Ph.D.-FSTC-10-2010
Faculté des Sciences, de la Technologie
et de la Communication

UFR 907 Sciences et Techniques
Le2i, Laboratoire, Electronique, Informatique
et Image, UMR CNRS 5158

THÈSE

Présentée le 30/04/2010 à Luxembourg
en vue de l'obtention du grade académique de

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN INFORMATIQUE

DOCTEUR DE L'UNIVERSITÉ DE BOURGOGNE
EN INFORMATIQUE
Spécialité: Automatique

par

Nicolas BOIZOT
Né le 14 Juillet 1977 à Dijon, France

ADAPTIVE HIGH-GAIN EXTENDED KALMAN FILTER AND APPLICATIONS

Jury
Dr Jean-Paul Gauthier, Président
Professeur, Université de Toulon

Dr Eric Busvelle, Directeur de thèse
Professeur, Université de Bourgogne

Dr Jürgen Sachau, Directeur de thèse
Professeur, Université du Luxembourg

Dr Jean-Claude Vivalda, Rapporteur
Directeur de Recherches, INRIA Metz

Dr Gildas Besançon, Rapporteur
Maître de conférences HDR, INP Grenoble

Dr Martin Schlichenmaier, Examinateur
Professeur, Université du Luxembourg

Qu'est-ce qui nous prend, de mourir comme des cons,
nous qui pourtant savions vivre[†] *?*

à Richard Boizot

Ce jour là, j'ai bien cru tenir quelque chose et que ma vie
s'en trouverait changée.
Mais rien de cette nature n'est définitivement acquis.
Comme une eau, le monde vous traverse et pour un temps
vous prête ses couleurs.
Puis se retire, et vous replace devant ce vide que l'on porte en soi,
devant cette espèce d'insuffisance centrale de l'âme
qu'il faut bien apprendre à côtoyer, à combattre,
et qui, paradoxalement est peut-être
notre moteur le plus sûr[‡].

à Joëlle Boizot

[†]Frederic Dard à propos de Michel Audiard, *Libération, 30 Juillet 1985.*
[‡]Nicolas Bouvier, *L'usage du monde.*

Remerciements

Je remercie tout d'abord monsieur Jean-Paul Gauthier pour avoir accepté la présidence de ce jury de thèse, pour ses intuitions extrêmement pertinentes et pour m'avoir transmis le goût de la théorie du contrôle il y a de ça quelques années.

Monsieur Gildas Besançon a accepté le rôle de rapporteur de ce manuscrit. Lui présenter ce travail me tenait particulièrement à coeur. Les échanges que nous avons eu, alors que je n'en étais qu'au tout début m'ont encouragé à croire en l'approche adaptative, à explorer l'aspect temps réel du problème.

Monsieur Jean-Claude Vivalda m'a fait l'honneur d'être rapporteur de mon travail. J'ai beaucoup apprécié l'acuité de sa relecture et de ses questions portant sur ce lemme avec lequel je me suis battu pendant la rédaction pensant, naïf, que ce ne serait qu'une bagatelle.

I'm thankful to mister Martin Schlichenmaier for being part of the jury. I remember that it took us ages to organize a talk at the *doctorands seminar of the Mathematics Research Unit*: although I don't know if the audience did, I learned a lot that day.

Je remercie monsieur Jürgen Sachau d'avoir coencadré ce travail, de m'avoir accueilli au sein de son équipe, de m'avoir offert beaucoup d'autonomie et surtout de m'avoir soutenu dans mes choix.

Je te suis reconnaissant, Eric, de.... plein de choses en fait. Tout d'abord, de m'avoir proposé ce beau sujet, de m'avoir instillé un peu de tes qualités scientifiques, pour ton humanité, pour ta façon à motiver les troupes. Pour le *Poisson-scorpion*.

Merci à M. Ulrich Sorger pour avoir suivi l'évolution de ce travail, à Pascal Bouvry, Nicolas Guelfy et Thomas Engel pour m'avoir accueilli avec bienveillance dans leur unité (CSC) et laboratoires respectifs (LASSY et Com.Sys).

M. Massimo Malvetti a été d'une rare gentillesse et a beaucoup contribué à ce que la cotutelle soit mise en place dans de bonnes conditions.

Merci à M. Jean-Regis Hadji-Minaglou de s'être demandé ce que je pouvais fiche dans le labo et d'avoir écouté la (longue) réponse. Merci encore pour nos discussions du midi et pour la Bonne Adresse.

Nos secrétaires font preuve de patience pour gérer le flou artistique qui caractérise souvent une université, je dois avouer en avoir eu ma part. Merci à Ragga, Fabienne, Denise, Danièle, Steffie, Virginie et tout particulièrement à Mireille

Kies, témoin de cette période qui m'a vu m'exercer au violon après les heures de bureau.

Tout le monde n'a pas la chance de converser avec une incarnation de Nick Tosches. C'est systématique, je quitte Alexandre Ewen avec la furieuse envie de lire à nouveau *Héros oubliés du rock'n roll* ou *Confessions d'un chasseur d'opium*.

Les amis, la famille m'ont aussi apporté beaucoup de soutien dans cette histoire, qu'il me soit permis de les remercier ici.

A tout seigneur tout honneur.
Mulţumesc Laurici şi Kenneth. Vous avez quelque part initié le mouvement, fait route avec moi et même aidé à manier les rames.

Ma trinité se décompose en M'sieur Alex, M'sieur Alain et M'sieur Davy. Merci pour vos conseils avisés quand mes pensées arpentaient un ruban de Moebius.

Je souris à mes soeurs Hélène et Marie-Pierre. Je n'oublie pas la belle Emma, ses geôliers et ses vénérables ancêtres Monique et Christian, ainsi que Serge qui a troqué le soleil de la drôme pour la bise du Luxembourg le temps d'un week-end.

Parlant de famille, je pense de suite aux Arpendiens. Sans être un inconditionnel de la série, je ne vois d'autre choix que de suivre un format de type *friends*.
Merci à
– Celui qui, de la perfide Albion, passait sous le manteau des articles traitant de cinétique chimique – Celui qui enseignait la théorie de la relaxation par les chats – Celui qui savait démontrer que la vie est toujours plus chaotique que ce que l'on pense – Celui qui discutait jusqu'à deux-trois heures du matin même si le train pour Auxerre est à huit – Celui qui parlait peu, mais recevait comme un prince, sous le regard d'un Lion – Celui dont l'humour était dévastateur – Celui qui se la dorait au soleil –

Merci Céline de parler aux types qui jonglent aux arrêts de bus et pour ton amitié.

Paul, Dylan et Prisque Busvelle forment un comité d'accueil d'une qualité à forger les légendes.

Merci à la danseuse aux yeux aigue-marine, aux chanteurs, au *happy living room band*, aux mentors *violonistiques*, bouilleurs de cru distillant l'enthousiasme.

Il n'y a qu'un authentique *Spritz* pour effacer tout souvenir d'une campagne simulations désastreuse. Merci Maria.

Last but so far from least. Anne-Marie, dame de l'été aux reflets de jais, Amandine, dame de l'hiver aux reflets d'or et Pierrick, le géant modéhen forment le gang de la rue de Neudorf. Une équipe dont on peut douter de l'existence quand on sait qu'on lui demande de vivre avec quelqu'un qui passe une partie de son temps dans les astres, n'écoute pas quand lui parle et *crincrinte* usuellement entre vingt-deux heures et minuit.
Me croirez-vous ?
Ils existent !

Abstract

Keywords - nonlinear observers, nonlinear systems, extended Kalman filter, adaptive high-gain observer, Riccati equation, continuous-discrete observer, DC-motor, real-time implementation.

The work concerns the "observability problem" — the reconstruction of a dynamic process's full state from a partially measured state— for nonlinear dynamic systems. The Extended Kalman Filter (EKF) is a widely-used observer for such nonlinear systems. However it suffers from a lack of theoretical justifications and displays poor performance when the estimated state is far from the real state, e.g. due to large perturbations, a poor initial state estimate, etc...

We propose a solution to these problems, the Adaptive High-Gain (EKF).

Observability theory reveals the existence of special representations characterizing nonlinear systems having the observability property. Such representations are called observability normal forms. A EKF variant based on the usage of a single scalar parameter, combined with an observability normal form, leads to an observer, the High-Gain EKF, with improved performance when the estimated state is far from the actual state. Its convergence for any initial estimated state is proven. Unfortunately, and contrary to the EKF, this latter observer is very sensitive to measurement noise.

Our observer combines the behaviors of the EKF and of the high-gain EKF. Our aim is to take advantage of both efficiency with respect to noise smoothing and reactivity to large estimation errors. In order to achieve this, the parameter that is the heart of the high-gain technique is made adaptive. *Voila*, the Adaptive High-Gain EKF.

A measure of the quality of the estimation is needed in order to drive the adaptation. We propose such an index and prove the relevance of its usage. We provide a proof of convergence for the resulting observer, and the final algorithm is demonstrated via both simulations and a real-time implementation. Finally, extensions to multiple output and to continuous-discrete systems are given.

Résumé

Mots clefs - observateurs non linéaires, systèmes non linéaires, filtre de Kalman étendu, observateur à grand-gain adaptatif, équation de Riccati, observateur continu/discrèt, moteur DC, implémentation temps réel.

Le travail porte sur la problématique de "l'observation des systèmes" — la reconstruction de l'état complet d'un système dynamique à partir d'une mesure partielle de cet état. Nous considérons spécifiquement les systèmes non linéaires. Le filtre de Kalman étendu (EKF) est l'un des observateurs les plus utilisés à cette fin. Il souffre cependant d'une performance moindre lorsque l'état estimé n'est pas dans un voisinage de l'état réel. La convergence de l'observateur dans ce cas n'est pas prouvée.

Nous proposons une solution à ce problème : l'EKF à grand gain adaptatif.

La théorie de l'observabilité fait apparaître l'existence de représentations caractérisant les systèmes dit observables. C'est la forme normale d'observabilité. L'EKF à grand gain est une variante de l'EKF que l'on construit à base d'un paramètre scalaire. La convergence de cet observateur pour un système sous sa forme normale d'observabilité est démontrée pour toute erreur d'estimation initiale. Cependant, contrairement à l'EKF, cet algorithme est très sensible au bruit de mesure.

Notre objectif est de combiner l'efficacité de l'EKF en termes de lissage du bruit, et la réactivité de l'EKF grand-gain face aux erreurs d'estimation. Afin de parvenir à ce résultat nous rendons adaptatif le paramètre central de la méthode grand gain. Ainsi est constitué l'EKF à grand gain adaptatif.

Le processus d'adaptation doit être guidé par une mesure de la qualité de l'estimation. Nous proposons un tel indice et prouvons sa pertinence. Nous établissons une preuve de la convergence de notre observateur, puis nous l'illustrons à l'aide d'une série de simulations ainsi qu'une implémentation en temps réel dur. Enfin nous proposons des extensions au résultat initial : dans le cas de systèmes multi-sorties et dans le cas continu-discret.

Résumé Étendu

Dans ce travail nous abordons le problème de la synthèse d'observateurs pour les systèmes non linéaires.

Un système de ce type est défini par un ensemble d'équations de la forme:

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

où

- t est la variable de temps,
- $x(t)$ est la variable d'état, de dimension n,
- $u(t)$ est l'entrée du système, ou variable de contrôle, de dimension n_u,
- $y(t)$ est la sortie, ou mesure, de dimension n_y,
- f, h, sont des applications dont au moins une est non linéaire.

Un observateur est un algorithme qui assure la reconstruction, ou estimation, de la variable $x(t)$ sur la base de données partielles : la mesure $y(t)$ et l'entrée $u(t)$. Cet état estimé sera par la suite utilisé à des fins de contrôle du système, ou de supervision comme illustré en Figure 1, où $z(t)$ désigne l'état estimé.

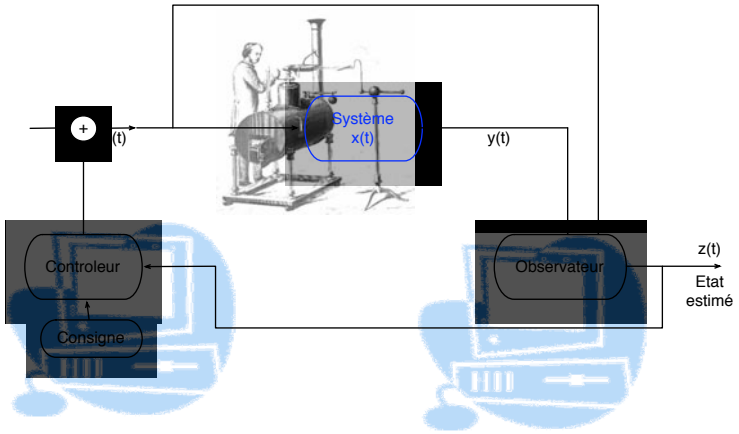

Figure 1: Boucle de contrôle

La mise au point et l'étude d'observateurs peuvent être décomposées en trois sous-problèmes :

– *Le problème d'observabilité* se penche sur les équations constituant le système. Ce modèle permet-il la reconstruction effective de l'état à partir de ses mesures ?
– *Le problème de convergence* se concentre sur l'observateur lui-même. L'état estimé converge-t-il vers l'état réel, et à quelle vitesse (i.e. exponentielle, polynomiale...) ?
– *Le problème de la fermeture de la boucle de contrôle* s'intéresse à la stabilité d'un algorithme de contrôle lorsque celui-ci s'appuie sur l'estimation que lui fournit un observateur.

Ces trois problèmes, bien connus dans le cas des systèmes linéaires, font l'objet d'une intense activité de recherche pour ce qui est des systèmes non linéaires. Le texte que nous présentons porte sur la mise au point d'un observateur et la résolution du second des problèmes énoncés plus tôt qui lui est associé.

Motivations

Notre observateur prototype est le filtre de Kalman étendu (EKF) pour lequel [43, 58, 68] constituent de bonnes entrées en matière. Travailler avec l'EKF est particulièrement intéressant car

– il est très répandu auprès des ingénieurs,
– sa mise en oeuvre pratique se fait de manière très naturelle,
– il est particulièrement adapté à une utilisation en temps réel,
– il possède de bonnes performance de lissage du bruit de mesure [101].

Un problème connu de ce filtre est son manque de garanties de stabilité. Sa convergence[1] n'est prouvée que localement, c'est-à-dire lorsque l'état estimé se trouve dans un voisinage de l'état réel.

Une garantie de convergence globale peut être obtenue en ayant recours à la méthodologie *grand gain* telle que proposée par Gauthier *et al.* [47, 54, 57]. Cette approche repose sur deux composantes:

1. l'usage d'une représentation du système non linéaire considéré, caractéristique de l'*observabilité*[2] du système,
2. une modification de l'algorithme EKF reposant sur l'usage d'un unique paramètre scalaire que l'on dénote usuellement θ.

L'observateur obtenu est appelé filtre de Kalman étendu grand gain (HGEKF). Pourvu que le paramètre θ soit fixé à une valeur assez grande, alors cet algorithme est convergent quelle que soit l'erreur d'estimation initiale. Il est de cette manière

[1] On entend par *convergence de l'observateur* le fait que l'état estimé tende asymptotiquement vers l'état réel.

[2] L'observabilité est une propriété (intrinsèque) d'un système indiquant qu'il est possible de distinguer deux état distinct à partir des sorties correspondantes.

robuste aux grandes perturbations que le système pourrait essuyer (par exemple un changement soudain de la variable d'état dû à une erreur de fonctionnement comme dans [64]). Ce paramètre permet de plus de régler la vitesse de convergence : plus il est choisi grand et plus la convergence est rapide. Cependant, comme θ est multiplié à y, la variable de sortie, il en amplifie d'autant le bruit de mesure. Il arrive donc que pour un signal de sortie fortement bruité il soit impossible d'utiliser efficacement le HGEKF.

Notre objectif est de proposer un observateur de type filtre de Kalman réunissant les avantages de l'EKF et du HGEKF. Nous allons faire appel à la structure adéquate en fonction de nos besoins. Pour ce faire, nous devons:

1. proposer une manière d'estimer à quel moment il est nécessaire de changer la configuration de l'observateur,
2. proposer un mécanisme d'adaptation,
3. prouver la convergence globale de l'observateur, montrer que celle-ci peut être réalisée en un temps arbitrairement court.

Forme normale d'observabilité entrées multiples, simple sortie

Nous situons notre approche dans le cadre de la théorie de l'observabilité déterministe de Gauthier *et al.*, présentée dans [57] et [40]. Une rapide revue des principales définitions et résultats de cette théorie est faite dans le Chapitre 2 de cette thèse. Nous nous contentons ici de définir la forme normale d'observabilité pour les systèmes à entrées multiples et simple sortie (MISO). Ce choix est fait de sorte à ce que l'exposé conserve toute sa clarté.

L'observateur, ainsi que les notions que nous présentons, restent valable pour les systèmes à sorties multiples (MIMO) pourvu que l'algorithme soit adapté en conséquence. En effet, contrairement aux systèmes MISO, pour qui la forme normale est essentiellement unique, il existe plusieurs formes d'observabilité MIMO. La description de l'observateur pour une représentation à sorties multiples est donnée au Chapitre 5 de ce texte.

Nous considérons les systèmes de la forme:

$$\left\{ \begin{array}{rcl} \frac{dx}{dt} & = & A\left(u\right)x + b\left(x,u\right) \\ y & = & C\left(u\right)x \end{array} \right. \tag{1}$$

où $x\left(t\right) \in \mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}$ compact, $y\left(t\right) \in \mathbb{R}$, et $u(t) \in \mathcal{U}_{\text{adm}} \subset \mathbb{R}^{n_u}$ est borné pour tout $t \geq 0$.

Les matrices $A(u)$ et $C(u)$ sont définies par:

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

$$C(u) = \begin{pmatrix} a_1(u) & 0 & \cdots & 0 \end{pmatrix},$$

où $0 < a_m \leq a_i(u) \leq a_M$ pour tout $u \in \mathcal{U}_{\text{adm}}$. Le champ de vecteurs $b(x,u)$ est supposé avoir une structure triangulaire de la forme

$$b(x,u) = \begin{pmatrix} b_1(x_1,u) \\ b_2(x_1,x_2,u) \\ \vdots \\ b_n(x_1,\ldots,x_n,u) \end{pmatrix}$$

et être à support compact. L_b est la borne supérieure de $b^*(x,u)$, la matrice Jacobienne de $b(x,u)$ dans le sens : $\|b^*(x,u)\| \leq L_b$. Comme $b(x,u)$ est à support compact et que u est borné, b est Lipschitz par rapport à x, uniformément par rapport à u : $\|b(x_1,u) - b(x_2,u)\| \leq L_b \|x_1 - x_2\|$.

Nous rappelons au lecteur qu'en dépit des apparences un tel système n'est pas une singularité. Il caractérise la propriété d'observabilité: pour tout système observable simple sortie, il existe un changement de coordonnées permettant de l'écrire sous la forme du système (1). Ainsi il semble naturel d'utiliser un observateur sur un système observable.

Filtre de Kalman étendu à grand gain adaptatif

Nous présentons ici le filtre de Kalman étendu à grand gain adaptatif (AEKF), de plus amples explications sont disponibles au Chapitre 3. Sa propriété de convergence est explicitée par le Théorème IV et les Lemmes II et V ci-dessous. Les définitions précises de l'EKF et du HGEKF sont données dans le Chapitre 2 de la thèse. Les théorèmes explicitant leurs propriétés de convergence y sont aussi rappelés. A la fin de ce même chapitre différentes stratégies de type adaptatives sont aussi passées en revue.

Définition I

Soient

- Q *une matrice réelle* $(n \times n)$ *symétrique définie positive et*
- R *et* θ *deux réels positifs, où* $\theta \geq 1$.

Nous définissons les matrices

$$\Delta = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\theta^{n-1}} \end{pmatrix},$$

$$Q_\theta = \theta \Delta^{-1} Q \Delta^{-1},$$

et

$$R_\theta = \theta^{-1} R.$$

Le ***filtre de Kalman étendu à grand gain adaptatif*** *est le système:*

$$\begin{cases} \frac{dz}{dt} &= A(u)z + b(z,u) - S^{-1}C'R_\theta^{-1}(Cz - y(t)) \\ \frac{dS}{dt} &= -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'R_\theta^{-1}C - SQ_\theta S \\ \frac{d\theta}{dt} &= \mathcal{F}(\theta, \mathcal{I}_d(t)) \end{cases} \quad (2)$$

avec pour conditions initiales : $z(0) \in \chi$, $S(0)$ symétrique définie positive, et $\theta(0) = 1$. La fonction $\mathcal{F}$ constitue le mécanisme d'adaptation de l'observateur. Nous la définissons par ses propriétés, récapitulées au Lemme V.

$\mathcal{I}_d$ nous sert à estimer la qualité de l'estimation rendue par l'observateur. Cette quantité est calculée comme suit:
*pour une "longueur d'horizon" $d > 0$, l'****innovation*** *est:*

$$\mathcal{I}_d(t) = \int_{t-d}^{t} \|y(t-d, x(t-d), \tau) - y(t-d, z(t-d), \tau)\|^2 \, d\tau \quad (3)$$

où $y(t_0, x_0, \tau)$ est la sortie du système (1) au temps τ avec pour condition initiale $x(t_0) = x_0$.

Par conséquent $y(t-d, x(t-d), \tau)$ n'est autre que $y(\tau)$, la sortie du système (1). Nous insistons sur le fait que $y(t-d, z(t-d), \tau)$ n'est pas la sortie de l'observateur.

L'importance de cette quantité est explicitée par le lemme suivant. Il apparaît aussi dans la preuve de convergence du Théorème IV (voir la preuve du Théorème 36 au Chapitre 3) que ce résultat est la pierre angulaire du mécanisme d'adaptation.

Lemme II

Soient $x_1^0, x_2^0 \in \mathbb{R}^n$, et $u \in \mathcal{U}_{\text{adm}}$. $y(0, x_1^0, \cdot)$ et $y(0, x_2^0, \cdot)$ sont les trajectoires de sortie du système (1) avec conditions initiales x_1^0 et x_2^0, respectivement. Alors la propriété suivante (dite "observabilité persistante") est vraie :

$$\forall d > 0, \exists \lambda_d^0 > 0 \text{ tel que } \forall u \in L_b^1(\mathcal{U}_{adm})$$

$$\|x_1^0 - x_2^0\|^2 \leq \frac{1}{\lambda_d^0} \int_0^d \|y\left(0, x_1^0, \tau\right) - y\left(0, x_2^0, \tau\right)\|^2 d\tau. \tag{4}$$

Remarque III

1. *Si l'on considère que* $x_1^0 = z(t-d)$, *et que* $x_2^0 = x(t-d)$ *alors le Lemme II nous indique que :*

$$\|z(t-d) - x(t-d)\|^2 \leq \frac{1}{\lambda_d^0} \int_{t-d}^{t} \|y(\tau) - y\left(t-d, z\left(t-d\right), \tau\right)\|^2 d\tau,$$

ou, de manière équivalente,

$$\|z(t-d) - x(t-d)\|^2 \leq \frac{1}{\lambda_d^0} \mathcal{I}_d(t).$$

c'est-à-dire que modulo la multiplication par un paramètre constant, l'innovation au temps t *est borne supérieure de l'erreur d'estimation au temps* $t-d$.

2. *L'innovation est définie de manière plus détaillée dans la Section 3.3 du Chapitre 3. Son implémentation est explicitée au Chapitre 4.*

Le théorème suivant est le résultat au coeur de cette thèse : le lemme précédent le sert, les différentes applications et extensions en découlent.

Théorème IV

Pour tout temps arbitraire $T^* > 0$ *et tout* $\varepsilon^* > 0$, *il existe* $0 < d < T^*$ *et une fonction d'adaptation* $\mathcal{F}(\theta, \mathcal{I}_d)$ *telle que décrite au Lemme V, de sorte que pour tout* $t \geq T^*$ *et n'importe quel couple de points* $(x_0, z_0) \in \chi^2$*:*

$$\|x(t) - z(t)\|^2 \leq \varepsilon^* \mathrm{e}^{-a\left(t-T^*\right)}$$

où $a > 0$ *est une constante (indépendante de* ε^**).*

Lemme V *(La fonction d'adaptation)*

Pour tout $\Delta T > 0$, *il existe une constante* $M(\Delta T)$ *telle que :*

- *pour tout* $\theta_1 > 1$, *et*
- *tout couple* $\gamma_1 \geq \gamma_0 > 0$,

il existe une fonction $\mathcal{F}(\theta, \mathcal{I})$ *de sorte que l'équation*

$$\dot{\theta} = \mathcal{F}\left(\theta, \mathcal{I}\left(t\right)\right), \tag{5}$$

où $1 \leq \theta(0) < 2\theta_1$, *et* $\mathcal{I}(t)$ *est une fonction positive et mesurable, a les propriétés suivantes:*

1. *il existe une unique solution* $\theta(t)$, *vérifiant* $1 \leq \theta(t) < 2\theta_1$, *pour tout* $t \geq 0$,
2. $\left|\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}\right| \leq M$,
3. *si* $\mathcal{I}(t) \geq \gamma_1$ *pour* $t \in [\tau, \tau + \Delta T]$ *alors* $\theta(\tau + \Delta T) \geq \theta_1$,
4. *tant que* $\mathcal{I}(t) \leq \gamma_0$, $\theta(t)$ *décroît vers* 1.

Résultats de simulation et implémentation temps réel

La mise en oeuvre pratique de l'observateur pour un moteur à courant continu et connecté en série est décrite en détail au Chapitre Chapitre 4. Le modèle est obtenu en réalisant d'une part l'équilibre des forces électriques – i.e. à partir de la représentation sous forme de circuit électrique, voir Figure 2 – et d'autre part l'équilibre mécanique – i.e. loi de Newton. L'état du système obtenu est de dimension 2, l'entrée et la sortie sont de dimension 1:

$$\left\{ \begin{array}{ccc} \begin{pmatrix} L\dot{I} \\ J\dot{\omega}_r \end{pmatrix} & = & \begin{pmatrix} u - RI - L_{af}\omega_r I \\ L_{af}I^2 - B\omega_r - T_l \end{pmatrix} \\ y & = & I \end{array} \right. \tag{6}$$

où

- I et ω_r sont les variables d'état : l'intensité du courant et la vitesse de rotation respectivement,
- $u(t)$ est la quantité de courant en entrée,
- R est la somme des résistances du circuit,
- L est la somme des inductances du circuit,
- L_{af} est l'inductance mutuelle,
- J est l'inertie du système,
- B est le coefficient de frottement de l'axe du moteur, et
- T_l est le couple de charge.

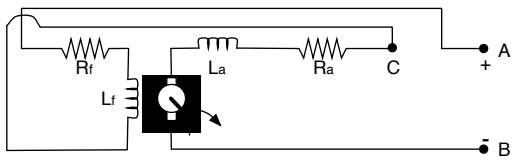

Figure 2: Moteur DC : circuit équivalent

La mise en oeuvre de l'observateur requiert trois étapes :

- l'analyse d'observabilité et la mise en lumière de la forme canonique,
- la définition de la fonction d'adaptation,
- le réglage des différents paramètres.

Observabilité

L'analyse d'observabilité par la méthode différentielle (Cf. Section 4.1.2) montre que si I ne s'annule pas, alors ce système est observable[3]. Il est de plus possible,

[3]notez que si I est nul, cela veut dire qu'il n'y a pas de courant et que le moteur est soit à l'arrêt, soit en phase d'arrêt.

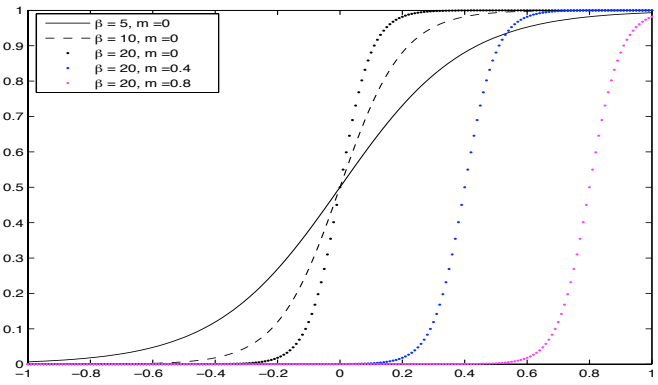

Figure 3: Influence de β et m sur la forme de la sigmoïde.

tout en conservant l'observabilité, d'estimer T_l, qui est une perturbation non mesurée. Pour ce faire, le modèle est étendu à trois équations en utilisant le modèle trivial : $\dot{T}_l = 0$.

Le changement de coordonnées:

$$\begin{array}{ccc} \mathbb{R}^{*+} \times \mathbb{R} \times \mathbb{R} & \rightarrow & \mathbb{R}^{*+} \times \mathbb{R} \times \mathbb{R} \\ (I, \omega_r, T_l) & \hookrightarrow & (x_1, x_2, x_3) = (I, I\omega_r, IT_l) \end{array}$$

transforme le système en la forme normale d'observabilité

$$\begin{pmatrix} \dot{x_1} \\ \dot{x_2} \\ \dot{x_3} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{L_{af}}{L} & 0 \\ 0 & 0 & -\frac{1}{J} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \frac{1}{L}(u(t) - Rx_1) \\ -\frac{B}{J}x_2 + \frac{L_{af}}{J}x_1^3 - \frac{L_{af}}{L}\frac{x_2^2}{x_1} + \frac{u(t)}{L}\frac{x_2}{x_1} - \frac{R}{L}x_2 \\ -\frac{L_{af}}{L}\frac{x_2 x_3}{x_1} + \frac{u(t)}{L}\frac{x_3}{x_1} - \frac{R}{L}x_3 \end{pmatrix}. \tag{7}$$

La fonction d'adaptation

L'observateur présenté dans la section précédente est complété en exprimant explicitement sa fonction d'adaptation:

$$\mathcal{F}(\theta, \mathcal{I}_d) = \mu(\mathcal{I}_d)\mathcal{F}_0(\theta) + (1 - \mu(\mathcal{I}_d))\lambda(1 - \theta) \tag{8}$$

où

- $\mathcal{F}_0(\theta) = \begin{cases} \frac{1}{\Delta T}\theta^2 & pour \theta \leq \theta_1 \\ \frac{1}{\Delta T}(\theta - 2\theta_1)^2 & pour \quad \theta > \theta_1 \end{cases}$,
- $\mu(\mathcal{I}) = \left[1 + e^{-\beta(\mathcal{I} - m)}\right]^{-1}$ est une fonction sigmoïde paramétrée par β et m (Cf. Figure 3).

Réglage des paramètres

La procédure d'adaptation implique l'usage de nombreux paramètres de réglages. La première impression est celle d'une certaine confusion face aux choix à réaliser. Il est en fait possible de régler tous ces paramètres un à un, selon un ordre logique. Cet ordre est illustré par la Figure 4. La procédure complète est détaillée dans la Section 4.2.3, nous en donnons un aperçu ci-dessous.

L'ensemble des paramètres se divise en deux groupes : les paramètres relatifs à la performance de chacun des deux modes de l'observateur, et ceux relatifs à la procédure d'adaptation.

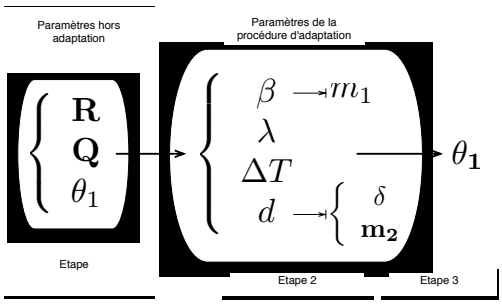

Figure 4: Gras : paramètres cruciaux.

Paramètres de performance

Tout l'intérêt de l'approche adaptative que nous développons est de pouvoir découpler l'influence de Q et R d'un côté, et de θ de l'autre, sur la performance de l'observateur. Dans un filtre de Kalman étendu à grand gain, les matrices Q et R perdent leur sens premier car elles sont, de manière définitive, changées à cause de leur multiplication par θ. Lorsque l'observateur n'est pas en mode grand gain, ces deux matrices retrouvent leur sens classique. Il faut donc les régler de sorte à ce que le bruit de mesure soit filtré de manière optimale. Nous pouvons pour cela faire appel à des méthodes classiques [58].

Il en va de même pour θ, le paramètre de grand gain : seule sa performance de convergence nous intéresse. θ est choisi comme il est habituel de le faire pour les observateurs de type grand gain. On prendra toutefois soin de limiter les *"overshoots"* au maximum.

Paramètres d'adaptation

Parmi l'ensemble des paramètres d'adaptation seuls d, la longueur de la fenêtre de calcul de l'innovation et m, le second paramètre de la sigmoïde, doivent être modifiés d'un observateur à l'autre. Les autres pourront être fixés une fois pour toutes quel que soit le procédé étudié.

Il faut être conscient de l'influence du bruit de mesure sur le calcul de l'innovation. En effet, lorsque l'observateur estime parfaitement l'état du système le calcul de l'innovation n'est constitué que de l'intégration du bruit de mesure. Conséquemment si la procédure d'adaptation devait être déclenchée dès que l'innovation est non nulle alors θ serait toujours grand. La sigmoïde est décalée sur la droite par l'usage du paramètre m. Nous proposons de calculer m à partir d'une estimation de la déviation standard du bruit de mesure. Une méthode de calcul efficace est donnée en Section 4.2.3.

Si d est trop petit alors l'information contenue dans l'innovation sera noyée dans le bruit. Si d est au contraire trop grand alors le temps de calcul devient rédhibitoire.

Démonstration de performance

Les Figures 5, 6, 7 donnent un aperçu de la performance de l'observateur lors d'un scénario simple. Ces figures sont commentées en Section 4.2.4, où les résultats donnés par un scénario plus complexe sont aussi montrés. Une série de courbes renseigne sur les effets d'un mauvais réglage du paramètre m.

On remarquera que le comportement hybride recherché, lissage du bruit équivalent à celui d'un filtre de Kalman étendu (courbe bleu foncé) et vitesse de convergence comparable à un filtre de Kalman étendu à grand gain (courbe bleu ciel), est atteint.

