mapsto \gamma(x, w)$ is an immersion on $\mathcal{C}_x$, i-e for all w in $\mathcal{C}_w$, and all x in $\mathcal{C}_x$, $\dfrac{\partial \gamma}{\partial x}(x, w)$ is full-rank, then ρ is linear and $z \mapsto \psi(z, w)$ is Lipschitz on $\mathbb{R}^q$, uniformly in w.

Proof : The proof of the existence of ρ follows exactly that of Lemma A.3.2, but adding in the max defining ρ_0, $w \in \mathcal{C}_w$. Since it is a compact set, ρ is well defined and the same ρ can then be used for any w in $\mathcal{C}_w$. Applying Lemma A.3.3 to every $x \mapsto \gamma(x, w)$ gives the result since it is shown there that the extensions admit all the same modulus of continuity $c\rho$ for some $c > 0$ depending only on the norm chosen on $\mathbb{R}^r$.

Now suppose that $x \mapsto \gamma(x, w)$ is full-rank for all w in $\mathcal{C}_w$. Let Δ be the function defined on $\mathcal{C}_x \times \mathcal{C}_x \times \mathcal{C}_w$ by

$$\Delta(x_a, x_b, w) = \gamma(x_a, w) - \gamma(x_b, w) - \frac{\partial \gamma}{\partial x}(x_b, w)(x_a - x_b) \ .$$

Since $\frac{\partial \gamma}{\partial x}(x, w)$ is full-rank by assumption, the function

$$P(x, w) = \left(\frac{\partial \gamma}{\partial x}(x, w)^\top \frac{\partial \gamma}{\partial x}(x, w)\right)^{-1} \frac{\partial \gamma}{\partial x}(x, w)^\top$$

is well-defined and continuous on $\mathcal{C}_x \times \mathcal{C}_w$, and for any (x_a, x_b, w) in $\mathcal{C}_x \times \mathcal{C}_x \times \mathcal{C}_w$, we have

$$|x_a - x_b| \leq P_m(|\gamma(x_a, w) - \gamma(x_b, w| + |\Delta(x_a, x_b, w)|)$$

with $P_m = \max_{\mathcal{C}_x \times \mathcal{C}_w} |P(x, w)|$. Besides, the function $\frac{|\Delta(x_a, x_b, w)|}{|x_a - x_b|^2}$ is defined and continuous on $\mathcal{C}_x \times \mathcal{C}_x \times \mathcal{C}_w$, thus there exists $L_\Delta > 0$ such that

$$|\Delta(x_a, x_b, w)| \leq L_\Delta |x_a - x_b|^2 \leq \frac{1}{2P_m}|x_a - x_b|$$

for any (x_a, x_b) in $\mathcal{C}_x^2$ such that $|x_a - x_b| \leq 2r$ with $r = \frac{1}{4P_m L_\Delta}$, and for any w in $\mathcal{C}_w$. Now, define the set

$$\Omega = \{(x_a, x_b) \in \mathcal{C}_x^2 \,|\, |x_a - x_b| \geq 2r\}$$

which is a closed subset of the compact set $\mathcal{C}_x^2$ and therefore compact. The function $(x_a, x_b, w) \mapsto \frac{|x_a - x_b|}{|\gamma(x_a, w) - \gamma(x_b, w)|}$ is defined and continuous on $\Omega \times \mathcal{C}_w$ since $\gamma(\cdot, w)$ is injective for any w in $\mathcal{C}_w$. Thus, it admits a maximum M on the compact set $\Omega \times \mathcal{C}_w$.

Finally, take any (x_a, x_b) in $\mathcal{C}_x^2$ and any w in $\mathcal{C}_w$. There are two cases :

-either $(x_a, x_b) \notin \Omega$, i-e $|x_a - x_b| < 2r$, and

$$|x_a - x_b| \leq \frac{P_m}{2} |\gamma(x_a, w) - \gamma(x_b, w)| \ .$$

-or $(x_a, x_b) \in \Omega$, and

$$|x_a - x_b| \leq M |\gamma(x_a, w) - \gamma(x_b, w)| \ .$$

We conclude that ρ can be chosen linear with rate $L = \max\{\frac{P_m}{2}, M\}$. ■

Appendix B

Proofs of Chapter 10

B.1 Proof of Lemma 10.1.1

The compact K_0 being globally asymptotically attractive and interior to E which is forward invariant, E is globally attractive. It is also stable due to the continuity of solutions with respect to initial conditions uniformly on compact time subsets of the domain of definition. So it is globally asymptotically stable. It follows from [Wil69, Theorem 3.2] that there exist C^∞ functions $V_K : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ and $V_E : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ which are proper on $\mathbb{R}^m$ and a class $\mathcal{K}_\infty$ function α satisfying

$$\begin{gathered}
\alpha(d(z,K_0)) \leq V_K(z)\ , \quad \alpha(d(z,E)) \leq V_E(z) \qquad \forall\, z \in \mathbb{R}^m\ , \\
V_K(z) = 0 \quad \forall z \in K_0 \quad , \quad V_E(z) = 0 \quad \forall\, z \in E\ , \\
\frac{\partial V_K}{\partial z}(z)\,\chi(z) \leq -V_K(z)\ , \quad \frac{\partial V_E}{\partial z}(z)\,\chi(z) \leq -V_E(z) \qquad \forall\, z \in \mathbb{R}^m\ .
\end{gathered}$$

With $\overline{d}$ an arbitrary strictly positive real number, the notations

$$v_E \ = \sup_{z \in \mathbb{R}^m :\, d(z,E) \leq \overline{d}} V_K(z) \quad , \qquad \mu = \frac{\alpha(\overline{d})}{2 v_E}\ ,$$

and since α is of class $\mathcal{K}_\infty$, we obtain the implications

$$V_E(z) + \mu V_K(z) = \alpha(\overline{d}) \quad \Rightarrow \quad \alpha(d(z,E)) \leq V_E(z) \leq \alpha(\overline{d})$$
$$\Rightarrow \quad d(z,E) \leq \overline{d} \quad \Rightarrow \quad V_K(z) \leq v_E\ .$$

With our definition of μ, this yields also

$$\alpha(\overline{d}) - \mu\, V_K(z) = V_E(z) \quad \Rightarrow \quad 0 \ < \ \frac{\alpha(\overline{d})}{2} \leq V_E(z) \quad \Rightarrow \quad 0 < d(z,E) \leq \overline{d}\ .$$

On the other hand, with the compact notation $\mathcal{V}(z) = V_E(z) + \mu V_K(z)$, we have $\frac{\partial \mathcal{V}}{\partial z}(z)\,\chi(z) \leq -\mathcal{V}(z)$, for all $z \in \mathbb{R}^m$. All this implies that the sublevel set $\mathcal{E} = \{z \in \mathbb{R}^m : \mathcal{V}(z) < \alpha(\overline{d})\}$ is contained in $\{z \in \mathbb{R}^m :\, d(z,E) \in [0, \overline{d}]\}$ and that $\mathtt{cl}(E)$ is contained in $\mathcal{E}$. Besides, $\mathcal{E}$ verifies property $\mathfrak{C}$ with the vector field χ and the function $\kappa = \mathcal{V} - \alpha(\overline{d})$.

B.2 Proof of Lemma 10.2.1

We use the following notations:
The complementary, closure and boundary of a set S are denoted S^c, $\mathtt{cl}(S)$ and ∂S, respectively. The Hausdorff distance d_H between two sets A and B is defined by :

$$d_H(A,B) \;=\; \max\left\{ \sup_{z_A\in A}\inf_{z_B\in B}|z_A-z_B|\,,\, \sup_{z\in A}\inf_{z_B\in B}|z_A-z_B| \right\}\;.$$

$Z(z,t)$ denotes the (unique) solution, at time t, to $\dot z = \chi(z)$ going trough z at time 0 and $\Sigma_\varepsilon \;=\; \bigcup_{t\in[0,\varepsilon]} Z(\partial E,t)$.

Lemma B.2.1 *Let E be an open strict subset of $\mathbb{R}^m$ verifying $\mathfrak{C}$, with a C^s vector field χ and a C^s mapping κ. There exists a strictly positive (maybe infinite) real number ε_∞ such that, for any ε in $[0,\varepsilon_\infty[$, there exists a C^s-diffeomorphism ϕ: $\mathbb{R}^m \to E$, such that*

$$\phi(z)=z \quad \forall z\in E_\varepsilon = E\cap(\Sigma_\varepsilon)^{\mathtt{c}} \quad , \qquad d_H(\partial E_\varepsilon,\partial E) \;\le\; \varepsilon \sup_z|\chi(z)|\;.$$

Proof : According to Condition $\mathfrak{C}$, χ is bounded and K_0 is a compact subset of the open set E. It follows that there exists a strictly positive (maybe infinite) real number ε_∞ such that

$$Z(z,t)\notin K_0 \qquad \forall (z,t)\in \partial E\times[0,2\varepsilon_\infty[\;.$$

In the following ε is a real number in $[0,\varepsilon_\infty[$.

We introduce the notations

$$\Sigma_{2\varepsilon} = \bigcup_{t\in[0,2\varepsilon]} Z(\partial E,t) \quad , \qquad E_{2\varepsilon} = E\cap(\Sigma_{2\varepsilon})^{\mathtt{c}}$$

and establish some properties.

– *E is forward invariant for χ.* This is a direct consequence of points $\mathfrak{C}$.1 and $\mathfrak{C}$.3.

– *$\Sigma_{2\varepsilon}$ is closed.* Take a sequence (z_k) of points in $\Sigma_{2\varepsilon}$ converging to z^*. By definition of $\Sigma_{2\varepsilon}$, there exists a sequence (t_k), such that :

$$t_k\in[0,2\varepsilon] \qquad \text{and} \qquad Z(z_k,-t_k)\in\partial E \qquad \forall k\in\mathbb{N}\;.$$

Since $[0,2\varepsilon]$ is compact, one can extract a subsequence $(t_{\sigma(k)})$ converging to t^* in $[0,2\varepsilon]$, and by continuity of the function $(z,t)\mapsto Z(z,-t)$, $(Z(z_{\sigma(k)},t_{\sigma(k)}))$ tends to $Z(z^*,-t^*)$ which is in ∂E, since ∂E is closed. Finally, because t^* is in $[0,2\varepsilon]$, z^* is in $\Sigma_{2\varepsilon}$ by definition.

– *$\Sigma_{2\varepsilon}$ is contained in $\mathtt{cl}(E)$.* Since, E is forward invariant for χ, and so is $\mathtt{cl}(E)$ (see [Hah67, Theorem 16.3]). This implies

$$\partial E \;\subset\; \Sigma_{2\varepsilon} = \bigcup_{t\in[0,2\varepsilon]} Z(\partial E,t) \;\subset\; \mathtt{cl}(E) \;=\; E\cup\partial E\;.$$

At this point, it is useful to note that, because $\Sigma_{2\varepsilon}$ is a closed subset of $\mathtt{cl}(E)$ and E is open, we have $\Sigma_{2\varepsilon}\cap E \;=\; \Sigma_{2\varepsilon}\backslash\partial E$. This implies :

$$E\backslash E_{2\varepsilon} = (E_{2\varepsilon})^{\mathtt{c}}\cap E = (E^{\mathtt{c}}\cup\Sigma_{2\varepsilon})\cap E = \Sigma_{2\varepsilon}\cap E = \Sigma_{2\varepsilon}\backslash\partial E, \tag{B.1}$$

and $E = E_{2\varepsilon}\underset{\neq}{\cup}(\Sigma_{2\varepsilon}\backslash\partial E)$.

With all these properties at hand, we define now two functions $\mathfrak{t}$ and θ. The assumptions of global attractiveness of the closed set K_0 contained in E open, of transversality of χ to ∂E, and the property of forward-invariance of E, imply that, for all z in $E^{\mathtt{c}}$, there exists a unique non negative real number $\mathfrak{t}(z)$ satisfying:

$$\kappa\left(Z(z,\mathfrak{t}(z))\right)=0 \quad\Longleftrightarrow\quad Z(z,\mathfrak{t}(z))\in\partial E.$$

The same arguments in reverse time allow us to see that, for all z in $\Sigma_{2\varepsilon}$, $\mathfrak{t}(z)$ exists, is unique and in $[-2\varepsilon,0]$. This way, the function $z\to\mathfrak{t}(z)$ is defined on $(E_{2\varepsilon})^{\mathtt{c}}$. Next, for all z in $(E_{2\varepsilon})^{\mathtt{c}}$, we define :

$$\theta(z)=Z(z,\mathfrak{t}(z)).$$

Thanks to the transversality assumption, the Implicit Function Theorem implies the functions $z\mapsto\mathfrak{t}(z)$ and $z\mapsto\theta(z)$ are C^s on $(E_{2\varepsilon})^{\mathtt{c}}$.

Now we evaluate $\mathfrak{t}(z)$ for z in $\partial\Sigma_{2\varepsilon}$. Let z be arbitrary in $\partial\Sigma_{2\varepsilon}$ and therefore in $\Sigma_{2\varepsilon}$ which is closed. Assume its corresponding $\mathfrak{t}(z)$ is in $]-2\varepsilon$ÃŁ$,0[$. The Implicit Function Theorem shows that $z\mapsto\mathfrak{t}(z)$ and $z\mapsto\theta(z)$ are defined and continuous on a neighborhood of z. Therefore, there exists a strictly positive real number r satisfying

$$\forall y\in B_r(z)\;,\;\exists t_y\in]-2\varepsilon,0[:\; Z(y,t_y)\in\partial E\;.$$

This implies that the neighborhood $B_r(z)$ of z is contained in $\Sigma_{2\varepsilon}$, in contradiction with the fact that z is on the boundary of $\Sigma_{2\varepsilon}$. This shows that, for all z in $\partial\Sigma_{2\varepsilon}$, $\mathfrak{t}(z)$ is either 0 or -2ε. We write this as

$$(\partial\Sigma_{2\varepsilon})_i = \{z \in \Sigma_{2\varepsilon} \,:\, \mathfrak{t}(z) = -2\varepsilon\} \quad , \qquad \partial\Sigma_{2\varepsilon} = \partial E \cup (\partial\Sigma_{2\varepsilon})_i \ .$$

Now we want to prove $\partial E_{2\varepsilon} \subset (\partial\Sigma_{2\varepsilon})_i$. To obtain this result, we start by showing :

$$\partial E_{2\varepsilon} \cap \partial E = \emptyset \quad \text{and} \quad \partial E_{2\varepsilon} \subset \partial\Sigma_{2\varepsilon} \ . \tag{B.2}$$

Suppose the existence of z in $\partial E_{2\varepsilon} \cap \partial E$. z being in ∂E, its corresponding $\mathfrak{t}(z)$ is 0. By the Implicit Function Theorem, there exists a strictly positive real number r such that,

$$\forall y \in B_r(z) \ , \ \exists t_y \in \,]-\varepsilon, \varepsilon[\ : Z(y, t_y) \in \partial E \ .$$

But, by definition, any y, for which there exists t_y in $]-\varepsilon, 0]$, is in $\Sigma_{2\varepsilon}$. If instead t_y is strictly positive, then necessarily y is in E^{c}, because E is forward-invariant for χ and a solution starting in E cannot reach ∂E in positive finite time. We have obtained : $B_r(z) \subset \Sigma_{2\varepsilon} \cup E^{\mathsf{c}} = (E_{2\varepsilon})^{\mathsf{c}}$. $B_r(z)$ being a neighborhood of z, this contradicts the fact that z is in the boundary of $E_{2\varepsilon}$.

At this point, we have proved that $\partial E_{2\varepsilon} \cap \partial E = \emptyset$, and, because $E_{2\varepsilon}$ is contained in E, this implies $\partial E_{2\varepsilon} \subset E$. With this, (B.2) will be established by proving that we have $\partial E_{2\varepsilon} \subset \partial\Sigma_{2\varepsilon}$. Let z be arbitrary in $\partial E_{2\varepsilon}$ and therefore in E which is open. There exists a strictly positive real number r such that we have :

$$\emptyset \ \neq \ B_r(z) \cap E_{2\varepsilon} \ = \ B_r(z) \ \cap \ (E \cap (\Sigma_{2\varepsilon})^{\mathsf{c}}) \ , \quad \emptyset \ \neq \ B_r(z) \cap (E_{2\varepsilon})^{\mathsf{c}} \ = \ B_r(z) \ \cap \ (E^{\mathsf{c}} \cup \Sigma_{2\varepsilon}) \ , \quad B_r(z) \subset E \ .$$

This implies $B_r(z) \ \cap \ (\Sigma_{2\varepsilon})^{\mathsf{c}} \ \neq \ \emptyset$ and $B_r(z) \ \cap \ \Sigma_{2\varepsilon} \ \neq \ \emptyset$ and therefore that z is in $\partial\Sigma_{2\varepsilon}$.

We have established $\partial E_{2\varepsilon} \cap \partial E \ = \ \emptyset$, $\partial E_{2\varepsilon} \subset \partial\Sigma_{2\varepsilon}$ and $\partial\Sigma_{2\varepsilon} \ = \ \partial E \cup (\partial\Sigma_{2\varepsilon})_i$. This does imply :

$$\partial E_{2\varepsilon} \ \subset \ (\partial\Sigma_{2\varepsilon})_i \ = \ \{z \in E \,:\, \mathfrak{t}(z) = -2\varepsilon\} \ . \tag{B.3}$$

This allows us to extend by continuity the definition of $\mathfrak{t}$ to $\mathbb{R}^m$ by letting

$$\mathfrak{t}(z) \ = \ -2\varepsilon \qquad \forall z \in E_{2\varepsilon} \ .$$

All the properties we have established for $\Sigma_{2\varepsilon}$ and $E_{2\varepsilon}$ hold also for Σ_ε and E_ε. In particular, we have

$$\mathfrak{t}(z) \in [-2\varepsilon, -\varepsilon] \qquad \forall z \in E_\varepsilon \setminus E_{2\varepsilon} \ . \tag{B.4}$$

Thanks to all these preparatory steps, we are finally ready to define a function $\phi : \mathbb{R}^m \to E$. Let $\nu : \mathbb{R} \to \mathbb{R}$ be a function such that the function $t \mapsto \nu(t) - t$ is a C^s (decreasing) diffeomorphism from $\mathbb{R}$ onto $]0, +\infty[$ mapping $[-\varepsilon, +\infty[$ onto $]0, \varepsilon]$ and being "minus" identity between $]-\infty, -\varepsilon]$ and $[\varepsilon, +\infty[$, i.e.

$$\nu(t) - t \ = \ -t \quad \forall t \leq -\varepsilon \ .$$

We have

$$\nu(t) > t \qquad \forall t \in \mathbb{R} \quad , \qquad \nu(\mathfrak{t}(z)) \ = \ 0 \qquad \forall z \in E_\varepsilon \setminus E_{2\varepsilon} \ . \tag{B.5}$$

We let :

$$\phi(z) = \begin{cases} Z\left(z, \nu(\mathfrak{t}(z))\right), & \text{if } z \in (E_{2\varepsilon})^c \ , \\ z, & \text{if } z \in E_{2\varepsilon} \ . \end{cases}$$

The image of ϕ is contained in E since we have (B.5), $E_{2\varepsilon} \ \subset \ E$ and :

$$Z(z, \mathfrak{t}(z)) \ \in \ \partial E \quad , \qquad Z(z,t) \ \in \ E \qquad \forall (z,t) \in \partial E \times \mathbb{R}_{>0} \ .$$

Like the functions Z, ν, and $\mathfrak{t}$, the function ϕ is C^s on the interior of $(E_{2\varepsilon})^{\mathsf{c}}$. Also, since (B.5) implies

$$\phi(z) \ = \ z \qquad \forall z \in E_\varepsilon \setminus E_{2\varepsilon} \ , \tag{B.6}$$

ϕ is trivially C^s on E_ε and therefore on $(E_{2\varepsilon})^{\mathsf{c}} \cup E_\varepsilon = \mathbb{R}^m$.

We now show that ϕ is invertible. Because of (B.6), this is trivial on E_ε. Let y be arbitrary in $E \cap (E_{2\varepsilon})^{\mathsf{c}} = E \cap \Sigma_{2\varepsilon}$. To y corresponds $\mathfrak{t}(y)$ in the interval $[-2\varepsilon, 0[$. Thus, $-\mathfrak{t}(y)$ is in $]0, 2\varepsilon]$, image of $[-2\varepsilon, +\infty[$ by the C^s diffeomorphism $t \mapsto \nu(t) - t$. Hence there exists $\mathfrak{s}(y)$ in $[-2\varepsilon, +\infty[$ satisfying

$$\nu(\mathfrak{s}(y)) \ - \ \mathfrak{s}(y) \ = \ -\mathfrak{t}(y) \ . \tag{B.7}$$

Moreover, (B.4) implies that for y in $E_\varepsilon \setminus E_{2\varepsilon}$ subset of $E \cap (E_{2\varepsilon})^{\mathsf{c}}$, we have

$$\mathfrak{s}(y) \ = \ \mathfrak{t}(y)$$

So with letting

$$\mathfrak{s}(y) \;=\; \mathfrak{t}(y) \;=\; -2\varepsilon \qquad \forall y \in E_{2\varepsilon}$$

we have defined a function $\mathfrak{s} : E \to [-2\varepsilon, +\infty[$, which thanks to the implicit function theorem, is C^s and satisfies (B.7).

This allows us to define properly $\phi^{-1} : \mathbb{R}^m \to E$ as :

$$\phi^{-1}(y) = Z\left(y, -\nu(\mathfrak{s}(y))\right) \ .$$

By composition, this function is C^s and it is an inverse of ϕ in particular because, with (B.7), we have

$$\mathfrak{t}(Z(y, -\nu(\mathfrak{s}(y)))) = \mathfrak{t}(Z(y, \mathfrak{t}(y) - \mathfrak{s}(y))) = \mathfrak{s}(y) \qquad \forall y \in E \ .$$

This gives

$$\phi(Z(y, -\nu(\mathfrak{s}(y))) = Z(Z(y, -\nu(\mathfrak{s}(y))), \nu(\mathfrak{t}(Z(y, -\nu(\mathfrak{s}(y)))))) = Z(Z(y, -\nu(\mathfrak{s}(y))), \nu(\mathfrak{s}(y))) = y$$

All this implies ϕ is a C^s-diffeomorphism from $\mathbb{R}^m$ to E.