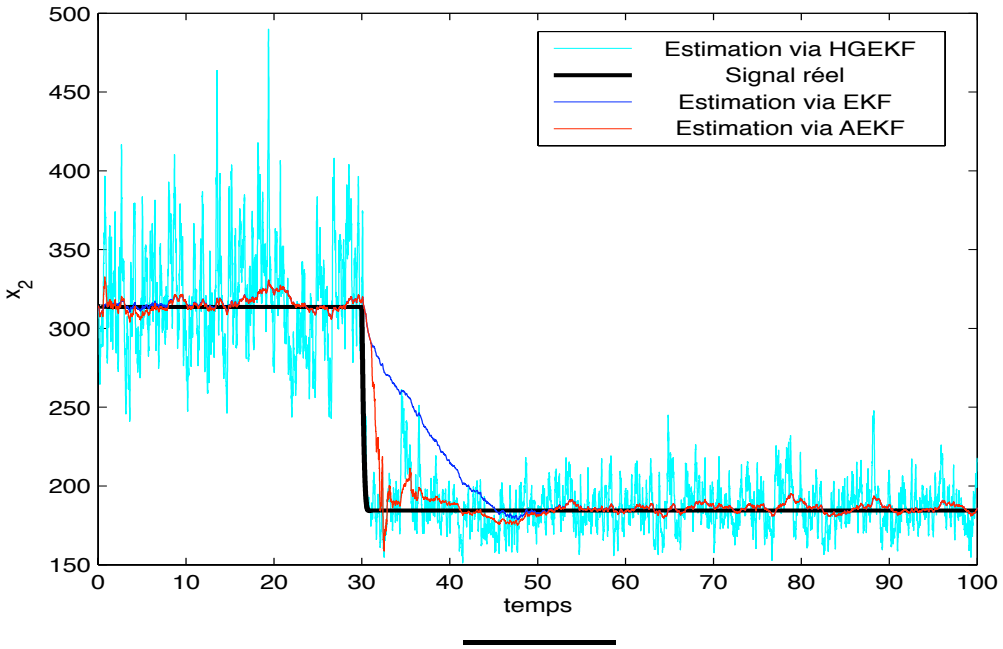

Figure 5: Estimation de la vitesse de rotation.

La seconde moitié du Chapitre 4 est consacrée à la description détaillée de la programmation de l'observateur dans un environnement temps réel dur. Un tel

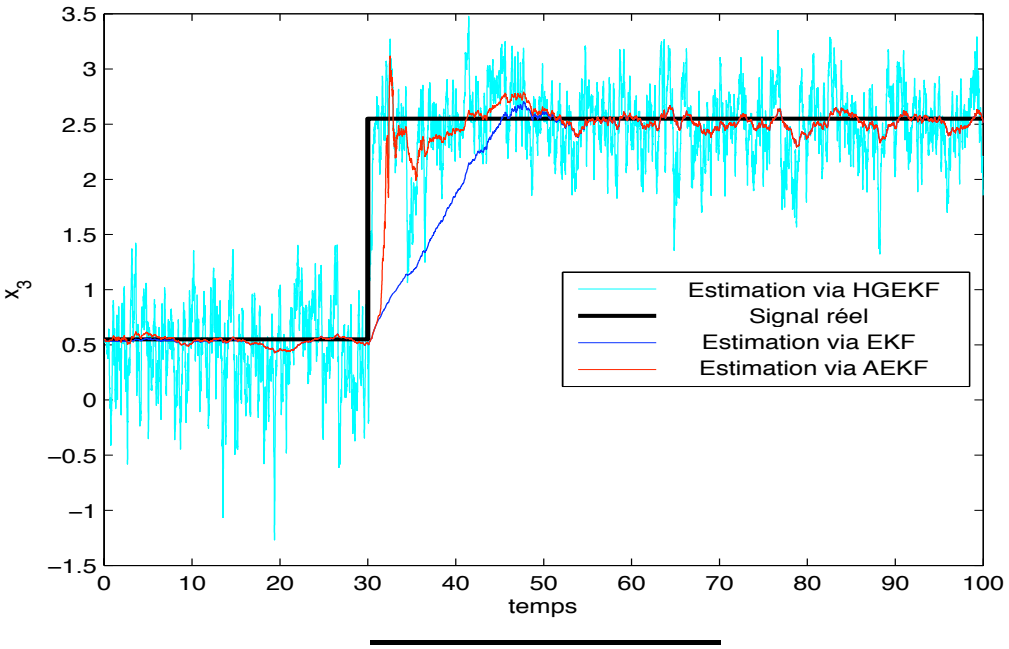

Figure 6: Estimation du couple de charge.

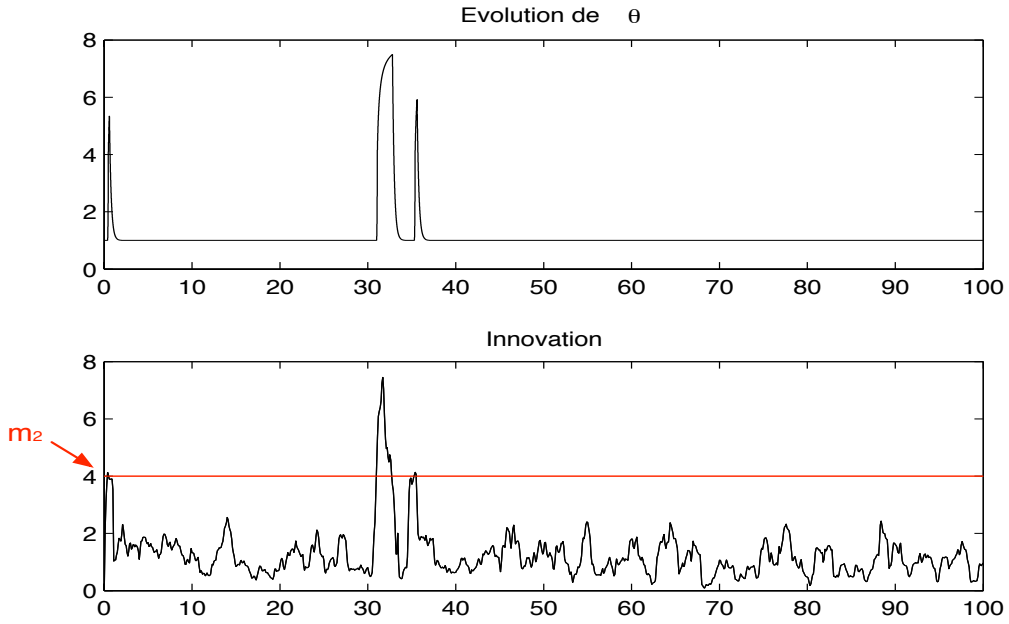

Figure 7: Grand gain et Innovation.

système d'exploitation force l'utilisateur à programmer son algorithme de sorte à respecter les contraintes de temps de calcul qui lui sont imposées – i.e. les calculs non terminés sont interrompus. Nous y démontrons que l'observateur peut être utilisé dans un environnement contraignant, sur un moteur réel : il fonctionne de belle manière à une vitesse d'échantillonnage de 100 Hz.

Extensions

Pour conclure nous proposons deux extensions au théorème principal.

La première concerne les systèmes à sorties multiples. Il n'existe pas de forme d'observabilité unique pour ce type de systèmes, ainsi l'observateur doit être adapté à chaque nouvelle forme rencontrée. Nous avons choisi une généralisation du système 1 pour laquelle nous avons décrit en détail les modifications à apporter. Ces modifications sont tout à fait cohérentes avec la définition originale de l'observateur dans le sens où si la sortie se résume à 1 variable alors la configuration redevient celle donnée initialement. Nous montrons en détails les points de la preuve conduisant à une modification de l'algorithme de sorte à inspirer de nouvelles structures pour systèmes à sorties multiples.

La seconde extension porte sur les systèmes de type continus/discrets. Ces systèmes constituent une description du procédé plus proche de la réalité dans le sens où la dynamique est décrite de façon continue et les mesures sont modélisées par un processus discret. Nous simulons ainsi le système de capteurs de manière réaliste. Nous proposons un observateur adapté et montrons que l'une des hypothèses sur la fonction d'adaptation peut être relâchée. Ceci nous permet de concevoir l'usage de fonctions algébriques au lieu de fonctions différentielles, évacuant ainsi la question du retard dû au temps de montée du paramètre grand gain à dynamique continue. Pour mener à bien la preuve de convergence de cet observateur, nous avons été amenés à développer une preuve de l'existence de bornes pour l'équation de Riccati, qui était jusqu'ici absente de la littérature.

En guise de conclusion, nous précisons que ce travail a permis la rédaction:

- d'un article de journal, accepté pour publication dans *Automatica* [22],
- d'un chapitre de livre, écrit à l'occasion de l'école d'été d'automatique de Grenoble 2007 [21],
- de communications lors de trois conférences internationales [23–25],
- d'une intervention pour un workshop *Linux Realtime* [26].

Contents

List of Figures

List of Tables

Nomenclature

χ	a compact subset of the state space
δ_t	sampling time
$\dot{x}(t)$	the time derivative of the time variable $x(t)$: $\frac{dx}{dt}(t)$
$\mathcal{F}$	adaptation function
$\mathcal{I}_d(t)$	innovation for a window of length d, at time t (continuous time framework)
$\mathcal{I}_{d,k}$	innovation for a window of length $d\delta_t$, at epoch k (discrete time framework)
θ, $\theta(t)$, θ_k	high-gain parameter
$A^{'}$	transpose of the matrix A
$diag(v)$	a $dim(v) \times dim(v)$ matrix, such that $(diag(v))_{i,j} = \delta_{ij} v_i$
$L_b^1(\mathcal{U}_{adm})$	the set of integrable (thus measurable), bounded functions having their values in $\mathcal{U}_{adm}$.
n	dimension of the state space
n_u	dimension of the input space
n_y	dimension of the output space
$P(t)$	$S^{-1}(t)$
$S(t)$	Riccati matrix
t	time variable
$u(t)$	input space
X	a n-dimensional analytic differentiable manifold
$x(t)$	state space
$y(t)$	output space
$z(t)$	estimated state (continuous time framework)

z_k	estimated state (discrete time framework)
AEKF	adaptive high-gain extended Kalman filter
EKF	extended Kalman filter
HGEKF	high-gain extended Kalman filter
MISO	multiple input multiple output
MISO	multiple input single output
SISO	single input single output
w.r.t.	with respect to

Chapter 1

Introduction

Juggling. Balls fly in thin air, tracing parabolas. The juggler sees, and reacts. Eyes move, muscles follow, and the improbable act of using only two hands to keep three, four, five or even more balls in the air continues, non-stop.

It goes without saying that the juggler has both eyes open. Why, though? With one eye closed, the juggling becomes harder, as depth perception becomes di?cult and field of view more limited. Juggling with one eye closed can be done, but only at the cost of training and perseverance.

What about juggling blindfolded? Only the best jugglers can do this by inferring the position of each ball from their hands and the accuracy of their throws. This is because they know what the parabola *should* look like.

When we rely on partial measurements to reconstruct the state of a physical process, we are like blindfolded jugglers. We only have limited information to tell us the di?erence between what we think the state is and what it actually is. In process control, the tool dedicated to the state reconstruction task is known as an *observer*.

An observer is a mathematical algorithm that estimates the state of a system. What makes the observer special is that it does not guess. It infers, based on a prediction of what it expects to measure and a correction driven by the difference between the *predicated* and the *measured*.

The concept of observer is introduced in this chapter with only little mathematical formalism. We provide an idea of the ma jor issues of the field and put forward the motivations underlying the present work.

1.1 Context

We can model any physical process by considering it as a system. In our framework, a set of relevant variables describes the evolution of the state of the system with time. These are called the *state variables*. The system interacts with the outside world in three different ways:

- *input, or control, variables* are quantities that have an effect on the system behavior and that can be set externally, they are denoted $u(t)$,
- *output variables, or measurements,* are quantities that are monitored, generally they are a subset or a transformation of the state variables, we denote them by $y(t)$,
- *perturbations* are variables that have an effect on the system behavior and that cannot be controlled; most of the time they cannot be measured[1].

The state variables are represented by a multidimensional vector, denoted $x(t)$. The evolution of $x(t)$ with time is accounted for by an ordinary differential equation[2]:

$$\frac{dx(t)}{dt} = f(x(t), u(t), t).$$

The relation between state and output variables, i.e. the measurement step, is described by an application:

$$y(t) = h(x(t), u(t), t).$$

A system is therefore defined as a set of two equations of the form:

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

The state estimate rendered by the observer can be used for monitoring purposes, by a control algorithm as schematized in Figure 1.1, or processed off-line (e.g. in prototyping assessment as in [21], section 3.6).

For example, in the juggling experiment, the state variables are the 3D position and speed of the balls. The input variables are the impulses the hands apply to the balls.

With eyes wide open, the output variables are the balls position. In the one eyed juggling experiment the output variables are an imperfect knowledge of the balls position, e.g. because of the hindered depth perception. In the blind juggling experiment, tactile signals are the output variable.

In all these three cases the model is the same: the one of a falling apple.

[1] From an observer's point of view, measured perturbations and control are both inputs. The controller's point of view is different since it uses the controls to stabilize the system and reject perturbations.

[2] We do not consider either partial differential equations or algebraic differential equations.

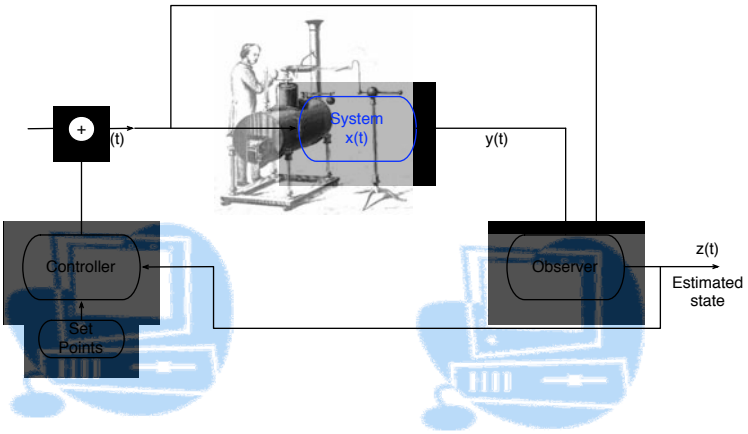

Figure 1.1: A control loop.

The study of observers is divided into three subproblems:

1. *the observability problem* is the study of a given mathematical model in order to determine to which extent it can be useful to estimate state variables,

2. *the convergence problem* focuses on the observer algorithm, the diminution of the error between the real and estimated state is studied and the convergence to zero assessed (see Figure 1.2),

3. *the loop closure problem* addresses the stability of closed control loops, when the state estimated by an observer is used by a controller[3] (see Figure 1.1).

The juggler solves the two first problems when he is capable of catching the balls having one or both eyes closed. He solves the third problem when he goes on juggling.

For systems that are described by linear equations, the situation is theoretically well known and answers to all those problems have already been provided (see for example [74]). We therefore concentrate on the nonlinear case.

There is a large quantity of observers available for nonlinear systems, derived from practical and/or theoretical considerations:

– *extended observers* are adaptations of classic linear observers to the nonlinear case (e.g. [58]),

– *adaptive observers* estimate both the state of the system and some of the model parameters; a linear model can be turned into one that is nonlinear in this configuration (e.g [98]),

– *moving horizon observers* see the estimation procedure as an optimization problem (e.g. [12]),

[3] A controller calculates input values that stabilize the physical process around a user defined set point or trajectory. The model used in the observer and the possible model used in the controller may not be the same.

– *interval observers* address models with uncertain parameters. Unlike adaptive observers they do not estimate the model parameters. They use the fact that any parameter p lives in an intervals of the form $[p_{min}, p_{max}]$ (e.g. [91, 105]).

The present work focuses on *high-gain observers*, which are a refinement of *extended observers*, allowing us to prove that the estimation error converges to zero in a global sense. The design of such observers relies on a general theory of observability for nonlinear systems [57]. It is proven that the convergence is exponential. This means that the estimation error is bounded on the upper limit by a decreasing exponential of time, whose rate of convergence can be set by the user. The loop closure problem is studied in [57], Chapter 7, and a *weak separation principle* is stated in [113] (see also [21]).

Our prototype observer is the Kalman filter. In order to further develop our motivations, we provide a review of high-gain observer algorithms in the next section. A complete discusion is the object of Section 2.5.

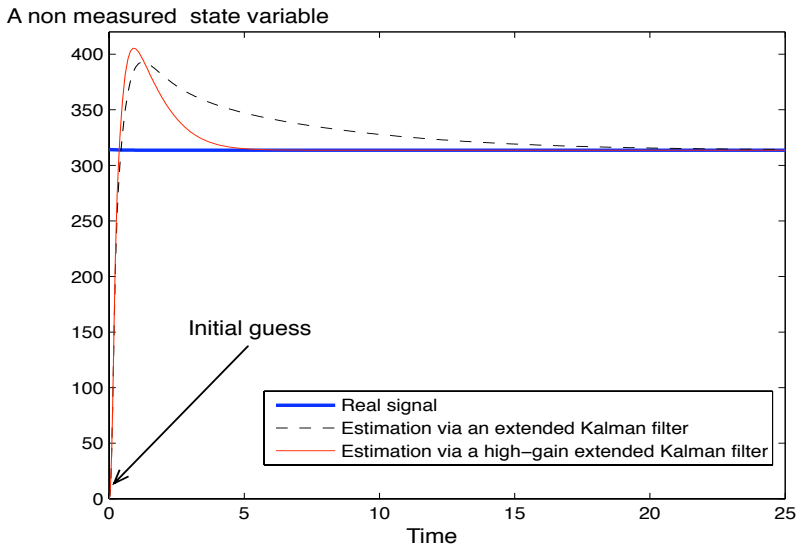

Figure 1.2: Estimation of a non measured variable:
The initial estimated state is wrong. With time, the observer reduces the estimation error.

1.2 Motivations

The *extended Kalman filter* is a practical answer to the observation problem for nonlinear systems. Although global convergence is not theoretically proven, this observer works well in practice. High-gain observers are based on extended observers, and result from the grouping of two main ingredients:

1. the use of a special representation of the nonlinear system, provided by the observability theory (see Chapter 2) and,

2. the use of a variant of extended observers.

A story is best told when started from the beginning. Thus we begin by introducing the Kalman filter in the linear case. A linear system is given by:

$$\begin{cases} \frac{dx(t)}{dt} &= Ax(t) + Bu(t), \\ y(t) &= Cy(t), \end{cases}$$

where A, B and C are matrices (having the appropriate dimensions) that may or may not depend on time. The two archetypal observers for such a system were proposed in the 1960's by D. G. Luenberger [86], R. E. Kalman. and B. S. Bucy [75, 76]. They are known as the Luenberger observer and the Kalman-Bucy filter respectively.

The leading mechanism in those two algorithms is the *prediction correction scheme*. The new estimated state is obtained by means of:

- a prediction based on the model and the old estimated state, and
- the correction of the obtained values by the measurement error weighted by a gain matrix.

We denote the estimated state by $z(t)$. The corresponding equation for the estimated state is:

$$\frac{dz(t)}{dt} = Az(t) + Bu(t) - K\left(Cz(t) - y(t)\right).$$

In the Luenberger observer, the matrix K is computed once and for all. The real part of all the eigenvalues of $(A - KC)$ have to be strictly negative.

In the Kalman filter, the gain matrix is defined as $K(t) = S^{-1}(t)C^{'}R^{-1}$ where S is the solution of the differential equation:

$$\frac{d}{dt}S(t) = -A^{'}S(t) - S(t)A - S(t)QS(t) + C^{'}R^{-1}C.$$

This equation is a Riccati equation of matrices and is referred to as *the Riccati equation*. The matrices A, B and C are expected to be time dependent (i.e. $A(t)$, $B(t)$ and $C(t)$). Otherwise, the Kalman filter is equivalent to the Luenberger observer.

The Kalman filter is the solution to an optimization problem where the matrices Q and R play the role of weighting coefficients (details can be found in [75]). These matrices must be symmetric definite positive. According to this definition, the Kalman filter is an optimal solution to the observation problem. We will see below that the Kalman filter has a stochastic interpretation that gives sense to those Q and R matrices.

Those two observers are well known algorithms, i.e. convergence can be proven. However, when the matrices A, B and C are time dependent, the Kalman filter has to be used since the gain matrix is constantly being updated.

Observability in the linear case is characterized by a simple criterion. The loop closure problem has an elegant solution called *the separation principle*: controller and observer can be designed independently and the overall loop remains stable. Interesting and detailed exposes on the Kalman filter can be found in [43, 45, 58].

Not all processes can be represented by linear relationships. Let us consider nonlinear systems of the form:

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

where f and h are nonlinear applications. The prediction correction strategy is applied as in the linear case. Since the correction gain is linearly computed via matrices, a linearization of the system is used.

This mathematical operation has to be done at some point in the state space. Since the real state of the system is unknown, it is difficult to pick such a point. The estimated state is, therefore, used. Consequently the correction gain matrix has to be constantly updated. The Kalman filter equations are considered[4]. Let us denote:

- z, as the estimated state,
- A, as the partial derivative of f with respect to x, at point (z,u,t) and,
- C, as the partial derivative of h with respect to x, at point (z,u,t).

This observer is then defined as:

$$\begin{cases} \frac{dz(t)}{dt} &= f(z,u,t) - S^{-1}C'R^{-1}\left(Cz-y\right) \\ \frac{d}{dt}S &= -A'S - SA - SQS + C'R^{-1}C \end{cases}.$$

This algorithm is called the *extended Kalman filter*.

The extended Kalman filter is an algorithm widely used in engineering sciences [58]. It works well in practice, which explains its popularity. There is, however, a lack of theoretical justification concerning its effectiveness. Indeed, except for small initial errors, there is no analytical proof of convergence for this observer. This is due to the linearization, which is performed along the estimated trajectory. The resulting matrices used in the Riccati equation are not correct enough outside of the neighborhood of the real trajectory. The system is poorly approximated and convergence cannot be proved.

A modification of this observer, the *high-gain extended Kalman filter*, solves the problem. Provided that the system is under a representation specific to the *observability property*[5], the observer converges exponentially. In other words, the estimation error is upper bounded by an exponential of time.

The main results of the observability theory for nonlinear systems are introduced in Chapter 2. The definitions of several high-gain observers are also included there.

Corruption of signals by noise is a major issue in engineering. Signals processed by an observer are derived from a sensor, which inherently implies the presence of noise. This influence appears both on the state and on the output variables. In the stochastic setting systems are then represented by equations of the form:

$$\begin{cases} dX(t) &= f(X(t),u(t),t)dt + Q^{\frac{1}{2}}dW(t) \\ dY(t) &= h(X(t),u(t),t)dt + R^{\frac{1}{2}}dV(t) \end{cases}$$

[4]The Lunenberger observer can be used for nonlinear systems, but necessitates a special representation of the nonlinear system that will be introduced in the next chapter.

[5]This observability property is related to the first problem, i.e. the observability problem.

where:

- $X(0)$ is a random variable,
- $X(t)$ and $Y(t)$ are random processes,
- $V(t)$ and $W(t)$ are two independent Wiener processes, also independent from $x(0)$ (refer to Appendix C.5).

In this context, Q is the covariance matrix of the state noise, and R is the covariance matrix of the measurement noise.

We consider again the linear case together with the two assumptions:

1. $x(0)$ is a gaussian random variable, and
2. state and output noises are gaussian white noises.

In this setting the Kalman filter is an estimator of the conditional expectation of the state, depending on the measurements available so far[6]. When the Q and R matrices of the observer defined above are set to the Q and R covariance matrices of the system, the noise is optimally filtered (refer to [43, 45]).

In practice noise characteristics are not properly known. Therefore the matrices Q and R of the observer are regarded as tuning parameters. They are adjusted in simulation.

In the nonlinear case the analysis in the stochastic setting consists of the computation of the conditional density of the random process $X(t)$. A solution to this problem is given by an equation known as the Duncan-Mortensen-Zakaï equation [80, 108]. It is rather complicated and cannot be solved analytically. Numerically, we make use of several approximations. One of the methods is called *particle filtering* (e.g. see [21, 42] for details). When we obfuscate the stochastic part of the problem, we obtain an observer with an elegant formulation: the extended Kalman filter. Let us focus on its stochastic properties.

As demonstrated in [101], the extended Kalman filter displays excellent noise rejection properties. However, proof of the convergence of the estimation error can be obtained only when the estimated state comes sufficiently close to the real state.

On the other hand, the high-gain extended Kalman filter is proven to globally converge (in a sense explained in subsequent chapters). However, it behaves poorly from the noise rejection point of view. Indeed, it has the tendency to amplify noise, thus resulting in a useless reconstructed signal. This problem is illustrated[7] in Figure 1.3.

1.3 Contributions

This dissertation concentrates on the fusion between the extended Kalman filter and the high-gain extended Kalman filter by means of an adaptation strategy. Our purpose is to merge the advantages of both structures:

[6] $z(t) = \mathbb{E}\left[x(t)/\mathcal{F}_t\right]$, where $\mathcal{F}_t = \sigma\left(y(s), s \in [0,t]\right)$, i.e. the $\sigma-$algebra generated by the output variables for $t \in [0;t]$.

[7] These two curves are obtained with the same observers as in Figure 1.2, the values of the tuning parameter are unchanged. The level of noise added to the output signal is the same in the two simulations.

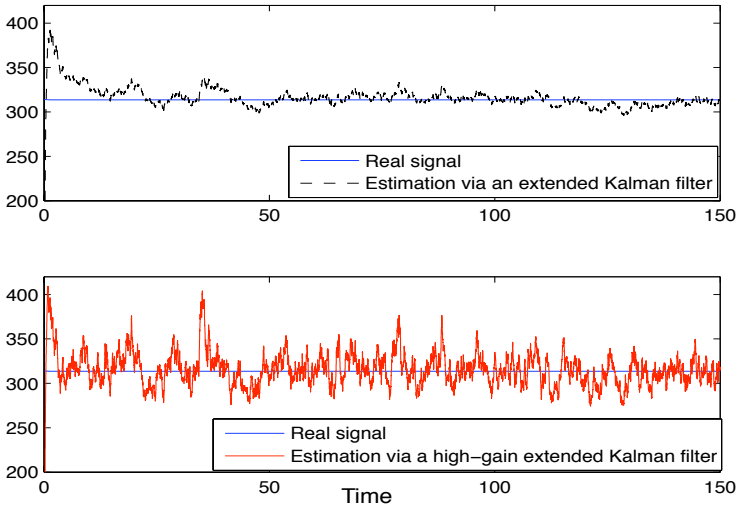

Figure 1.3: Estimation with noise.

– the noise smoothing properties of the extended Kalman filter when the estimated state is sufficiently close to the real state;

– the global convergence property of the high-gain extended Kalman filter when the estimated state is far from the real state.

We propose an adaptation strategy in order to navigate between high-gain and non high-gain modes. The adaptation requires knowledge of a quantity that reflects the quality of the estimation. Such a quality measurement is proposed and its usage is justified by an important lemma, i.e. Lemma 33 in Chapter 3. The convergence of the resulting observer is analytically proven. Extensions to the initial result are proposed.

Practical aspects of the implementation of the algorithm are also considered. The use of an adaptation strategy implies the introduction of several new parameters. Guidelines to the choice of those parameters are given and a methodology is proposed.

1.4 Dissertation Outline and Comments to the Reader

The reminder of the manuscript is organized as follows:

– **Chapter 2** contains a summary of the theory of observability in the deterministic setting. Observability normal forms are introduced and high-gain observers defined. A review of nonlinear observers having adaptive strategies for their correction gain closes the chapter.

– The adaptive high-gain extended Kalman filter is defined in **Chapter 3**. We define our strategy and justify it with an important lemma. The complete proof of convergence is

detailed here.

– **Chapter 4** is dedicated to the implementation of the observer. An example is used to progress through the different steps. An experiment on a real process illustrates the behavior of the observer in a real-time environment.

– Finally **Chapter 5** includes extensions to the original result.

Complementary material is included in the appendices.

Readers Interested in a practical approach of the adaptive high-gain extended Kalman filter...

...should read Chapter 4.

...should take a look to Section 3.1 (definition of the MISO normal form), and to Definition 30 (of the observer). If they are not familiar with high-gain observers, Definition 15 (high-gain EKF) may be of help.

...will find the main theoretical results in Lemma 33 (lemma on innovation) and Theorem 36 (convergence of the observer).

Readers interested only in the definition of the observer and its proof of convergence...

...can read directly Chapter 3.

...shall read Chapter 2 in order to understand the motivations for the use of the observability form.

Readers that want to explore a bit further than the initial continuous MISO normal form...

...should go to Chapter 5. Some of the informations of this chapter are also useful for standard high-gain EKFs: dealing with multiple outputs systems, understanding the proof of the properties of the continuous discrete Riccati equation,...

Chapter 2

Observability and High-gain Observers

Contents

The work presented in this dissertation follows the framework of the *theory of deterministic observation* developed by J-P Gauthier, I. Kupka, and coworkers. This theory has a very long history the beginning of which can be found in the articles of R. Herman, A. J. Krener [65], and H. J. Sussmann [110]. In the book [57], which is in itself a summary of several papers [53–56, 70], J-P. Gauthier and I. Kupka exposed well established definitions for observability and several important subsequent theorems.

An important result of the theory is that there exist representations of nonlinear systems that characterize observability (those are denoted by observability normal forms in the literature). The article [52], from J-P Gauthier and G. Bornard, often referenced, contains early results.

The construction of high-gain observers, either of the *Luenberger* (J-P. Gauthier, H. Hammouri and S. Othman [54]) or of the *Kalman* (F. Deza, E. Busvelle *et al.* [47]) style, rests upon the theory that normal forms are crucial in order to establish the convergence of such observers. Here convergence means that the estimation error decreases to zero. As we will see in Chapter 3, and in Theorems 14, 16 and 29, the estimation error decays at an exponential rate. Those observers are also called *exponential observers*.

In the present chapter, the main concepts of observability theory in the deterministic setting are presented together with more recent results such as those from E. Busvelle and J-P. Gauthier [38–40].

The main contribution of this work is the construction and the proof of convergence of a high-gain observer algorithm, whose high-gain parameter is time varying. Hence a review of such adaptive-gain observers is proposed: F. Allgőwer and E. Bullinger (1997), M. Boutayeb *et al.* (1999), E. Busvelle and J-P. Gauthier (2002), L. Praly *et al.*(2004 and 2009) and a recent paper from H. K. Khalil (2009).

2.1 Systems and Notations

A system in the state space representation is composed of two time dependent equations[1]:

$$(\Sigma)\begin{cases} \frac{dx(t)}{dt} &= f(x(t),u(t),t) \\ y(t) &= h(x(t),u(t),t) \\ x(0) &= x_0 \end{cases}$$

where

- $x(t)$ denotes the state of the system, belonging to $\mathbb{R}^n$, or more generally to a n-dimensional analytic differentiable manifold X,
- $u(t)$ is the control (or input) variable with $u(t) \in \mathcal{U}_{adm} \subset \mathbb{R}^{n_u}$,
- $y(t)$ denotes the measurements (or outputs) and takes values in a subset of $\mathbb{R}^{n_y}$,
- f is a u-parameterized smooth nonlinear vector field, and
- the observation mapping $h : X \times \mathcal{U}_{adm} \to \mathbb{R}^{n_y}$ is considered to be smooth and possibly nonlinear.

[1]Later on, the time dependency will be omitted for notation simplicity.

The inputs are time functions defined on open intervals of the form $[0, T[$ (with the possibility that $T = +\infty$). The functions $u(.)$ are assumed to be measurable and bounded almost everywhere on any compact sub-interval of $[0, T[$. The corresponding function set is $L^\infty(\mathcal{U}_{adm})$.

The outputs are also functions of time defined on open intervals of the form $[0, T(u)[$. This notation takes into account that for a given input function, defined for a maximum time T, the system might become unstable (i.e. *explode toward infinity*). The explosion time is likely to be less than T. Therefore we have $T(u) \leq T$. Output functions are also measurable and bounded almost everywhere on any compact sub-interval of $[0, T(u)[$. The corresponding function set is $L^\infty(\mathbb{R}^{n_y})$.

The set of all systems of the form (Σ) is denoted $\mathcal{S} = \{\Sigma = (f, h)\}$. The genericity (or non-genericity) property of observable systems is considered with respect to the set $\mathcal{S}$.

The topologies associated to those sets are

- the $\mathcal{C}^\infty$ Whitney topology for the set $\mathcal{S}$ (see, e.g. [66])[2]. Two important features of that topology are
  1. it is *not metrizable* and,
  2. it has the *Baire property*[3],
- either the topology of uniform convergence or the weak-$*$ topology for the sets $L^\infty(\mathcal{U}_{adm})$ and $L^\infty(\mathbb{R}^{n_y})$,
- we will also use the topology of the euclidean norm when dealing with subspaces of $\mathbb{R}^q$, $q = n, n_u$ or n_y.

2.2 The Several Definitions of Observability

Observability is the notion that translates the property of a system to permit the reconstruction of the full state vector from the knowledge of the input and output variables. In other words: *considering any input function $u(.)$, can any two distinct initial states x_0^1, x_0^2 be distinguished from one another?*

Definition 1

- *The **state-output mapping** of the system (Σ) is the application:*

$$\begin{array}{rccc} PX_{\Sigma,u} & : & X & \to & L^\infty(\mathbb{R}^{n_y}) \\ & & x_0 & \mapsto & y(.) \end{array}$$

- *A system (Σ) is **uniformly observable** (or just observable) w.r.t. a class $\mathcal{C}$ of inputs if for each $u(.) \in \mathcal{C}$, the associated state-ouput mapping is injective.*

[2] A basic neighborhood of a system $\Sigma = (f, h)$ in the $\mathcal{C}^j$ Whitney topology is determined by a set of functions $\epsilon(z) > 0$, and formed by the systems $\tilde{\Sigma} = (\tilde{f}, \tilde{g}) \in \mathcal{S}$ such that the derivatives, up to the order j, of $(f - \tilde{f}, h - \tilde{h})$, *w.r.t* all the variables, have their norm at point $z = (x, u)$ less than $\epsilon(z)$.

[3] Baire property: a countable intersection of open dense subsets is dense.

This definition appears as the most natural one[4], but as injectivity is not a stable property observability is difficult to manipulate from a topological point of view [5]. In order to render the notion of observability more tractable, Definition 1 is modified by considering the first order approximation of the state-output mapping. Let us begin with the definition of the first order approximation of a system:

Definition 2

– *The **first (state) variation** of* (Σ) *(or **lift of** (Σ) **on** TX) is given by:*

$$(TX_\Sigma)\begin{cases} \frac{dx(t)}{dt} &= f(x,u) \\ \frac{d\xi(t)}{dt} &= D_x f(x,u)\xi \\ \hat{y} &= d_x h(x,u)\xi \end{cases} \tag{2.1}$$

where

- $(x,\xi) \in TX$ *(or* $\mathbb{R}^n \times \mathbb{R}^n$*) is the state of* (TX_Σ)*,*
- $d_x h$ *is the differential of* h *w.r.t. to* x*, and*
- $D_x f$ *is the tangent mapping to* f *(represented by the Jacobian matrices of* h *and* f *w.r.t.* x*).*

– *the **state-output mapping** of* TX_Σ *is denoted* $PTX_{\Sigma,u}$*. This mapping is in fact the differential of* $PX_{\Sigma,u}$ *w.r.t.* x_0 *(i.e. its first order approximation,* $TPX_{\Sigma,u}|_{x_0}$ *).*

The second part of Definition 1 is adapted to this new state-output mapping in a very natural way.