Finally, we note that, for any point z_ε in ∂E_ε, there exists a point z in ∂E satisfying :

$$|z_\varepsilon - z| \;=\; \left| \int_0^\varepsilon \chi(Z(z,s))ds \right| \;\leq\; \varepsilon \sup_\zeta |\chi(\zeta)| \ .$$

And conversely, for any z in ∂E, there exist z_ε in ∂E_ε satisfying :

$$|z_\varepsilon - z| \;=\; \left| \int_0^\varepsilon \chi(Z(z,s))ds \right| \;\leq\; \varepsilon \sup_\zeta |\chi(\zeta)| \ .$$

It follows that, with ε as small as needed,

$$d_H(\partial E_\varepsilon, \partial E) \;\leq\; \varepsilon \sup_\zeta |\chi(\zeta)| \qquad (B.8) \qquad \blacksquare$$

Lemma 10.2.1 is a direct consequence of Lemma B.2.1 if we pick ε_∞, maybe infinite, satisfying

$$Z(z,t) \notin K \qquad \forall (z,t) \in \partial E \times [0, 2\varepsilon_\infty[\ .$$

ε_∞ can be chosen strictly positive since $d(K, \partial E)$ is non zero and χ is bounded.

B.3 Proof of case b) of Theorem 10.1.1

To complete the proof of Theorem 10.1.1, we use another technical result.

Lemma B.3.1 (Diffeomorphism extension from a ball) *Consider a C^2 diffeomorphism $\lambda : B_R(0) \to \lambda(B_R(0)) \subset \mathbb{R}^m$, with R a strictly positive real number. For any real number ε in $]0,1[$, there exists a diffeomorphism $\lambda_e : \mathbb{R}^m \to \mathbb{R}^m$ satisfying*

$$\lambda_e(z) = \lambda(z) \qquad \forall z \in \mathtt{cl}(B_{\frac{R}{1+\varepsilon}}(0)) \ .$$

Proof : It sufficient to prove that [Hir76, Theorem 8.1.4] applies. We let

$$U = B_{\frac{R}{1+\frac{\varepsilon}{2}}}(0) \quad , \qquad A = \mathtt{cl}(B_{\frac{R}{1+\varepsilon}}(0)) \quad , \qquad I = \left]-\frac{\varepsilon}{2}, 1+\frac{\varepsilon}{2}\right[\ ,$$

and, without loss of generality we may assume that $\lambda(0) = 0$.

Then, consider the function $F : U \times I \to \mathbb{R}^m$ defined as

$$F(z,t) = \left(\frac{\partial \lambda}{\partial z}(0)\right)^{-1} \frac{\lambda(zt)}{t} \ , \ \forall t \in I \setminus \{0\} \ , \quad F(z,0) = z \ .$$

We start by showing that F is an isotopy of U.

- For any t in I, the function $z \mapsto F_t = F(z,t)$ is an embedding from U onto $F_t(U) \subset \mathbb{R}^m$. Indeed, for any pair (z_a, z_b) in U^2 satisfying $F(z_a,t) = F(z_b,t)$, we obtain $\lambda(z_a t) = \lambda(z_b t)$ where $(z_a t, z_b t)$ is in U^2. The function λ being injective on this set, we have $z_a = z_b$ which establishes F_t is injective. Moreover, we have:

  $$\frac{\partial F_t}{\partial z}(z) = \left(\frac{\partial \lambda}{\partial z}(0)\right)^{-1} \frac{\partial \lambda}{\partial z}(zt) \quad \forall t \in I \setminus \{0\} \quad , \qquad \frac{\partial F_0}{\partial z}(z) = \mathsf{Id}.$$

  Hence, F_t is full rank on U and therefore an embedding.

- For all z in U, the function $t \mapsto F(z,t)$ is C^1. This follows directly from the fact that, the function λ being C^2, and $\lambda(0) = 0$, we have

$$\frac{\lambda(zt)}{t} = \frac{\partial \lambda}{\partial z}(0)z + z' \left(\frac{\partial^2 \lambda}{\partial z \partial z}(0) \right) z\frac{t}{2} + \circ(t) \ .$$

  In particular, we obtain $\frac{\partial F}{\partial t}(z,t) = \left(\frac{\partial \lambda}{\partial z}(0)\right)^{-1} \rho(z,t)$ where

$$\rho(z,t) = \frac{1}{t^2}\left[\frac{\partial \lambda}{\partial z}(zt)zt - \lambda(zt)\right] \quad \forall t \in I \setminus \{0\} \ , \quad \rho(z,0) = \frac{1}{2}z' \left(\frac{\partial^2 \lambda}{\partial z \partial z}(0) \right) z \ .$$

  Moreover, for all t in I, the function $z \mapsto \frac{\partial F}{\partial t}(z,t)$ is locally Lipschitz and therefore gives rise to an ordinary differential equation with unique solutions.

  Also the set $\bigcup_{(z,t)\in U\times I}\{(F(z,t),t)\}$ is open. This follows from Brouwer's Invariance theorem since the function $(z,t) \mapsto (F(z,t),t)$ is a diffeomorphism on the open set $U \times I$. With [Hir76, Theorem 8.1.4], we know there exists a diffeotopy F_e from $\mathbb{R}^m \times I$ onto $\mathbb{R}^m$ which satisfies $F_e = F$ on $A \times [0,1]$. Thus, the diffeomorphism $\lambda_e = F_e(.,1)$ defined on $\mathbb{R}^m$ onto $\mathbb{R}^m$ verifies $\lambda_e(z) = F_e(z,1) = F(z,1) = \lambda(z)$ for all $z \in A$. ■

We now place ourselves in the case b) of Theorem 10.1.1, namely we suppose that τ_a^* is C^2 and $\mathcal{S}$ is C^2-diffeomorphic to $\mathbb{R}^m$. Let $\phi_1 : \mathcal{S} \to \mathbb{R}^m$ denote the corresponding diffeomorphism. Let R_1 be a strictly positive real number such that the open ball $B_{R_1}(0)$ contains $\phi_1(K)$. Let R_2 be a real number strictly larger than R_1. With Lemma 10.2.1 again, and since $B_{R_2}(0)$ verifies property $\mathfrak{C}$, there exists of C^2-diffeomorphism $\phi_2 : \mathbb{R}^m \to B_{R_2}(0)$ satisfying $\phi_2(z) = z$ for all z in $B_{R_1}(0)$. At this point, we have obtained a C^2-diffeomorphism $\phi = \phi_2 \circ \phi_1 : \mathcal{S} \to B_{R_2}(0)$. Consider $\lambda = \tau_a^* \circ \phi^{-1} : B_{R_2}(0) \to \tau_a^*(\mathcal{S}) \quad (= \lambda(B_{R_2}(0)))$. According to Lemma B.3.1, we can extend λ to $\lambda_e : \mathbb{R}^m \to \mathbb{R}^m$ such that $\lambda_e = \tau_a^* \circ \phi^{-1}$ on $B_{R_1}(0)$. Finally, consider $\tau_e^* = \lambda_e \circ \phi_1 : \mathcal{S} \to \mathbb{R}^m$. Since, by construction of ϕ_2, $\phi = \phi_1$ on $\phi_1^{-1}(B_{R_1}(0))$ which contains K, we have $\tau_e^* = \tau_a^*$ on K.

Appendix C

Proof of Theorem 13.1.1

In this appendix, we prove that Theorem 13.1.1, namely that System (13.2) is an observer for System (13.1). We have seen that to do so, it is enough to show that

Lemma C.0.2.

Consider a strictly positive real number Φ and a function $\omega : [0, +\infty) \to \mathbb{R}$ such that there exists $\underline{\omega}_0 > 0$, $\overline{\omega}_0 > 0$ and $\overline{\omega}_1 > 0$ such that for all t in $[0, +\infty)$

$$\underline{\omega}_0 \leq \omega(t) \leq \overline{\omega}_0 \quad , \quad \dot{\omega}(t) \leq \overline{\omega}_1 \ .$$

Then, $(\Phi, 0, \Phi)$ is an asymptotically stable equilibrum point of the dynamics

$$\left\{ \begin{array}{rcl} \dot{\hat{X}}_d & = & \omega \hat{X}_q - 2q\hat{X}_d \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2 \right) \\ \dot{\hat{X}}_q & = & -\omega \hat{X}_d + \omega\Phi - 2q\hat{X}_q \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2 \right) \\ \dot{\hat{\Phi}} & = & q\,\hat{\Phi} \left(\hat{X}_d^2 + \hat{X}_q^2 - \hat{\Phi}^2 \right) \end{array} \right. \tag{C.1}$$

with basin of attraction containing the forward invariant set $\Omega = \mathbb{R}^2 \times (0, +\infty)$.

C.1 Lyapunov function

Our first step for the convergence analysis is to look for a Lyapunov function. To facilitate this task, we do another change of coordinates aiming at getting the dynamics in a triangular form, the so-called feedback form. Our motivation is that for this specific form, we have the backstepping methodology allowing us in particular to build Lyapunov functions. This task is easily achieved after noticing that we have, $\hat{\Phi}$ being non zero when the solution is in Ω,

$$\begin{array}{rcl} \dot{\hat{X}}_d + 2q\hat{X}_d \dfrac{\dot{\hat{\Phi}}}{\hat{\Phi}} & = & \omega \hat{X}_q \\ \dot{\hat{X}}_q + 2q\hat{X}_q \dfrac{\dot{\hat{\Phi}}}{\hat{\Phi}} & = & -\omega \hat{X}_d + \omega\Phi \end{array}$$

and therefore

$$\begin{array}{rcl} \dot{\widehat{\hat{X}_d\hat{\Phi}^2}} & = & \omega \hat{X}_q \hat{\Phi}^2 \\ \dot{\widehat{\hat{X}_q\hat{\Phi}^2}} & = & -\omega \hat{X}_d \hat{\Phi}^2 + \omega \hat{\Phi}^2 \Phi \end{array}$$

This leads us to the second set of coordinates

$$\begin{aligned}\hat{x}_1 &= \hat{X}_d\hat{\Phi}^2\\ \hat{x}_2 &= \hat{X}_q\hat{\Phi}^2\\ \hat{x}_3 &= \hat{\Phi}^4\ .\end{aligned}$$

As desired, the dynamics take the following feedback form

$$\begin{aligned}\dot{\hat{x}}_1 &= \omega\hat{x}_2\\ \dot{\hat{x}}_2 &= -\omega\hat{x}_1+\omega\Phi\sqrt{\hat{x}_3}\\ \dot{\hat{x}}_3 &= -4q\left(\hat{x}_3^{\frac{3}{2}}-(\hat{x}_1^2+\hat{x}_2^2)\right)\end{aligned}$$

which we can write compactly as :

$$\begin{aligned}\dot{\hat{x}}_{12} &= f_{12}(\hat{x}_{12},\hat{x}_3)\\ \dot{\hat{x}}_3 &= -4q\left(\hat{x}_3^{\frac{3}{2}}-(\hat{x}_1^2+\hat{x}_2^2)\right)\end{aligned}$$

with $\hat{x}_{12}=(\hat{x}_1,\hat{x}_2)$. Now, a necessary condition to have a Lyapunov function V such that

$$\frac{\partial V}{\partial\hat{x}_{12}}(\hat{x}_{12},\hat{x}_3)f_{12}(\hat{x}_{12},\hat{x}_3)-4q\,\frac{\partial V}{\partial\hat{x}_3}(\hat{x}_{12},\hat{x}_3)\left(\hat{x}_3^{\frac{3}{2}}-(\hat{x}_1^2+\hat{x}_2^2)\right)\le 0$$

is to have (just pick $\hat{x}_3^{\frac{3}{2}}=\hat{x}_1^2+\hat{x}_2^2$)

$$\frac{\partial V}{\partial\hat{x}_{12}}(\hat{x}_{12},(\hat{x}_1^2+\hat{x}_2^2)^{\frac{2}{3}})f_{12}(\hat{x}_{12},(\hat{x}_1^2+\hat{x}_2^2)^{\frac{2}{3}}))\le 0\ .$$

This suggests to find first a Lyapunov function for the system

$$\begin{aligned}\dot{\hat{x}}_1 &= \omega\hat{x}_2\\ \dot{\hat{x}}_2 &= -\omega\hat{x}_1+\omega\Phi(\hat{x}_1^2+\hat{x}_2^2)^{\frac{1}{3}}\ .\end{aligned}$$

The latter system admits periodic orbits which are level sets of

$$V_1(\hat{x}_1,\hat{x}_2)=\frac{3}{4}(\hat{x}_1^2+\hat{x}_2^2)^{2/3}-\Phi\hat{x}_1+\frac{\Phi^4}{4}$$

which is positive, 0 only at $(\hat{x}_1,\hat{x}_2)=(\Phi^3,0)$ and proper in $(\hat{x}_1,\hat{x}_2)$.

Then, inspired by the backstepping methodology (see [PdNC91]), we look for a Lyapunov function in the form

$$V(\hat{x})=V_1(\hat{x}_1,\hat{x}_2)+V_2(\hat{x}_3,r)$$

with

$$V_2(\hat{x}_3,r)=\int_{r^{2/3}}^{\hat{x}_3}\varphi(s,r)ds\quad,\qquad r=\hat{x}_1^2+\hat{x}_2^2$$

where φ is a C^1 function satisfying

$$\varphi(\hat{x}_3,r)\left(\hat{x}_3^{\frac{3}{2}}-r\right)\ >\ 0\qquad\forall r\neq\hat{x}_3^{\frac{3}{2}}\ . \tag{C.2}$$

Along the solutions, we obtain

$$\begin{aligned}\dot{V} &= \frac{\partial V_1}{\partial\hat{x}_{12}}(\hat{x}_{12},\hat{x}_3)f_{12}(\hat{x}_{12},\hat{x}_3)-4q\frac{\partial V_2}{\partial\hat{x}_3}(\hat{x}_3,r)\left(\hat{x}_3^{\frac{3}{2}}-r\right)+\frac{\partial V_2}{\partial r}(\hat{x}_3,r)2\Phi\omega\hat{x}_2\sqrt{\hat{x}_3}\\ &= \Phi\omega\hat{x}_2\left(\frac{\sqrt{\hat{x}_3}}{r^{1/3}}-1\right)\ +\ \left[\int_{r^{2/3}}^{\hat{x}_3}\frac{\partial\varphi}{\partial r}(s,r)ds\right]2\Phi\omega\hat{x}_2\sqrt{\hat{x}_3}\ -\ 4q\varphi(\hat{x}_3,r)\left(\hat{x}_3^{\frac{3}{2}}-r\right)\end{aligned}$$

In view of (C.2), $\dot{V}$ is non positive if we select the function φ satisfying (C.2) and

$$\left[\int_{r^{2/3}}^{\hat{x}_3} \frac{\partial \varphi}{\partial r}(s,r)ds\right] 2\Phi\omega\hat{x}_2\sqrt{\hat{x}_3} \;=\; -\Phi\omega\hat{x}_2\left(\frac{\sqrt{\hat{x}_3}}{r^{1/3}}-1\right)$$

and thus

$$\left[\int_{r^{2/3}}^{\hat{x}_3} \frac{\partial \varphi}{\partial r}(s,r)ds\right] \;=\; \frac{1}{2}\left(\frac{1}{\sqrt{\hat{x}_3}}-\frac{1}{r^{1/3}}\right) \;.$$

It is necessary to have

$$\frac{\partial \varphi}{\partial r}(\hat{x}_3,r) \;=\; -\frac{1}{4}\frac{1}{\hat{x}_3^{3/2}}$$

and finally

$$\varphi(\hat{x}_3,r) \;=\; \frac{1}{4}\left[1-\frac{r}{\hat{x}_3^{3/2}}\right] \;.$$

This gives us

$$V_2(\hat{x}_3,r) \;=\; \frac{1}{4}\left[\hat{x}_3 - r^{2/3} + 2\left(\frac{r}{\hat{x}_3^{1/2}} - r^{2/3}\right)\right]$$

and

$$V = V_1 + V_2 = \frac{1}{4}\hat{x}_3 + \frac{1}{2}\frac{r}{\sqrt{\hat{x}_3}} - \Phi x_1 + \frac{\Phi^4}{4} \;.$$

Since for any r, the function $\varphi(\cdot,r)$ is strictly positive for $\hat{x}_3 > r^{2/3}$ and strictly negative for $\hat{x}_3 < r^{2/3}$, $V_2(\cdot,r)$ is positive and $V_2(\hat{x}_3,r)=0 \Leftrightarrow \hat{x}_3 = r^{2/3}$. We deduce that V is positive and

$$V=0 \Leftrightarrow V_1=0 \text{ and } V_2=0 \Leftrightarrow (\hat{x}_1,\hat{x}_2)=(\Phi^3,0) \text{ and } \hat{x}_3=r^{2/3} \Leftrightarrow (\hat{x}_1,\hat{x}_2,\hat{x}_3)=(\Phi^3,0,\Phi^4) \;.$$

In the original coordinates, the expression of V is

$$V(\hat{X}_d,\hat{X}_q,\hat{\Phi}) = \frac{\hat{\Phi}^4}{4} + \frac{1}{2}\hat{\Phi}^2(\hat{X}_d^2+\hat{X}_q^2) - \Phi\hat{\Phi}^2\hat{X}_d + \frac{\Phi^4}{4} \;.$$

For any $(\hat{X}_q,\hat{\Phi})$, $\hat{X}_d \mapsto (\hat{X}_d^2+\hat{X}_q^2-2\Phi\hat{X}_d)$ reaches its minimum for $\hat{X}_d=\Phi$. Thus,

$$V \geq \frac{1}{2}\hat{\Phi}^2\hat{X}_q^2 + \frac{1}{4}(\hat{\Phi}^2-\Phi^2)^2 \geq 0$$

and V vanishes only at the equilibrium. It satisfies

$$\dot{V} = -q\,\hat{\Phi}^2(\hat{\Phi}^2-(\hat{X}_d^2+\hat{X}_q^2))^2 \leq 0 \;. \tag{C.3}$$

C.2 Analysis of convergence

Consider a solution $(\hat{X}_q(t),\hat{X}_d(t),\hat{\Phi}(t))$ maximally defined on $[0,\overline{t}[$ in Ω. Because of (C.3), V is bounded on $[0,\overline{t}[$ when evaluated along the solution. V being proper in $\hat{\Phi}$, $\hat{\Phi}$ is also bounded on $[0,\overline{t}[$, let's say by Φ_m. Besides,

$$\begin{aligned}\dot{\overbrace{\hat{X}_d^2+\hat{X}_q^2}} \;&=\; -4q(\hat{X}_d^2+\hat{X}_q^2)(\hat{X}_d^2+\hat{X}_q^2-\hat{\Phi}^2) + 2\omega\Phi X_q \\ &\leq\; -4q(\hat{X}_d^2+\hat{X}_q^2)^2 + 4q\Phi_m^2(\hat{X}_d^2+\hat{X}_q^2) + 2\overline{\omega}_0\Phi\sqrt{\hat{X}_d^2+\hat{X}_q^2} \;.\end{aligned}$$

The negative term dominates for large values of $(\hat{X}_d^2+\hat{X}_q^2)$, which implies that $(\hat{X}_d,\hat{X}_q)$ is also bounded on $[0,\overline{t}[$.

Now assume that $\bar{t}$ is finite. Since the solution is bounded, it tends to the boundary of Ω when t tends to $\bar{t}$, i-e $\hat{\Phi}$ tends to 0 (in finite time). But this is impossible, because of uniqueness of solution, knowing that $\hat{\Phi} = 0$ is a solution. Therefore $\bar{t}$ is infinite and any solution is defined on $[0, +\infty)$. Let us now show that it converges to $(\Phi, 0, \Phi)$.

Note that (C.1) is time-varying because of ω and LaSalle invariance principle may not apply. But since V decreases and is lower-bounded, it converges. Besides, the solution and ω being bounded $\ddot{V}$ is bounded. It follows according to Barbalat's lemma that

$$\lim_{t\to+\infty} \dot{V} = \lim_{t\to+\infty} \hat{\Phi}(\hat{\Phi}^2 - (\hat{X}_d^2 + \hat{X}_q^2)) = 0 \ .$$

Using again Barbalat's lemma on $\ddot{V}$ ($\dot{V}$ converges and $V^{(3)}$ is bounded because $\dot{\omega}$ is bounded) gives $\lim_{t\to+\infty} \ddot{V} = 0$ and thus

$$\lim_{t\to+\infty} \omega\Phi\hat{\Phi}\hat{X}_q = 0 \ ,$$

which yields since ω is lower-bounded away from zero,

$$\lim_{t\to+\infty} \hat{\Phi}\hat{X}_q = 0 \ .$$

Finally, applying again Barbalat's lemma to the derivative of this function, we end up with

$$\lim_{t\to+\infty} \omega\hat{\Phi}(\hat{X}_d - \Phi) = 0 \ ,$$

and again since ω is lower-bounded,

$$\lim_{t\to+\infty} \hat{\Phi}(\hat{X}_d - \Phi) = 0 \ .$$

To sum up, we have established the following three limits

$$\lim_{t\to+\infty} \hat{\Phi}(\hat{\Phi}^2 - (\hat{X}_d^2 + \hat{X}_q^2)) = 0 \quad , \quad \lim_{t\to+\infty} \hat{\Phi}\hat{X}_q = 0 \quad , \quad \lim_{t\to+\infty} \hat{\Phi}(\hat{X}_d - \Phi) = 0 \ .$$

This is not enough to conclude since we could have a priori $\liminf_{t\to+\infty} \hat{\Phi} = 0$. However, the following points give the result :

1. The time function $(\hat{X}_d, \hat{X}_q, \hat{\Phi})$ is bounded and continuous. It follows that for any sequence (t_n) such that $\lim_{n\to\infty} t_n = +\infty$, the sequence $(\hat{X}_d(t_n), \hat{X}_q(t_n), \hat{\Phi}(t_n))$ admits at least one accumulation point.