Definition 3

The system (Σ) *is said **uniformly infinitesimally observable**[6] w.r.t. a class* $\mathcal{C}$ *of inputs if for each* $u(.) \in \mathcal{C}$ *and each* $x_0 \in X$*, all the* (x_0 *parameterized) tangent mappings* $TPX_{\Sigma,u|_{x_0}}$ *are injective.*

Since the state-output mapping considered in this definition is linear then the injectivity property has been topologically stabilized. Finally a third definition of observability has been proposed by using the notion of $k-$jets[7].

[4] An equivalent definition, based on the notion of *indistinguishability* can be found in [19] (Definitions 2 and 3).

[5] e.g. $(x \mapsto x^3)$ is injective, but for all $\epsilon > 0$, $(x \mapsto x^3 - \epsilon x)$ isn't.

[6] Infinitesimal observability can also be considered *"at a point* $(u,x) \in L^\infty \times X$*"*, or only *"at a point* $u \in L^\infty$*"*

[7] The k-jets $j^k u$ of a smooth function u at the point $t = 0$ are defined as

$$j^k u = \left(u(0), \dot{u}(0), ..., u^{(k-1)}(0)\right).$$

Then for a smooth function u and for each $x_0 \in X$, the k-jets $j^k y = \left(y(0), \dot{y}(0), \ldots, y^{(k-1)}(0)\right)$ is well defined: this is the k-jets extension of the state-output mapping.
See [9] for details. The book is though not so easy to find. R. Abraham's webpage can be of help.

Definition 4

– (Σ) *is said to be* ***differentially observable*** *of order k, if for all $j^k u$ the extension to k-jets mapping*

$$\begin{array}{rccc} \Phi_k^\Sigma & : & X & \to & \mathbb{R}^{kn_y} \\ & & x_0 & \mapsto & j^k y \end{array}$$

is injective.

– (Σ) *is said to be* ***strongly differentially observable*** *of order k if for all $j^k u$, the extension to k-jets mapping*

$$\begin{array}{rccc} \Phi_{k,j^k u}^\Sigma & : & X & \to & \mathbb{R}^{kn_y} \\ & & x_0 & \mapsto & j^k y \end{array}$$

is an injective immersion[8]*.*

From Definition 4, strong differential observability implies differential observability. The first component of the application $(\Phi_{k,j^k u}^\Sigma)$ corresponds to the state-ouput mapping (i.e. Definition 1). When the control variable belongs to $\mathcal{C}^\infty$, (strong) differential observability implies $\mathcal{C}^\infty$ observability. It may seem that little is gained by adding those extra definitions. In fact the notion of differential observability can be checked in a really practical way as it is explained at the end of this section.

In order to prove differential observability we need the controls to be sufficiently differentiable so that differentiations can be performed. But since in practice the control variable is not likely to be differentiable, we are looking forward to L^∞ observability. We consider now the following theorem:

Theorem 5 ***([57], 4.4.6 page 56)***

For an analytic system (Σ) (i.e. when f and g are analytic, or $\mathcal{C}^\omega$ functions) the following properties are equivalent:

1. *(Σ) is observable for all $\mathcal{C}^\omega$ inputs,*
2. *(Σ) is observable for all L^∞ inputs.*

This means that for an analytic system both strong differential observability and differential observability imply L^∞ observability (i.e. observability for the class of $L^\infty(\mathcal{U}_{adm})$ inputs).

Another consequence of the theory (as explained in [40, 57], Chapters 3.1 and 4.4) is that, for analytic systems, uniform infinitesimal observability implies observability of systems (Σ) restricted to small open subsets of X, the union of which is dense in X.

As we will see with the theorems of next paragraph, although observability or uniform observability is not an easy property to establish in itself, a good method is to find a coordinate transformation that puts it under an observability form.

Another method comes from the definition of differential observability:

[8] Immersion: all the tangent mappings $T_{x_0}\Phi_{k,j^k u}^\Sigma$ to the map$\Phi_{k,j^k u}^\Sigma$ have full rank n at each point.

– Consider that the output $y(t)$ is known for all times $t \geq 0$.

– Compute the successive time derivatives of the outputs until there exists a $k > 0$ such that the state $x(t)$ can be uniquely computed, for all times, from the equations of $(\dot{y}(t), \ddot{y}(t), ...)$ where $y(t), u(t)$ are known time varying parameters.

– We have found a k-jets extension of (Σ) that is injective and therefore it is *at least* differentially observable.

– Assuming that (Σ) is analytic, then it is L^∞ observable.

An illustration of this method is given in Chapter 4, Section 4.1.2. The process under consideration is a series-connected DC machine (see also [21], part 3.5).

2.3 Observability Normal Forms

The main significance of the theory is the existence of two distinct situations, which depend on the number of outputs with respect to the number of inputs. In one case observability is a generic property[9], and it is not a generic property in the other case. These two specific situations are explained in the subsections below.

2.3.1 First case: $n_y > n_u$

The first case occurs when the number of outputs is greater than the number of inputs, i.e. $n_y > n_u$. The situation is defined by two theorems. The first states the genericity property of the set of observable systems in $\mathcal{S}$; the second theorem introduces the (generic) observability normal form.

Theorem 6 ***([57], 4.2.2 and 4.2.4 page 40)***

1. *The set of systems that are strongly differentially observable of order* $2n+1$ *is residual in* $\mathcal{S}$.

2. *The set of analytic strongly differentially observable systems (of order* $2n+1$*) that are moreover* L^∞*-observable is dense in* $\mathcal{S}$.

Theorem 7 ***([40])***

The following is a generic property on $\mathcal{S}$. *Set* $k = 2n+1$. *For all sufficiently smooth* $u(.)$, *denote* $j^k u(t) = \big(u(t), \dot{u}(t), ..., u^{(k-1)}(t)\big)$. *Choose an arbitrarily large, relatively compact*[10] *open subset* Γ *of* X. *Consider also an arbitrary bound on the control and its first* k *derivatives (i.e.* $u, \dot{u}, .., u^{(k)}$*). Then the mappings*

$$\begin{array}{rccl} \Phi^{\Sigma}_{k,j^k u} & : & X & \to & \mathbb{R}^{kn_y} \\ & & x(t) & \mapsto & \big(y(t), \dot{y}(t), ..., y^{(k-1)}(t)\big) \end{array}$$

[9] A subset is said to be *generic* if it contains a residual subset. A subset is said to be *residual* if it is a countable intersection of open dense subsets.

[10] A subset is said *relatively compact* if its closure is compact.

are smooth injective immersions that map the trajectories of the system (Σ) *(restricted to* Γ*) to the trajectories of the system:*

$$\begin{cases} y &= z_1 \\ \dot{z_1} &= z_2 \\ \vdots & \\ \dot{z_{k-1}} &= z_k \\ \dot{z_k} &= \phi_k\left(z_1, z_2, ..., z_k, u, \dot{u}, ..., u^{(k)}\right) \end{cases} \tag{2.2}$$

The form of (2.2) is called a *phase variable representation* where all the z_i components of the state vectors are of dimension n_y. Since $k = 2n + 1$ then the total dimension of the state space of the phase variable representation is $(2n + 1)n_y$.

Let $v = \left(u(t), \dot{u}(t), \ldots, u^{(k)}(t)\right)$ denote the controls of the phase variable representation. Then since all the mappings$\Phi_{k,j^k u}^{\Sigma}$ are smooth injective immersions restricted to the subset Γ of Theorem 7, systems of the form (2.2) are observable, strongly differentially observable and uniformly infinitesimally observable. That is to say that the set of systems that can be embedded in a phase variable representation is contained into the set of observable systems.

Therefore, in the case $n_y > n_u$, observable systems are generic in $\mathcal{S}$ and can generically be embedded into a phase variable representation.

2.3.2 Second case: $n_y \leq n_u$

In the previous section we depicted observability as a generic property when there are more outputs than inputs $(n_y > n_u)$. This is no longer the case when $n_y \leq n_u$. We restrict the exposure in this instance to single output systems, i.e. $n_y = 1$. We also restrict the exposure to analytic systems within the set $\mathcal{S}$. The study of observability in this case is done with the help of a tool called the *canonical flag of distributions*. We define it below.

As we will see in the theorems to come, the study of observability is done by establishing existence or non-existence of a canonical flag of distributions.

Definition 8

– *Consider a system* $\Sigma = (f, h) \in \mathcal{S}$*. We define the* ***canonical flag of distributions*** $D(u)$ *as*

$$\begin{cases} D(u) &= \left\{D^0(u) \supset D^1(u) \supset ... \supset D^{n-1}(u)\right\} \\ D^0(u) &= Ker(d_x h) \\ D^{k+1}(u) &= D^k(u) \cap Ker(d_x L_f^{k+1} h) \end{cases} \tag{2.3}$$

where $L_f h$ *is the Lie derivative of* h *with respect to the vector field* f*.* Ker *denotes the kernel of an application. The control* $u(t)$ *being considered as fixed.*

– *If the distributions* $D^i(u)$ *have constant rank* $n - i - 1$*, and are independent of* $u(.)$*, then* $D(u)$ *is denoted as a* ***uniform canonical flag****.*

The property

$$\left((\Sigma) \text{ has a uniform canonical flag }\right)$$

is highly non-generic (it has co-dimension ∞, see [57]). The non-genericity of observability is given by the two following theorems.

Theorem 9 ***([38], Pg 22)***

The system (Σ) *has a uniform canonical flag if and only if for all* $x^0 \in X$*, there is a coordinate neighborhood of* x^0*,* (V_{x^0}, x)*, such that in those coordinates the restriction of* (Σ) *to* V_{x^0} *can be written as*

$$\begin{cases} y &= h(x_1, u) \\ \dot{x}_1 &= f_1(x_1, x_2, u) \\ \dot{x}_2 &= f_2(x_1, x_2, x_3, u) \\ &\vdots \\ \dot{x}_{n-1} &= f_{n-1}(x_1, x_2, ..., x_n, u) \\ \dot{x}_n &= f_n(x_1, x_2, ..., x_n, u) \end{cases} \tag{2.4}$$

where $\frac{\partial h}{\partial x_1}$ *and* $\frac{\partial f_i}{\partial x_{i+1}}$*,* $i = 1, ..., n-1$*, never equal zero on* $V_{x^0} \times \mathcal{U}_{adm}$*.*

As was the case for the form (2.2) of the previous section, a system under the form (2.4) is infinitesimally observable, observable and differentially observable of order n. Therefore if a system (Σ) has a uniform canonical flag then, when restricted to neighborhoods of the form $V_{x^0} \times \mathcal{U}_{adm}$ an be summarized in t

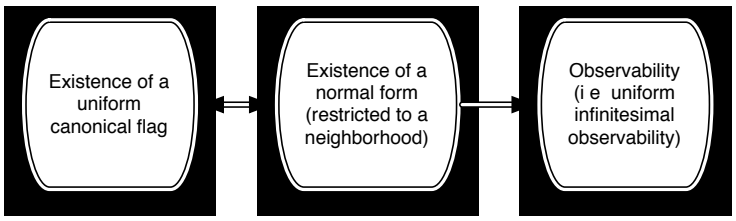

Figure 2.1: Observability Equivalence Diagram.

Infinitesimal observability still needs to be related to the normal form (2.4) in order to obtain a complete equivalence diagram. This is done with a second theorem.

Theorem 10 ***([38], Pg 24-25)***

If (Σ) *is uniformly infinitesimally observable, then, on the complement of a sub-analytic subset of* X *of co-dimension* 1*,* (Σ) *has a uniform canonical flag.*

The combination of those two theorems closes the Diagram 2.1 which means that the normal form (2.4) characterizes uniform infinitesimal observability.

In the **control affine case**, the situation above can be rewritten in a stronger way. We first recall that such a control affine system is

$$\begin{cases} \dot{x} &= f(x) + \sum_{i=1}^{p} g_i(x) u_i \\ y &= h(x). \end{cases} \tag{2.5}$$

We then define the function

$$\begin{array}{rccl} \Phi & : & X & \to & \mathbb{R}^n \\ & & x & \mapsto & \left(h(x), L_f h(x), ..., L_f^{n-1} h(x)\right) \end{array}$$

for which we can establish Lemma 11 below.

Lemma 11

(Σ) is observable $\Rightarrow$ Φ is of maximum rank (i.e. n) on an open dense subset V of X.

Let us choose any subset $W \subset X$ such that the restriction of Φ to W is, as said in the lemma, a diffeomorphism. Then the theorems above are modified into:

Theorem 12 ***([38], Pg 26-27)***

Assume that (Σ) is observable. Then the restriction $\Phi|_W$ maps (Σ) into a system of the form:

$$\left\{\begin{array}{rcl} y & = & x_1 \\ \dot{x}_1 & = & x_2 + \sum_{i=1}^{p} g_{1,i}(x_1) u_i \\ \dot{x}_2 & = & x_3 + \sum_{i=1}^{p} g_{2,i}(x_1, x_2) u_i \\ & \vdots & \\ \dot{x}_2 & = & x_n + \sum_{i=1}^{p} g_{n-1,i}(x_1, x_2, ..., x_{n-1}) u_i \\ \dot{x}_n & = & \psi(x) + \sum_{i=1}^{p} g_{n,i}(x_1, x_2, ..., x_{n-1}, x_n) u_i. \end{array}\right. \tag{2.6}$$

.

Conversely, if a system is under the form (2.6), on an open subset $\Omega \subset \mathbb{R}^n$, then it is observable.

In conclusion, we want to emphasize the representations (2.2), (2.4) and (2.6) as they translate observability. Their general form is $\dot{x} = Ax + b(x, u)$ where A is an upper diagonal matrix and $b(x, u)$ is a triangular vector field. As a consequence the proof of the convergence of high-gain observers is carried out considering such normal forms. In addition, as we will see with the proof of Chapter 3, both the structure of the matrix A and of the vector field $b(x, u)$ will be very useful.

A generalization of the normal form (2.6) to multiple output systems is used in Chapter 5 in order to extend the definition of the observer to systems with more than one output variables. This form is distinct from the phase variable representation given in equation (2.2).

2.4 Single Output Normal Form

The normal form introduced in this section is the one we use in order to define and prove the convergence of the observer for multiple-input, single-output (MISO) systems. We use

the single-output assumption for simplicity and clarity of the exposure only. Up to a few modifications, the theorems can be proven in the multiple output case. Indeed, there is no unique normal form when $n_y > 1$, therefore the definition of the observer has to be changed according to each specific case. In Chapter 5 a block wise generalization of the MISO normal form is considered, and the differences between the single output and the multiple output case are explained.

As usual the system is represented by a set of two equations:

- an ordinary differential equation that drives the evolution of the state,
- an application that models the sensor measurements.

Those two equations are of the form:

$$\left\{ \begin{array}{rcl} \frac{dx}{dt} & = & A(u)\,x + b(x,u) \\ y & = & C(u)\,x, \end{array} \right. \tag{2.7}$$

where

- $x(t) \in \mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}$ compact,
- $y(t) \in \mathbb{R}$,
- $u(t) \in \mathcal{U}_{\text{adm}} \subset \mathbb{R}^{n_u}$ bounded.

The matrices $A(u)$ and $C(u)$ are defined by:

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

$$C(u) = \begin{pmatrix} a_1(u) & 0 & \cdots & 0 \end{pmatrix}$$

with $0 < a_m \leq a_i(u) \leq a_M$ for any $u(t)$ in $\mathcal{U}_{\text{adm}}$ and $i = 1, ..., n$. We assume that the vector field $b(x,u)$ is compactly supported and has the triangular structure:

$$b(x,u) = \begin{pmatrix} b_1(x_1,u) \\ b_2(x_1,x_2,u) \\ \vdots \\ b_n(x_1,\ldots,x_n,u) \end{pmatrix}.$$

The Jacobian matrix $b^*(x,u)$ of $b(x,u)$ is considered to be upper bounded and the vector field $b(x,u)$ has the Lipschitz property. Those constants will be defined more precisely in Chapter 3 since we do not manipulate them here.

The observability part of this Chapter comes to its end. In the following nonlinear observers are highlighted. In a first section the definitions of the two main high-gain observers are recalled and the corresponding convergence theorems displayed. In a final section we give a detailed look at observers with varying or adaptive high-gain that can be found in the literature.

2.5 High-gain Observers

Early descriptions of high-gain observers can be found in multiple references including J-P. Gauthier *et al.*, [54], F. Deza *et al.*, [46, 47], A. Tornambé, [111], H. K. Khalil *et al.* and [50].

The first high-gain observer construction that we present in this section is the Luenberger style prototype algorithm of [54]. Afterwards we define the high-gain extended Kalman filter. Its structure is quite similar to that of the extended Kalman filter, which is well known and well used in engineering [58, 60]. In our case, it has been adapted to the observability normal form.

High-gain observers are embedded with a structure based on a fixed scalar parameter (denoted θ) which allows us to prove that the estimation error decays to zero, exponentially at a rate that depends on the value of θ. When the high-gain parameter is taken equal to 1, high-gain observers reduce to their initial nonlinear version.

Definition 13

We suppose that all the $a_i(u)$ coefficients of the normal form (2.7) are equal to 1. The classical (or Luenberger) ***high-gain observer*** *is defined by the equation*

$$\frac{dz}{dt} = Az + b(z,u) - K_\theta\,(Cz - y(t)) \tag{2.8}$$

where $K_\theta = \Delta K$ with[11] *$\Delta = diag\left(\{\theta, \theta^2, ..., \theta^n\}\right)$ and K is such that $(A - KC)$ is Hurwicz stable.*

It is in fact not important that the $a_i(u)$ coefficients equal 1 or any other non zero constant since a change of coordinates brings us back to the situation where $a_i = 1$.

On the other hand, it is of the utmost importance that the coefficients do not depend on either $u(t)$ or time:

- the correction gain K is computed off-line (i.e. for $u = u^*$), $(A(u) - KC(u))$ may not remain stable for $u \neq u^*$ and the convergence of the observer is not guaranteed anymore.
- in full generality: $(A(t) - K(t)C(t)$ stable $\forall t > 0)$, $\not\Rightarrow$ (the system is stable).

The convergence of the high-gain Luenberger observer is expressed in the following theorem.

Theorem 14

For any $a > 0$, there is a large enough $\theta > 1$ such that $\forall (x_0, z_0) \in (\chi \times \chi)$, we have

$$\|z(t) - x(t)\|^2 \leq k(a)e^{-at}\|z_0 - x_0\|$$

for some polynomial k of degree n.

In Kalman style observers, the correction gain is computed at the same time as the estimated state, and therefore constantly updated. It is the solution of a Riccati equation

[11] Here, *diag* denotes the square matrix filled with zeros except for the diagonal that is composed of the adequate vector.

(of matrices). The corresponding matrix (denoted S or P) is called the Riccati matrix. It is a symmetric and positive definite matrix[12]. Since S is a $(n \times n)$ symmetric square matrix, we only need to compute the upper or the lower part of the matrix (i.e. there are $\frac{n(n+1)}{2}$ equations to solve).

Definition 15

*The **high-gain extended Kalman filter** is defined by the two equations below:*

$$\left\{ \begin{array}{rcl} \frac{dz}{dt} & = & A(u)z + b(z,u) - S^{-1}C'R^{-1}(Cz - y) \\ \frac{dS}{dt} & = & -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'R^{-1}C - SQ_\theta S. \end{array} \right. \tag{2.9}$$

The matrices Q and R are originally the covariance matrices of the state and output noise respectively, and therefore are expected to be symmetric and positive definite. Since this observer is developed within the frame of the deterministic observation theory, those two matrices will be used as tuning parameters. Q_θ is defined as $Q_\theta = \theta^2 \Delta^{-1} Q \Delta^{-1}$ where $\theta > 1$ is a fixed parameter and

$$\Delta = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\theta^{n-1}} \end{pmatrix}.$$

In most cases, normal form representations are not used when implementing an extended Kalman filter (case $\theta = 1$). The Jacobian matrices of f and h (computed with respect to the variable x) are used in the Riccati equation:

$$\frac{dS}{dt} = -\left(\frac{\partial f}{\partial x}_{|x=z}\right)' S - S\left(\frac{\partial f}{\partial x}_{|x=z}\right) + \left(\frac{\partial h}{\partial x}_{|x=z}\right)' R^{-1} \left(\frac{\partial h}{\partial x}_{|x=z}\right) - SQ_\theta S.$$

Refer to Chapter 1 for a more detailed explanation.

Theorem 16 *([47, 57])*

For θ large enough and for all $T > 0$, the high-gain extended Kalman filter (2.9) satisfies the inequality below for all $t > \frac{T}{\theta}$:

$$\|z(t) - x(t)\|^2 \leq \theta^{n-1} k(T) \|z(\frac{T}{\theta}) - x(\frac{T}{\theta})\|^2 e^{-(\theta\omega(t) - \mu(T))(t - \frac{T}{\theta})}$$

for some positive continuous functions $k(T)$, $\omega(T)$ and $\mu(T)$.

2.6 On Adaptive High-gain Observers

In this section, we review a few adaptation strategies for the high-gain parameter that can be found in the literature. Strategies that are based on Luenberger like[13] observers are

[12] The fact that the solution of the Riccati equation of the observer (2.9) remains definite positive is not obvious and must be proven. Such a proof can be found in [57], Lemma 6.2.15 and Appendix B in the continuous discrete case.

[13] Here *Luenberger like* has to be understood in a broad sense. That is: observers with a correction gain matrix that is not computed online and, most of the time, computed following a pole placement-like scheme.

presented in Subsections 2.6.1, 2.6.2 and 2.6.3. In the case of Kalman like observers, recall that the high-gain modification is used to provide a structure to the Q and R matrices. Therefore *adaptation of the high-gain parameter* can be understood as an *adaptation of the covariance matrices of the state and measurement noise*. This topic has been the object of quite extensive studies and a large bibliography on the subject is available, both in the linear and nonlinear cases. Subsection 2.6.4 provides references to such works while Subsections 2.6.5 and 2.6.6 focuses on high-gain constructions specifically.

The adaptation of the high-gain parameter is, as said in Chapter 1, motivated by the need to combine the antagonistic behaviors of an observer that filters noise efficiently and of an observer that is able to converge quickly when large perturbations or jumps in the state are detected. We want to emphasize the usefulness of this approach with three examples:

1. In [64] S. Haugwitz and coworkers describe the model of a chemical reactor coupled with a highly efficient heat exchanger. The reactor is expected to be used to process a highly exothermic chemical reaction, and the temperature measured at specific spots constitutes the multidimensional output variable. Although the inlet concentrations are supposed to be known and fixed, it may happen that the apparatus meant to blend the reactants fails. The inlet concentration is then no longer the one expected, thus provoking a non-measured large perturbation. The authors use, quite successfully, an extended Kalman filter that takes into account this possibility of failure in the same spirit as when we estimate the load torque of the DC machine in Chapter 4. An adaptive high-gain extended Kalman filter can be really efficient for this kind of application by increasing the performance of the observer at perturbation times.

2. In vehicle navigation, data from an inertial navigation system (INS) — a 3-axis accelerometer coupled with a gyroscope — and data from a global navigation satellite systems (GPS, GALILEO, GLONASS) are fusioned. The first type of sensors is very precise at time 0 but with an error domain that grows with time. Sensors of the second type have an error domain larger that the one of the INS at time 0 but that is stable. The purpose here is to know the position as precisely as possible with an error domain as small as possible. An observer that filters the measurement noise is needed but estimation error may increase with time because of sudden changes of direction, sudden changes in the topology of the road or the loss of the GPS signal because of tunnels or urban canyons. The estimation of the covariance matrices Q and R of Kalman like filters is the subject of the articles [96, 115] (linear case) or [37] (nonlinear case). The book of M. S. Greywal [60] provides a solid introduction to this topic.

3. In a refinery, changes of the processed crude oil are perturbations. Starting from the atmospheric column, the disturbance propagates along the refinery. The speed of propagation of the disturbance front depends on many parameters (the several processes have low time constants, crude oil can be retained,...), and is not accurately known. An EKF[14] like observer is useful when there are no such changes, and an adaptive high-gain observer would be of use in order to detect the feed change [38, 47, 113].

Finally we want to cite a few techniques used in order to render an observer's gain adaptive that we have not considered:

[14] Extended Kalman filter.

- statistical methods [115],
- genetic algorithms based observers [97],
- Neural networks based observers [109],
- fuzzy logic approach [72].

Those observers are based on empirical methods. As a result, very little can be demonstrated or proven with respect to their convergence properties.

2.6.1 E. Bullinger and F. Allőgower

In a 1997 paper, E. Bullinger and F. Allgőwer proposed a high-gain observer having a varying high-gain parameter. Their observer is inspired by the structure proposed by A. Tornambè in [111]. It is a Luenberger like observer for a system of the form (2.7) except that $\alpha_i(u) = 1$ for all $i \in 1, ..., n$. Only the last component of the vector field $b(x, u)$ is not equal to zero (see definition below). The control variable u may be of the form $u = \left(u, \dot{u}, u^{(2)}, \ldots, u^{(n)}\right)$ which is not one of the assumptions of system (2.7). As it appears below, the main difference between this observer and a classic high-gain is that the influence of the vector field to the model dynamic is neglected.

Definition 17

Consider a ***single input, single output system*** *of the form*

$$\left\{\begin{array}{rcl} \dot{x_1} & = & x_2 \\ \dot{x_2} & = & x_3 \\ & \vdots & \\ \dot{x_{n-1}} & = & x_n \\ \dot{x_n} & = & \phi(x,u) \\ y & = & x_1 \end{array}\right. \tag{2.10}$$

Then define a high-gain obsever

$$\dot{z} \quad = \quad Az - \Delta K(z_1 - y) \tag{2.11}$$

with ΔK defined in the same manner as for the observer (2.8):

- $\Delta = diag\left(\{\theta, \theta^2, \ldots, \theta^n\}\right)$, *and*
- $\left(A - K\left[\begin{array}{cccc} 1 & 0 & \ldots & 0 \end{array}\right]\right)$ *is Hurwicz stable.*

Then:

- *select a strictly increasing sequence of elements of* $\mathbb{R}$*:* $\{1, \theta_1, \theta_2, ...\}$,
- *choose* $\lambda > 0$*, a small positive scalar, and*

– *consider the adaptation function:*

$$\dot{s} = \begin{cases} \gamma|y(t) - z_1(t)|^2 & for \quad |y(t) - z_1(t)| > \lambda \\ 0 & for \quad |y(t) - z_1(t)| \leq \lambda. \end{cases} \tag{2.12}$$

The parameter θ is adapted according to the rule:

– *when $t = 0$ set $\theta = 1$*

– *when $s(t) = \theta_i$, set $\theta = \theta_i$ (or in other words $\theta = \sup\{\theta_i \leq s(t)\}$).*

It is clear from the definition of the adaptation procedure that the parameter θ cannot decrease which makes this observer a solution to a tuning problem rather than the noise reduction problem. The convergence of the observer is established in the following theorem.

This observer comes from a paper published in 1997 (i.e. [35]), we therefore checked for updates of the strategy in more recent paper. In [36], the authors address λ-tracking problems[15]. The observer they use is the one described in this section.

Theorem 18 *([35])*

Consider the system (2.10) above, together with the assumptions

1. *the system exhibits no finite escape time, and*
2. *the nonlinearity $\phi(x, u)$ is bounded.*

Then for any $\lambda > 0$, $\gamma > 0$, $\beta > 0$ and any S_0:

the total length of time for which the observer output error is larger than λ is finite.

That is to say:

$$\exists T_{max} < \infty : \int_{\mathcal{T}} dt < T_{max} \text{ where } \mathcal{T} = \{t | \|y(t) - z_1(t)\| \geq \lambda\}.$$

In another theorem of the same paper (i.e. Theorem 3 of [35]), an upper bound for the estimation error for high values of t is provided. But in the original article from which this observer is inspired, [111], the estimation error is not bounded by an exponentially decreasing function of the time.

The main differences between these works and with the work proposed here is that firstly that our observer is a Kalman based observer which, implies taking into account the evolution of the Riccati matrix. The second distinction is that here we address the problem of noise reduction when the estimation is sufficiently good.

2.6.2 L. Praly, P. Krishnamurthy and coworkers

The second observer with a dynamically sized high-gain parameter proposed that we will expose in this review is one from L. Praly, K. Krishnamurthy and coworkers. Descriptions of

[15] As explained in the article, λ-tracking is used for processes for which we know for certain that asymptotic stabilization can be achieved only by approximation. The objective is therefore modified into one of stabilizing the state within a sphere of radius λ centered on the set point.

their observer construction can be found in articles like [16, 78, 79, 103][16]. In the following we present the observer defined in [16], as it complies with the latest updates of their theory.

Definition 19

*The **system dynamics** are:*

$$\left\{\begin{array}{rcl} \dot{x} & = & \begin{pmatrix} 0 & a_2(y) & 0 & ... & 0 \\ 0 & 0 & a_3(y) & ... & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & ... & ... & ... & a_n(y) \\ 0 & ... & ... & ... & 0 \end{pmatrix} x + \begin{pmatrix} f_1(u,y) \\ f_2(u,y,x_2) \\ f_3(u,y,x_2,x_3) \\ \cdots \\ f_n(u,y,x_2,...,x_n) \end{pmatrix} + \begin{pmatrix} \delta_1(t) \\ \delta_2(t) \\ \delta_3(t) \\ \cdots \\ \delta_n(t) \end{pmatrix} \\ y & = & x_1 + \delta_y(t). \end{array}\right. \tag{2.13}$$

It is also denoted

$$\left\{\begin{array}{rcl} \dot{x} & = & A(y)x + b(y,x_2,..,x_n,,u) + \Delta(\,t) \\ y & = & x_1 + \delta_y(t) \end{array}\right. \tag{2.14}$$

where

- *y is the measured output,*
- *the functions a_i are locally Lipschitz,*
- *u is a vector in $\mathbb{R}^{n_u}$ representing the known inputs and a finite number of their derivatives,*
- *the vector $\Delta(t)$ represents the unknown inputs, and*
- *δ_y is the measurement noise.*

The construction of the observer is based on some additional knowledge concerning the system:

1. a function α of the output y such that $0 < \rho < \alpha\,(y)$, and
   $0 < \alpha_{up} < \frac{a_i(y)}{\alpha(y)} < \alpha_{down}$ for all $y \in \mathcal{R}$, all i (where ρ, α_{up}, α_{down} are positive constants), and
2. the vector fields $f_i(u,y,x_2,...,x_i)$ are such that:

   $$|f_i(u,y,\hat{x}_2,...,\hat{x}_i) - f_i(u,y,x_2,...,x_i)| \leq \Gamma(u,y)\left(1+\sum_{j=2}^{n}|\hat{x}_j|^{v_j}\right)\sum_{j=2}^{i}|\hat{x}_j - x_j| + C\sum_{j=2}^{i}|\hat{x}_j - x_j|^{\frac{1-D(n-i-1)}{1-D(n-j)}}$$

   where C is a positive number, the v_j are in $[0, \frac{1}{j-1}[$, $D \in [0, \frac{1}{n-1}[$ and$\Gamma(\,u,y)$ is a continuous function.

[16] In [78] the high-gain observer is coupled with a controller. The authors show that there exists a family of adaptation functions for the high-gain parameter such that the observer-controler closed loop is stable.

Those conditions, imposed on the vector fields f_i, appear to be very technical, but are in fact necessary in order to prove the convergence of the observer. Moreover, the importance of the function$\Gamma(\ u, y)$ has not to be underestimated as it is of the utmost importance in the definition of the update law of the high-gain parameter.

Definition 20

*The **high-gain observer with updated gain** is defined by the set of equations:*

$$\left\{\begin{array}{rcl} \dot{z} & = & A(y)x + b(y, z_2, .., z_n, , u) + \theta\mathcal{L}A(y)K\left(\frac{z_1-y}{\theta^b}\right) \\ \dot{\theta} & = & \theta\left[\varphi_1(\varphi_2 - \theta) + \varphi_3\mathbf{\Gamma(u, y)}\left(\mathbf{1} + \sum_{\mathbf{j=2}}^{\mathbf{n}} |\mathbf{\hat{x}_j}|^{\mathbf{v_j}}\right)\right] \end{array}\right. \tag{2.15}$$

where $\mathcal{L} = diag(L^b, L^{b+1}, ..., L^{b+n-1})$.

Theorem 21

Consider the system (2.13) and the associated observer (2.15). It is is then possible to choose the parameters φ_1, φ_2, φ_3 *(with* φ_2, φ_3 *high enough) such that for any* $L(0) \geq \varphi_2$ *the estimation error* $e(t) = z(t) - x(t)$ *is bounded as follows:*

$$|\mathcal{L}^{-1}e(t)| \leq \beta_1(\mathcal{L}(0)^{-1}e(0), t) + sup_{s\in[0,t[}\gamma_1\left(\left|\begin{pmatrix} \frac{\delta(s)}{\varphi_2} \\ \alpha(y(s))\delta_y(s) \end{pmatrix}\right|\right)$$

for all $t \in [0, T_u[$. *Moreover* L *statisfies the relation:*

$$L(t) \leq 4\varphi_2 + \beta_2\left(\begin{pmatrix} e(0) \\ L(0) \end{pmatrix}, t\right) + sup_{s\in[0,t[}\gamma_2\left(\left|\begin{pmatrix} \frac{\delta(s)}{\varphi_2} \\ \alpha(y(s))\delta_y(s) \\ \Gamma(u(s), y(s)) \\ x(s) \end{pmatrix}\right|\right)$$

where β_1 *and* β_2 *are* $\mathcal{KL}$ *functions, and* γ_1, γ_2 *are functions of class* $\mathcal{K}$.

This work uses a special form of the phase variable representation (2.2) of J-P. Gauthier *et al.* and therefore applies to observable systems in the sense of Section (2.2) above. The observer (2.15) has roughly the same structure as a Luenberger high-gain observer except for the fact that the correction gain is given as a function of the output error. The update function is determined by a function that bounds the incremental rate of the vector field $b(y, x, u)$(this is the part written in bold in (2.15)).

The strategy adopted here isn't based on a global Lipschitz vector field, which implies the off-line search and tuning of the upper bound (or the value) of the high-gain parameter θ. Instead, the observer tunes itself as a consequence of the adaptation function. However the function that drives the adaptation may not be that easy to find.

The idea is quite similar to that of E. Bullinger and F. Allgőwer [35], with the difference being that in their case the adaptation is driven by the output error and in the case of L. Praly *et al.* the adaptation is model dependent.