2. Let $P^* = (\hat{X}_d^*, \hat{X}_q^*, \hat{\Phi}^*)$ be such an accumulation point. Because of the limits we have established, it verifies :

$$\begin{aligned} \hat{\Phi}^*(\hat{\Phi}^{*^2} - (\hat{X}_d^{*^2} + \hat{X}_q^{*^2})) &= 0 \\ \hat{\Phi}^*\hat{X}_q^* &= 0 \\ \hat{\Phi}^*(\hat{X}_d^* - \Phi) &= 0 \ . \end{aligned}$$

   Thus P^* is either of the type $(\hat{X}_d^*, \hat{X}_q^*, 0)$ with $(\hat{X}_d^*, \hat{X}_q^*)$ in $\mathbb{R}^2$ (type I) or equal to $P_0 = (\Phi, 0, \Phi)$.

3. Assume that P_0 is not an accumulation point. Then, any accumulation point is of type I and the only possible accumulation value for $\hat{\Phi}$ is 0. Thus

$$\lim_{t\to+\infty} \hat{\Phi}(t) = 0 \ .$$

To ease the notations let us denote the vector $\hat{X} = (\hat{X}_d, \hat{X}_q)$. Solving the differential equation ruling $\hat{\Phi}^2$, there exists a_0, b_0 strictly positive such that

$$\hat{\Phi}^2 = \frac{\phi(t)}{1 + b_0 + 2q \int_0^t \phi(s)ds}$$

with

$$\phi(t) = a_0 b_0 \exp\left(2q \int_0^t |\hat{X}(s)|^2 ds\right) .$$

Since $\hat{\Phi}$ tends to 0, for any $\eta > 0$, there exists $\bar{t} > 0$ such that for all $t \geq \bar{t}$, we have $\hat{\Phi}^2(t) \leq \frac{\eta}{2}$. This means that for all $t \geq \bar{t}$,

$$\phi(t) \leq \underbrace{\frac{\eta}{2}\left(1 + b_0 + 2q \int_0^{\bar{t}} \phi(s)ds\right)}_{c_0} + \eta q \int_{\bar{t}}^t \phi(s)ds$$

and by Gronwall's lemma

$$\phi(t) \leq c_0 \exp\left(\eta q (t - \bar{t})\right) .$$

We conclude that for any $\eta > 0$, there exists $\bar{t} > 0$ such that

$$\int_0^t |\hat{X}(s)|^2 ds \leq \frac{\eta}{2}(t - \bar{t}) + \frac{1}{2q} \log\left(\frac{c_0}{a_0 b_0}\right) . \tag{C.4}$$

But we are going to prove the existence of t_0, $\eta > 0$, and a sequence (t_k) such that $\lim_{k\to\infty} t_k = +\infty$ and

$$\int_{t_0}^{t_k} |\hat{X}(s)|^2 ds \geq \eta(t_k - t_0) , \tag{C.5}$$

which contradicts (C.4). Indeed, consider the dynamics of $\hat{X}$ with inputs ω and $\hat{\Phi}$. $\omega\Phi$ being lower-bounded by $\underline{\omega}\Phi > 0$, there exist a and $\bar{b}$ such that the conditions of Lemma C.2.1 given below are satisfied for

$$x = \hat{X} \quad , \quad \underline{b} = \frac{\underline{\omega}\Phi}{2} \quad , \quad \bar{t}_+ = +\infty \quad , \quad v = \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

First, assume that there exists $t_0 > 0$ such that for all $t \geq t_0$, $|\hat{X}(t)| \geq \frac{a}{2}$, then (C.5) is true for any sequence (t_k), and $\eta = \frac{a^2}{4}$. Assume now that this is not the case, i-e for any $\bar{t}$, there exists $t \geq \bar{t}$ such that $|\hat{X}(t)| \leq \frac{a}{2}$. Then, according to Lemma C.2.1, one can build sequences $(t_{k,1}), (t_{k,2}), (t_{k,3})$, each tending to $+\infty$ such that for all k :

$$\begin{gathered} t_{k,1} < t_{k,2} < t_{k,3} \\ |\hat{X}(t)| \leq \tfrac{a}{2} \quad \forall t \in [t_{k-1,3}, t_{k,1}] \\ \tfrac{a}{2} \leq |\hat{X}(t)| \leq a \quad \forall t \in [t_{k,1}, t_{k,2}] \\ |\hat{X}(t_{k,2})| = a \\ |\hat{X}(t)| \geq \tfrac{a}{2} \quad \forall t \in [t_{k,2}, t_{k,3}] \end{gathered}$$

and

$$\frac{3a}{\underline{b}} \;\geq\; t_{k,2} - t_{k-1,3} \quad , \qquad t_{k,2} - t_{k,1} \;\geq\; \frac{a}{2\bar{b}} .$$

We denote

$$\bar{t}_{k,1} = t_{k,1} - t_{k-1,3} \quad , \qquad \bar{t}_{k,2} = t_{k,2} - t_{k,1} \quad , \qquad \bar{t}_{k,3} = t_{k,3} - t_{k,2}$$

and $\bar{t}_k = t_{k,3} - t_{k-1,3} = \bar{t}_{k,1} + \bar{t}_{k,2} + \bar{t}_{k,3}$ the duration of the cycle k. Then, the mean over a cycle

$$\frac{1}{\bar{t}_k}\int_{t_{k-1,3}}^{t_{k,3}} |\hat{X}(s)|^2 ds \geq \frac{a^2}{4}\frac{\bar{t}_{k,2}+\bar{t}_{k,3}}{\bar{t}_{k,1}+\bar{t}_{k,2}+\bar{t}_{k,3}} \geq \frac{a^2}{4}\frac{\frac{a}{2\bar{b}}+\bar{t}_{k,3}}{\frac{3a}{\underline{b}}+\bar{t}_{k,3}} \geq \frac{a^2}{4}\min\left(\frac{\underline{b}}{6\bar{b}}, 1\right) .$$

is lower-bounded. Thus, (C.5) holds with $\eta = \frac{a^2}{4}\min\left(\frac{\underline{b}}{6\bar{b}}, 1\right)$ and $t_k = t_{k,3}$.
Finally, with (C.4) and η given above, there exists $\bar{t}$ such that

$$\eta(t_k - t_0) \leq \frac{\eta}{2}(t_k - \bar{t}) + \frac{1}{2q}\log\left(\frac{c_0}{a_0 b_0}\right)$$

for all k greater than some k_0. This is impossible. Thus, P_0 is accumulation point.

4. Assume that $P_1 = (\hat{X}_d^*, \hat{X}_q^*, 0)$ is another accumulation point. There exists an increasing sequence of times (t_n) such that

$$\begin{aligned} \lim_{n\to\infty} t_n &= +\infty \\ \left|(\hat{X}_d(t_{2k}), \hat{X}_q(t_{2k}), \hat{\Phi}(t_{2k})) - P_0\right| &< \frac{\Phi}{2} \\ \left|(\hat{X}_d(t_{2k+1}), \hat{X}_q(t_{2k+1}), \hat{\Phi}(t_{2k+1})) - P_1\right| &< \frac{\Phi}{2} . \end{aligned}$$

Since $|P_0 - P_1| \geq \Phi$, by continuity of the solution, for all $k \geq 0$, there exists $\tau_k > t_{2k}$ such that

$$\left|(\hat{X}_d(\tau_k), \hat{X}_q(\tau_k), \hat{\Phi}(\tau_k)) - P_0\right| = \frac{\Phi}{2} .$$

But then, the sequence $(\hat{X}_d(\tau_k), \hat{X}_q(\tau_k), \hat{\Phi}(\tau_k))$ admits an accumulation point P' verifying $|P' - P_0| = \frac{\Phi}{2}$. This is impossible because P' should be P_0 or of type I, and for any P of type I, $|P - P_0| \geq \Phi$. Therefore, P_0 is the only accumulation point and the solutions converge to P_0.

To complete the proof, it remains to show the following technical lemma :

Lemma C.2.1.

Let a, $\underline{b}$ and $\bar{b}$ be three strictly positive real numbers, v be a unit vector in $\mathbb{R}^n$ and f be a continuous function such that[1]:

$$v^\top f(x,t) \geq \underline{b} \quad , \qquad \bar{b} \geq |f(x,t)| \qquad \forall (x,t) \in \overline{B_a(0)} \times \mathbb{R} .$$

Let $x(t)$ be a solution of

$$\dot{x} = f(x,t)$$

defined on $(\bar{t}_-, \bar{t}_+)$ with values in $\mathbb{R}^n$. If there exists t_0 in $(\bar{t}_-, \bar{t}_+)$ such that $|x(t_0)| \leq \frac{a}{2}$, then there exist t_1 and t_2 both in $(\bar{t}_-, \bar{t}_+)$ such that

$$|x(t_1)| = \frac{a}{2} \quad , \qquad |x(t_2)| = a \quad , \qquad \frac{3a}{\underline{b}} \geq t_2 - t_0 \quad , \qquad t_2 - t_1 \geq \frac{a}{2\bar{b}}$$

and $|x(t)| \geq \frac{a}{2}$ for all t in $[t_1, t_2]$.

Proof : Let $t_2 < \bar{t}_+$ be the maximum time such that $x(t)$ is in $B_a(0)$ for all t in $[t_0, t_2[$. We have

$$v^\top \dot{x}(t) \geq \underline{b} \quad , \qquad \forall t \in [t_0, t_2[$$

[1] We denote $B_a(0)$ the open ball of $\mathbb{R}^n$ centered at the origin with radius a and $\overline{B_a(0)}$ its closure.

and

$$v^\top x(t_0) \;\geq\; -|x(t_0)| \;\geq\; -\frac{a}{2}\;.$$

This yields

$$|a| \;>\; |x(t)| \;\geq\; v^\top x(t) \;\geq\; -\frac{a}{2} \;+\; \underline{b}\,[t-t_0] \qquad \forall t \in [t_0, t_2[\;.$$

Thus t_2 is finite and by continuity,

$$|x(t_2)| \;=\; a \quad , \qquad \frac{3a}{2\underline{b}} \;\geq\; t_2 - t_0\;.$$

By continuity of solutions, there also exists t_1 in $[t_0, t_2[$, satisfying :

$$|x(t_1)| \;=\; \frac{a}{2} \quad , \qquad \frac{a}{\underline{b}} \;\geq\; t_1 - t_0\;.$$

But we also have

$$x(t_2) = x(t_1) + \int_{t_1}^{t_2} f(x(t), t)dt$$

so that

$$|a| \;=\; |x(t_2)| \;\leq\; \frac{a}{2} \;+\; \overline{b}\,[t_2 - t_1]$$

and therefore

$$t_2 - t_1 \;\geq\; \frac{a}{2\overline{b}}\;.$$

■

Appendix D

Proofs of Chapter 14

In this appendix, we prove most of the results presented in Chapter 14.

D.1 About observability

D.1.1 Proof of Theorem 14.1.1

Consider a solution (x, x_3) to System (14.2) verifying for all t

$$0 = y(t) = |x(t) - Li(t)|^2 - \Phi^2 \ .$$

x is necessarily of the form

$$x(t) = x_0 + \int_0^t u(\tau)d\tau - x_3 \int_0^t i(\tau)d\tau$$

with

$$\dot{x}_0 = 0 \quad , \quad \dot{x}_3 = 0 \ ,$$

and finding (x, x_3) is equivalent to finding (x_0, x_3). It follows that for all t

$$\begin{aligned} 0 &= |x(t) - Li(t)|^2 - |x_0 - Li(0)|^2 \\ &= [x(t) - x_0 - L(i(t) - i(0)]^\top [x(t) + x_0 - L(i(t) + i(0)] \\ &= \tilde{\eta}(x_3, t)^\top [2(x_0 - Li(0)) + \tilde{\eta}(x_3, t)] \end{aligned}$$

where we have defined

$$\tilde{\eta}(x_3, t) = \int_0^t u(\tau)d\tau - x_3 \int_0^t i(\tau)d\tau - L(i(t) - i(0)) \ . \tag{D.1}$$

We deduce that for any time t,

$$2\tilde{\eta}(x_3, t)^\top (x_0 - Li(0)) = -\tilde{\eta}(x_3, t)^\top \tilde{\eta}(x_3, t) = -|\tilde{\eta}(x_3, t)|^2 \ .$$

Therefore, unless x_3 makes $\tilde{\eta}(x_3, t_1)$ and $\tilde{\eta}(x_3, t_2)$ colinear for any (t_1, t_2), there exits at most one possible value of x_0 for each x_3.

Take x_3 such that $\tilde{\eta}(x_3, .)$ is not constant. There exists t_1 such that $\tilde{\eta}(x_3, t_1) \neq 0$. Thus, for some $t_2 \neq t_1$, $\tilde{\eta}(x_3, t_2)$ colinear to $\tilde{\eta}(x_3, t_1)$ implies that there exists λ such that $\tilde{\eta}(x_3, t_2) = \lambda\tilde{\eta}(x_3, t_1)$. But then,we have

$$2\lambda\tilde{\eta}(x_3, t_1)^\top (x_0 - Li(0)) = -\lambda^2 |\tilde{\eta}(x_3, t_1)|^2 = -\lambda|\tilde{\eta}(x_3, t_1)|^2$$

and necessarily $\lambda = 1$ or $\lambda = 0$, i-e $\tilde{\eta}(x_3, t_2) = \tilde{\eta}(x_3, t_1)$ or $\tilde{\eta}(x_3, t_2) = 0$. But, since $\tilde{\eta}(x_3, .)$ is continuous and not constant, there exists t_2 such that $\tilde{\eta}(x_3, t_2) \neq \tilde{\eta}(x_3, t_1)$ and $\tilde{\eta}(x_3, t_2) \neq 0$.

Actually, still by continuity, we can even say that there exist two intervals I_1 and I_2 such that for all (t_1, t_2) in $I_1 \times I_2$, we have $\tilde{\eta}(x_3, t_1) \neq 0$, $\tilde{\eta}(x_3, t_2) \neq 0$ and $\tilde{\eta}(x_3, t_2) \neq \tilde{\eta}(x_3, t_1)$, i-e such that $\tilde{\eta}(x_3, t_1)$ and $\tilde{\eta}(x_3, t_2)$ are not colinear. For each such couple (t_1, t_2), x_0 is uniquely determined by the value of x_3. Indeed, denoting

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} ,$$

we have $\tilde{\eta}(x_3, t_1)^\top J \tilde{\eta}(x_3, t_2) \neq 0$ and necessarily

$$x_0 = Li(0) + \frac{1}{2} \frac{J\left(\tilde{\eta}(x_3, t_1)\ ,\ \tilde{\eta}(x_3, t_2)\right)}{\tilde{\eta}(x_3, t_1)^\top J \tilde{\eta}(x_3, t_2)} \begin{pmatrix} |\tilde{\eta}(x_3, t_2)|^2 \\ -|\tilde{\eta}(x_3, t_1)|^2 \end{pmatrix} .$$

Replacing this expression in the constraint

$$|x_0 - Li(0)|^2 - \Phi^2 = 0 ,$$

we obtain a polynomial of degree 6 in x_3 for each couple (t_1, t_2) in $I_1 \times I_2$. In order to deduce that there are at most 6 solutions x_3 making $\tilde{\eta}(x_3, .)$ not constant, we need to prove that at least one of these polynomials is not a constant. It is possible to show that the coefficient of highest degree is given by $I(t_1)^2 I(t_2)^2 (I(t_1) - I(t_2))$ with $I(t) = \int_0^t i(\tau) d\tau$. If $I(t_1) = 0$ for all t_1 in I_1, then i is zero on I_1 which is excluded by assumption, thus there exists t_1 in I_1 such that $I(t_1) \neq 0$. Now assume $I(t_2) = I(t_1)$ or $I(t_2) = 0$ for all t_2 in I_2. Again this means that i is zero on I_2, which is impossible. We conlude that there exists (t_1, t_2) in $I_1 \times I_2$ such that the corresponding polynomial is "truly" of order 6 (i-e with a nonzero coefficient of order 6) and therefore, there are at most 6 solutions x_3 making $\tilde{\eta}(x_3, .)$ not constant, and for each of these values, there is a unique corresponding x_0. This characterizes at most 6 solutions (x, x_3).

Now take x_3 such that $\tilde{\eta}(x_3, .)$ is constant. Since $\tilde{\eta}(x_3, 0) = 0$, $\tilde{\eta}(x_3, t) = 0$ for all t. It follows that any x_0 verifying $|x_0 - Li(0)| = \Phi$ is solution, and there exists an infinity of solutions associated to this value of x_3.

The only remaining question is therefore : does there exist a value of x_3 making $\tilde{\eta}(x_3, .)$ constant ? Assume such a value exists. Differentiating $\tilde{\eta}(x_3, .)$ with respect to time, we get for all t

$$u(t) - x_3 i(t) - L\dot{\widehat{i}}(t) = 0 .$$

But differentiating (14.3) with respect to time, we also know that

$$u(t) - Ri(t) = L\dot{\widehat{i}}(t) + \omega\Phi \begin{pmatrix} -\sin(\theta(t)) \\ \cos(\theta(t)) \end{pmatrix}$$

so that, necessarily, by combining the two equations and multiplying by $\mathcal{R}(-\theta(t))$, the following system must be satisfied for all t :

$$\begin{aligned} (R - x_3) i_d(t) &= 0 \\ (R - x_3) i_q(t) &= -\omega(t)\Phi . \end{aligned} \tag{D.2}$$

We distinguish the following cases :

- if $\omega(t) = 0$ for all t, there exists at least one constant value of x_3 solution to System (D.2) for all t. Thus, $\tilde{\eta}(x_3, \cdot)$ is constant and there is an infinite number of solutions (x, x_3).
- if for all t such that $\omega(t) \neq 0$, $i_d(t) = 0$, $i_q(t) \neq 0$, and $\frac{\omega}{i_q}$ is constant, there exists a constant value of x_3 solution to System (D.2) for all t and thus an infinity of solutions (x, x_3).
- otherwise, there exist no solutions to System (D.2). Therefore, $\tilde{\eta}(x_3, \cdot)$ cannot be constant and there are at most 6 solutions (x, x_3) to our observability problem.

D.1.2 Proof of Theorem 14.1.3

Consider a solution to System (14.2) verifying $y(t) = 0$ for all t. According to Corollary (14.1.1), since $y(t) = \dot{y}(t) = \ddot{y}(t) = 0$, it is necessarily (Ψ, R) or (Ψ_δ, R_δ) where Ψ_δ is given by :

$$\Psi_\delta(t) = L\, i(t) + \frac{|\eta(R_\delta, t)|^2}{\eta(R_\delta, t)^\top J \dot{\eta}(R_\delta, t)} J\eta(R_\delta, t) \ ,$$

with η defined in (14.7). It remains to show that (Ψ_δ, R_δ) is a solution to System (14.2).

Using (14.8), we get

$$|\eta(x_3, t)|^2 = \omega^2\Phi^2 \ + \ 2(R - x_3)\Phi\omega\, i_q \ + \ (R - x_3)^2\, |i|^2$$

and thus,

$$|\eta(R_\delta, t)| = \omega\Phi \ .$$

It follows that there exists θ_δ such that

$$\eta(R_\delta, t) = \omega\Phi \begin{pmatrix} -\sin\theta_\delta \\ \cos\theta_\delta \end{pmatrix} = -\omega\Phi J z_\delta$$

where we denote

$$z_\delta = \begin{pmatrix} \cos\theta_\delta \\ \sin\theta_\delta \end{pmatrix} \ .$$

We deduce according to (14.8) that

$$\begin{aligned} -\omega\Phi J z_\delta &= -\omega\Phi J z + (R - R_\delta) i \\ &= -\omega\Phi J z - \frac{2\Phi\omega\, i_q}{|i|^2}\, i \end{aligned}$$

and after a rotation of angle $-\theta$, we have

$$J \begin{pmatrix} \cos(\theta_\delta - \theta) \\ \sin(\theta_\delta - \theta) \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \frac{2\, i_q}{|i|^2} i_{dq} \ . \tag{D.3}$$

Therefore, $\theta_\delta - \theta$ is a constant and

$$w_\delta = \dot{\theta_\delta} = w \ .$$

It follows that x_δ defined by

$$x_\delta = L\, i + \Phi z_\delta$$

verifies the dynamics :

$$\dot{x_\delta} = L\, \dot{\widehat{i}} - \Phi\omega_\delta J z_\delta = L\, \dot{\widehat{i}} - \Phi\omega J z_\delta = L\, \dot{\widehat{i}} + \eta(R_\delta, t) = u - R_\delta\, i$$

and

$$0 = y = |x_\delta - L\, i|^2 - \Phi \ .$$

Thus, (x_δ, R_δ) is a solution to (14.2), which must appear among $\{(\Psi, R), (\Psi_\delta, R_\delta)\}$, i-e it is necessarily (Ψ_δ, R_δ). Therefore, (Ψ, R) and (Ψ_δ, R_δ) are the only two indistinguishable solutions.

D.1.3 Proof of Theorem 14.1.4

According to (D.3),

$$\begin{aligned}\cos(\theta_\delta - \theta) &= 1 - 2i_q^2 \\ \sin(\theta_\delta - \theta) &= 2i_q i_d \ .\end{aligned}$$

But after a rotation of $-\theta_\delta$ instead of θ, we would have obtained :

$$\begin{pmatrix} 0 \\ -1 \end{pmatrix} = J \begin{pmatrix} \cos(\theta - \theta_\delta) \\ \sin(\theta - \theta_\delta) \end{pmatrix} + \frac{2\, i_q}{|i|^2} i_{dq,\delta}$$

i-e

$$\begin{aligned}\cos(\theta_\delta - \theta) &= 1 + 2i_q i_{q,\delta} \\ \sin(\theta_\delta - \theta) &= 2i_q i_{d,\delta} \ ,\end{aligned}$$

which gives the result.