Theorem 21 states[17] that the observer (2.15) together with its adaptation function gives, at least for bounded solutions, an estimation error converging to a ball centered at the origin with a radius that depends on the asymptotic L^∞-norm of the disturbances δ and δ_y. This

[17]The precise and complete theorem appear in the article [16], Theorem 1.

consequently means that provided the disturbances vanish, the estimation error converges to 0. This observer is not based on any quality metric of the estimation and the evolution of the high-gain doesn't depend of the quality convergence. Therefore the situation may arise such that the observer has already converged and that the high-gain is still high, which would therefore amplify the noise.

The work herein, contrary to this section's observer, is set in the global Lipschitz setting. Further, it aims to provide an observer for which the high-gain parameter decreases to 1 (or the the lowest value allowed by the user) when the local convergence[18] of the algorithm can be used (i.e. the high-gain is not needed anymore).

2.6.3 Ahrens and Khalil

This paper deals with a closed loop control strategy that comprises an observer having a high-gain switching scheme. We focus on the observer's definition together with the switching strategy used, and only give a simplified version of the system the authors consider, refer to [11] for details.

Definition 22

*The **simplified version** of the system used in [11] is*

$$\left\{ \begin{array}{rcl} \dot{x} & = & Ax + B\phi(x, d, u) \\ y & = & Cx + v \end{array} \right. \tag{2.16}$$

where

- $x \in \mathbb{R}^n$ *is the state variable,*
- $y \in \mathbb{R}$ *is the output,*
- $d(t) \in \mathcal{R}^p$ *is a vector of exogenous signals,*
- $v(t) \in \mathbb{R}$ *is the measurement noise, and*
- $u(t)$ *is the control variable.*

The matrices A, B and C are:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

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

The set of assumptions for this system is:

[18] The convergence of the extended Kalman filter can be proven for small initial errors only (see the proof of Chapter 3).

- *$d(t)$ is continuously differentiable, and takes its values in a compact subset of $\mathbb{R}^p$,*
- *both $d(t)$ and $\frac{d}{dt}d(t)$ are bounded,*
- *$v : \mathbb{R}^+ \mapsto \mathbb{R}$ is a measurable function of t and is bounded (i.e. $\exists \mu > 0$, $|v(t)| \leq \mu$), and*
- *ϕ is a locally Lipschitz function in x and u, uniformly in d, over the domain of interest. The Lipschitz constant is independent of $d(t)$.*

The observer proposed for such a system comes from earlier works such as [50].

Definition 23

*Let us denote by z the estimated state. The **observer** is*

$$\dot{z} = Az + B\phi(z, d, u) - H_i(Cz - y)$$

where for $i = 1, 2$

$$H_i' = \left(\begin{array}{cccc} \frac{\alpha_{1,i}}{\theta_i} & \frac{\alpha_{2,i}}{\theta_i^2}, ..., \frac{\alpha_{n,i}}{\theta_i^n} \end{array} \right).$$

The θ_i's are small positive parameters such that $0 < \theta_1 < \theta_2$, and the α_i's are chosen in such a way that the roots of the polynomial:

$$s^n + \alpha_{1,i}s^{n-1} + \alpha_{2,i}s^{n-2} + ... + \alpha_{n,i} = 0$$

have negative real parts.

This observer is defined with two values for the high-gain parameter:

- θ_1 corresponds to the fast state reconstruction mode,
- θ_2 makes the observer much more efficient *w.r.t.* noise filtering.

The two sets of α_i parameters can be chosen such that they are distinct from one another. The switching scheme between the two values of θ has two main restrictions:

- the value of θ should change whenever an excessively large estimation error is detected, and
- the value of θ should not change because of overshoots. During an overshoot, the situation may arise when the estimated trajectory crosses the real trajectory, but has not yet converged. Switching from θ_1 to θ_2, in this case, is not desirable.

Definition 24 ***(Switching scheme)***

Let us define $\delta > 0$ and $T_d > 0$, two constant parameters and $\mathcal{D} = [-\delta; \delta]$. The value of θ is changed whenever[19] *$(z_1 - y_1)$ exits or enters $\mathcal{D}$. When an overshoot occurs, the estimation error may enter and exit quickly the domain $\mathcal{D}$: convergence is not achieved yet. The large value of the high-gain parameter is still needed. Those situations are handled by the use of a delay timer. Priority is given to the high-gain mode (see Figure 2.2):*

[19] Matrix C gives us $(Cz - y) = (z_1 - y_1)$.

- *whenever* $|y_1 - z_1| > \delta$ *then* $\theta = \theta_1$*, and the delay timer is reset,*
- *when* $|y_1 - z_1| < \delta$*, we don't know wether it is an overshoot or not:* $\theta = \theta_1$ *and the delay timer is started,*
- *when* $|y_1 - z_1| < \delta$ *and the delay timer is equal to* T_d*, estimation is satisfactory and* $\theta = \theta_2$*.*

The authors consider a control strategy that stabilizes the system provided that the full state is known. They include a fixed high-gain observer and consider the closed loop system. They propose a set of assumptions such that the *system-observer-controler* ensemble is uniformly asymptotically stable[20], (see [11], Theorem 1). The last step is the demonstration that stability remains when the high-gain of the observer is switched between two well defined values (see [11], Theorem 2 and example of Section 4).

The Luemberger observer doesn't have the same local properties as the extended Kalman filter, namely good filtering properties and analytically guaranteed convergence for small initial errors. We therefore expect a high-gain extended Kalman filter having a varying θ parameter to be more efficient with respect to the noise filtering issue.

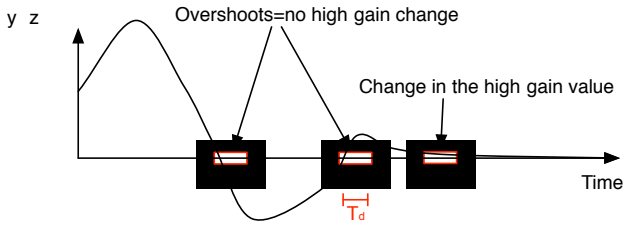

Figure 2.2: Switching strategy to deal with peaking.

2.6.4 Adaptive Kalman Filters

We now consider the problem of the adaptation of the high-gain parameter of Kalman filters. Recall that the high-gain parameter is used to provide the Q and R matrices with a specific structure. Therefore, the adaptation of the high-gain parameter may be seen as the modification of those two matrices[21]. There is nonetheless a big difference when the high-gain structure is not considered: there is no proof of convergence of the extended Kalman filter when the estimated state doesn't lie in a neighborhood of the real state. The situation could be even worse. Examples of systems for which the filter doesn't converge can be found in [100, 101].

The adaptation problem, when viewed as the adaptation of the Q and R matrices, has been the object of quite a few publications both in linear and nonlinear cases. In the linear

[20] The Theorem demonstrates that all the trajectories are bounded (ultimately with time) and that the trajectory (under output feedback i.e. using the observer) is close to that of the state feedback (i.e. controller with full state knowledge).

[21] Recall that when the system is modeled using stochastic differential equations, Q and R represent the covariances of the state and output noise, respectively.

part of the world, early references may be found in the book edited by A. Gelb[58], the two volumes book of P. S. Maybeck [89, 90], Chapter 10 of the second one in particular, the book from C. K. Chui and G. Chen [43]. Recent papers can be found in the INS/GPS community, such as [96, 115] or the book from M. S. Grewal [60]. The review article of R. K. Mehra [92] is warmly advised as an introduction to the topic. We do not expatiate on the subject since 1) most of the techniques developed in those papers are statistical methods, 2) the main bottleneck in the analysis is due to the linearization of the model in order to use the Riccati equation which is specific to the nonlinear case, 3) when switching based methods are used we have a better time explaining them directly in the nonlinear setting.

In the non linear case a vast majority of strategies are proposed for discrete-time systems. We describe a subset of those strategies below.

Definition 25

*A **discrete time system** is defined by a set of two equations of the form:*

$$\left\{ \begin{array}{rcl} x_{k+1} & = & f(x_k, u_k) \\ y_{k+1} & = & h(x_{k+1}) \end{array} \right. \tag{2.17}$$

with the usual notation $x_k = x(k\delta_t)$, $\delta_t > 0$ being the sample time. At least one of the two functions f and h is nonlinear.

The discrete extended Kalman filter associated with this system is given by the set of equations:

$$Prediction \left\{ \begin{array}{rcl} z^-_{k+1} & = & f(z_k, u_k) \\ P^-_{k+1} & = & A_k P_k A_k^{'} + Q_k, \end{array} \right.$$

$$Correction \left\{ \begin{array}{rcl} z_{k+1} & = & z^-_{k+1} + L_{k+1}(y_{k+1} - h(z^-_{k+1})) \\ P^+_{k+1} & = & (I_d - L_{k+1}C)P^-_{k+1} \\ L_{k+1} & = & P^-_{k+1}C^{'}(CP^-_{k+1}C^{'} + R)^{-1}, \end{array} \right.$$

where

- *z_k denotes the estimated state, and P_0 is a symmetric positive definite matrix,*
- *A is the jacobian matrix of f, computed along the estimated trajectory,*
- *C is the jacobian matrix of h, computed along the estimated trajectory.*

The first strategy we present was proposed by M. G. Pappas and J. E. Doss in their article [99] (1988). Although they do not consider the observability issue of the system, it is nonetheless an underlying concern of their work:

- when the system is at steady state, the system is *less observable* and a slow observer is required to give an accurate estimate, noise smoothing being an additional derived benefit,

– when the state of the system changes (different operating point, modification of the physical characteristics of the process,...), it becomes *more observable*. A faster observer can be considered and is needed to efficiently track the rapidly changing state variables.

The implementation they propose consists of two observers in parallel. Changes in the state of the system are seen as faults. They are detected via a modified fault detection algorithm [62]. The original algorithm is decomposed into three steps.

1. The fault input sequence is formed as

$$w_k = \alpha_1 w_{k-1} + (z_k - z_{k-1})$$

where w_k is an estimation of the direction of the parameter change. At steady state, $(z_k - z_{k-1})$ corresponds to the measurement noise and therefore its sign is expected to change frequently. Depending on the value of α_1 so is the case of w_k.

2. The fault test sequence is

$$s_k = sign\left((z_k - z_{k-1})\, w_{k-1}\right).$$

   – s_k negative indicates that the parameter estimate variation is changing direction: no fault occurs,

   – s_k positive over many successive tests, indicates that the variation is constant and a fault is detected.

3. The fault sequence is filtered with the equation:

$$r_k = \alpha_2 r_{k-1} + (1 - \alpha_2) s_k$$

where α_2 determines the speed of fault detection and the rate of false alarms[22].

This strategy is modified in the three following ways:

1. in order to reduce the influence of small noise components in the signal, the number of significant figures used in the calculations of $(z_k - z_{k-1})$ is truncated,

2. in order to cope with the situation when $(z_k - z_{k-1}) = 0$, the sequence s is modified as

$$s_k = sign(-10^8 + (z_k - z_{k-1}) w_{k-1}).$$

When the parameter estimations doesn't change, the sign function remains negative and no fault is detected,

3. the sequence $r(t)$ is clamped at a minimum value of -0.5, preventing excessively large drifts at steady state.

[22] For high values of α_2 we obtain a high pass filter like behavior, and conversely for α_2 small.

In this strategy, α_2 is a very important parameter as it is used to decide if the sign is considered to be either constant or if it is *constantly changing*.

The ideas behind **the second strategy** come directly from articles like the one of M. S. Mehra [92] or the book of P. S. Maybeck [90]. References for this method are [69] (sensor fusion in robotics), [83] (visual motion estimation via camera sensor), and [37, 102] (in flight orientation). This method aims at estimating Q, the process state noise covariance matrix. It is viewed as a measure of the uncertainty in the state dynamics between two consecutive updates of the observer. An observation of Q, denoted Q^* is given by the equation (see [37]):

$$Q^* = (z_{k+1} - z_{k+1}^-)(z_{k+1} - z_{k+1}^-)' + P_{k+1}^- - P_{k+1} - Q_k \ ,$$

which can be rewritten

$$\begin{aligned} Q^* &= (z_{k+1} - z_{k+1}^-)(z_{k+1} - z_{k+1}^-)' - (P_{k+1} - (P_{k+1}^- - Q_k)) \\ &= (z_{k+1} - z_{k+1}^-)(z_{k+1} - z_{k+1}^-)' - (A_k P_k A_k' - Q_k)). \end{aligned}$$

The new value for the matrix Q, denoted $\hat{Q}_{k+1}$ is obtained using a moving average (or low pass filter) process:

$$\hat{Q}_{k+1} = \hat{Q}_k + \frac{1}{L_Q}\left(Q^* - \hat{Q}_k\right).$$

L_Q is the size of the window that sets the number of updates being averaged. In this procedure, L_Q is a performance parameter that has to be tuned. The quantity $(z_{k+1} - z_{k+1}^-) = K_k(y_k - h_k(z_k^-))$ plays a key role in this strategy. It is denoted as innovation, and contains specific information on the quality of the estimation. In our work, we use a modified definition of innovation, and prove that it is a quality measurement of the estimation error[23].

A **third strategy** consists of designing of a set of nonlinear observers with different values for Q and R. They are used in parallel and the final estimated state is chosen among all the estimates available. A selection criteria has to be defined: minimization of innovation is the most straightforward method that can be used (see the observer of Subsection 2.6.6). (We refer the reader to the algorithm of K. J. Bradshaw, I. D. Reid and D. W. Murray [32], and references therein.) Every estimate is associated with a probability density computed with a maximum likelihood algorithm. The state estimate is then obtained as a combination of all the observers' outputs weighted by their associated probability.

Notice that for nonlinear systems, $\mathbb{E}[u(x)]$ is distinct from $u(\mathbb{E}[x])$. An interesting feature of this strategy, is that the first quantity can be computed quite naturally.

2.6.5 Boutayeb, Darouach and coworkers

In their research M. Boutayeb, M. Darouach and coworkers proposed several types of observers, including extended Kalman filter based obsevers [30, 31], observers based on the differential mean value theorems [116, 117], $\mathcal{H}_\infty$ filtering [13], or for systems facing bounded

[23] Cf. Lemma 33 of Chapter 3

disturbances [18]. This section deals with the observer described in [31], and in particular with the adaptive scheme proposed in [30]. The analysis they propose is set in the discrete time setting.

First of all, note that the approach we described in the first part of this Chapter, and the approach followed in the articles cited above are different, in the sense that the systems considered are not expected to display the same observability property. Indeed in the present case, the authors *only* need the system to be *N-locally uniformly observable* and do not perform any change of variables. This implies that the class of nonlinear systems for which the observer is proven to converge is bigger than the one considered in the present work (see the numerical examples displayed in [30], for example). This observer can be used for systems that cannot be put into a canonical observability form. The drawback to this approach then, is that the observer converges locally and asymptotically (i.e. the state error is not upper bounded by an exponential term).

Definition 26

1. *We consider a discrete, nonlinear, system as in Definition 25 where* $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^{n_u}$ *and* $y_k \in \mathbb{R}^{n_y}$. *The maps* f *and* h *are assumed to be continuously differentiable with respect to the variable* x.

2. *The observer is defined as:*

$$\begin{cases} z_{k+1/k} &= f(z_k, u_k) \\ P_{k+1/k} &= F_k P_k F_k^{'} + Q_k \end{cases}$$
$$\begin{cases} z_{k+1/k+1} &= z_{k+1/k} - K_{k+1}(h(z_{k+1/k}) - y_{k+1}) \\ P_{k+1/k+1} &= (I_n - K_{k+1} H_{k+1}) P_{k+1/k} \end{cases} \tag{2.18}$$

where

$$K_{k+1} = P_{k+1/k} H_{k+1}^{'} (H_{k+1} P_{k+1/k} H_{k+1}^{'} + R_{k+1})^{-1},$$

and

$$F_k = \frac{\partial f(x, u_k)}{\partial x}|_{x=z_k},$$
$$H_k = \frac{\partial h(x, u_k)}{\partial x}|_{x=z_{k+1/k}}.$$

This definition is that of a discrete extended Kalman filter. It is completed by a set of assumptions that appear in the statement of the convergence theorem. Since the extended Kalman filter is known to converge when the estimated state is very close to the real state, the following theorem increases the size of this region.

Theorem 27 *([30])*
We assume that:

1. *the system defined in equation (2.17) is N-locally uniformly rank observable, that is to*

say that there exists an integer $N \geq 1$ *such that*

$$rank \frac{\partial}{\partial x} \begin{pmatrix} h(x,u_k) \\ h(.,u_k) \circ f(x,u_k) \\ \vdots \\ h(.,u_{k+N-1}) \circ h(.,u_{k+N-2}) \circ h(.,u_k) \circ f(x,u_k) \end{pmatrix}_{|_{x=x_k}=n}$$

for all $x_k \in K$, *and* N*-tuple of controls* $(u_k, \ldots, u_{k+N-1}) \in \mathcal{U}$ *(where* K *and* $\mathcal{U}$ *are two compact subsets of* $\mathbb{R}^n$ *and* $(\mathbb{R}^{n_u})^N$, *respectively),*

2. F_k, H_k *are uniformly bounded matrices, and* F_k^{-1} *exists.*

Let us define:

1. *the time varying matrices* α *and* β *by:*

$$\begin{aligned} (x_{k+1} - z_{k+1/k}) &= \beta_k F_k (x_k - z_k) \\ \alpha_{k+1} e_{k+1} &= H_{k+1}(x_{k+1} - z_{k+1/k}), \end{aligned}$$

2. *the weighting matrices* R_k *and* Q_k *such that, there exists a parameter* $0 < \zeta < 1$ *such that*

$$\sup_{i=1,\ldots,n_y} |(\alpha_{k+1})_i - 1| \leq \left(\frac{\underline{\sigma}(R_{k+1})}{\overline{\sigma}(H_{k+1} P_{k+1/k} H_{k+1}^{'} + R_{k+1})} \right)^{\frac{1}{2}},$$

and that[24]

$$\sup_{j=1,\ldots,n} |(\beta_k)_j| \leq \left(\frac{(1-\zeta)\underline{\sigma}(F_k P_k F_k^{'} + Q_k)}{\overline{\sigma}(F_k^{'})\overline{\sigma}(P_k)\overline{\sigma}(F_k)} \right)^{\frac{1}{2}}.$$

Then, the observer (2.18) ensures local asymptotic convergence:

$$\lim_{k \to \infty} (x_k - z_k) = 0.$$

In [30] the authors propose to choose $Q_k = \gamma e_k^{'} e_k I_n + \delta I_n$ where $e_k = h(z_{k/k-1}) - y_k$ is the innovation. γ is taken sufficiently large, and δ sufficiently small such that the inequalities of Assumption (4) are met for all values of e_k. Special attention must be given to the fact that Q_k should not be set to a high value when the innovation is small[25].

The adaptation strategy we propose in Chapter 3 is based on the same kind of strategy but uses *the innovation over a sliding horizon* as the quality measurement for the estimation. By doing this we can link the adaptive scheme to the proof of convergence of the observer, which is not done for the observer of this section.

[24] $\overline{\sigma}$ and $\underline{\sigma}$ denotes respectively the maximum and minimum singular values.

[25] The article of L. Z. Guo and Q. M. Zhu [61], propose a hybrid strategy based on this subsection's observer. They use the structure proposed by M. Boutayeb *et al* together with a neural network approach.

2.6.6 E. Busvelle and J-P. Gauthier

The article [38] of E. Busvelle and J-P Gauthier propose an observer that is high-gain at time 0 and then decreases toward 1. In a nutshell, the observer evolves from a pure high-gain mode that ensures convergence to an extended Kalman filter configuration that efficiently smooths the noise. To achieve this, the high-gain parameter is allowed to decrease and the convergence is proven. This article is, in some sense, the starting point of the present Ph.D. work.

Definition 28

*The **high-gain and non high-gain extended Kalman filter**, for a system as defined in Section 2.4, is given by the three equations:*

$$\left\{\begin{array}{rcl} \frac{dz}{dt} & = & A(u)z + b(z,u) - S(t)^{-1}C'R^{-1}(Cz - y(t)) \\ \frac{dS}{dt} & = & -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'R^{-1}C - SQ_\theta S \\ \frac{d\theta}{dt} & = & \lambda(1-\theta) \end{array}\right. \tag{2.19}$$

where $Q_\theta = \theta^2\Delta^{-1}Q\Delta^{-1}$, $\Delta = diag(\{1, \frac{1}{\theta}, ..., (\frac{1}{\theta})^{n-1}\})$, Q *and* R *as in (2.9), and* λ *is a positive parameter.*

If $\theta(0) = 1$, *then* $\theta(t) \equiv 1$ *and (2.19) is nothing else than a classical extended Kalman filter applied in a canonical form of coordinates. Therefore it may not converge, depending on the initial conditions of the system.*

If $\lambda = 0$ *and* $\theta(0) = \theta_0$ *are large, then* $\theta(t) \equiv \theta_0$ *remains large and (2.19) is a high-gain extended Kalman filter as defined above in Equation (2.9).*

The idea in (2.19) is to set $\theta(0)$ to a sufficiently large value, and to set λ to a sufficiently small value such that the observer converges exponentially quickly at the beginning. The estimated state reaches the vicinity of the real trajectory before θ becomes too small and the local convergence of the extended Kalman filter guarantees that it will remain close to the state of the system.

Theorem 29

For any $\varepsilon^* > 0$, *there exists* λ_0 *such that for all* $0 \leq \lambda \leq \lambda_0$, *for all* θ_0 *large enough, for all* $S_0 \geq c\ Id$, *for all* $\chi \subset \mathbb{R}^n$, χ *a compact subset, for all* $\varepsilon_0 = z_0 - x_0$,*with* $(z_0, x_0) \in \chi^2$, *with* $x_0 \in \chi$ *the following estimation holds for all* $t \geq 0$:

$$||\varepsilon(t)||^2 \leq ||\varepsilon(0)||^2\varepsilon^* e^{-at}.$$

Moreover the short term estimate

$$||\varepsilon(t)||^2 \leq ||\varepsilon(0)||^2\theta(t)^{2(n-1)}e^{-(a_1\theta(T)-a_2)t}$$

holds for all $T > 0$ *and for all* $0 \leq t \leq T$, *for all* θ_0 *sufficiently large. The scalars* a_1 *and* a_2 *are positive constants.*

This theorem demonstrates that the observer converges for any initial error. Nevertheless it is clearly not a persistent observer since after some time it is more or less equivalent to an extended Kalman filter because $\theta(t)$ is close to one. In order to make it persistent, the authors propose to use several such observers, each of them being initialized at different times

in such a way that at any moment at least one of the observers has θ large and at least one observer has θ close to 1. The state of the observer having the shortest output error[26] is then selected. As shown in [38], this observer performs well in our applications. It is successfully applied for identification purposes in [40]. Nevertheless this procedure is not very reasonable since:

- Even if the overall construction gives a persistent observer, it is a time-dependant observer,
- It can be time consuming since it requires at least five (empirical value) observers in parallel,
- The parameter λ has to be chosen sufficiently small, which means that after a perturbation, we can not return as quickly as we would like to a classical extended Kalman filter, even if the observer performs well,
- The choice of the criteria used to select the best prediction between our observers, is not theoretically justified.

This second chapter, which focused on the theoretical framework that encompasses our work, ends with the analysis of this last observer. In this chapter, we also provided insight into several adaptation strategies. In the next chapter we introduce the adaptive high-gain extended Kalman filter and develop the full proof of convergence.

[26] The output error $(Cz - y)$ is the equivalent of innovation for continuous time systems.

Chapter 3

Adaptive High-gain Observer: Definition and Convergence

Contents

Up to this point, we have already studied the system, i.e., the system is of the normal form, either naturally or after a change of variables. According to theory, in order to reconstruct the state of this system, we can use any of the exponentially converging observers of Chapter 2. However, we wont.

In this chapter, we solve the convergence part of the observability problem. Our goal is to define an observer that combines the antagonistic behaviors of the extended Kalman filter (EKF) and the high-gain extended Kalman filter (HG-EKF).

The EKF is extensively used, 1) because of its attractive filtering properties (as explained in articles such as [101]), and 2) because it actually performs well in practice. However, a proof of convergence for this algorithm is known only for small initial estimation errors (as it can be seen in [17, 38] or within the proof of the main theorem below). Additionally, from a practical point of view, the EKF handles large perturbations with difficulty, as has been observed in simulations and experiments.

Contrarily, **the HG-EKF** possesses improved global properties [47]. It converges regardless of the initial guess and/or independently of large perturbations. On the other hand, it is rather sensitive with respect to noise.

Recall that the high-gain structure uses a single parameter denoted θ ($\theta > 1$), and referred to as the high-gain parameter. The HG-EKF does its global job if and only if θ is sufficiently large. When θ is set to 1, it is formally equivalent to the standard EKF.
The idea here is to make the parameter θ adaptive. Thus,

- when the estimated state is far from the real state, θ is made sufficiently large such that the observer converges for any initial guess,
- when the estimation is sufficiently close to the real state we allow θ to decrease. Once this condition is satisfied, the local convergence of the extended Kalman filter is applicable and the noise is more efficiently smoothed.

It is natural to perform the adaptation under the guise of a differential equation of the form

$$\dot{\theta} = \mathcal{F}(\theta, \mathcal{I}), \tag{3.1}$$

where $\mathcal{I}$ is some quantity reflecting the amplitude of the estimation error: the smaller $\mathcal{I}$ the smaller the error.

We introduce a simple and natural concept of "innovation" for the quantity $\mathcal{I}$. This innovation concept is different from the one that is usually used[1]. It allows us to reflect the estimation error more precisely.

The convergence of this observer is established in the continuous time setting for multiple inputs, single output systems[2]. This choice is made for the sake of maintaining the simplicity of the exposure, because a few modifications have to be made in order to cope with multiple outputs systems. Such modifications are explained in Chapter 5.

[1] Most of the time innovation is defined as

- $\mathcal{I} = y - h(z, u)$ for the continuous case, with the notations of Definition 15,
- $\mathcal{I} = y - h(z_k^-, u_k)$ for the discrete case, with the notations of Definition 25.

[2] We have the generic case when $n_u > 1$ and the non-generic case when $n_u = 1$, C.f. Chapter 2.

3.1 Systems Under Consideration

We consider multiple input, single output nonlinear systems as in Section 2.4. We properly define all the constants:

$$\left\{ \begin{array}{rcl} \frac{dx}{dt} & = & A(u)\,x + b(x,u) \\ y & = & C(u)\,x \end{array} \right. \tag{3.2}$$

where $x(t) \in \mathcal{X} \subset \mathbb{R}^n$, $\mathcal{X}$ compact, $y(t) \in \mathbb{R}$, and $u(t) \in \mathcal{U}_{\text{adm}} \subset \mathbb{R}^{n_u}$ is bounded for all times. The matrices $A(u)$ and $C(u)$ are:

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

$$C(u) = \begin{pmatrix} a_1(u) & 0 & \cdots & 0 \end{pmatrix},$$

with[3] $0 < a_m \leq a_i(u) \leq a_M$ for any u in $\mathcal{U}_{\text{adm}}$. The vector field $b(x,u)$ is assumed to be compactly supported and to have the following triangular structure:

$$b(x,u) = \begin{pmatrix} b_1(x_1,u) \\ b_2(x_1,x_2,u) \\ \vdots \\ b_n(x_1,\ldots,x_n,u) \end{pmatrix}.$$

We denote L_b the bound on the Jacobian matrix $b^*(x,u)$ of $b(x,u)$ (i.e. $\|b^*(x,u)\| \leq L_b$). Since $b(x,u)$ is compactly supported and u is bounded, b is Lipschitz *w.r.t.* x, and uniform *w.r.t.* u: $\|b(x_1,u) - b(x_2,u)\| \leq L_b \|x_1 - x_2\|$.

3.2 Observer Definition

Let

- Q be a $(n \times n)$ symmetric positive definite matrix, and
- R and θ be strictly positive real numbers, $\theta \geq 1$.

Set

$$\Delta = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\theta^{n-1}} \end{pmatrix},$$

$$Q_\theta = \theta \Delta^{-1} Q \Delta^{-1},$$

and

$$R_\theta = \theta^{-1} R.$$

[3] The crucial point here is that $a_i(u)$ must remain suitably *distant from zero*. The condition $-a_M < a_i < a_m < 0$ is also valid.

Definition 30

*The **adaptive high-gain extended Kalman filter** is the system:*

$$\begin{cases} \frac{dz}{dt} &= A(u)z + b(z,u) - S^{-1}C'R_\theta^{-1}(Cz - y(t)) \\ \frac{dS}{dt} &= -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'R_\theta^{-1}C - SQ_\theta S \\ \frac{d\theta}{dt} &= \mathcal{F}(\theta, \mathcal{I}_d(t)). \end{cases} \tag{3.3}$$

The functions $\mathcal{F}$ and $\mathcal{I}_d$ will be defined later, in Section 3.3 and Lemma 42. The function $\mathcal{I}_d$ is called the innovation. The initial conditions are $z(0) \in \chi$, $S(0)$ is symmetric positive definite, and $\theta(0) = 1$.

The function $\mathcal{F}$ has only to satisfy certain requirements stated precisely in Lemma 42. Therefore, several different choices for an adaptation function are possible.

Roughly speaking, $\mathcal{F}(\theta, \mathcal{I}_d(t))$ should be such that if the estimation $z(t)$ is far from $x(t)$ then $\theta(t)$ increases (high-gain mode). Contrarily, if $z(t)$ is close to $x(t)$, θ goes to 1 (Kalman filtering mode). As it is clear from the proof of Theorem 36, this observer makes sense only when $\theta(t) \geq 1$, for all $t \geq 0$. This is therefore another requirement that $\mathcal{F}(\theta, \mathcal{I}_d)$ has to meet.

The achievement of this behavior requires that we evaluate the quality of the estimation. This is the object of the next section.

Remark 31

1. *Readers familiar with high-gain observers may notice that the matrices R_θ and Q_θ are not exactly the same as in earlier articles such as [38, 47, 57]. The definitions developed here can also be substituted into those previous works without consequence.*

2. *The hypothesis $\theta(0) = 1$ may appear a bit atypical as compared to the results in [38, 57] for instance.*

   - *For technical reasons, the result of Lemma 39 depends on the initial value of θ, and $\theta(0) = 1$ has no impact on α and β.*
   - *Secondly, note that in the case of large perturbations, θ will increase. In these instances, the initial value of θ is of little importance as we show in Lemma 42 that the adaptation function can be chosen in such a way that θ reaches any large value in an arbitrary small time.*
   - *Finally, in the ideal case of no initial error, θ doesn't increase which saves us from useless noise sensitivity due to a large high-gain initial value.*

3.3 Innovation

The innovation $\mathcal{I}_d$ is a measurement of the quality of the estimation. It is different[4] from the standard concept of innovation, which is based on a linearization around the estimated

[4]The same definition of innovation is used for moving horizon observers where the estimated state is the solution of a minimization problem. The cost function used here represents the proximity to the real state. It is the innovation we use, ([12, 95]).
In related publications, observability is defined by the inequality of Lemma 33. In our case, the inequality is a consequence of the observability theory.

trajectory.

Definition 32

*For a "forgetting horizon" $d > 0$, the **innovation** is:*

$$\mathfrak{I}_d(t) = \int_{t-d}^{t} \|y(t-d, x(t-d), \tau) - y(t-d, z(t-d), \tau)\|^2 \, d\tau \tag{3.4}$$

where $y(t_0, x_0, \tau)$ denotes the output of the system (3.2) at time τ with $x(t_0) = x_0$.
Hence $y(t-d, x(t-d), \tau)$ denotes $y(\tau)$, the output of the process. Notice that $y(t-d, z(t-d), \tau)$ is not the output of the observer.

For a good implementation, it is important to understand the significance of this definition. Figure 3.1 illustrates the situation at time t. Innovation is obtained as the square of the L^2 distance between the black (plain) and the red (dot and dashed) curves. They respectively represent the output of the system on the time interval $[t-d, t]$, and the prediction performed with $z(t-d)$ as initial state.

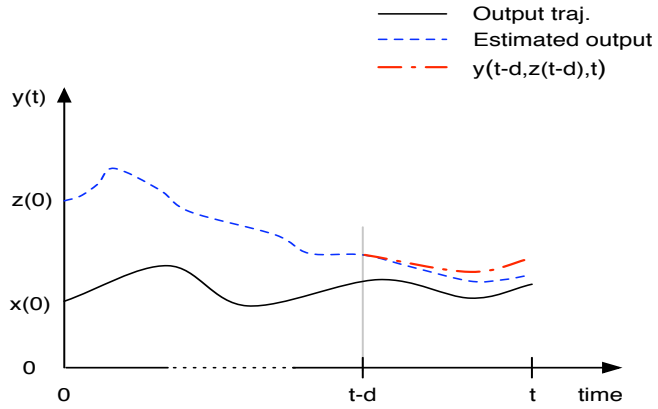

Figure 3.1: The Computation of Innovation.

The importance of innovation in this construction is explained by the following lemma. As we will see in Section 3.8, this is the cornerstone of the proof.

Lemma 33

Let $x_1^0, x_2^0 \in \mathbb{R}^n$, and $u \in \mathcal{U}_{\text{adm}}$. Let us consider the outputs $y(0, x_1^0, \cdot)$ and $y(0, x_2^0, \cdot)$ of system (3.2) with initial conditions respectively x_1^0 and x_2^0. Then the following property (called persistent observability) holds:

$$\forall d > 0, \exists \lambda_d^0 > 0 \text{ such that } \forall u \in L_b^1(\mathcal{U}_{adm})$$

$$\|x_1^0 - x_2^0\|^2 \leq \frac{1}{\lambda_d^0} \int_0^d \|y(0, x_1^0, \tau) - y(0, x_2^0, \tau)\|^2 d\tau. \tag{3.5}$$

Let us set $x_1^0 = z(t-d)$, and $x_2^0 = x(t-d)$ then Lemma 33 gives:

$$\|z(t-d) - x(t-d)\|^2 \leq \frac{1}{\lambda_d^0} \int_{t-d}^{t} \|y(\tau) - y(t-d, z(t-d), \tau)\|^2 \, d\tau,$$

or, equivalently,

$$\|z(t-d) - x(t-d)\|^2 \leq \frac{1}{\lambda_d^0} \mathfrak{I}_d(t).$$

This is to say, that up to a multiplicative constant, innovation at time t upper bounds the estimation error at time $t-d$.

Remark 34

One could think that the adaptation scheme is likely to react with a delay time d when the estimation error is large. However, as is explained in Chapter 4, Remark 45, it may not always be the case in practice.

Proof.

Let $x_1(t) = x_{x_1^0,u}(t)$ and $x_2(t) = x_{x_2^0,u}(t)$ be the solutions of (3.2) with $x_i(0) = x_i^0$, $i = 1, 2$. For any $a \in [0,1]$,

$$\begin{aligned} & b(a\,x_2 + (1-a)\,x_1, u) \\ &\quad = b(x_1, u) + \int_0^a \frac{\partial}{\partial \alpha} b(\alpha\,x_2 + (1-\alpha)\,x_1, u)\, d\alpha \\ &\quad = b(x_1, u) + \int_0^a \frac{\partial}{\partial x} b(\alpha\,x_2 + (1-\alpha)\,x_1, u)\, d\alpha\,(x_2 - x_1). \end{aligned}$$

Hence for $a = 1$

$$\begin{aligned} b(x_2, u) - b(x_1, u) &= \left(\int_0^1 \frac{\partial b}{\partial x}(\alpha\,x_2 + (1-\alpha)\,x_1, u)\, d\alpha \right)(x_2 - x_1) \\ &= B(t)(x_2 - x_1), \end{aligned}$$

where $B(t) = (b_{i,j})_{(i,j)\in\{1,..,n\}}$ is a lower triangular matrix since

$$b(x, u) = (b(x_1, u), b(x_1, x_2, u), \ldots, b(x, u))'.$$

Set $\varepsilon = x_1 - x_2$, and consider the system:

$$\begin{cases} \dot{\varepsilon} &= A(u)\,x_1 + b(x_1, u) - A(u)\,x_2 - b(x_2, u) \\ &= [A(u) + B(t)]\,\varepsilon \\ y_\epsilon &= C(u)\,\varepsilon = a_1(u)\,\varepsilon_1. \end{cases}$$

It is uniformly observable[5] as a result of the structure of $B(t)$. Let us consider $\Psi(t)$, the resolvent of the system, and the Gramm observability matrix G_d:

$$G_d = \int_0^d \Psi(v)' C' C \Psi(v)\, dv.$$

[5] See [57] for instance, or compute the observability matrix

$$\phi_\circ = \left[C' | (CA)' | \ldots | (CA^n)' \right]',$$

and check the full rank condition for all inputs.

Since $\|B(t)\| \leq L_b$ each $b_{i,j}(t)$ can be interpreted as a bounded element of $L^\infty_{[0,d]}(\mathbb{R})$. We identify $\left(L^\infty_{[0,d]}(\mathbb{R})\right)^{\frac{n(n+1)}{2}}$ to $L^\infty_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right)$ and consider the function:

$$\begin{array}{rccc} \Lambda: & L^\infty_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right) \times L^\infty_{[0,d]}(\mathbb{R}^{n_u}) & \longrightarrow & \mathbb{R}^+ \\ & (b_{i,j})_{(j\leq i)\in\{1,..,n\}}, u^\varsigma & \to & \lambda_{\min}(G_d) \end{array}$$

where $\lambda_{\min}(G_d)$ is the smallest eigenvalue of G_d. Let us endow $L^\infty_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right) \times L^\infty_{[0,d]}(\mathbb{R}^{n_u})$ with the weak-* topology[6] and $\mathbb{R}$ has the topology induced by the uniform convergence. The weak-* topology on a bounded set implies uniform continuity of the resolvent, hence Λ is continuous[7].
Since control variables are supposed to be bounded,

$$\Omega_1 = \left\{L^\infty_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right); \|B\| \leq L_b\right\}$$

and

$$\Omega_2 = \left\{u \in L^\infty_{[0,d]}(\mathbb{R}^n); \|u\| \leq M_u\right\}$$

are compact subsets. Therefore $\Lambda(\Omega_1 \times \Omega_2)$ is a compact subset of $\mathbb{R}$ *which does not contain* 0 since the system is observable for any input. Thus G_d is never singular. Moreover, for M_u sufficiently large, $\left\{u \in L^\infty_{[0,d]}(\mathbb{R}^n); \|u\| \leq M_u\right\}$ includes $L^\infty_{[0,d]}(\mathcal{U}_{\text{adm}})$.

Hence, there exists λ_d^0 such that $G_d \geq \lambda_d^0 Id$ for any u and any matrix $B(t)$ as above. Since

$$y\left(0, x_1^0, \tau\right) - y\left(0, x_2^0, \tau\right) = C\Psi(\tau) x_1^0 - C\Psi(\tau) x_2^0,$$

then

$$\left\|y\left(0, x_1^0, \tau\right) - y\left(0, x_2^0, \tau\right)\right\|^2 = \left\|C\Psi(\tau) x_1^0 - C\Psi(\tau) x_2^0\right\|^2,$$

and finally

$$\begin{array}{rcl} \displaystyle\int_0^d \left\|y\left(0, x_1^0, \tau\right) - y\left(0, x_2^0, \tau\right)\right\|^2 d\tau & = & \left(x_1^0 - x_2^0\right)' G_d \left(x_1^0 - x_2^0\right) \\ & \geq & \lambda_d^0 \left\|x_1^0 - x_2^0\right\|^2. \end{array} \tag{3.6}$$

■

Remark 35

As is clear from the proof, we could have used both linearizations along the trajectories x_1 and x_2 in order to define $\mathfrak{I}$. However, that definition would lead to the same inequality. In addition our definition is more practical to implement.

The solution of the Riccati equation can also not be used to obtain an information equivalent to innovation. Here, to compute our innovation, we make an exact prediction (without the correction term that could disturb the estimation).

[6] The definition of the weak-* topology is given in Appendix A.
[7] This property is explained in Appendix A.

3.4 Main Result

The exponential convergence of the adaptive high-gain extended Kalman filter is expressed in the theorem below.

Theorem 36

For any time $T^* > 0$ and any $\varepsilon^* > 0$, there exist $0 < d < T^*$ and a function $\mathcal{F}(\theta, \mathcal{I}_d)$ such that, for all times $t \geq T^*$ and any initial state couple $(x_0, z_0) \in \chi^2$:

$$\|x(t) - z(t)\|^2 \leq \varepsilon^* \mathrm{e}^{-a(t-T^*)}$$

where $a > 0$ is a constant (independent from ε^*).

This theorem can be expressed in two different ways: with or without a term $\|\epsilon_0\|^2$ in the upper bound. The bound of Theorem 36 was presented in [23] and [22][8]. The expression we use here should be interpreted to mean that the square of the error can be made arbitrarily small in an arbitrary small time. For the sake of completeness, we develop the other inequality in Remark 44.

The proof is a Lyapunov stability analysis which requires several preliminary computations and additional results. In order to facilitate comprehension, we divide the proof into several parts:

1. the computation of several preliminary inequalities, in particular the expression of the Lyapunov function we want to study,
2. the derivation of the properties of the Riccati matrix S,
3. the statement of several intermediary lemmas, among which is the lemma that states the existence of eligible adaptive functions, and finally
4. the articulation of the proof.

We begin with the computation of some preliminary inequalities in Section 3.5.

3.5 Preparation for the Proof

Remember[9] that $\theta \geq 1$, for all $t \geq 0$.
We denote by z the time dependent state variable of the observer.
The estimation error is $\varepsilon = z - x$.
We consider the change of variables $\tilde{x} = \Delta x$, and

- $\tilde{z} = \Delta z$, and $\tilde{\varepsilon} = \Delta \varepsilon$,
- $\tilde{S} = \Delta^{-1} S \Delta^{-1}$,

[8] The other bound is

$$\|x(t) - z(t)\|^2 \leq \|\varepsilon_0\|^2 \varepsilon^* \mathrm{e}^{-\alpha q_m (t-\tau)}.$$

[9] This is one of the requirements $\mathcal{F}(\theta, \mathcal{I}_d)$ have to meet. Existence of such a function is shown in Lemma 42.

– $\tilde{b}(.,u) = \Delta b(\Delta^{-1}.,u)$,

– $\tilde{b}^*(\cdot,u) = \Delta b^*\left(\Delta^{-1}\cdot,u\right)\Delta^{-1}$.

Since $C = \begin{pmatrix} a_1(u) & 0 & \dots & 0 \end{pmatrix}$, and $\Delta = diag\left(\left\{1,\theta^{-1},...,\theta^{-(n-1)}\right\}\right)$ then $C\Delta = C$. We have the following identity for the $A(u)$ andΔ:

$$A(u)\Delta = \begin{pmatrix} 0 & a_2(u) & 0 & \cdots & 0 \\ & 0 & a_3(u) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ & & & 0 & a_n(u) \\ 0 & & \cdots & & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\theta} & 0 & & \vdots \\ 0 & 0 & \frac{1}{\theta^2} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \frac{1}{\theta^{n-1}} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & \frac{a_2(u)}{\theta} & 0 & \cdots & 0 \\ & 0 & \frac{a_3(u)}{\theta^2} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ & & & 0 & \frac{a_n(u)}{\theta^{n-1}} \\ 0 & & \cdots & & 0 \end{pmatrix} = \frac{1}{\theta}\Delta A(u).$$

This relation leads to the set of equalities:

$$\begin{array}{llll} (a) & \Delta A = \theta A\Delta, & (b) & A'\Delta = \theta\Delta A', \\ (c) & A\Delta^{-1} = \theta\Delta^{-1}A, & (d) & \Delta^{-1}A' = \theta A'\Delta^{-1}. \end{array} \tag{3.7}$$

Because we want to express the time derivative of $\tilde{\varepsilon}$ we need to know the time derivative of Δ, as θ is time dependent. We simply write

$$\frac{d\Delta}{dt} = \begin{pmatrix} \frac{d(1)}{dt} & 0 & \cdots & 0 \\ 0 & \frac{d}{dt}\left(\frac{1}{\theta}\right) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{d}{dt}\left(\frac{1}{\theta^{n-1}}\right) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & -\frac{\dot{\theta}}{\theta^2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{(n-1)\dot{\theta}}{\theta^n} \end{pmatrix},$$

which can be rewritten as a multiplication of matrices with the use of $N = diag\left(\{0,1,2,...,n-1\}\right)$. We obtain the two identities[10]:

$$(a) \quad \frac{d}{dt}(\Delta) = -\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta}N\Delta \quad (b) \quad \frac{d}{dt}\left(\Delta^{-1}\right) = \frac{\mathcal{F}(\theta,\mathcal{I})}{\theta}N\Delta^{-1}. \tag{3.8}$$

The dynamics of the error are given by:

$$\dot{\varepsilon} = \dot{z} - \dot{x} = \left(A(u) - S^{-1}C'R_\theta^{-1}C\right)\varepsilon + b(z,u) - b(x,u),$$

and the error dynamics after the change of variables are:

$$\begin{array}{rcl} \frac{d\tilde{\varepsilon}}{dt} = & \frac{d\Delta}{dt}\varepsilon + \Delta\dot{\varepsilon} & \\ = & -\frac{\dot{\theta}}{\theta}N\Delta\varepsilon + \Delta\left(A(u) - S^{-1}C'R_\theta^{-1}C\right)\varepsilon + \Delta(b(z,u) - b(x,u)) & \\ = & -\frac{\dot{\theta}}{\theta}N\tilde{\varepsilon} + \theta A(u)\tilde{\varepsilon} - \theta\Delta S^{-1}\Delta\Delta^{-1}C'R^{-1}C\Delta^{-1}\Delta\varepsilon & \\ & \qquad + \Delta b\left(\Delta^{-1}\tilde{z},u\right) - \Delta b\left(\Delta^{-1}\tilde{x},u\right) & \\ = & \theta\left[-\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}N\tilde{\varepsilon} + A\tilde{\varepsilon} - \tilde{S}^{-1}C'R^{-1}C\tilde{\varepsilon} + \frac{1}{\theta}\left(\tilde{b}(\tilde{z},u) - \tilde{b}(\tilde{x},u)\right)\right]. & \end{array} \tag{3.9}$$

[10] remember that $\dot{\theta} = \mathcal{F}(\theta,\mathcal{I})$.

The Riccati equation turns into

$$\begin{aligned}
\tfrac{d\tilde{S}}{dt} = & \tfrac{\dot{\theta}}{\theta} N\Delta^{-1}S\Delta^{-1} + \tfrac{\dot{\theta}}{\theta}\Delta^{-1}S\Delta^{-1}N - \Delta^{-1}\left(A + b^*(z,u)\right)' S\Delta^{-1} \\
& -\Delta^{-1}S\left(A + b^*(z,u)\right)\Delta^{-1} + \Delta^{-1}C' R_\theta^{-1} C\Delta^{-1} \\
& -\Delta^{-1}SQ_\theta S\Delta^{-1} \\
= & \tfrac{\dot{\theta}}{\theta}\left(N\tilde{S} + \tilde{S}N\right) - \left(A\Delta^{-1} + b^*(z,u)\,\Delta^{-1}\right)' \Delta\tilde{S} \\
& -\tilde{S}\Delta\left(A\Delta^{-1} + b^*(z,u)\,\Delta^{-1}\right) + C' R_\theta^{-1} C - \tilde{S}\Delta Q_\theta \Delta\tilde{S} \\
= & \theta\left[\tfrac{\dot{\theta}}{\theta^2}\left(N\tilde{S} + \tilde{S}N\right) - \left(A'\tilde{S} + \tilde{S}A\right) + C' R^{-1} C - \tilde{S}Q\tilde{S}\right. \\
& \left. -\tilde{S}\Delta b^*(z,u)\,\Delta^{-1} - \Delta^{-1} b^{*\prime}(z,u)\,\Delta\tilde{S}\right] \\
= & \theta\left[\tfrac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}\left(N\tilde{S} + \tilde{S}N\right) - \left(A'\tilde{S} + \tilde{S}A\right) + C' R^{-1} C - \tilde{S}Q\tilde{S}\right. \\
& \left. -\tfrac{1}{\theta}\tilde{S}\tilde{b}^*(\tilde{z},u) - \tfrac{1}{\theta}\tilde{b}^{*\prime}(\tilde{z},u)\,\tilde{S}\right].
\end{aligned} \tag{3.10}$$

The derivative of the Lyapunov function $\widetilde{\varepsilon}' \widetilde{S}\widetilde{\varepsilon}$ is

$$\begin{aligned}
\tfrac{d\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}}{dt} = & \theta\left[-\tfrac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2} N\tilde{\epsilon} + A\tilde{\epsilon} - \tilde{S}^{-1} C' R^{-1} C\tilde{\epsilon} + \tfrac{1}{\theta}\left(\tilde{b}(\tilde{z},u) - \tilde{b}(\tilde{x},u)\right)\right]' \tilde{S}\tilde{\epsilon} \\
& +\theta\tilde{\epsilon}'\left[\tfrac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}\left(N\tilde{S} + \tilde{S}N\right) - \left(A'\tilde{S} + \tilde{S}A\right) + C' R^{-1} C - \tilde{S}Q\tilde{S}\right. \\
& \left. -\tfrac{1}{\theta}\left(\tilde{S}\tilde{b}^*(\tilde{z},u) + \tilde{b}^{*\prime}(\tilde{z},u)\,\tilde{S}\right)\right]\tilde{\epsilon} \\
& +\theta\tilde{\epsilon}'\tilde{S}\left[-\tfrac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2} N\tilde{\epsilon} + A\tilde{\epsilon} - \tilde{S}^{-1} C' R^{-1} C\tilde{\epsilon} + \tfrac{1}{\theta}\left(\tilde{b}(\tilde{z},u) - \tilde{b}(\tilde{x},u)\right)\right] \\
= & \theta\left[-\tilde{\epsilon}' C' R^{-1} C\tilde{\epsilon} - \tilde{\epsilon}'\tilde{S}Q\tilde{S}\tilde{\epsilon} + \tfrac{2}{\theta}\tilde{\epsilon}'\tilde{S}\left(\tilde{b}(\tilde{z},u) - \tilde{b}(\tilde{x},u) - \tilde{b}^*(\tilde{z},u)\,\tilde{\epsilon}\right)\right].
\end{aligned} \tag{3.11}$$

We consider that $Q \geq q_m\, Id$, and since $\widetilde{\varepsilon}' C' R^{-1} C\widetilde{\varepsilon} \geq 0$

$$\frac{d}{dt}\left(\widetilde{\varepsilon}'\widetilde{S}\widetilde{\varepsilon}\right) \leq -\theta q_m \widetilde{\varepsilon}'\widetilde{S}^2\widetilde{\varepsilon} + 2\widetilde{\varepsilon}'\widetilde{S}\left(\tilde{b}(\widetilde{z},u) - \tilde{b}(\widetilde{x},u) - \tilde{b}^*(\widetilde{z},u)\,\widetilde{\varepsilon}\right). \tag{3.12}$$

The theorem is proven using the inequality (3.12), which requires some knowledge of the properties of S. We can note that the equality (3.10) has a strong dependency on θ, which we must remove in order to derive properties for S. Indeed, for the moment we don't know which values θ should reach during runtime. To determine these values, we introduce the time reparametrization $d\tau = \theta(t)\, dt$, or equivalently $\tau = \int_0^t \theta(\nu)\, d\nu$. We denote:

- $\bar{x}(\tau) = \tilde{x}(t)$, $\bar{z}(\tau) = \tilde{z}(t)$, and $\bar{\varepsilon}(\tau) = \widetilde{\varepsilon}(t)$, and
- $\bar{\theta}(\tau) = \theta(t)$, $\bar{u}(\tau) = u(t)$, and $\overline{S}(\tau) = \widetilde{S}(t)$.

We obtain the time derivative with respect to the time scale, τ, using the simple calculation:

$$\frac{d\bar{\varepsilon}}{d\tau} = \frac{d\tilde{\varepsilon}(t)}{dt}\frac{dt}{d\tau} = \frac{1}{\theta(t)}\frac{d\tilde{\varepsilon}(t)}{dt}.$$

Therefore

$$\frac{d\bar{\epsilon}}{d\tau} = -\frac{\mathcal{F}(\theta,\mathcal{I})}{\bar{\theta}^2} N\bar{\varepsilon} + A\bar{\varepsilon} - \bar{S}^{-1} C' R^{-1} C\bar{\varepsilon} + \frac{1}{\bar{\theta}}\left(\tilde{b}(\bar{z},\bar{u}) - \tilde{b}(\bar{x},\bar{u})\right)$$

such that

$$\frac{d\bar{S}}{dt} = \frac{\mathcal{F}(\theta,\mathcal{I})}{\bar{\theta}^2}\left(N\bar{S}+\bar{S}N\right) - \left(A'\bar{S}+\bar{S}A\right) + C'R^{-1}C - \bar{S}Q\bar{S} - \frac{1}{\bar{\theta}}\left(\bar{S}\tilde{b}^*(\bar{z},\bar{u}) + \tilde{b}'^*(\bar{z},\bar{u})\bar{S}\right). \tag{3.13}$$

We complete the description of the observer in the τ time scale with the equation

$$\frac{d\bar{\theta}}{d\tau} = \frac{\mathcal{F}(\bar{\theta},\bar{\mathcal{I}})}{\bar{\theta}}.$$

This last set of equations is established in order to investigate the properties of the Riccati matrix, particularly the fact that it is bounded (Cf. Section 3.6).

Remark 37

The Lipstchitz constant of the vector field $b(.,u)$ is the same in the $x(t)$, $\tilde{x}(t)$ and $\bar{x}(\tau)$ coordinates. It occurs quite naturally in the single output case but implies a special definition of the observer in the multiple output case. Consequently this fact is proven in Chapter 5, Lemma 51, for systems having multiple outputs.

3.6 Boundedness of the Riccati Matrix

The Riccati matrix $S(t)$ has some very important properties:

- $S(t)$ can be upper and lower bounded by matrices of the form $c.Id$, and
- if $S(0)$ is symmetric definite positive, then $S(t)$ is also symmetric definite positive for all times.

Those properties are established in the book [57], with the difference being that the term $\left|\frac{\mathcal{F}(\bar{\theta},\bar{\mathcal{I}})}{\bar{\theta}^2}\right|$ doesn't appear in the Riccati equation (Cf. Section 2.4 of the 6^{th} chapter of the book), and that the result there is also only valid for times $t \geq T^*$, for some $T^* > 0$. As we will see in the present section, when $\theta(0)$ is set to 1 we can say a little bit more.

Lemma 38 *([57])*

Let us consider the Riccati equation (3.13). If the functions $a_i(u(t))$, $\left|\widetilde{b}^*_{i,j}(\overline{z},\overline{u})\right|$, $\left|\frac{\mathcal{F}(\bar{\theta},\bar{\mathcal{I}})}{\bar{\theta}^2}\right|$ are smaller than $a_M > 0$, and if $a_i(u(t)) > a_m > 0$ then for all $\tau_0 > 0$ there exist two constants $0 < \widetilde{\alpha} < \widetilde{\beta}$ (depending on τ_0, a_M, a_m) such that, for all $\tau \geq \tau_0$, the solution of the Riccati equation satisfies the inequality

$$\widetilde{\alpha}\, Id \leq \overline{S}(\tau) \leq \widetilde{\beta}\, Id.$$

This first relation is extended to all times $t > 0$ via a second lemma.

Lemma 39

We still consider the Riccati equation (3.13) with $S(0) = S_0$ being a symmetric definite positive matrix taken in a compact subset of the form $aId \leq S_0 \leq bId$, $0 < a < b$ and $\theta(0) = 1$. Then there exist two constants $0 < \alpha < \beta$ such that the solution of the equation satisfies $\alpha\, Id \leq \overline{S}(\tau) \leq \beta\, Id$ for all $\tau > 0$ (and therefore $\alpha\, Id \leq \widetilde{S}(t) \leq \beta\, Id$ for all $t \geq 0$).

Proof.

We denote by $|.|$, the Frobenius norm of matrices: $|A| = \sqrt{Trace(A'A)}$. Recall that for two symmetric semi-positive matrices A, B such that $0 \leq A \leq B$ we have[11]: $0 \leq |A| \leq| B|$.

Choose $\tau_0 > 0$ and apply Lemma 38 to obtain $\widetilde{\alpha}$ and $\widetilde{\beta}$ such that $\widetilde{\alpha}\, Id \leq \overline{S}(\tau) \leq \widetilde{\beta}\, Id$ for all $\tau \geq \tau_0$. In order to extend the inequality for $0 \leq \tau \leq \tau_0$, we start from:

$$\begin{aligned}\overline{S}(\tau) &= \overline{S_0} + \int_0^\tau \frac{d\overline{S}(v)}{d\tau} dv \\ &= \overline{S_0} + \int_0^\tau \left[-\left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right)' \overline{S} \right. \\ &\qquad \left. -\overline{S}\left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right) + C'R^{-1}C - \overline{S}Q\overline{S} \right] dv.\end{aligned}$$

As $\theta(0) = 1$, then $\overline{S}(0) = S(0) = S_0$, which together with $\overline{S}Q\overline{S} > 0$ (symmetric semi-positive) leads to

$$\begin{aligned}\overline{S}(\tau) \leq\ & S_0 + \int_0^\tau \left[-\left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right)' \overline{S} \right. \\ &\left. -\overline{S}\left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right) + C'R^{-1}C \right] dv,\end{aligned}$$

and

$$\left|\overline{S}(\tau)\right| \leq |S_0| + \int_0^\tau 2\left(A_M + B + \left| \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} \right| |N| \right) \left|\overline{S}\right| + \left| C'R^{-1}C \right| dv$$

with $A_M = \sup_{[0;\tau_0]} (|A(u(\tau))|)$ and $\left|\widetilde{b}^*(\overline{z},\overline{u})\right| \leq B$. Then

$$\left|\overline{S}\right| \leq |S_0| + \left| C'R^{-1}C \right| \tau_0 + \int_0^\tau 2s \left|\overline{S}\right| dv,$$

with $s = a_M |N| + A_M + B$. Applying Gronwall's lemma gives us for all $0 \leq \tau \leq \tau_0$,

$$\begin{aligned}\left|\overline{S}\right| &\leq \left(|S_0| + \left| C'R^{-1}C \right| \tau_0 \right) e^{2s\tau} \\ &\leq \left(|S_0| + \left| C'R^{-1}C \right| \tau_0 \right) e^{2s\tau_0} \\ &\leq \left(b\sqrt{n} + \left| C'R^{-1}C \right| \tau_0 \right) e^{2s\tau_0} \\ &= \beta_1.\end{aligned} \tag{3.14}$$

In the same manner we denote $\overline{P} = \overline{S}^{-1}$, and use the equation

$$\begin{aligned}\frac{d\overline{P}}{dt} = \overline{P}\left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right)' + \left(A(u) + \frac{\widetilde{b}^*(\overline{z},\overline{u})}{\overline{\theta}} - \frac{\mathcal{F}(\overline{\theta},\overline{\jmath})}{\overline{\theta}^2} N \right) \overline{P} \\ - \overline{P}C'R^{-1}C\overline{P} + Q\end{aligned}$$

[11] See Appendix B.1 for details of establishing this fact.

to obtain, for all $0 \leq \tau \leq \tau_0$, and with $\widetilde{s} > 0$:

$$\begin{array}{rcl} \left|\overline{P}\right| & \leq & \left(|P_0| + |Q|\,\tau_0\right) e^{2\widetilde{s}\tau_0} = \frac{1}{\alpha_1} \\ & \leq & \left(\frac{1}{a}\sqrt{n} + |Q|\,\tau_0\right) e^{2\widetilde{s}\tau_0} = \frac{1}{\alpha_1}. \end{array} \tag{3.15}$$

From the inequalities (3.14) and (3.15) we deduce that[12] for all $0 \leq \tau \leq \tau_0$

$$\alpha_1\, Id \leq \overline{S}\,(\tau) \leq \beta_1\, Id.$$

We now define

$$\alpha = \min\left(a,\alpha_1,\widetilde{\alpha}\right) \text{ and } \beta = \max\left(b,\beta_1,\widetilde{\beta}\right)$$

such that for all $\tau \geq 0$

$$\alpha\, Id \leq \overline{S}\,(\tau) \leq \beta\, Id.$$

This relation is therefore true also in the t time scale.

■

3.7 Technical Lemmas

Three lemmas are proposed in the following section. The two first are purely technical and are used in the very last section of the present chapter. They are from [38]. Their respective proofs are reproduced in Appendix B.2.

The third lemma concerns the adaptation function. It basically shows that the set of canditate adaptation functions for our adaptive high-gain observer is not empty. The proof is constructive: we display such a function.

Lemma 40 *([38])*

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

$$\frac{dx(t)}{dt} \leq -k_1 x + k_2 x\sqrt{x},$$

[12] Trivially, we have

$$\|S\| \leq \beta_1 \Rightarrow S \leq \beta_1 Id.$$

However

$$\alpha_1 \leq \|S\| \not\Rightarrow \alpha_1 Id \leq S.$$

This is the reason why we need the relation:

$$\|P\| \leq \frac{1}{\alpha_1} \Rightarrow \|P\| \leq \frac{1}{\alpha_1}$$

and then we use the following matrix property (see Appendix B.1):

$$(P \geq Q > 0) \Rightarrow \left(Q^{-1} \geq P^{-1} > 0\right).$$

in order to end up with

$$\alpha_1 Id \leq S.$$

for almost all $t > 0$, *for* k_1, $k_2 > 0$. *Then, if* $x(0) < \frac{k_1^2}{4k_2^2}$, *we have*

$$x(t) \leq 4\, x(0)\, e^{-k_1 t}.$$

Lemma 41 *([38])*

Consider $\widetilde{b}(\tilde{z}) - \widetilde{b}(\tilde{x}) - \widetilde{b}^*(\tilde{z})\,\tilde{\varepsilon}$ *as in the inequality (3.12) (omitting to write* u *in* $\tilde{b}$*) and suppose* $\theta \geq 1$. *Then* $\left\| \widetilde{b}(\tilde{z}) - \widetilde{b}(\tilde{x}) - \widetilde{b}^*(\tilde{z})\,\tilde{\varepsilon} \right\| \leq K\theta^{n-1} \|\tilde{\varepsilon}\|^2$, *for some* $K > 0$.

Lemma 42 ***(adaptation function)***

For any $\Delta T > 0$, *there exists a positive constant* $M(\Delta T)$ *such that:*

- *for any* $\theta_1 > 1$,*and*
- *any* $\gamma_1 > \gamma_0 > 0$,

there is a function $\mathcal{F}(\theta, \mathcal{I})$ *such that the equation*

$$\dot{\theta} = \mathcal{F}(\theta, \mathcal{I}(t)), \tag{3.16}$$

for any initial value $1 \leq \theta(0) < 2\theta_1$, *and any measurable positive function* $\mathcal{I}(t)$, *has the properties:*

1. *that there is a unique solution* $\theta(t)$ *defined for all* $t \geq 0$, *and this solution satisfies* $1 \leq \theta(t) < 2\theta_1$,
2. $\left| \frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2} \right| \leq M$,
3. *if* $\mathcal{I}(t) \geq \gamma_1$ *for* $t \in [\tau, \tau + \Delta T]$ *then* $\theta(\tau + \Delta T) \geq \theta_1$,
4. *while* $\mathcal{I}(t) \leq \gamma_0$, $\theta(t)$ *decreases to* 1.

Remark 43

The main property is that if $\mathcal{I}(t) \geq \gamma_1$, $\theta(t)$ *can reach any arbitrarily large* θ_1 *in an arbitrary small time* ΔT, *and that this property can be achieved by a function satisfying* $\mathcal{F}(\theta, \mathcal{I}) \leq M\theta^2$ *with* M *independent from* θ_1 *(but dependant from* ΔT*).*

Proof.

Let $\mathcal{F}_0(\theta)$ be defined as follows:

$$\mathcal{F}_0(\theta) = \begin{cases} \frac{1}{\Delta T}\theta^2 & \text{if } \theta \leq \theta_1 \\ \frac{1}{\Delta T}(\theta - 2\theta_1)^2 & \text{if } \theta > \theta_1 \end{cases}$$

(the choice $2\theta_1$ is more or less arbitrary) and let us consider the system

$$\begin{cases} \dot{\theta} &= \mathcal{F}_0(\theta) \\ \theta(0) &= 1 \end{cases}.$$

Simple computations give the solution:

$$\theta(t) = \begin{cases} \dfrac{\Delta T}{\Delta T - t} & \text{while} \quad \theta \leq \theta_1 \\ 2\theta_1 - \dfrac{\theta_1 \Delta T}{\theta_1 t + (2 - \theta_1)\Delta T} & \text{when} \quad \theta > \theta_1. \end{cases}$$

Therefore $\theta(t)$ reaches θ_1 at time $t < \Delta T$. This holds a fortiori[13] whatever the value of $\theta(0) \in [1, 2\theta_1[$. Let us remark also that $\mathcal{F}_0$ is Lipschitz. Now, let us define

$$\mathcal{F}(\theta, \mathcal{I}) = \mu(\mathcal{I})\,\mathcal{F}_0(\theta) + (1 - \mu(\mathcal{I}))\,\lambda(1-\theta)$$

for a $\lambda > 0$. The function μ is such that[14]:

$$\mu(\mathcal{I}) = \begin{cases} 1 & \text{if } \mathcal{I} \geq \gamma_1 \\ \in [0,1] & \text{if } \gamma_0 \leq \mathcal{I} \leq \gamma_1 \\ 0 & \text{if } \mathcal{I} \leq \gamma_0. \end{cases}$$

We claim that all properties are satisfied.

If $\mathcal{I} \geq \gamma_1$, $\mathcal{F}(\theta, \mathcal{I}) = \mathcal{F}_0(\theta)$ ensuring *Property 3*, (refer to the beginning of the proof). Conversely, if $\mathcal{I} \leq \gamma_0$, $\mathcal{F}(\theta, \mathcal{I}) = \lambda(1-\theta)$ then *Property 4* is fulfiled. Moreover, because $\mathcal{F}(\theta, \mathcal{I})$ is Lipschitz, *Property 1* is verified. Let us check *property 2*:

$$\left|\frac{\mathcal{F}(\theta, \mathcal{I})}{\theta^2}\right| \leq \left|\frac{\mathcal{F}_0(\theta)}{\theta^2}\right| + \left|\frac{\lambda(1-\theta)}{\theta^2}\right|. \tag{3.17}$$

The first term satisfies:

- $\theta \leq \theta_1$, $\left|\frac{\mathcal{F}_0(\theta)}{\theta^2}\right| = \frac{1}{\Delta T}$, and
- $\left|\frac{\mathcal{F}_0(\theta)}{\theta^2}\right| = \frac{1}{\Delta T}\left(\frac{\theta - 2\theta_1}{\theta}\right)^2 \leq \frac{1}{\Delta T}$ if $\theta \geq \theta_1$ (and $\theta < 2\theta_1$).

The second term satisfies:

$$\begin{aligned} \left|\frac{\lambda(1-\theta)}{\theta^2}\right| = \lambda\frac{\theta - 1}{\theta^2} &= \lambda\left(\frac{1}{4} - \frac{\frac{\theta^2}{4} - \theta + 1}{\theta^2}\right) \\ &= \lambda\left(\frac{1}{4} - \left(\frac{\frac{\theta}{2} - 1}{\theta}\right)^2\right) \leq \frac{\lambda}{4}. \end{aligned}$$

Property 2 is satisfied because of (3.17) with $M = \frac{1}{\Delta T} + \frac{\lambda}{4}$.

■

[13] When $1 < \theta_0 \leq \theta_1$ the solution to equation (5.11) is:

$$\begin{cases} \theta(t) &= \frac{\Delta T \theta_0}{\Delta T - \theta_0 t} & \text{when} \quad \theta(t) \leq \theta_1 \\ \theta(t) &= 2\theta_1 - \frac{\Delta T \theta_0 \theta_1}{\theta_0 \theta_1 t + (2\theta_0 - \theta_1)\Delta T} & \text{when} \quad \theta(t) > \theta_1. \end{cases}$$

And for $\theta_1 < \theta_0 < 2\theta_1$ the solution of (5.11) is:

$$\theta(t) = 2\theta_1 - \frac{\Delta T(2\theta_1 - \theta_0)}{\Delta T + t(2\theta_1 - \theta_0)}.$$

The conclusion remains the same: if $\theta_0 < \theta_1$, θ_1 is reached in a time smaller than$\Delta\, T$, and remains below $2\theta_1$ in all cases.

[14] Such a function is explicitely defined in Section 4.2.1 of Chapter 4.

3.8 Proof of the Theorem

First of all let us choose a time horizon d (in $\mathcal{I}_d(t)$) and a time T such that $0 < d < T < T^*$. Set$\Delta T = T - d$. Let λ be a strictly positive number and $M = \frac{1}{\Delta T} + \frac{\lambda}{4}$ as in Lemma 42. Let α and β be the bounds from Lemma 39.