Now, if $\hat{R} = R_\delta$, one can find R by computing

$$R = \hat{R} - \frac{2\Phi\omega\, i_q}{|i|^2} = \hat{R} + \frac{2\Phi\hat{\omega}\,\hat{i_q}}{|i|^2}$$

and if $\hat{R} = R$, one can find R_δ by computing

$$R_\delta = \hat{R} + \frac{2\Phi\omega\, i_q}{|i|^2} = \hat{R} + \frac{2\Phi\hat{\omega}\,\hat{i_q}}{|i|^2} \ .$$

This means that whatever the value of $\hat{R}$ is, the value of the other candidate is $\hat{R} + \frac{2\Phi\hat{\omega}\,\hat{i_q}}{|i|^2}$. Similarly, if $\hat{\theta} = \theta_\delta$, then $\theta = \hat{\theta} + \Delta$ with Δ defined by

$$\begin{aligned}\cos(\Delta) &= 1 - 2i_q^2 = 1 - 2\hat{i_q}^2 \\ \sin(\Delta) &= -2i_q i_d = 2\hat{i_q}\hat{i_d}\end{aligned}$$

and if $\hat{\theta} = \theta$, then $\theta_\delta = \hat{\theta} - \Delta$ with $-\Delta$ defined by

$$\begin{aligned}\cos(-\Delta) &= 1 - 2i_q^2 = 1 - 2\hat{i_q}^2 \\ \sin(-\Delta) &= 2i_q i_d = 2\hat{i_q}\hat{i_d} \ .\end{aligned}$$

Therefore, whatever the value of $\hat{\theta}$, the value of the other solution is $\hat{\theta} + \arctan_2(2\hat{i_q}\hat{i_d}, 1 - 2\hat{i_q}^2)$.

D.2 Observer design

D.2.1 Proof of Lemma 14.2.2

The quantity

$$\mu_\lambda(x_3, t) = -(c_\lambda + x_3\, b_\lambda + 2\lambda L i) \tag{D.4}$$

verifies

$$\dot{\mu_\lambda(x_3, t)} = -\lambda(\mu_\lambda(x_3, t) - 2\,\eta(x_3, t)) \tag{D.5}$$

This means that μ_λ is a filtered version of 2η. Besides, denoting

$$\mu = \begin{pmatrix} \mu_{\lambda_1}^\top \\ \mu_{\lambda_2}^\top \\ \mu_{\lambda_3}^\top \end{pmatrix}$$

we have

$$\Lambda\,\mu(x_3,t) = -(\Lambda c + x_3\,\Lambda b + 2Lm_\lambda i^\top)$$

and thus since $M_\lambda m_\lambda = 0$,

$$\mathcal{M}(x_3,t) = -M_\lambda \Lambda\,\mu(x_3,t)\ . \tag{D.6}$$

Let us further investigate the link between μ and η. According to the dynamics (D.5), μ_λ can be developed in the following way :

$$\mu_\lambda = 2\eta - \frac{2\dot\eta}{\lambda} + \frac{2r_\mu}{\lambda^2} \tag{D.7}$$

where the rest r_μ follows the dynamics :

$$\dot r_\mu = -\lambda(r_\mu - \ddot\eta)\ . \tag{D.8}$$

Thus, we have the development

$$\mu_\lambda = 2\eta - \frac{2\dot\eta}{\lambda} + O\left(\frac{1}{\lambda^2}\right)$$

so that μ_λ can be approximated by $2\eta - \dfrac{2\dot\eta}{\lambda}$ for large values of λ. Therefore,

$$\begin{aligned}
\mathcal{M} &= -2M_\lambda\Lambda \begin{pmatrix} \eta - \frac{\dot\eta}{\lambda_1} + O\left(\frac{1}{\lambda_1^2}\right) \\ \eta - \frac{\dot\eta}{\lambda_2} + O\left(\frac{1}{\lambda_2^2}\right) \\ \eta - \frac{\dot\eta}{\lambda_3} + O\left(\frac{1}{\lambda_3^2}\right) \end{pmatrix} \\
&= -2M_\lambda \begin{pmatrix} \lambda_1\eta - \dot\eta\lambda_1 + O\left(\frac{1}{\lambda_1}\right) \\ \lambda_2\eta - \dot\eta\lambda_1 + O\left(\frac{1}{\lambda_2}\right) \\ \lambda_3\eta - \dot\eta\lambda_1 + O\left(\frac{1}{\lambda_3}\right) \end{pmatrix} \\
&= -2\begin{pmatrix} \lambda_1\lambda_2(\lambda_2-\lambda_1) & \lambda_1^2-\lambda_2^2 \\ \lambda_2\lambda_3(\lambda_3-\lambda_2) & \lambda_2^2-\lambda_3^2 \end{pmatrix}\begin{pmatrix} \eta^\top \\ \dot\eta^\top \end{pmatrix} + O(\lambda)\ ,
\end{aligned} \tag{D.9}$$

and straightforward computations give (14.25).

Let us now develop $\chi(x_3,t)$ with respect to λ. To do that, we define ρ_λ with

$$\rho_\lambda(x_3,t) = -(z_\lambda - a_\lambda x_3 - d_\lambda x_3^2 + \lambda L\mu_\lambda^\top i + \lambda^2 L^2|i|^2 - \lambda^2\Phi^2)\ , \tag{D.10}$$

which follows the dynamics

$$\dot\rho_\lambda = -\lambda(\rho_\lambda - \mu_\lambda^\top\eta)\ . \tag{D.11}$$

In other words, by denoting

$$\rho = \begin{pmatrix} \rho_{\lambda_1} \\ \rho_{\lambda_2} \\ \rho_{\lambda_3} \end{pmatrix}\ ,$$

we define

$$\rho(x_3,t) = -(Z(t) - a(t)x_3 - d(t)x_3^2 + L\Lambda\mu\, i + m_\lambda L^2|i|^2 - m_\lambda\Phi^2)\ . \tag{D.12}$$

When $\mathcal{M}$ is invertible, by definition of χ in (14.22), we have

$$\mathcal{M}(x_3,t)\chi(x_3,t) = M_\lambda(Z(t) - a(t)x_3 - d(t)x_3^2)$$

and thus since $M_\lambda m_\lambda = 0$ and (D.6),

$$\mathcal{M}(x_3,t)(\chi(x_3,t) - Li) = -M_\lambda\,\rho\ .$$

As we did above for μ, it is possible to develop ρ thanks to (D.11). Indeed, it is straightforward to check that

$$\rho_\lambda = 2\eta^\top\eta + \frac{r_\rho}{\lambda} = 2|\eta|^2 + \frac{r_\rho}{\lambda}$$

with

$$\dot r_\rho = -\lambda(r_\rho + 6\eta^\top\dot\eta)\ .$$

In other words

$$\rho_\lambda = 2|\eta|^2 + O\left(\frac{1}{\lambda}\right) \tag{D.13}$$

and we have with (D.9)

$$\begin{array}{rcl}
\begin{pmatrix} \eta^\top \\ \dot\eta^\top \end{pmatrix}(\chi(x_3,t) - Li) &=& \frac{1}{2}\begin{pmatrix} \lambda_1\lambda_2(\lambda_2-\lambda_1) & \lambda_1^2-\lambda_2^2 \\ \lambda_2\lambda_3(\lambda_3-\lambda_2) & \lambda_2^2-\lambda_3^2 \end{pmatrix}^{-1} M_\lambda\,\rho + O\left(\frac{1}{\lambda}\right) \\
&=& \begin{pmatrix} \lambda_1\lambda_2(\lambda_2-\lambda_1) & \lambda_1^2-\lambda_2^2 \\ \lambda_2\lambda_3(\lambda_3-\lambda_2) & \lambda_2^2-\lambda_3^2 \end{pmatrix}^{-1} M_\lambda \begin{pmatrix} |\eta|^2 \\ |\eta|^2 \end{pmatrix} + O\left(\frac{1}{\lambda}\right) \\
&=& \begin{pmatrix} 0 \\ -1 \end{pmatrix}|\eta|^2 + O\left(\frac{1}{\lambda}\right)\ .
\end{array}$$

Finally, according to the definitions (14.19), (D.4) and (D.12), it is straightforward to check that

$$Z(t) - T(\chi(x_3,t),x_3,t) = -m_\lambda\left(|\chi(x_3,t) - Li|^2 - \Phi^2\right) + \Lambda\,\mu(x_3,t)\,(\chi(x_3,t) - Li) - \rho(x_3,t)\ .$$

It follows that for x_3 and t making $\Big(\eta(x_3,t)\,,\,\dot\eta(x_3,t)\Big)$ invertible

$$\begin{array}{rcl}
J(x_3,t) &=& -(\lambda_1^4+\lambda_2^4+\lambda_3^4)\left(|\chi(x_3,t) - Li|^2 - \Phi^2\right) + O(\lambda^3) \\
&=& (\lambda_1^4+\lambda_2^4+\lambda_3^4)\left(\dfrac{P(x_3,t)}{\det\Big(\eta(x_3,t)\,,\,\dot\eta(x_3,t)\Big)^2} + O\left(\frac{1}{\lambda}\right)\right) + O(\lambda^3) \\
&=& (\lambda_1^4+\lambda_2^4+\lambda_3^4)\,\dfrac{P(x_3,t)}{\det\Big(\eta(x_3,t)\,,\,\dot\eta(x_3,t)\Big)^2} + O(\lambda^3)\ .
\end{array}$$

D.2.2 Proof of Theorem 14.2.2

Assume $\Big|\det\Big(\eta(x_3,t)\,,\,\dot\eta(x_3,t)\Big)\Big| \geq \underline{d}$. In order to deduce from (14.25) that $\Big|\det\Big(\mathcal{M}(x_3,t)\Big)\Big| \geq \underline{\delta}$ for all t, if α is sufficiently large, we need to bound the term $O(\lambda^4)$ uniformly in time. Coming back to (D.7)-(D.8), since $\ddot\eta$ is a polynomial in x_3 with coefficients depending on the bounded signals $(\ddot u, \ddot{\widehat{i}}, i^{(3)})$, r_μ is too. Following this term in (D.9), and then in its determinant (14.25), the reader can easily check that $O(\lambda^4)$ in (14.25) is also a polynomial in x_3 with bounded (in time) coefficients. Thus, there exists a polynomial $\mathfrak{R}$ (time and α independent) such that for all t and all α

$$\frac{1}{\alpha^5}\Big|\det\Big(\mathcal{M}(x_3,t)\Big)\Big| \geq 4\tilde\lambda_2^2(\tilde\lambda_1-\tilde\lambda_2)(\tilde\lambda_2-\tilde\lambda_3)(\tilde\lambda_3-\tilde\lambda_1)\,\underline{d} - \frac{\mathfrak{R}(|x_3|)}{\alpha} \tag{D.14}$$

which is strictly positive when α is sufficiently large.

In particular, according to (14.11),

$$\left|\det\Big(\eta(R,t)\,,\,\dot\eta(R,t)\Big)\right| = |\omega^3\Phi^2| \geq \underline{\omega}^3\Phi^2 > 0$$

for all t, hence the result.

D.2.3 Proof of Theorem 14.2.3

The first point of the result is a direct consequence of Theorem 14.2.2.