From the preparation for the proof, inequality (3.12) can be written, using Lemma 39 (i.e. using $\tilde{S} \geq \alpha\, Id$), and omitting the control variable u

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \theta \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + 2\tilde{\varepsilon}'\tilde{S}\left(\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right). \tag{3.18}$$

From (3.18) we can deduce two inequalities: the first one, local, will be used when $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)$ is small, whatever the value of θ. The second one, global, will be used mainly when $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)$ is not in a neighborhood of 0 and θ is large.
Using

$$\left\|\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right\| \leq 2L_b \|\tilde{\varepsilon}\|,$$

together with $\alpha\, Id \leq \tilde{S} \leq \beta\, Id$ (Lemma 39), (3.18) becomes the "global inequality"

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq \left(-\alpha q_m \theta + 4\frac{\beta}{\alpha}L_b\right)\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t). \tag{3.19}$$

Because of Lemma 41, we obtain the"local inequality" as follows:

$$\left\|\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right\| \leq K\theta^{n-1}\|\tilde{\varepsilon}\|^2.$$

Since $1 \leq \theta \leq 2\theta_1$, inequality (3.18) implies

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + 2K(2\theta_1)^{n-1}\left\|\tilde{S}\right\|\|\tilde{\varepsilon}\|^3.$$

Since $\|\tilde{\varepsilon}\|^3 = \left(\|\tilde{\varepsilon}\|^2\right)^{\frac{3}{2}} \leq \left(\frac{1}{\alpha}\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)\right)^{\frac{3}{2}}$, the inequality becomes

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \leq -\alpha q_m \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + \frac{2K(2\theta_1)^{n-1}\beta}{\alpha^{\frac{3}{2}}}\left(\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)\right)^{\frac{3}{2}}. \tag{3.20}$$

Let us apply[15] Lemma 40 which states that if

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) \leq \frac{\alpha^5 q_m^2}{16\,K^2(2\theta_1)^{2n-2}\beta^2},$$

then, for any $t \geq \tau$,

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \leq 4\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau)\,\mathrm{e}^{-\alpha q_m(t-\tau)}.$$

[15]This lemma cannot be applied if we use Q_θ and R instead of Q_θ and R_θ in the definition of the observer as it is done in [38]. This is due to the presence of a $\frac{\mathcal{F}}{\theta}$ term that prevents parameters k_1 and k_2 to be positive for all times.

Consequently, provided there is a real γ such that

$$\gamma \leq \frac{1}{(2\theta_1)^{2n-2}} \min\left(\frac{\alpha\varepsilon^*}{4}, \frac{\alpha^5 q_m^2}{16\,K^2\beta^2}\right), \tag{3.21}$$

then $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(\tau) \leq \gamma$ implies, for any $t \geq \tau$,

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(t) \leq \frac{\alpha\varepsilon^*}{(2\theta_1)^{2n-2}} \mathrm{e}^{-\alpha q_m (t-\tau)}. \tag{3.22}$$

Note that excluding the change of variables, we have arrived at the end result.

From (3.19):

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(T) \leq \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(0)\,\mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha}L_b\right)T},$$

and if we suppose $\theta \geq \theta_1$ for $t \in [T, T^*]$, $T^* > T$, using (3.19) again:

$$\begin{aligned} \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(T^*) &\leq \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(0)\,\mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha}L_b\right)T}\mathrm{e}^{\left(-\alpha q_m\theta_1 + 4\frac{\beta}{\alpha}L_b\right)(T^*-T)} \\ &\leq M_0\mathrm{e}^{-\alpha q_m T}\mathrm{e}^{4\frac{\beta}{\alpha}L_b T^*}\mathrm{e}^{-\alpha q_m\theta_1(T^*-T)}, \end{aligned}$$

where

$$M_0 = \sup_{x,z\in X} \varepsilon' S\varepsilon\,(0)\,. \tag{3.23}$$

Now, we choose θ_1 and γ for

$$M_0\mathrm{e}^{-\alpha q_m T}\mathrm{e}^{4\frac{\beta}{\alpha}L_b T^*}\mathrm{e}^{-\alpha q_m\theta_1(T^*-T)} \leq \gamma \tag{3.24}$$

and (3.21) to be satisfied simultaneously, which is possible since $\mathrm{e}^{-\mathrm{cte}\times\theta_1} < \frac{\mathrm{cte}}{\theta_1^{2n-2}}$ for θ_1 large enough. Let us chose a function $\mathcal{F}$ as in Lemma 42 withΔ $T = T - d$ and $\gamma_1 = \frac{\lambda_d^0\gamma}{\beta}$.

We claim that there exists $\tau \leq T^*$ such that $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(\tau) \leq \gamma$.

Indeed, if $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(\tau) > \gamma$ for all $\tau \leq T^*$ because of Lemma 33:

$$\gamma < \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(\tau) \leq \beta\,\|\tilde{\varepsilon}\,(\tau)\|^2 \leq \beta\,\|\varepsilon\,(\tau)\|^2 \leq \frac{\beta}{\lambda_d^0}\mathfrak{I}_d\,(\tau + d)\,.$$

Therefore, $\mathfrak{I}_d\,(\tau + d) \geq \gamma_1$ for $\tau \in [0, T^*]$ and hence $\mathfrak{I}_d\,(\tau) \geq \gamma_1$ for $\tau \in [d, T^*]$. Thus, we have $\theta\,(t) \geq \theta_1$ for $t \in [T, T^*]$ which provides a contradiction (i.e. $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(T^*) \leq \gamma$) thanks to (3.8) and (3.24).

Finally, for $t \geq \tau$, using (3.22)

$$\begin{aligned} \|\varepsilon\,(t)\|^2 &\leq (2\theta_1)^{2n-2}\,\|\tilde{\varepsilon}\,(t)\|^2 \\ &\leq \frac{(2\theta_1)^{2n-2}}{\alpha}\tilde{\varepsilon}'\,\tilde{S}\tilde{\varepsilon}\,(t) \\ &\leq \varepsilon^* e^{-\alpha q_m (t-\tau)} \end{aligned} \tag{3.25}$$

which proves the theorem.

Remark 44 ***(alternative result)***

We propose a modification of the end of the proof that allows $\|\varepsilon(0)\|^2$ *to appear in the final inequality.*

Consider equation (3.21) and replace it with:

$$\gamma \leq \frac{1}{(2\theta_1)^{2n-2}} \min\left(\frac{\alpha\varepsilon^*}{4\beta}, \frac{\alpha^5 q_m^2}{16\,K^2\beta^2}\right). \tag{3.26}$$

Then equation (3.22) becomes:

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}\,(t) \leq \frac{\alpha\varepsilon^*}{\beta\,(2\theta_1)^{2n-2}} \mathrm{e}^{-\alpha q_m(t-\tau)}. \tag{3.27}$$

Now replace the definition of M_0 *given in equation (5.18) by* $M_0 = \max\left(\sup_{x,z\in X} \varepsilon' S\varepsilon\,(0), 1\right)$. *Finally consider the very last inequality (3.25). It can also be developed as follows (with* $M_0 \geq 1$*):*

$$\begin{aligned}
\|\varepsilon\,(t)\|^2 &\leq (2\theta_1)^{2n-2}\,\|\tilde{\varepsilon}\,(t)\|^2 \leq \frac{(2\theta_1)^{2n-2}}{\alpha}\tilde{\varepsilon}'\,\tilde{S}\tilde{\varepsilon}\,(t) \\
&\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha}\tilde{\varepsilon}'\,\tilde{S}\tilde{\varepsilon}\,(\tau)\,\mathrm{e}^{-\alpha q_m(t-\tau)} \\
&\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha}\left[\tilde{\varepsilon}'\,\tilde{S}\tilde{\varepsilon}\,(0)\,\mathrm{e}^{\left(-\alpha q_m+4\frac{\beta}{\alpha}L_b\right)T}\mathrm{e}^{\left(-\alpha q_m\theta_1+4\frac{\beta}{\alpha}L_b\right)(T^*-T)}\right]\mathrm{e}^{-\alpha q_m(t-\tau)} \\
&\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha}\,\beta.\,\|\tilde{\varepsilon}\,(0)\|^2\left[\mathrm{e}^{\left(-\alpha q_m+4\frac{\beta}{\alpha}L_b\right)T}\mathrm{e}^{\left(-\alpha q_m\theta_1+4\frac{\beta}{\alpha}L_b\right)(T^*-T)}\right]\mathrm{e}^{-\alpha q_m(t-\tau)} \\
&\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha}\,\beta.\,\|\tilde{\varepsilon}\,(0)\|^2\left[M_0\mathrm{e}^{\left(-\alpha q_m+4\frac{\beta}{\alpha}L_b\right)T}\mathrm{e}^{\left(-\alpha q_m\theta_1+4\frac{\beta}{\alpha}L_b\right)(T^*-T)}\right]\mathrm{e}^{-\alpha q_m(t-\tau)}
\end{aligned}$$

From (3.24) we know that the bracketed expression is smaller than γ. *Equations (3.26) and (3.27),* $\theta(0) = 1$ *lead to:*

$$\begin{aligned}
\|\varepsilon\,(t)\|^2 &\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha}\,\beta\,\|\tilde{\varepsilon}\,(0)\|^2\left[\frac{\alpha\varepsilon^*}{4\beta\,(2\theta_1)^{2n-2}}\right]\mathrm{e}^{-\alpha q_m(t-\tau)} \\
&\leq \|\varepsilon_0\|^2\,\varepsilon^*\mathrm{e}^{-\alpha q_m(t-\tau)}.
\end{aligned}$$

Since $\tau \leq T^*$, $e^{-\alpha q_m(\tau-T^*)} \geq 1$ *and we can obtain the inequality*

$$\|\epsilon(t)\|^2 \leq \|\varepsilon_0\|^2\,\varepsilon^*\mathrm{e}^{-\alpha q_m(t-T^*)}.$$

3.9 Conclusion

In this chapter, the adaptive high-gain extended Kalman filter was introduced for a multiple input, single output system. The exponential convergence of the algorithm has been proven. This proof has been decomposed in a series of significant lemmas:

1. the innovation at time t places an upper bound on the error at time $t - d$,
2. the Riccati matrix S is bounded from above and below, for all times t, independently from θ,
3. the set of candidate adaptation functions is non empty.

In the next chapter, the description of the observer is completed with an adaptation function and the analysis of an example. This study is done both in a simulation and in a real process. We also propose guidelines to the tuning of the several parameters of the observer.

Chapter 4

Illustrative Example and Hard real-time Implementation

Contents

In this chapter, we focus now on the implementation of the adaptive high-gain extended Kalman filter that was introduced in Chapter 3. We provide a full definition of the observer, in which an adaptation function is explicitly given. A methodology is advanced for tuning the several parameters. Our goal is to demonstrate that the adaptive high-gain extended Kalman filter can be used in practice even in the case of a relatively fast process, e.g. 100 Hz.

The process we consider is a series-connected DC motor, modeled via a nonlinear SISO[1] system when current and voltage are the only observables. This process has been used in previous studies, allowing us to compare our results here with those from the earlier works (see [87, 93]). Moreover, the machine itself is readily available and experiments can be considered quite realistic. Although the process is quite simple, the implementation of the observer in a real-time environment raises interesting questions that a simulation does not.

The modeling itself of the process and the observability study are investigated in Section 4.1. The implementation of the process in a simulation is the subject of Section 4.2. The methodology for the tuning the parameters is also developed in this section. Finally, a set of real experiments performed using an actual machine using a *hard real-time operating system* is detailed in Section 4.3.

4.1 Modeling of the Series-connected DC Machine and Observability Normal Form

Basically, an electric motor converts electrical energy into mechanical energy. In a DC motor, the stator (also denoted field) is composed of an electromagnet, or a permanent magnet, that immerses the rotor in a magnetic field. The rotor (also denoted armature) is made of an electromagnet that once supplied with current creates a second magnetic field. The stator is kept fixed while the rotor is allowed to move — i.e., rotate. The attraction/repelling behavior of magnets generates the rotative motion.

In order to make the rotative motion permanent, one of the two magnetic fields has to be switched at appropriate moments. The magnetic field created by the stator remains fixed. The rotor windings are connected to a commutator causing the direction of the current flowing through the armature coils to switch during the rotation. This reverses the polarity of the armature magnetic field. Successive commutations then maintain the rotating motion of the machine.

A DC motor whose field circuit and armature circuit are connected in series, and therefore fed by the same power supply, is referred to as a *series-connected DC motor* [77].

4.1.1 Mathematical Model

The model of the series-connected DC motor is obtained from the equivalent circuit representation shown in Figure 4.1. We denote I_f as the current flowing through the field part of the circuit (between points A and C), and I_a as the current flowing through the armature circuit (between points C and B). When the shaft of the motor is turned by an external force, the motor acts as a generator and produces an electromotive force. In the case of the DC

[1] Single input single output.

motor, this force will act against the current applied to the circuit and is then denoted *back or counter electromotive force* (BEMF or CEMF). The electrical balance leads to

$$L_f \dot{I}_f + R_f I_f = V_{AC}$$

for the field circuit, and to

$$L_a \dot{I}_a + R_a I_a = V_{CB} - E$$

for the armature circuit. The notations are:

- L_f and R_f for the inductance and the resistance of the field circuit,
- L_a and R_a for the inductance and the resistance of the armature circuit,
- E for the Back EMF.

Kirchoff's laws give us the relations:

$$\begin{cases} I = I_a = I_f \\ V = V_{AC} + V_{CB}. \end{cases}$$

The total electrical balance is

$$L\dot{I} + RI = V - E,$$

where $L = L_f + L_a$ and $R = R_f + R_a$. The field flux is denoted byΦ . We have $\Phi = f(I_f) = f(I)$, and $E = K_m \Phi \omega_r$ where K_m is a constant and ω_r is the rotational speed of the shaft.

The second equation of the model is given by the mechanical balance of the shaft of the motor using Newton's second law of motion. We consider that the only forces applied to the shaft are the electromechanical torque T_e, the viscous friction torque and the load torque T_l leading to

$$J\dot{\omega_r} = T_e - B\omega_r - T_l$$

where J denotes the rotor inertia, and B the viscous friction coefficient. The electromechanical torque is given by $T_e = K_e \Phi I$ with K_e denoting a constant parameter. We consider that the motor is operated **below saturation** [93]. In this case, the field flux can be expressed by the linear expression $\Phi = L_{af} I$, where L_{af} denotes the mutual inductance between the field and the rotating armature coils. To conclude with the modeling of the DC, motor we impose the ideal hypothesis of 100% efficiency of conservation of energy, which is expressed as $K = K_m = K_e$. For simplicity in the notation, we write L_{af} instead of KL_{af}. The voltage

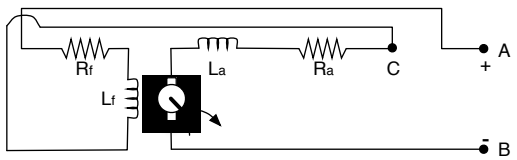

Figure 4.1: Series-connected DC Motor: equivalent circuit representation.

is the input of the system, $u(t)$, and the current, $I(t)$, is the measured output. The resulting following SISO model for the series-connected DC motor is:

$$\left\{ \begin{array}{ccc} \left(\begin{array}{c} L\dot{I} \\ J\dot{\omega}_r \end{array} \right) & = & \left(\begin{array}{c} u - RI - L_{af}\omega_r I \\ L_{af}I^2 - B\omega_r - T_l \end{array} \right) \\ y & = & I \end{array} \right. \tag{4.1}$$

This model is used to simulate the DC motor by means of a Matlab/Simulink S-function.

4.1.2 Observability Cannonical Form

Before implementing the observer to reconstruct the state vector of this system, we test the system's observability property. We use the *differentiation approach*, i.e. we check differential observability (which implies observability):

- if $I(t)$ is known with time, then $\dot{I} = (u - R.I - L_{af}\omega_r I)/L$ is known and as long as $u(t)$, R, L_{af} and L are known then ω_r can be computed,
- because $\omega_r(t)$ is known, $\dot{\omega_r} = (L_{af}I^2 - B\omega_r - T_l)/J$ can also be computed. From the knowledge of $I(t)$, L_{af}, B,and J, then T_l can be estimated.

We conclude that a third variable can be added to the state vector in order to reconstruct the load torque applied to the shaft of the motor along with the state of the system. We assume that the load torque is constant over time. Sudden changes of the load torque then, are interpreted as non modeled perturbations. The estimation of the load torque is made possible including the constraint $\dot{T}_l = 0$ in equation (4.1). We now need to find the coordinate transformation that puts this systems into the observability canonical form.

From the equation $y = I$, we choose $x_1 = I$ and then

$$\dot{x_1} = \frac{1}{L}(u(t) - RI - L_{af}I\omega_r),$$

which by setting $x_2 = I\omega_r$ becomes

$$\dot{x_1} = -\frac{L_{af}}{L}x_2 + \frac{1}{L}(u(t) - Rx_1) = \alpha_2(u)x_2 + b_1(x_1, u). \tag{4.2}$$

We now compute the time derivative of x_2:

$$\dot{x_2} = \dot{I}\omega_r + I\dot{\omega_r} = -\frac{1}{J}T_l I - \frac{B}{J}I\omega_r + \frac{L_{af}}{J}I^3 - \frac{L_{af}}{L}\omega_r^2 I + \frac{u(t)}{L}\omega_r - \frac{R}{L}\omega_r I$$

provided that $I > 0$ (i.e. $x_1 > 0$). This constraint represents a reasonable assumption since when I, the current of the circuit, equals zero there is no power being supplied to the engine and therefore there is nothing to observe. We have $\omega_r = \frac{x_2}{x_1}$, and $x_3 = T_l I$. The above equation then becomes

$$\begin{array}{rcl} \dot{x_2} & = & -\frac{1}{J}x_3 - \frac{B}{J}x_2 + \frac{L_{af}}{J}x_1^3 - \frac{L_{af}}{L}\frac{x_2^2}{x_1} + \frac{u(t)}{L}\frac{x_2}{x_1} - \frac{R}{L}x_2 \\ & = & \alpha_3(u)x_3 + b_2(x_1, x_2, u) \end{array} \tag{4.3}$$

again provided that $I > 0$ (i.e. $x_1 > 0$). This leads us to the expression $T_l = \frac{x_3}{x_1}$. Recall that $\dot{T}_l = 0$, then

$$\dot{x_3} = -\frac{L_{af}}{L}\frac{x_2 x_3}{x_1} + \frac{u(t)}{L}\frac{x_3}{x_1} - \frac{R}{L}x_3 = b_3(x_1, x_2, x_3, u). \tag{4.4}$$

Thus the application:

$$\begin{array}{ccc} \mathbb{R}^{*+} \times \mathbb{R} \times \mathbb{R} & \rightarrow & \mathbb{R}^{*+} \times \mathbb{R} \times \mathbb{R} \\ (I, \omega_r, T_l) & \hookrightarrow & (I, I\omega_r, IT_l) \end{array}$$

is a change of coordinates that puts the system (4.1), into the observer canonical form[2] defined by equations (4.2), (4.3), (4.4).

The inverse application is:

$$(x_1, x_2, x_3) \hookrightarrow \left(x_1, \frac{x_2}{x_1}, \frac{x_3}{x_1}\right).$$

Computations of the coefficients of the matrix b^* that appears in the Riccati equation of the observer — Cf. next section — are left to the reader.

4.2 Simulation

4.2.1 Full Observer Definition

We now recall the equations of the adaptive high-gain extended Kalman filter. As we want to minimize the computational time required to invert the matrix S, let us define $P = S^{-1}$. The identity $\frac{dP}{dt} = \frac{dS^{-1}}{dt} = S^{-1}\frac{dS}{dt}S^{-1}$ allows us to rewrite the observer as:

$$\left\{ \begin{array}{rcl} \frac{dz}{dt} & = & A(u)z + b(z,u) + PC'R_\theta^{-1}(Cz - y(t)) \\ \frac{dP}{dt} & = & P(A(u) + b^*(z,u))' + (A(u) + b^*(z,u))P - PC'R_\theta^{-1}CP + Q_\theta \\ \frac{d\theta}{dt} & = & \mu(\mathcal{I}_d)\mathcal{F}_0(\theta) + (1 - \mu(\mathcal{I}_d))\lambda(1-\theta) \end{array} \right. \tag{4.5}$$

where

- $R_\theta = \theta^{-1}R$,
- $Q_\theta = \theta\Delta^{-1}Q\Delta^{-1}$,
- $\Delta_\theta = diag\left(\left\{1, \theta, \theta^2, \ldots, \theta^{n-1}\right\}\right)$,
- $\mathcal{F}_0(\theta) = \left\{ \begin{array}{ll} \frac{1}{\Delta T}\theta^2 & if \theta \leq \theta_1 \\ \frac{1}{\Delta T}(\theta - 2\theta_1)^2 & if \quad \theta > \theta_1 \end{array} \right.$,
- $\mu(\mathcal{I}) = \left[1 + e^{-\beta(\mathcal{I}-m)}\right]^{-1}$ is a β and m parameterized sigmoid function (Cf. Figure 4.10),

[2] One could ask about the compact subset required by Theorem 36. In the present situation, a compact subset would be a collection of three closed and bounded intervals. The problem arises from the exclusion of 0 as a possible value for x_1. This is solved by picking any small $\epsilon > 0$ and considering that for $I = x_1 < \epsilon$ the motor is running too slowly to be of any practical use. Those trajectories are now in a compact subset of the state space.

– the innovation, $\mathcal{I}_d$, is defined by the formula:

$$\mathcal{I}_d(t) = \int_{t-d}^{t} \|y(s) - \hat{y}_{t-d}(s)\|^2 \, ds \tag{4.6}$$

where

– $y(s)$ is the output of the DC machine (the current),

– $\hat{y}_{t-d}(s) = x_1\,(t-d, z(t-d), s)$ with $x\,(t-d, z(t-d), s)$ is the solution of the normal form equations (4.2), (4.3), (4.4) over the time window $[t-d, t]$, with the initial condition that $x(t-d) = z(t-d)$, is the estimated state at time $t-d$.

4.2.2 Implementation Considerations

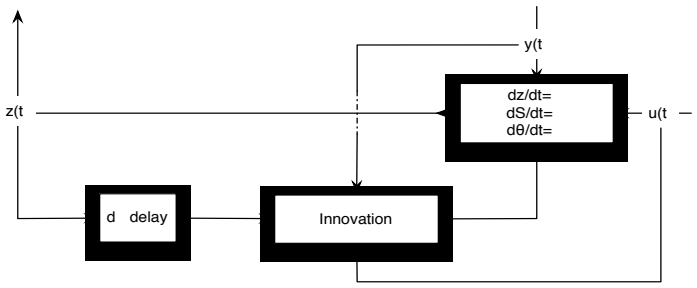

Figure 4.2: Observer Structure.

The simulation of the DC motor is straightforward. Thus, here we only comment on the implementation of the observer. The computation of a solution for the algorithm above is decomposed in two main parts (see Figure 4.2):

1. the computation of the innovation, $\mathcal{I}_d(t)$,
2. the update of the estimated state, the elements of Riccati matrix and the variable θ.

The latter part is common to all continuous time Kalman style observers. It presents no particular difficulties[3], and is illustrated in Figure 4.3.

In order to compute the innovation (refer to Figure 4.4) we need to:

[3]From theory, we know that the Riccati matrix is symmetric. Therefore we can solve the equation for only either the upper triangular part or the lower triangular part of the matrix.The introduction of this artifice has two advantages:

– it saves computational time (the solution of $\frac{n(n+1)}{2}$ equations are needed instead of n^2),

– because of tiny machine inaccuracies, the situation may arise when the coefficients, which are supposed to be equal, are determined to be unequal. This situation violates the *symmetric* requirement of the matrix and algorithm breaks down.

To solve the matrix equations, the implementation of the observer requires the use of a small utility that transforms $(n \times n)$ square matrices into vectors, and vice-versa.
Another device, which can be used when considering discrete-time systems, is to work directly with the square root of the Riccati matrix. This method is detailed in [43], Chapter 7.2.

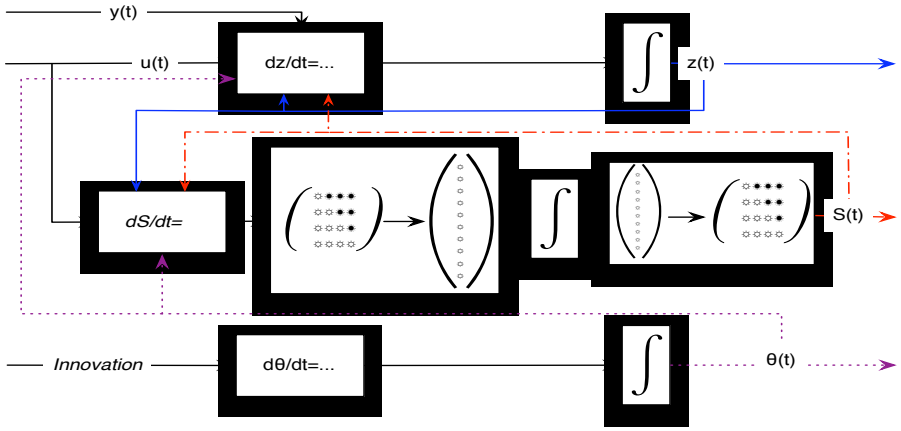

Figure 4.3: Computation of the main equations.

- store in memory the output signal over a time window of length d,
- store the observer's output trajectory in memory over a time window of length d in order to compute the prediction,
- compute a prediction of the trajectory over the time window $[t-d, t]$ (which implies storing of the input signal),
- and compute the integral (e.g. by a trapezoidal method).

The simulation is done using the Matlab/Simulink environment and particularly *level-1 S-functions*. Following our decomposition of the observer algorithm into two processes, we used two such *S-functions*. In the present example the sampling time of the measurements is taken sufficiently small, such that we may consider the estimation as a continuous process. However, we compute the innovation at discrete time time intervals (i.e. using a discrete S-function) because:

- the integral is computed by means of a fixed step trapezoidal method,
- we must store in memory the input and output trajectories over a time interval [0;d] where d is the delay of equation (4.6). A fixed step process simplifies the storage.

This sampling period constitutes an extra parameter that must be tuned.

4.2.3 Simulation Parameters and Observer Tuning

The values of the parameters of the DC machine model are displayed in Figure 4.5. The output signal is corrupted by an additive noise generated by an Orstein-Ulhenbeck process (refer to Appendix C.5). Perturbations may come from two different sources:

1. bad initialization of the observer or
2. a sudden change in the load torque (e.g. braking).

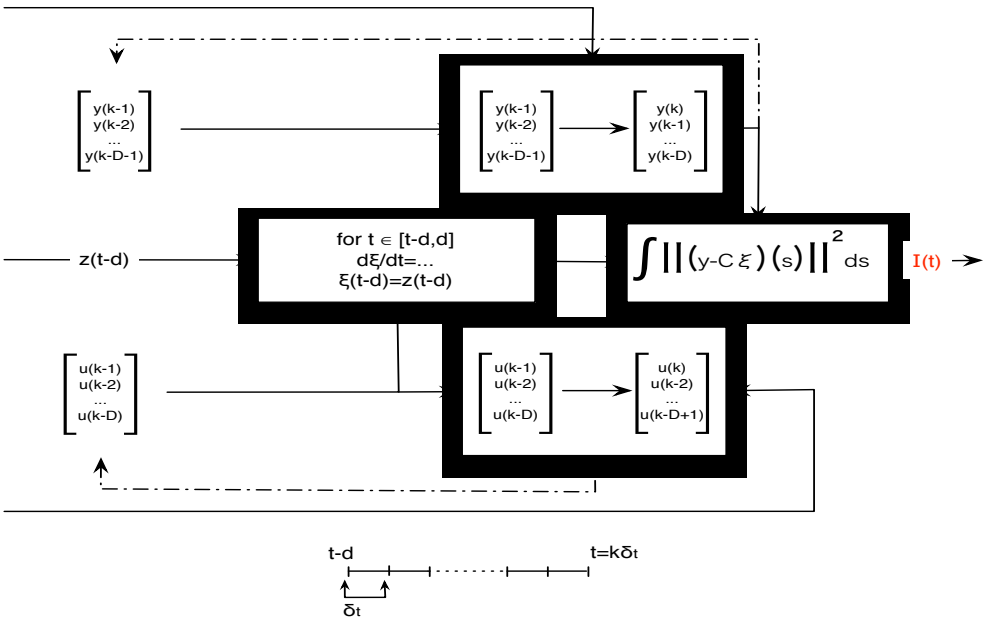

Figure 4.4: Computation of the innovation (at time $t = k\delta_t$).

Parameter	Value	Unit
L	1.22	H
Res	5.4183	Ω
L_{af}	0.0683	$N.m.Wb^{-1}A^{-1}$
J	0.0044	$kg.m^{-2}$
B	0.0026	$N.m.s^{-1}.rad^{-1}$

Figure 4.5: Simulation Parameters.

In practice, when we tune an observer, we do not compute the bounds as they appear in the proof of the theorem. Indeed, they are defined from the uniform Lipschitz properties of the vector field $b(z, u)$ and the α, β constants that bound the solution of the Riccati equation. The importance of the theorem, which was proven in the previous section, is to confirm the existence of configurations such that the observer converges. That is to say, that the algorithm is consistent and actually works. The theorem also gives us some qualitative insight into how the tuning of the variables should be handled.

We tune the parameters to achieve best performance. This is done by simulating and/or experimenting on the process. Although the methodology we describe below may appear complicated, it is in fact a succession of well-defined steps. Similar considerations concerning the tuning of a high-gain observer may also be found in the article [38], part 5.2.2.

An adaptive observer can achieve optimal performances both with and without large perturbations. This property greatly simplifies the tuning. We use different parameters to manage state perturbation rejection and noise filtering. Therefore each step focuses on a

single objective contrary to the multi-objective tuning required for non adaptive observers. The parameters are split into two categories (see Figure 4.6) :

- the ones defining the performance of the system with respect to noise and perturbations:
  - matrices Q and R are meant to provide decent noise filtering[4],
  - θ_1 characterizes the performance of the high-gain mode,
- the parameters related to the adaptation procedure:
  - computation of innovation: d, the delay, and δ the time discretization used to calculate $\mathcal{I}_d$,
  - the rising time of θ: ΔT (ΔT appears in the function $\mathcal{F}_0(\theta)$ of the observer (4.5)),
  - the sigmoid function parameters: β, $m = m_1 + m_2$,
  - the speed of the decay of $\theta(t)$ when innovation is small: λ.

Let us propose a methodology for the tuning of our set of parameters.

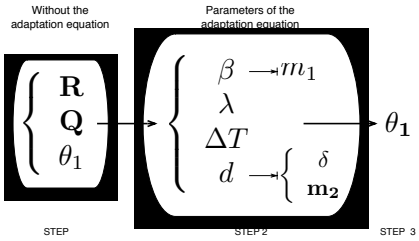

Figure 4.6: Bold: crucial parameters.

1. Non adaptive parameter tuning (Q, R and θ_1).

   At this stage simulations/experiments are done using the observer with $\mathcal{F}(\theta, \mathcal{I}_d) = 0$ (i.e. it is an extended Kalman filter, high-gain if $\theta(0) > 1$).

   First, we tune the classical EKF by making some simulations with noise, or by choosing some experimental data sets without large perturbations. Our goal here is to achieve acceptable smoothing of the noise. The input signal of the DC machine is set to $u(t) = 120 + 12\sin(t)$. This condition makes the system state oscillate and eases the burden of the graphical analyses[5]. The observer's initial state is taken to be equal to the system's initial state whenever the initial state is known. For all these simulations, the load torque is kept constant equal to 0.55. We refer to this situation as the *first scenario*.

   The matrices Q and R are required to be symmetric definite positive. Usually, in the continuous case, they are considered diagonal with positive coefficients:

[4] In the stochastic setting, Q and R are the covariance matrices of the state (resp. output) noise. In our deterministic point of view, they constitute additional tuning parameters.

[5] This is not a requirement at all. It's only here to provide a non stationary signal as the output. This step can also be performed with a stationary system.

- R is set in order to reflect the measurement noise covariance,
- the diagonal coefficients of Q are made larger for the state variables, which are unknown or for which the model is less accurate (in our case, this would be the load torque).

We decided to set $R = 1$ and attempted several different configurations for Q. Figure 4.7 shows a plot of our estimate of the second state variable (red line) as compared to the real values (black line) for several simulations. Comparing all the results in the figure, we consider that the tuning in the third panel (from the top) provided the observer with optimal noise smoothing properties. The values of Q used in the subsequent experiments (fourth and fifth panels from the top) do not improve the performance, while also making the observer really slow.

Contrary to what is normally done (see for example the application part of [38]), we choose values for the Q matrix, which are much larger for the measured variable than those that are unknown. By choosing the values in such a way, the observer is retarded with respect to the observable, which mean that those parameters would converge quite slowly, as compared to a bad initialization and/or sudden changes in the load torque. The high-gain mode is meant to cope with those situations.

In order to determine an initial value for θ_1, we simulate large disturbances in a *second scenario*. The input variable is kept constant[6] ($V = 120$), and the load torque is increased from 0.55 to 2.55 at the time step=30. The initial state is still considered as being equal to the actual system state. Q and R are set to be the values chosen previously. The high-gain parameter, $\theta(0) = \theta_0$, is then chosen in order to achieve the best observer time response during this disturbance. The performance with respect to noise is ignored. The data corresponding to the estimate of the load torque (third state variable) are plotted in Figure 4.9. We expect the noise to amplify the overshoot problem. Thus, we therefore try to select θ_0 such that offsets are avoided as much as possible.

In the definition of the function $\mathcal{F}_0$ of the observer (4.5), θ_1 has to be set such that $2\theta_1 = \theta_0$ where θ_0 denotes the value we just found.

2. Sigmoid function, innovation and adaptive procedure.

   We now consider the fully implemented observer with $\dot{\theta} = \mathcal{F}(\theta, \mathcal{I}_d)$ *and* $\theta(0) = 1$.

   Several parameters can be set in this case regardless of the application:

   - β and m: recall that according to Lemma 42 of Chapter 3 the function $\mu(\mathcal{I}_d)$ should possess the following features:

$$\mu(\mathcal{I}_d) = \begin{cases} 1 & if\gamma_1 \leq \mathcal{I}_d \\ \in [0,1] & if\gamma_1 \leq \mathcal{I}_d < \gamma_0 \\ 0 & if\ \mathcal{I}_d < \gamma_0 \end{cases}$$

[6] This not a requirement at al. The input variable may still be considered to be varying when changes in the load torque are made.

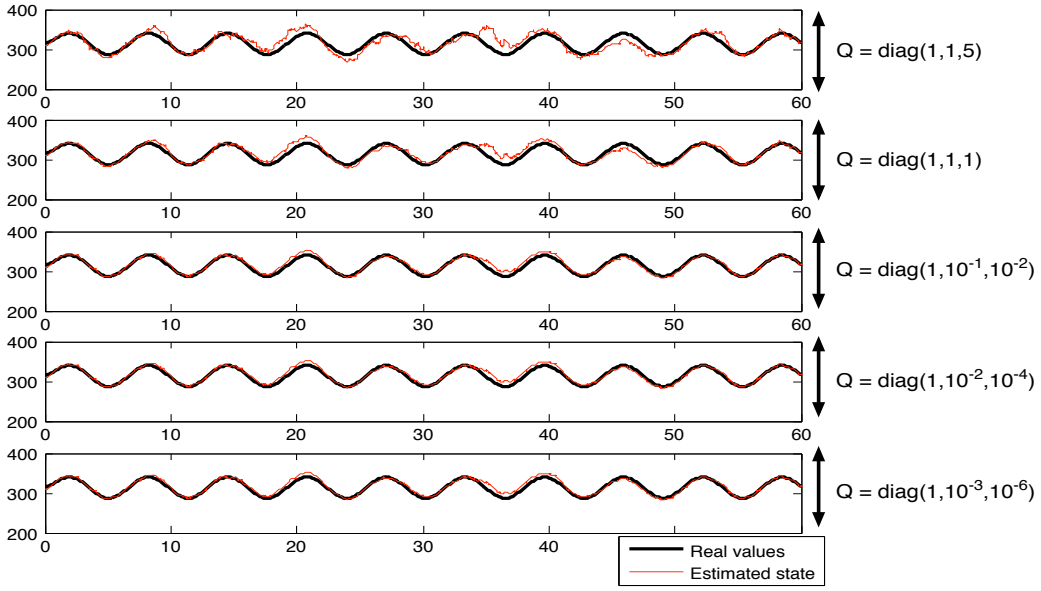

Figure 4.7: Tuning of the Q and R matrices.

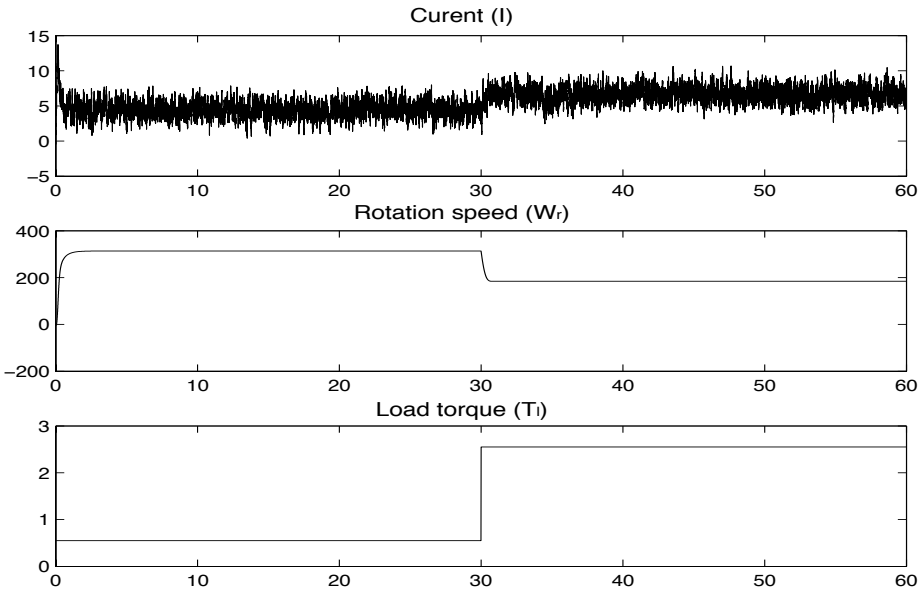

Figure 4.8: Sample of a simulation (scenario 1).

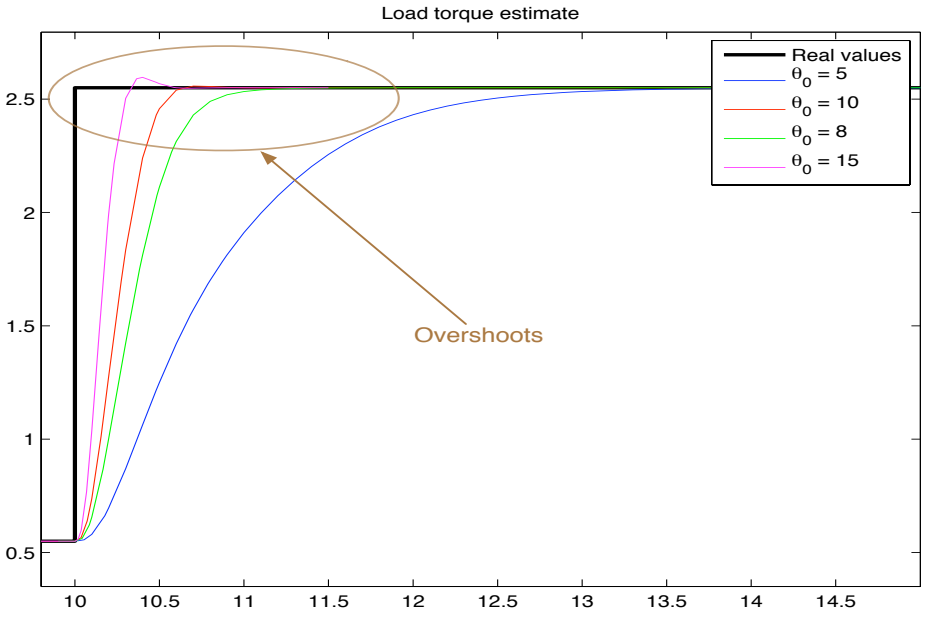

Figure 4.9: Choice of a value for $2\theta_1$.

We chose a sigmoid function whose equation appears in the definition of the observer (4.5). A graphical representation is displayed in Figure 4.10. The parameters β and m play the same role as the bounding parameters, γ_0 and γ_1, in the properties of the innovation shown above. The first parameter, β, controls the duration of the transition part of the sigmoid. The higher β is, the smaller is the transition. In practice, the best results are obtained for a small transition time (i.e. a large value of β)[7].

When $m = 0$, $\mu(0) = 0.5$. The role of the parameter m is to pull the sigmoid to the right. m is actually divided into two components, i.e., $m = m_1 + m_2$. We want m_1 is to be such that $\mu(\mathcal{I}_d) \approx 0$ when $\mathcal{I}_d$ is around 0. Because of the presence of noise in the measured signal, we need to add m_2 to this value. The procedure is explained below. The choice of m_1 can be made either via a graphical method or by solving a nonlinear equation[8].

– λ: Contrary to the observer of [38] "λ small enough" is no longer required, since θ increases when estimation error becomes too large. However, the situation may arise when the innovation oscillates up and down after some large disturbance.

[7]Note that the need for a small transition time is consistent with the requirements expressed in Theorem 36 of Chapter 3, i.e., when neither θ is high-gain nor the estimation error is sufficiently small, the state of the system is unclear. We want to reach one of those two domains.

[8]Let us choose $0 < \varepsilon_1 < 1$, small and l, a transition length. We keep $m = 0$ and solve the equation $\mu(l/2) - \mu(-l/2) = (1 - \varepsilon_1) - \varepsilon_1$ in order to find the corresponding β.
Now choose $\varepsilon_2 > 0$, sufficiently small, is used as the *zero value* (the function $\mu(x) = 0$ has no solution). Set β to the value found previously and solve for m using the equation $\mu(0) = \varepsilon_2$.

With a high value of λ, θ oscillates as well. In order to give some resilience to θ in this case, λ should be not be set too high. A value between 1 and 10 seems to be sufficient.

– ΔT: (in the adaptation function of the observer (4.5)). The smaller$\Delta\ T$, the shorter the rising time of θ. We take$\Delta\ T = 0.1$. This is sufficiently small that the equation $\mathcal{F}_0(\theta)$ remains compatible with the ODE solver[9].

As explained at the end of the previous subsection (i.e. Subsection 4.2.2), innovation is considered as a discrete function of time. Therefore we need to set the sampling time of this process. We denote it δ. It depends on the measurement hardware and is not a critical parameter. Nevertheless, it seems intuitive that $\frac{d}{\delta}$ must be sufficiently high to at least reflect the rank of the observability of the system, i.e. $\frac{d}{\delta} \geq n-1$. Indeed, innovation is used as a direct measurement of distinguishability. A theoretical justification of this remark is provided in Section 5.2 of Chapter 5 where an adaptive continuous-discrete version of our observer is provided.

Parameters d and m_2 are closely related to the application, but in a very clear manner.

When d is too small, innovation is not sufficiently large to distinguish between an increase in the estimation error and the influence of noise. On the other hand, a value for the innovation that is too high increases the computation time as the prediction is made on a larger time interval. The value of d has to be chosen using our knowledge of the time constant of the system (simulations or data samples), i.e., some fraction (e.g. $\frac{1}{3}$ to $\frac{1}{5}$) of the smallest time constant appears to be a reasonable choice.

Remark 45

When the system encounters a high perturbation at time t, one would expect a delay of length d in the adaptation of θ. However, the inequality of Lemma 33 is valid for any delay $0 < d_1 < d$ with a new constant $\lambda^0_{d_1}$ (smaller if $d_1 < d$) . Therefore we have:

$$\begin{aligned} \mathcal{I}_d(t+d_1) &= \mathcal{I}_{d-d_1}(t) + \mathcal{I}_{d_1}(t+d_1) \\ \mathcal{I}_d(t+d_1) &\geq \mathcal{I}_{d-d_1}(t) + \tfrac{1}{\lambda^0_{d_1}}\|x(t)-z(t)\|^2. \end{aligned}$$

If there have been no perturbations before the time t, then $\mathcal{I}_{d-d_1}(t)$ is close to zero, but $\|x(t)-z(t)\|^2$ is not. Consequently, $\mathcal{I}_d(t+d_1)$ is greater than zero. Hence, provided that the perturbation is sufficiently large, adaptation is triggered.

As a consequence, the parameter d shouldn't be shortened for the purpose of only making the adaptation faster.

Parameter m_2 is one of the most important parameters here. Its role is to avoid θ increasing when the innovation does not vanish due to the influence of noise. Indeed, if we suppose that the observer estimates the state of the system perfectly, then the output trajectory predicted during the computation of the innovation is equal to the

[9]Notice that there is a test in the definition of $\mathcal{F}_0$.

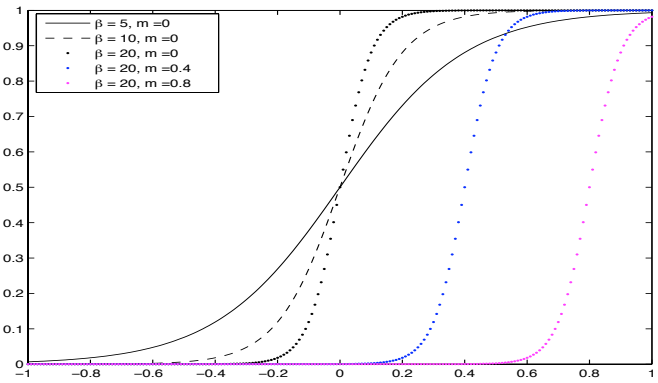

Figure 4.10: Effect of the parameters β and m on the shape of the sigmoid.

output signal without noise. In the case where the output signal is corrupted by noise $v(t)$, we have

$$y_{mes}(t) = y(t-d, x(t-d), \tau) + v(\tau).$$

Therefore with $x(t-d) = z(t-d)$:

$$\begin{aligned} \mathcal{I}_d(t) &= \int_{t-d}^{t} \|y(t-d, x(t-d), \tau) + v(\tau) - y(t-d, z(t-d), \tau)\|^2 d\tau \\ &= \int_{t-d}^{t} \|v(\tau)\|^2 d\tau \neq 0. \end{aligned} \tag{4.7}$$

We use σ to denote the standard deviation of $v(t)$. We estimate m_2 is the three-sigma, which seams reasonable and empirically sound. Then from equation (4.7), we obtain the relation $\mathcal{I}_d(t) \leq 9\sigma^2 d$. Therefore, $m_2 \approx 9\sigma^2 d$ appears to be a reasonable choice. However, practice demonstrates that m_2 computed in this way is over estimated. Although less common in engineering practice, we advise using a *one-sigma rule*[10]: $m_2 \approx \sigma^2 d$.

3. Final tuning

   All the parameters being set as before, we run a series of simulations with the output signal corrupted by noise. The parameter load torque is changed suddenly from 0.55 to 2.55 (scenario 2). We modify θ_1 in order to improve (shorten) convergence time of the observer when the system faces perturbations. Overshoots are kept as low as possible.

Remark 46

This methodology can also be applied for a hardware implementation. In the case where a complete simulator for the process is absent, the observer can be tuned in an open loop:

- *when the plant is operating, more or less, in steady state, in order to tune the parameters related to noise filtering,*
- *when a perturbation occurs in order to set parameters related to adaptation.*

[10] In the present example: $\sigma = 2$ and $d = 1$, thus $m_2 = 4$.

Parameter	Value	Role
Q	$diag(1, 10^{-1}, 10^{-2})$	Filtering
R	1	Filtering
θ_1	4	High-gain
β	$1664\frac{\pi}{e}$	Adaptation*
m_1	0.005	Adaptation*
m_2	4	Adaptation
λ	5	Adaptation*
ΔT	0.1	Adaptation*
d	1	Innovation
δ	0.1	Innovation*

Table 4.1: Final choice of parameters (*: Application-free parameters).

4.2.4 Simulation Results

The performance of the observer is accounted for via two scenarios:

- *scenario 2:* a single change in the load torque is performed at time 30,
- *scenario 3:* a series of changes are implemented every 20 units of time. The sequence of the values taken by T_l is $[0.55, 2.5, 1.2, 1.5, 3.0.8, 0.55]$.

The output of the system for each scenario is displayed in Figure 4.12. The estimation results are displayed in

- Figures 4.13 and 4.14, for scenario 2,
- Figures 4.15 and 4.16, for scenario 3.

In all the figures, the thick black line corresponds to the real values of the state variables. In each case the behaviors of several observers are shown:

- dark blue plot: estimation rendered by an extended Kalman filter, with Q and R matrices as given in Table 4.1,

Scenario 1	$u(t) = 120 + 12\sin(t)$, $T_l(t) = 0.55\ \forall t \geq 0$
Scenario 2	$u(t) = 120$, $T_l(t) = 0.55\ \forall t \in [0; 30[$ then $T_l(t) = 2.55\ \forall t \in [30; 100[$
Scenario 3	$u(t) = 120$, $T_l(0) = 0.55$, $T_l(k20)$, $k > 0$ changes according to the sequence $[0.55, 2.5, 1.2, 1.5, 3.0.8, 0.55]$

Figure 4.11: The several simulation scenarios.

– light blue curve: estimation rendered by a High-gain extended Kalman filter with $\theta = 7$,

– red line: estimation done by the adaptive high-gain Kalman filter.

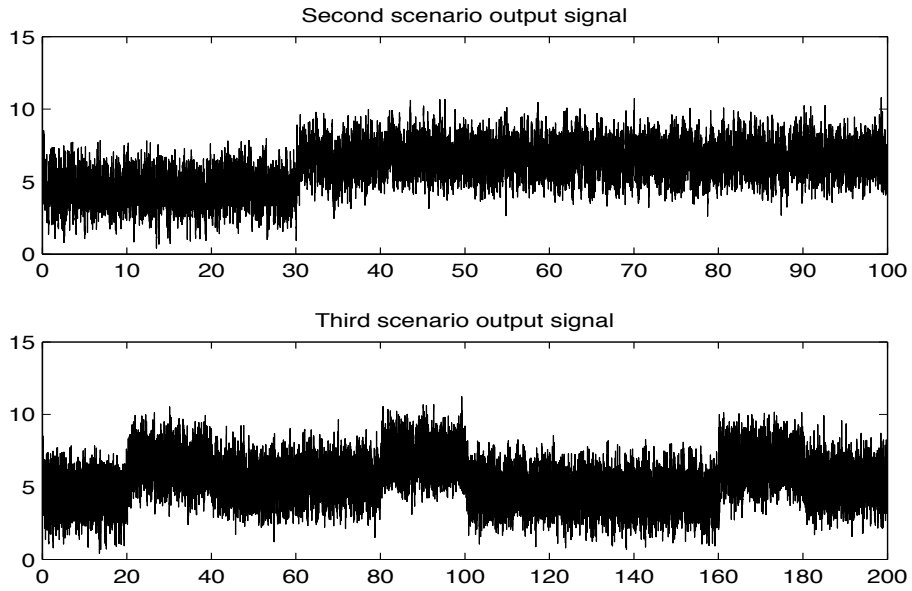

Figure 4.12: Output signal for the two scenarios.

In the first scenario, when the perturbation occurs (at time 30), we see that, as expected, the perturbation is detected by innovation, which crosses the $y = m_2$ red line of Figure 4.17. The high-gain parameter then increases and convergence is made more effective. We see that when no perturbation occurs, innovation remains less than m_2 and the behavior of the observer is the same as the one of the extended Kalman filter[11]. The speed of convergence after the perturbation is comparable to that of the high-gain extended Kalman filter modulo, i.e., a small delay that corresponds to the time needed:

1. for the perturbation to have an effect on the output,
2. for the perturbation to be detected,
3. and for the high-gain to rise.

Notice that this delay depends also on the parameter δ, the sample time set for the computation of the innovation. Since θ reacts on behalf of the innovation, its behavior can change only every δ period of time.

[11] Actually, the performance of the AEKF versus that of the EKF depends also on the level of the noise and on the system under consideration. For example, if the noise level is really high, one would probably set the matrix Q to a very low value thus rendering the EKF even slower. This very low value of Q has only a little effect on the AEKF behavior in high-gain mode. Therefore the AEKF would be as quick to respond as in the present situation and the EKF would be slower. In [21] and [22], the AEKF is used in some other examples thus providing additional insight into the differences between EKF and AEKF.

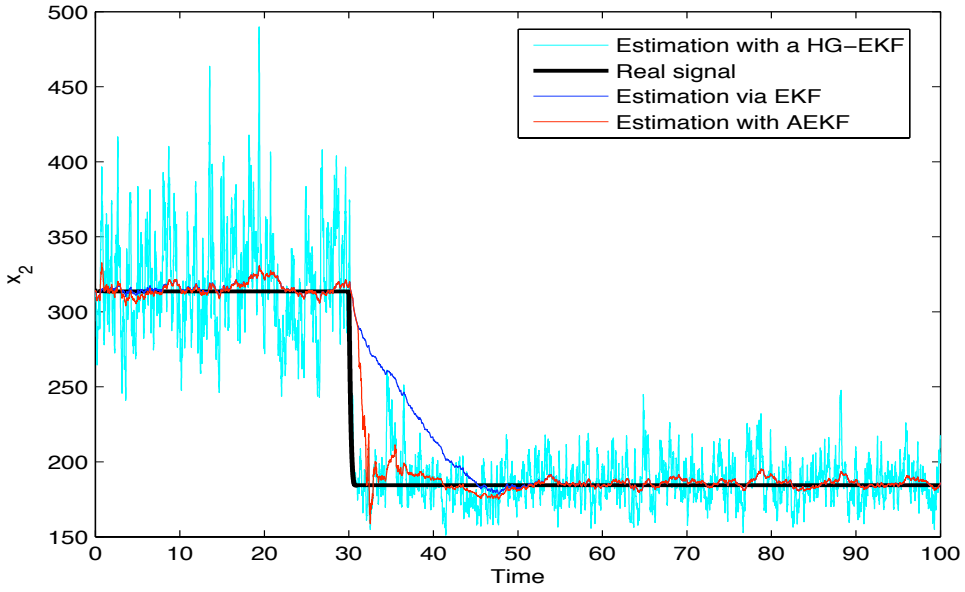

Figure 4.13: *Scenario 2:* Estimation of the rotation speed.

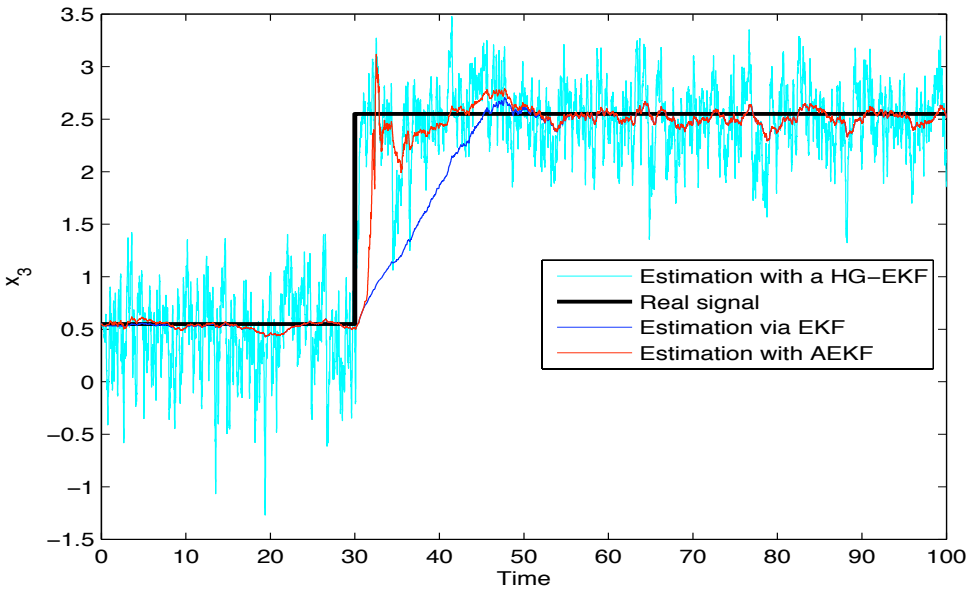

Figure 4.14: *Scenario 2:* Estimation of the torque load.

In the second scenario, we notice that at some unexpected moments the adaptive high-gain observer has the same behavior as that of the non high-gain observer (e.g. Figure 4.15). When we take a look at Figure 4.18, we see that θ didn't actually increase for $t \in [40; 80]$ and $t \in [100; 160]$. The explanation lies in Figure 4.12. The sudden change in the torque load wasn't sufficient enough to have a significant effect on the output signal.

As in the previous scenario, the adaptive observer presents two advantages with respect to the two other filters, namely that of improved noise rejection and that of increased speed of convergence in the event of perturbations.

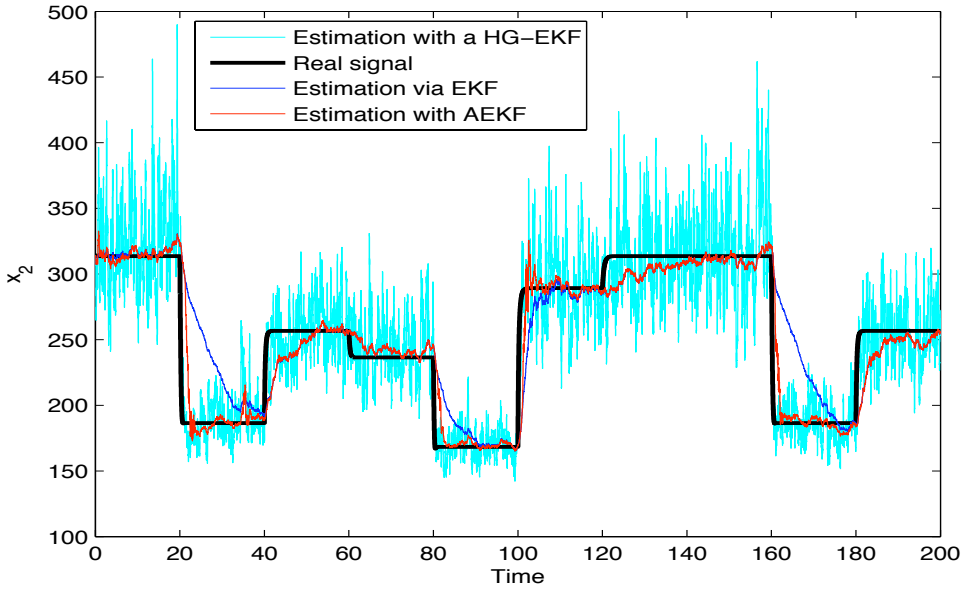

Figure 4.15: *Scenario 3:* Estimation of the rotation speed.

We conclude this section with Figure 4.19, which shows the estimation obtained from a adaptive high-gain observer with a poorly chosen value for m_2. Since it is too small, θ is increasing when it is not needed. The corresponding innovation plot is provided in Figure 4.20.

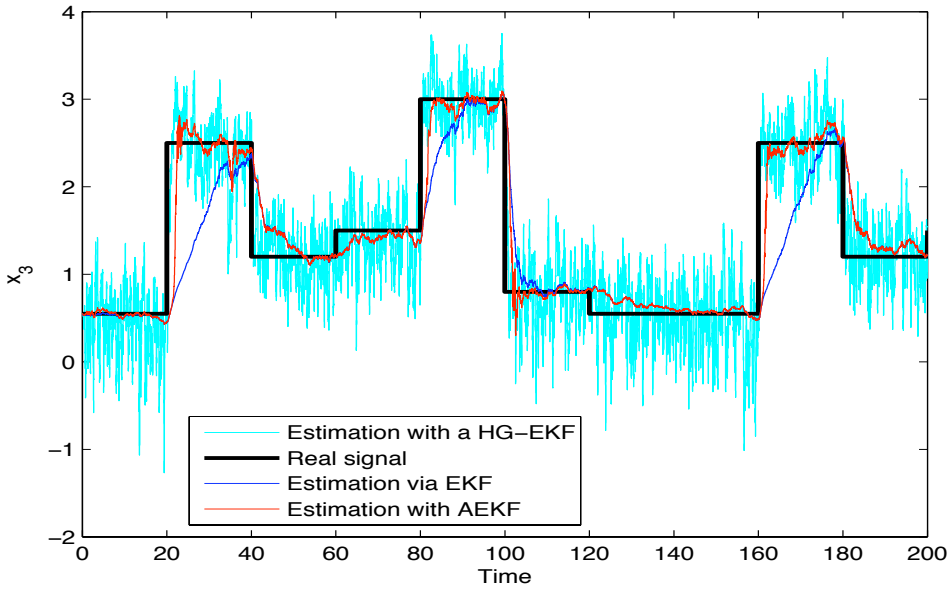

Figure 4.16: *Scenario 3:* Estimation of the torque load.

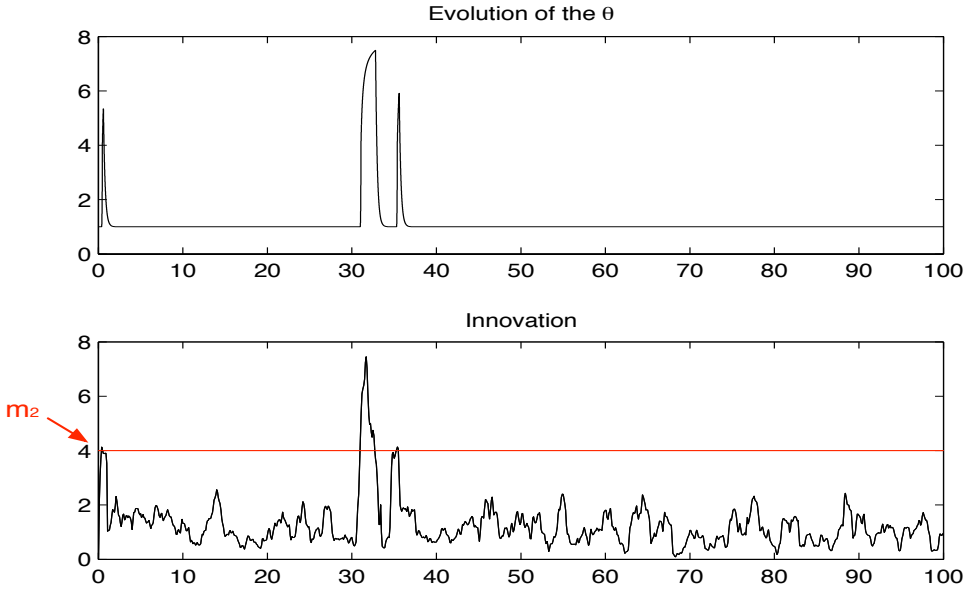

Figure 4.17: *Scenario 2:* Innovation.

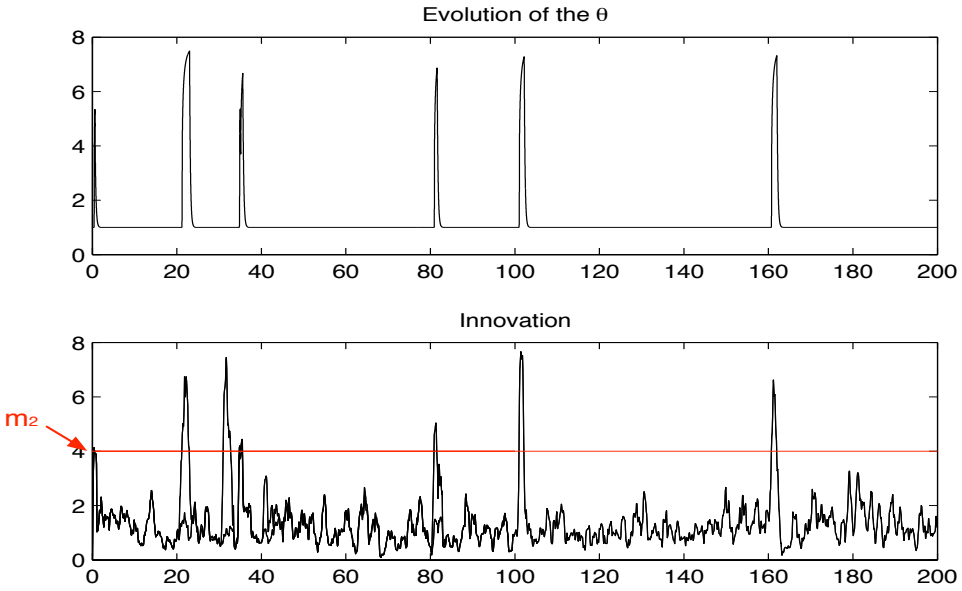

Figure 4.18: *Scenario 3:* Innovation.

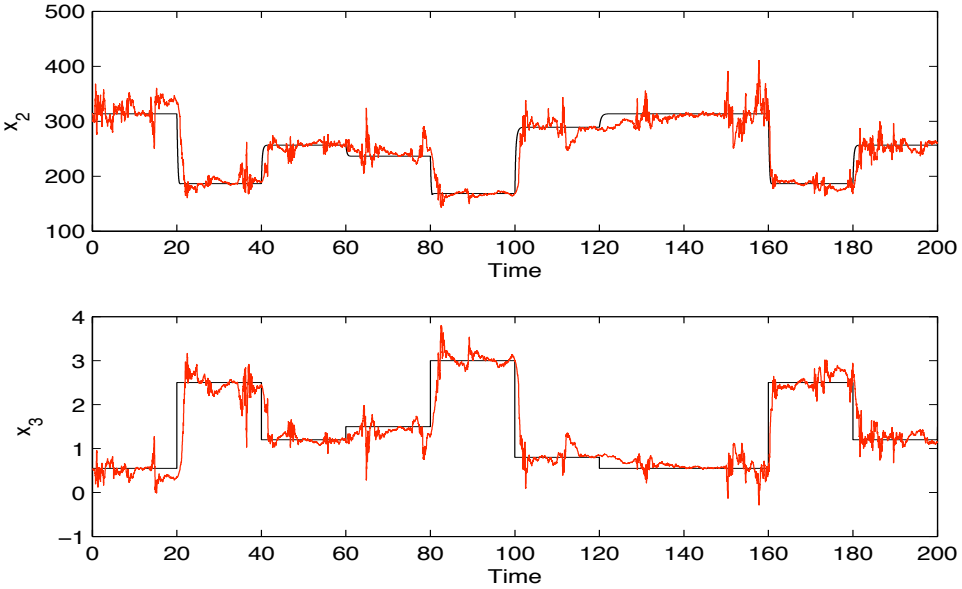

Figure 4.19: Estimation with a adaptive extended Kalman filter with a too low m_2 parameter.

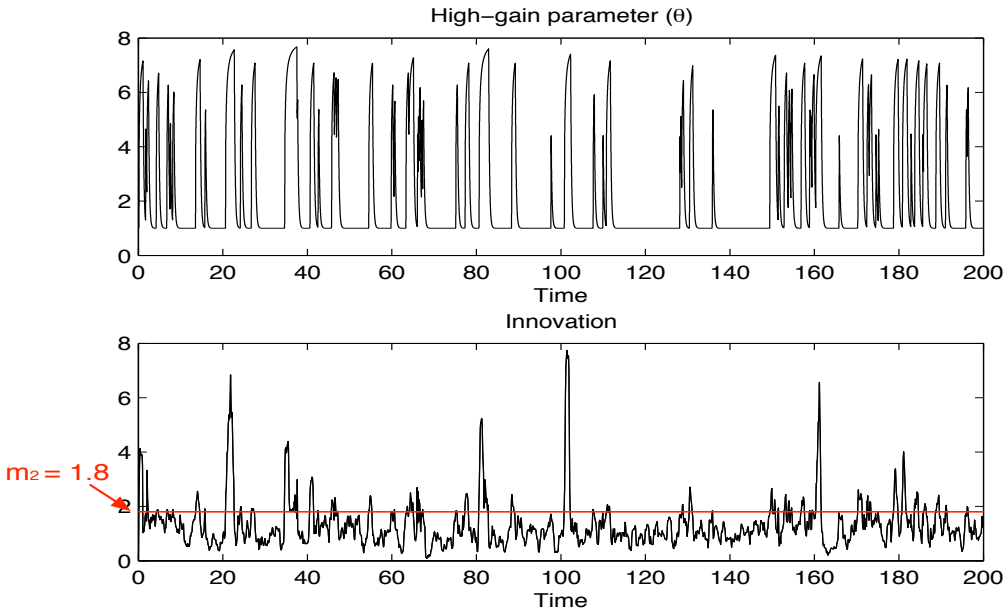

Figure 4.20: Innovation for a low m_2 parameter.

4.3 real-time Implementation

In order to investigate in depth the way the adaptive high-gain extended Kalman filter works, we implemented the filter on a DC machine in our laboratory, at the University of Luxemburg. The issues which concerned us when performing those experiments were:

- the study of the feasibility of a real time implementation, since:
  1. a solution of the Riccati equation requires the integration of $\frac{n(n+1)}{2}$ differential equations, where n denotes the dimension of the state space,
  2. the computation of innovation appears to be time intensive[12],
- the study of the influence of unknown and non-estimated modeling errors on the adaptation scheme and finding a way to handle them.

We considered the real-time part of this problem as being the major issue of this set of experiments. Therefore we decided to use a *hard real-time* operating system.