Then, remember that in the case where ω, i_d and i_q are constant, η, $\dot\eta$ and P are time independent polynomials such that (see (14.12))

$$\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big) = \omega^3\Phi^2\left(1+\frac{(R-x_3)}{\omega\Phi}2i_q+\frac{(R-x_3)^2}{\omega^2\Phi^2}|i|^2\right)$$

$$P(x_3) = -\Phi^2\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)^2\,\frac{(R-x_3)}{\omega\Phi}\,\left(2i_q+\frac{(R-x_3)}{\omega\Phi}|i|^2\right)\ .$$

Therefore, the roots of $\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)$ are the complex numbers $\frac{\Phi\omega}{|i|^2}(-i_q \pm j\,i_d)$, both situated on the circle with center R and radius $\frac{\Phi\omega}{|i|}$, and

$$Q(x_3) = \frac{P(x_3)}{\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)^2} = -\Phi^2\,\frac{(R-x_3)}{\omega\Phi}\,\left(2i_q+\frac{(R-x_3)}{\omega\Phi}|i|^2\right)$$

is a polynomial of order 2, with the two roots $(R,R_\delta) = \Big(R, R+\frac{2\omega\Phi i_q}{|i|^2}\Big)$ identified in Corollary (14.1.1).

Now, take any $\varepsilon>0$ and consider $\Gamma_{\underline{r_\varepsilon}}(R)$ and $\Gamma_{\overline{r_\varepsilon}}(R)$, the circles with center R and radius $\underline{r_\varepsilon}$ and $\overline{r_\varepsilon}$ respectively. The polynomial $\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)$ has no root on those circles so that it can be lower-bounded by some $\underline{d}>0$. Also, $\mathfrak{R}$ introduced in (D.14) is continuous on those compact sets and is bounded by some $\overline{\mathfrak{R}}$. Choosing $(\lambda_1,\lambda_2,\lambda_3)$ as suggested in the theorem and denoting $\gamma = 4\tilde\lambda_2^2(\tilde\lambda_1-\tilde\lambda_2)(\tilde\lambda_2-\tilde\lambda_3)(\tilde\lambda_3-\tilde\lambda_1)$, we have for all t and all x_3 in $\Gamma_{\overline{r_\varepsilon}}(R)\cup\Gamma_{\underline{r_\varepsilon}}(R)$,

$$\left|\frac{1}{\alpha^5}\det\Big(\mathcal{M}(x_3,t)\Big)-\gamma\,\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)\right| \leq \frac{\mathfrak{R}(|x_3|)}{\alpha} \leq \frac{\overline{\mathfrak{R}}}{\alpha} \leq \gamma\,\underline{d} \leq \left|\gamma\,\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)\right|$$

for α sufficiently large. Both functions being holomorphic (polynomials), according to Rouché's theorem, $\det\Big(\mathcal{M}(x_3,t)\Big)$ and $\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)$ have the same number of roots in[1] $B_{\underline{r_\varepsilon}}(R)$ (resp $B_{\overline{r_\varepsilon}}(R)$), namely no roots in $B_{\underline{r_\varepsilon}}(R)$ (resp 2 roots in $B_{\overline{r_\varepsilon}}(R)$). Since we know $\det\Big(\mathcal{M}(x_3,t)\Big)$ is a polynomial of order 2, we deduce that its only two roots are situated in the annulus $C(R,\underline{r_\varepsilon},\overline{r_\varepsilon})$. Besides, since $\det\Big(\eta(x_3)\,,\,\dot\eta(x_3)\Big)$ does not admit any real roots, its modulus is lower-bounded on the real axis and according to (D.14), one can make $\det\Big(\mathcal{M}(x_3,t)\Big)$ strictly positive for x_3 in the compact set $[R-\overline{r_\varepsilon},R-\underline{r_\varepsilon}]\cup[R+\underline{r_\varepsilon},R+\overline{r_\varepsilon}]$ by choosing α sufficiently large. In that case, its roots are in $C(R,\underline{r_\varepsilon},\overline{r_\varepsilon})$, but not in $C(R,\underline{r_\varepsilon},\overline{r_\varepsilon})\cap\mathbb{R}$: they are necessarily complex.

As for the third point of the result, it can be proved by applying Rouché's theorem on (14.27) with path $\Gamma_{\underline{r_\varepsilon}}(R)$. To do this, we need to lower-bound $|Q(x_3)|$ and upper-bound the term $O(\lambda^3)$ in (14.27). When $|i_q| \neq \frac{1-\varepsilon}{2}|i|$, Q does not have any root on $\Gamma_{\underline{r_\varepsilon}}(R)$ and $|Q|$

[1] $B_r(a)$ denotes the ball with center a and radius r.

can thus be lower-bounded in this set by some $\underline{d}_2 > 0$. As for the term $O(\lambda^3)$, by following the proof of Lemma 14.2.2, the reader can check that it is actually a rational function in x_3 whose coefficients are bounded in time and whose denominator is necessarily of the form $\det\Big(\mathcal{M}(x_3,t)\Big)^{k_1} \det\Big(\eta(x_3)\,,\,\dot{\eta}(x_3)\Big)^{k_2}$ (coming from the inversion of the corresponding matrices). On $\Gamma_{\underline{r_\varepsilon}}(R)$, $\det\Big(\eta(x_3)\,,\,\dot{\eta}(x_3)\Big)$ does not have any root and can be lower-bounded. The same thing holds for $\det\Big(\mathcal{M}(x_3,t)\Big)$ (uniformly in time) by taking α sufficiently large according to (D.14). Therefore, the term $O(\lambda^3)$ in (14.27) can be upper-bounded by a polynomial of $|x_3|$. More precisely, denoting $\gamma_2 = \tilde{\lambda}_1^4 + \tilde{\lambda}_2^4 + \tilde{\lambda}_3^4$, there exists a polynomial $\mathfrak{R}_2$ (time and α independent) such that for all t and all x_3 in $\Gamma_{\underline{r_\varepsilon}}(R)$,

$$\left|\frac{1}{\alpha^4}J(x_3,t) - \gamma_2\,Q(x_3)\right| \leq \frac{\mathfrak{R}_2(|x_3|)}{\alpha} \leq \frac{\overline{\mathfrak{R}}_2}{\alpha} \leq \gamma_2\,\underline{d}_2 \leq |\gamma_2\,Q(x_3)| \ .$$

Since $\mathcal{M}(\cdot,t)^{-1}$ is defined on $B_{\underline{r_\varepsilon}}(R)$ (where its determinant does not have any root), $J(\cdot,t)$ is holomorphic on that set and according to Rouché's theorem, it admits as many zeros as Q on $B_{\underline{r_\varepsilon}}(R)$, i-e either one or two depending on i_q.

Résumé

Contrairement aux systèmes linéaires, il n'existe pas de méthode systématique pour la synthèse d'observateurs pour systèmes non linéaires. Cependant, la synthèse peut être plus ou moins simple suivant les coordonnées choisies pour exprimer la dynamique. Des structures particulières, appelées formes normales, ont notamment été identifiées comme permettant la construction facile et directe d'un observateur. Une façon usuelle de résoudre le problème consiste donc à chercher un changement de coordonnées réversible permettant l'expression de la dynamique dans l'une de ces formes normales, puis à synthétiser l'observateur dans ces coordonnées, et enfin à en déduire une estimation de l'état du système dans les coordonnées initiales par inversion de la transformation. Cette thèse contribue à chacune de ces trois étapes.

Premièrement, nous montrons l'intérêt d'une nouvelle forme triangulaire avec des non linéarités continues (non Lipschitz). En effet, les systèmes observables pour toutes entrées, mais dont l'ordre d'observabilité différentielle est supérieur à la dimension du système, peuvent ne pas être transformables dans la forme triangulaire Lipschitz standard, mais plutôt dans une forme triangulaire "seulement continue". Le célèbre observateur grand gain n'est alors plus suffisant, et nous proposons d'utiliser plutôt des observateurs homogènes.

Une autre forme normale intéressante est la forme linéaire Hurwitz, qui admet un observateur trivial. La question de la transformation d'un système non linéaire dans une telle forme n'a été étudiée que pour les systèmes autonomes à travers les observateurs de Kazantzis-Kravaris ou de Luenberger. Nous montrons ici comment cette synthèse, consistant à résoudre une EDP, peut être étendue aux systèmes instationnaires/commandés.

Quant à l'inversion de la transformation, cette étape est loin d'être triviale en pratique, surtout lorsque les espaces de départ et d'arrivée ont des dimensions différentes. En l'absence d'expression explicite et globale de l'inverse, l'inversion numérique repose souvent sur la résolution d'un problème de minimisation couteux en calcul. C'est pourquoi nous développons une méthode permettant d'éviter l'inversion explicite de la transformation en ramenant la dynamique de l'observateur (exprimée dans les coordonnées de la forme normale) dans les coordonnées initiales du système. Ceci nécessite une extension dynamique, i-e l'ajout de nouvelles coordonnées et l'augmentation d'une immersion injective en un difféomorphisme surjectif.

Enfin, dans une partie totalement indépendante, nous proposons des résultats concernant l'estimation de la position du rotor d'un moteur synchrone à aimant permanent en l'absence d'informations mécaniques (sensorless) et lorsque des paramètres tels que la résistance ou le flux de l'aimant sont inconnus. Ceci est illustré par des simulations sur données réelles.

Mots Clés

observabilité, formes normales, observateur grand gain, observateur de Luenberger, extension dynamique, sensorless

Abstract

Unlike for linear systems, no systematic method exists for the design of observers for nonlinear systems. However, observer design may be more or less straightforward depending on the coordinates we choose to express the system dynamics. In particular, some specific structures, called normal forms, have been identified for allowing a direct and easier observer construction. It follows that a common way of addressing the problem consists in looking for a reversible change of coordinates transforming the exression of the system dynamics into one of those normal forms, design an observer in those coordinates, and finally deduce an estimate of the system state in the initial coordinates via inversion of the transformation. This thesis contributes to each of those three steps.

First, we show the interest of a new triangular normal form with continuous (non-Lipschitz) nonlinearities. Indeed, we have noticed that systems which are observable for any input but with an order of differential observability larger than the system dimension, may not be transformable into the standard Lipschitz triangular form, but rather into an "only continuous" triangular form. In this case, the famous high gain observer no longer is sufficient, and we propose to use homogeneous observers instead.

Another normal form of interest is the Hurwitz linear form which admits a trivial observer. The question of transforming a nonlinear system into such a form has only been addressed for autonomous systems with the so-called Lunberger or Kazantzis-Kravaris observers. This design consists in solving a PDE and we show here how it can be extended to time-varying/controlled systems.

As for the inversion of the transformation, this step is far from trivial in practice, in particular when the domain and image spaces have different dimensions. When no explicit expression for a global inverse is available, numerical inversion usually relies on the resolution of a minimization problem with a heavy computational cost. That is why we develop a method to avoid the explicit inversion of the transformation by bringing the observer dynamics (expressed in the normal form coordinates) back into the initial system coordinates. This is done by dynamic extension, i-e by adding some new coordinates to the system and augmenting an injective immersion into a surjective diffeomorphism.

Finally, in a totally independent part, we also provide some results concerning the estimation of the rotor position of a permanent magnet synchronous motors without mechanical information (sensorless) and when some parameters such as the magnet flux or the resistance are unknown. We illustrate this with simulations on real data.

Keywords

observability, normal forms, high gain observer, Luenberger observer, dynamic extension, sensorless