4.3.1 Softwares

In computer science, real-time computing is the study of hardware and software systems that are subject to operational deadlines from event to system response. The real-time framework can be divided into two parts: *soft* and *hard* real-time. A requirement is considered

[12]This question was also raised by M. Farza and G. Besançon at an early stage of this work, at the occasion of the conference [25].

to be *hard real-time* whenever the completion of an operation after its deadline is considered useless. When limited delays in the response time can be accepted, or in other words when we can afford to wait for the end of computations, the system is said to be *soft real-time*.

The most common approach consists of defining a real-time task and using a clock that sends signals to the system at a frequency set by the user. Each time a signal is sent by the clock, a real-time task is launched whatever the conclusion of the previous task. In other words, a hard real-time system does not make a task conditional with respect to the real-time constraints. The system forces the designer of the task to respect the real-time constraints. We chose a Linux based real-time engine provided with a full development suite: RTAI-Lab [34].

RTAI-Lab is composed of several software components including:

- RTAI [5]: RTAi is a user friendly RealTime Operating System.

  The linux O.S. suffers from a lack of real time support. To obtain real time behavior, it is necessary to change the kernel source code. RTAI is an add-on to the Linux Kernel core that provides it with the features of an industrial real-time operating system within a full non real-time operating system (access to TCP/IP, graphical display and windowing systems, etc...).

  Basically RTAI is a non intrusive interrupt dispatcher, it traps the peripheral interrupts and when necessary re-routes them to Linux. It uses a concept called Hardware Abstraction Layer to get information into and out of the kernel with only a few dependencies. RTAI considers Linux as a background task running when no real time activity occurs.

- Comedi [4]: Comedi is a collection of drivers for a variety of common data acquisition plug-in boards. The drivers are implemented as a core Linux kernel module.

- Scilab/Scicos [6–8, 41]: Scilab is a free scientific software package for numerical computation similar to Matlab. The software was initially developed at *INRIA* and is now under the guidance of the *Scilab consortium* (see the history section of [8]).

  Scicos is a a graphical dynamical system modeler and simulator (or Computer Aided Control System Design Software) developed by the group *METALAU* at INRIA. It provides a block oriented development environment that can be found either embedded into Scilab or in the distribution ScicosLab[13].

- RTAI-Lib [5, 34]: RTAI-Lib is a Scicos *palette*, i.e. a collection of blocks to use with Scicos. This palette is specific to the real-time issues RTAI is dealing with. These blocks can be used to generate a real-time task (which is not the case of the regular Scicos blocks).

- Xrtailab [34]: Xrtailab is a oscilloscope-like software that takes care of communications between the non real-time part of the platform (i.e. the Linux O.S, the graphic displays) and the real-time executable when it is active. With Xrtailab, it is possible to plot and record signals and change online the parameter values of the simulation blocks (e.g. PI and PID coefficients, activate braking).

[13] Notice that recent versions of Scilab (2010) doesn't seem to include Scicos anymore but some similar utility called *xcos*.

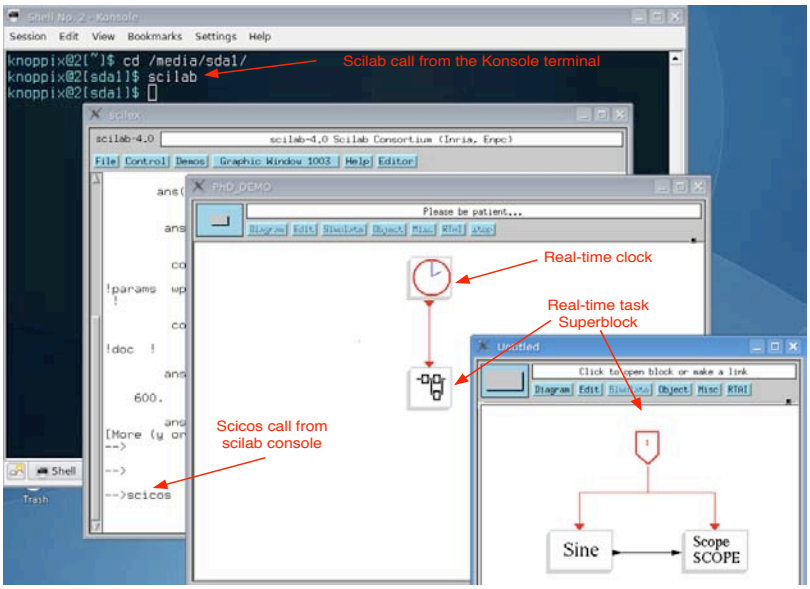

Figure 4.21: Graphical Implementation of a real-time task.

For portability and flexibility reasons, we used a Linux Live CD[14] comprising the RTAI-Lab suite: *RTAI-Knoppix*[15] [3, 94]. When we operate in realtime, that is to say when an observer is running, the RT tasks don't require any hard drive access. Consequently there is no difference between the linux live CD and a regular linux installation.

The development of a real-time executable is done from Scicos, launched from Scilab as shown on Figure 4.21. A Scicos diagram, which is meant to be compiled as a real-time application is composed of two blocks:

– an external clock (in red),

– a Scicos superblock that contains the whole real-time task (in black).

The only input to this block is the external clock signal. Communication between the system and the real-time task is done using specific blocks (signal generation, Scopes, Analog/Digital and Digital/Analog blocks [34]). They can be found in the *RTAI-Lib* palette[16].

The graphical program obtained is compiled into a real-time executable with the help of an automatic code generator. Figure 4.22 shows the three steps of the compilation:

[14] A Live CD is an O.S. that deploys directly from the CD. No specific installation is needed on the host machine. The programs and real-time tasks can be provided via an external storage source as a USB key. In short, as far as the various softwares are concerned, the only hardware devices required are the CD with RTAI-Knoppix and a USB key.

[15] The version we used was built on a Linux kernel, 2.6.17 (SMP enabled kernel is available) and embedded with

– RTAI version 3.4

– Scilab-4.0/Scicos CACSD platform.

[16] In Scicos language, a palette is a collection of predefined blocks.

– the selection of the compilation options,

– the setting of the real-time parameters (sample time of the real-time executable,...),

– the display of a successful compilation notification.

The program that was compiled above is called *Phd_DEMO*. This real-time executable is started from the system console with one of the two following command lines:

– *./Phd_DEMO -f xx*

– *./Phd_DEMO*

In the first case, the *-f* option specifies that a duration of execution is provided. The *xx* symbols have to be replaced by the desired length of time, in seconds. In the second case, no duration is given and the program runs until it is stopped by the user. In RTAI-Lab, there is only one way to properly stop a real-time executable while it is running, i.e., to use the Xrtailab application.

As for Scilab, Xrtailab can be launched from a system console. The use of Xrtailab is illustrated in Figure 4.23. The two first images give an account of the connection procedure. The last image demonstrates a few possibilities of Xrtailab such as the display of signals entering a *Scope* block or the ability to change blocks parameters on the fly.

A more detailed explanation on how the whole RTAI-lab suite works may be found in [34].

4.3.2 Hardware

The testbed is composed of

– a DC motor from *Lucas Nuelle* (ref. SE2665-5C) that can be connected in series. This machine is coupled

  – on one end with a tachometer[17] (ref.2662-5U),

  – on the other end with a propeller[18] (47.0 x 30.5 cm, together with a 5° pitch). The propeller is attached to a 52mm center hub. Those parts were manufactured by *Aero-naut* (ref.7234/97),

– a programmable DC source from *Delta Electronika* (SM-300-10D),

– an I/O card from *National Instruments* (6024E-DAQ) that allows for communications between the physical system and the control system.

A communication diagram showing the relationship between the different elements of the testbed is provided in Figure 4.24. The figure shows:

[17]The tachometer is only used in order to compare the estimated speed to the real one, and to calibrate the mathematical model (see Subsection 4.3.3) .

[18]The brake we had at our disposal was not working properly. We therefore decided to pursue the experiments without it. As it can be seen from the second equation of system (4.1), if the resistive torque is not sufficient, there are good chances for the machine to run wild. The role of the propeller is to provide the motor with a sufficiently high and stable resistive torque. We chose our propeller on behalf of the study [44].

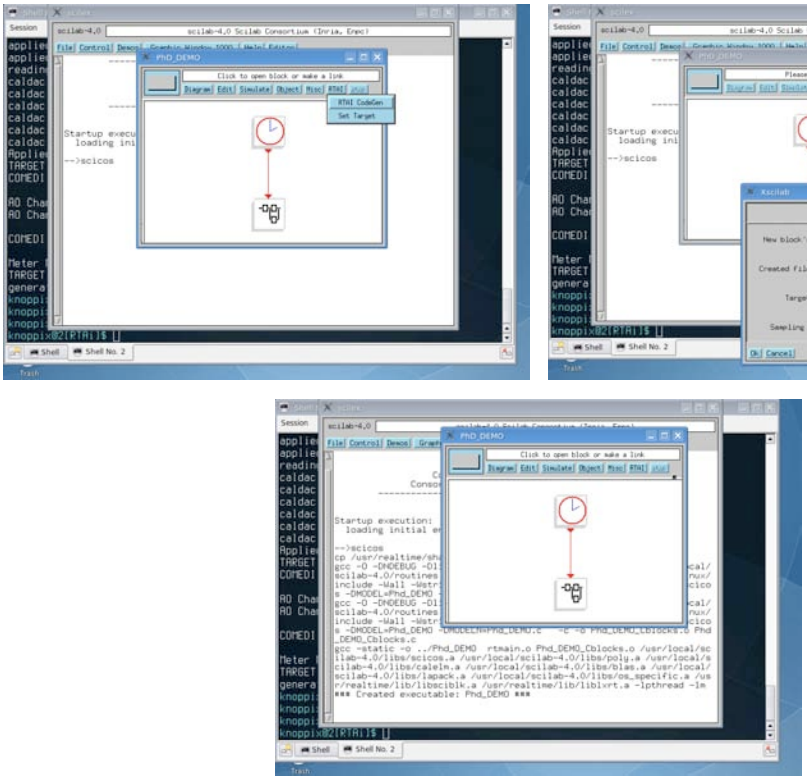

Figure 4.22: Compilation of a real-time task.

– that measurements of the current, voltage and speed are fed to the hard real-time O.S.,

– and that set values for the voltage are delivered to the power supply.

The hard real-time O.S. has therefore 3 inputs and 1 output signals.

Figure 4.25 displays both a picture of the testbed. Also shown in the photograph is the *the friction tool* $n°1$. The undetermined perturbations were produced by means of a hand applied braking friction on the (back) motor's shaft.

The computer was a Dell PC equipped with a P. IV, 3GHz processor and a 512 Mb DDR2 SDRAM memory.

4.3.3 Modeling

A model for the series-connected DC machine has been proposed in Section 4.1.1 above. We now need to adapt this model to the testbed as the presence of the propeller needs to be taken into account. This model can be written, in a short form,

$$\left\{ \begin{array}{rcl} \dot{I} & = & \frac{1}{L}(V - R.I - L_{af} I \omega_r) \\ \dot{\omega_r} & = & \frac{1}{J}(L_{af} I \omega_r - T_{res}) \end{array} \right.$$

where T_{res} is the overall resistive torque. The model from Section 4.1.1 is changed in the two following ways:

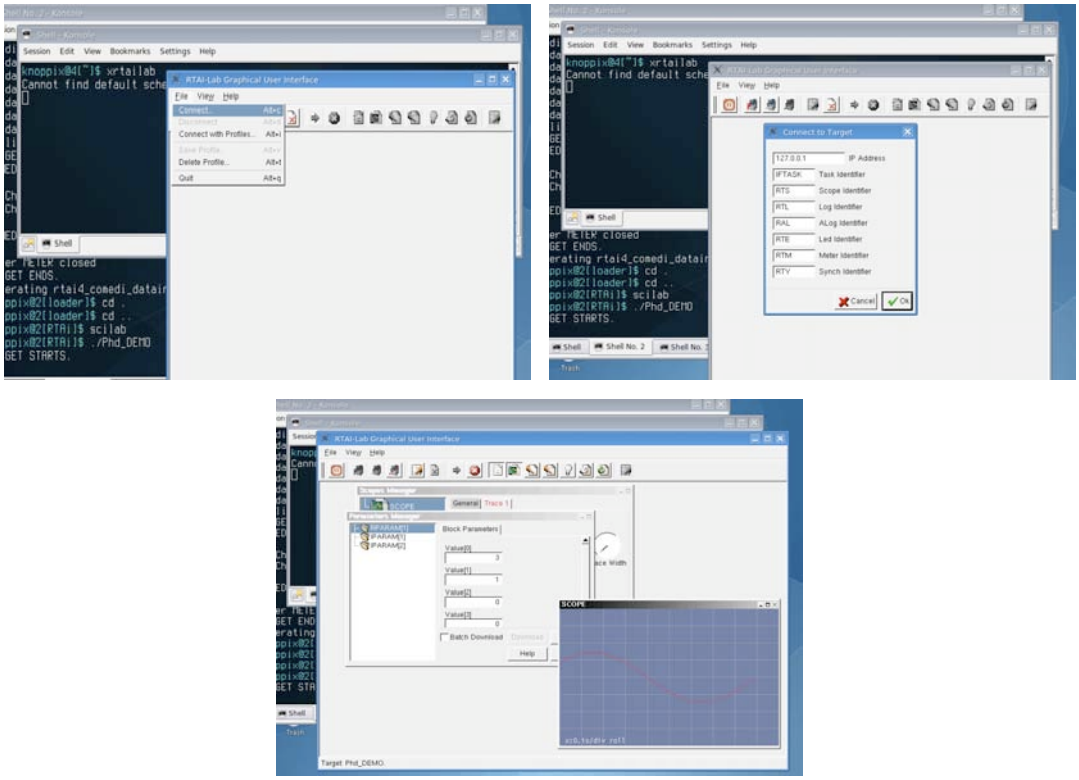

Figure 4.23: Xrtailab.

1. we cannot keep the *ideal efficiency of the machine* assumption any longer. Therefore, instead of considering a unique mutual inductance L_{af} we separate it into two distinct parameters: L_{af_1} and L_{af_2},
2. the resistive torque T_{res} is modeled as the sum of
   - a viscous friction torque generated by the contacts inside the motor: $B\omega_r$,
   - a torque due to the presence of the propeller and which is modeled according to the technical report [44] as $(p\omega_r^{2.08})$,
   - an unknown perturbation torque (i.e. braking), just as before: T_l.

The final model thus obtained is

$$\begin{cases} \dot{I} &= (V - R.I - L_{af_1} I \omega_r)/L \\ \dot{\omega_r} &= (L_{af_2} I \omega_r - B\omega_r - p\omega_r^{2.08} - T_l)/J. \end{cases}$$

Neither the observability analysis nor the change of variables that brought the model into its normal form are affected by these modifications. Therefore, in the same manner as before, we can estimate the load torque T_l by adding a third equation. The corresponding

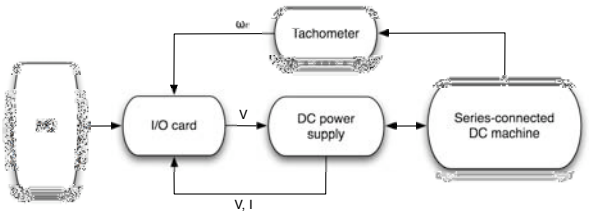

Figure 4.24: Connections Diagram.

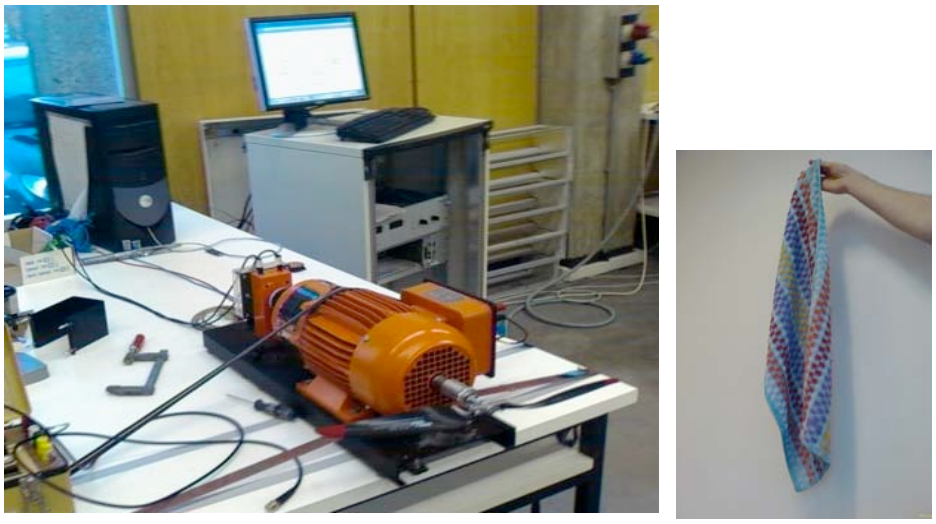

Figure 4.25: A view on the testbed (and, on the left, the friction tool).

observability normal form is then

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{L_{af_1}}{L} & 0 \\ 0 & 0 & -\frac{1}{J} \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} \frac{1}{L}\left(u(t) - Rx_1\right) \\ \frac{1}{L}\left(u(t)\frac{x_2}{x_1} - Rx_2 - Laf_1\frac{x_2^2}{x_1}\right) + \frac{1}{J}\left(L_{af_2}x_1^3 - Bx_2 - p\frac{x_2^{2.08}}{x_1^{1.08}}\right) \\ \frac{1}{L}\left(u(t)\frac{x_3}{x_1} - L_{af_1}\frac{x_2x_3}{x_1} - Rx_3\right) \end{pmatrix}.$$

This model needs to be calibrated, i.e., we need to find appropriate values for the parameters $R, L, B, J, L_{af_1}, L_{af_2}$ and p. We find the values following a standard approach. First, we applied a linear method in order to identify some of the parameters. Second, we used a nonlinear optimization technique to obtain more accurate values for the complete set of parameters.

1. From data samples (voltage, current and speed) an initial estimation is obtained via the least mean squares technique.

   Let us consider the initial model (not in normal form) at steady state. Suppose that $L = J = 1$ and that the load torque (T_l) is null. With the parameters B, L_{af}, p as

unknowns (i.e. $L_{af_1} = L_{af_2}$), the model becomes[19]:

$$\begin{pmatrix} V \\ 0 \end{pmatrix} = \begin{pmatrix} I & I\omega_r & 0 & 0 \\ 0 & I^2 & \omega_r & \omega_r^{2.08} \end{pmatrix} \begin{pmatrix} R \\ L_{af} \\ B \\ p \end{pmatrix}.$$

From a series of experiments we collected enough data to constitute three sets of values: $(V_1, ..., V_N)$, $(I_1, ..., I_N)$ and $((\omega_r)_1, ..., (\omega_r)_N)$ and found a mean least squares solution to the equation above.

2. At this stage a nonlinear optimization routine was used[20]. The solutions found above together with $L = J = 1$, and $L_{af} = L_{af_1} = L_{af_2}$ where taken as the initial values of the optimization routine. The cost function to minimize is taken as the distance between measured data and predicted trajectory

$$K(L, R, L_{af_1}, J, L_{af_2}, B, p) \mapsto \alpha_1 \int_0^{T^*} \|I(v) - \tilde{I}(v)\|^2 dv + \alpha_2 \int_0^{T^*} \|\omega_r(v) - \tilde{\omega}_r(v)\|^2 dv,$$

where

   - I and ω_r are some measured data that excite the dynamical modes of the process (in a pseudo random binary sequence manner [85]),
   - $\tilde{I}$ and $\tilde{\omega}_r$ are the predicted trajectory obtained from the model above using the measured input variables and $\tilde{I}(0) = I(0)$ and $\tilde{\omega}_r(0) = \omega_r(0)$.
   - and α_1, α_2 are weighting factors, that may be used to compensate for the difference of scale between the current and voltage amplitudes.

3. The situation may arise when the solution found by the optimization routine is a local minimum. In this case, the solution set of parameters doesn't match the experimental data while the search algorithm stops. A solution would be to slightly modify the output of the algorithm and relaunch the search.

   Because of the great number of parameters and the difficulty in analyzing the cost function, this re-initialization is rather venturesome. It is difficult to determine which parameters shall be modified, and how. We propose to compare the measured data (I and ω_r) to the predictions ($\tilde{I}$ and $\tilde{\omega}_r$), using the initial set of parameters, and to find p_1 (resp. p_2) such that $p_1 I \approx \tilde{I}$ (resp. $p_2\omega_r \approx \tilde{\omega}_r$). Then we reuse the initial model in order to determine how to modify the parameters, e.g.

$$L\dot{\tilde{I}} = V - R\tilde{I} - L_{af_1}\tilde{I}\tilde{\omega}_r$$

$$L(p_1\dot{I}) = V - R(p_1 I) - L_{af_1}(p_1 I)(p_2\omega_r)$$

$$(p_1 L)\dot{I} = V - (Rp_1)I - (L_{af_1}p_1p_2)I\omega_r.$$

   Although non standard, this method gives us new initial values for a new optimization search. In practice, this method produced excellent results.

[19] If we keep $L_{af_1} \neq L_{af_2}$, the least square solution of the second line is trivially $(0, 0, 0)$

[20] We kept it simple: the simplex search method (see, for example, [88, 104]).

4.3.4 Implementation Issues

Once the Live CD is loaded and the computer is running under the Linux distribution, the real-time environment has to be activated following three stages:

1. several RTAI and COMEDI modules are loaded with the following shell commands (Cf. [34], page 11)[21]:

```
insmod /usr/realtime/modules/rtai_hal.ko}
insmod /usr/realtime/modules/rtai_up.ko $\sharp$or rtailxrt.ko}
insmod /usr/realtime/modules/rtai_fifos.ko}
insmod /usr/realtime/modules/rtai_sem.ko}
insmod /usr/realtime/modules/rtai_mbx.ko}
insmod /usr/realtime/modules/rtai_msg.ko}
insmod /usr/realtime/modules/rtai_netrpc.ko ThisNode="127.0.0.1"}
insmod /usr/realtime/modules/rtai_shm.ko}
insmod /usr/realtime/modules/rtai_leds.ko}
insmod /usr/realtime/modules/rtai_signal.ko}
insmod /usr/realtime/modules/rtai_tasklets.ko}
modprobe ni_pcimio}
modprobe comedi}
modprobe kcomedilib}
modprobe comedi_fc}
insmod /usr/realtime/modules/rtai_comedi.ko }
```

2. the calibration of input/output signals is taken over by a routine provided with the Comedi drivers. The corresponding shell commands are:

```
comedi\_config /dev/comedi0 ni\_pcimio
comedi\_calibrate --no-calibrate -S ni6024e.calibration
chmod 666 /dev/comedi{*}
```

   During this calibration large amplitude step impulses are sent to the motor. In this part of the procedure, it is essential to safety that the power supply BE OFF.

   At the end of the calibration, when the routine stops, the voltage step IS UP (i.e. a high amount of voltage is delivered as soon as one turns the power supply on). This problem is addressed by implementing a (dummy) program[22] that generates any input signal and sends it to the power supply. The corresponding executable is then run for a few seconds. At the end of every RTAI-Lab generated program, output signals generated by the program are reset to zero. Consider that this program is named *end_of_load.cos*. The corresponding realtime executable, after compilation, is *end_of_load.exe*. The command lines for a full and safe calibration are then:

```
comedi\_config /dev/comedi0 ni\_pcimio
comedi\_calibrate --no-calibrate -S ni6024e.calibration
chmod 666 /dev/comedi{*}
./ end_of_load -f 5
```

[21] Without *Super User* rights, those commands are ignored. Employing the *sudo* command is sufficient.
[22] i.e. A clock, any signal (sine wave or step) and a Digital/Analog block.

3. the RTAI version (i.e. 3.4) that is embedded in the *RTAI-Knoppix* has an error in the application file that links the Scicos *Analog/Digital block* (Analog to digital) of the *RTAI-lib* palette to the real-time executable end file. The consequence of this error is that is that it is impossible to measure more than one signal. After a few discussions with the RTAI community, a solution was provided to us by R. Bucher in the form of a corrected version of the faulty file: *rtai4_comedi_datain.sci*. The correct code is given in Appendix C.1. The files to be replaced are located on the virtual file system UNIONFS deployed by the Live CD, in the following repositories:

```
/UNIONFS/usr/src/rtai-3.4/rtai-lab/scilab/macros/RTAI/
/UNIONFS/usr/local/scilab-4.0/macros/RTAI/
```

It is also necessary to recompile the file in the second repository using the shell command

```
scilab -comp rtai4_comedi_datain.sci
```

To do this, one need only to copy/paste the code of Appendix C.1 in any text editor, save this file with the name *rtai4_comedi_datain.sci* and perform the following commands (MY_USB_KEY has to be replaced with the correct path name):

```
sudo rm /UNIONFS/usr/local/scilab-4.0/macros/RTAI/rtai4_comedi_datain.sci
sudo rm /UNIONFS/usr/local/scilab-4.0/macros/RTAI/rtai4_comedi_datain.bin
sudo rm /UNIONFS/usr/src/rtai-3.4/rtai-lab/scilab/macros/RTAI/..
    rtai4_comedi_datain.sci
sudo cp MY_USB_KEY/rtai4_comedi_datain.sci /UNIONFS/usr/local/scilab-4.0/..
    macros/RTAI/rtai4_comedi_datain.sci
sudo cp MY_USB_KEY/rtai4_comedi_datain.sci /UNIONFS/usr/src/rtai-3.4/rtai..
    -lab/scilab/macros/RTAI/rtai4_comedi_datain.sci
sudo cd /UNIONFS/usr/local/scilab-4.0/macros/RTAI/
scilab -comp rtai4_comedi_datain.sci }
```

The real-time platform is now fully functional.

The I/O signal is sampled sufficiently fast as compared to the motor time constant. We implement the continuous version of the adaptive high-gain extended Kalman filter. The general structure of the observer is the same as before. It is displayed in Figure 4.26 together with the corresponding Scicos diagram:

- as before the observer consists of two user defined functions that work in quite the same way as Matlab S-functions,
- the update of the observer equations is performed continuously and the computation of the innovation is done in the discrete time framework,
- a delay block of length d guarantees that we have access at time t to the estimated state computed at time $t-d$ (with the default value being equal to the initial guess of the observer)
- two Analog/Digital blocks get the measured voltage (control variable of the machine) and current (output variable of the machine),

– the two last blocks are gain factors that adjust the measured signals to the appropriate scale[23].

The implementation of the two main blocks is done with the *RTAICblock* block that appears in the *RTAI-lib* palette. It is an adaptation of the *Cblock2* block of the palette *Others*[24].
Generally speaking, Scicos blocks are composed of two files: the *interfacing function* and the *computational function*. The role of the interfacing function is to link the computational function to Scicos. It defines how the computational function has to be interpreted and what the appearance of the Scicos block actually is (number of entry points, size and name of the block, etc.). The computational function is the core of the block. It defines what the block does. Most of the time a flag parameter is used to specify which part of the computational function has to be considered. As may be guessed from its name, the computational function of this block has to be written in C code. The structure is similar to that of a Matlab S-function (see [41], Chpt. 9, in particular Sec. 9.5.2).

The calculation of a solution for the Riccati equation requires several matrix multiplications. We examined the two following approaches:

1. Scicos is embedded into Scilab, which deals pretty smoothly with the multiplication of matrices. Scilab is built from several FORTRAN, C and C++ routines and is open source. This means that the original files are serviceable from Scilab source code. In order to use those routines in C, the header
   #include <routines/machine.h>
   is required. The two routines we need are

   (a) *extern int C2F(dmmul)();* that takes the matrices A, B, C as input parameters and outputs the matrix $C = A \times B$,

   (b) *extern int C2F(dmmul)();* that takes the matrices A, B, C as input parameters and outputs the matrix $C = C + A \times B$,

   Combinations of those two functions enable us to perform all the matrix multiplications required. In addition, recall that the Ricatti matrix of Kalman-like filters is square symmetric, i.e., for a $dim(n \times n)$ matrix, only $n(n+1)/2$ integrations (or updates) are required. A small program that transforms the square matrix into a corresponding column vector and vice versa, must be developed.

2. the matrices used have particular shapes:

   – $A(u)$ is an upper diagonal matrix,
   – Q and R are taken diagonal,
   – $b^*(z, u)$ is lower triangular.

[23] The I/O card delivers a digital input positive signal on the range $(0 - 5)$, while the maximum current supported by the DC supply is 10 A. The scaling factor is therefore 2. The scaling factor for the voltage is 60.

[24] The palette *Others* is a regular Scicos palette. Since those two functions need to interact with the real-time routine of the operating system and have to be taken from the *RTAI-lib* palette.

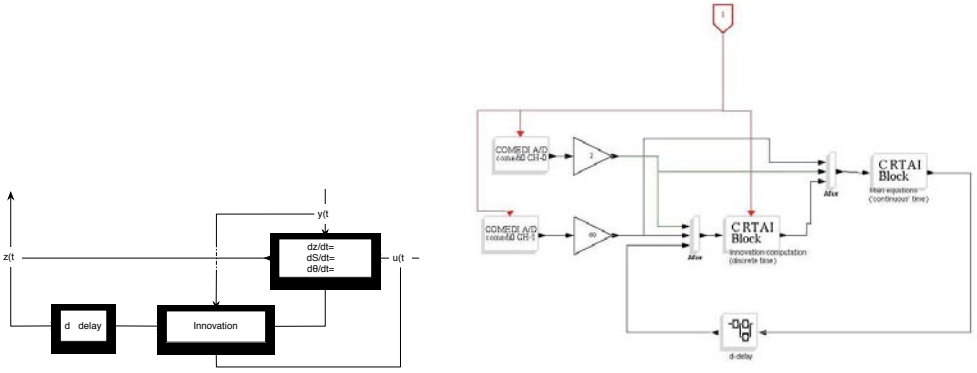

Figure 4.26: Decomposition of the adaptive-gain extended Kalman filter.
(development diagram and Scicos block of the real-time task)

Consequently, developing the equations on paper gives us simplified expressions. In addition, useless computations are avoided[25].

In the case of the DC machine, with a model made up of 3 equations, the second solution is definitely the one to use.

The computation of the innovation requires a solution of the model given in equation (4.3.3) over a time window of length d. This may also be done using an external routine, *lsode*, which already exists in Scilab source code. It is the simulator used both by Scilab and Scicos to cope with ordinary differential equations. Here again the *machine.h* header and a function call, identical to the ones described above, need to be included (for more information on how to use this routine see [73]). Unfortunately, unlike *dmmul* and *dmmul1*, the use of *lsode* is not supported by the real-time compiler. We instead implement a fourth order Runge-Kutta algorithm (see [104] for example).

The C code of those two computational functions is given in Appendices C.3 and C.4.

4.3.5 Experimental Results

We implemented the adaptive high-gain extended Kalman filter along with

- a Luenberger high-gain observer,
- an extended Kalman filter,
- and a high-gain extended Kalman filter.

Several simulations were performed in order to find and correct values for the high-gain parameters. The parameters were estimated at a low speed ($\omega_r = 180\ rad.s^{-1}$) and the observers tested according to the following scenario:

[25] In the present case $n = 3$ which is rather low and allows us to reasonably develop the equations by hand. In the case of systems with much higher dimensions or multiple output systems the use of formal calculation softwares such as *Mathematica*, *Mapple* or *Maxima* [2] shall be considered.

– the machine is fed $54V$ for 30 seconds

– at $t = 30$ the input voltage is switched to $42V$ for another 30 seconds

– the voltage is then raised to $66V$ and finally shut down after 30 seconds.

At each step, for a period of about 10 seconds, an undetermined frictional force is applied to the shaft of the motor (those perturbations are applied by hand, which explains why we operate at such low speeds).

Figure 4.27 shows the estimations provided by the Luenberger observer in an open loop[26]. The high-gain parameter was set to $\theta = 2.5$ and the application was run with a 0.001 seconds time sampling.

Because of the additional computations needed to dynamically solve the Ricatti equation, the high-gain Kalman filter is often seen as a very slow observer. Compared to a Luenberger filter this is indeed the case. Thus, we ran the real-time task at time samplings of 0.01 seconds. On the positive side, this observer is more efficient than a Luenberger when dealing with systems with a matrix A that depends on the input variable $u(t)$.

The results presented in Figure 4.28 show the estimation of the rotational speed computed by both an extended Kalman filter ($\theta = 1$) and its high-gain counterpart ($\theta = 2.5$). Since our perturbations are done by hand (and are thus not reproducible), the only way we can display such information is to run the two observers in parallel. We therefore feel that it is possible to tune the code of the extended Kalman filter to make it run efficiently with a sample time of 0.001 seconds. As expected, the non high-gain filter reacts more slowly to the applied perturbations. As the measurement noise is not large, it is difficult to distinguish the (bad) influence of a large value of θ on the estimation. This influence can still be noticed in the time windows $[5; 10]$ and $[25; 30]$.

The adaptive-gain extended Kalman filter is more demanding in terms of computational time even though this observer runs at the sample time 0.01 seconds. Table 4.2 shows the values selected for this experiment.

The results of this last experiment are displayed in Figure 4.29. Because this run was done with no other observer in parallel, a comparison is not easy to make. Still, we can see that, when dealing with perturbations, the observer has a speed of convergence comparable to that of the high-gain extended Kalman filter and responds in a nature that is as smooth as the one of the extended Kalman filter.

If we take a look at Figure 4.30 we see that the parameter θ reacts 9 times during the experiment: once for every modification plus one extra time corresponding to a bad initialisation. Those reactions correspond to:

– measured changes of the input voltage,

– non-measured changes in the load torque (i.e. perturbations).

It seems regular that changes occur in the second situation but not in first. In fact, this is due to modeling errors which means that the innovation may not vanish. This problem is solved by filtering the innovation[27] in the following way:

[26] During perturbations estimation is less precise but stays in a range of 10 to 15% of the real value. Remember that we had no sensor to measure the torque precisely when the model was calibrated.

[27] The convergence result can be proven with such a filtering process. Once the time d^* of Theorem 36 has been set, we have to design the filtering procedure in such a way that $\mathcal{I}$ doesn't vanish in a time less than d^*.

Parameter	Value	Role
Q	$diag(1, 10^{-1}, 10^{-2})$	Filtering
R	1	Filtering
θ_1	1.25	High-gain
β	$1664\frac{\pi}{e}$	Adaptation*
m_1	0.005	Adaptation*
m_2	0.004	Adaptation
λ	100	Adaptation*
ΔT	0.01	Adaptation*
d	0.1	Innovation
δ	0.1	Innovation*

Table 4.2: Final choice of parameters (*: Application-free parameters).

$$\begin{cases} \dot{\mathcal{I}}_f &= \alpha(\mathcal{I} - \mathcal{I}_f) \\ \mathcal{I}_{used} &= \mathcal{I} - \mathcal{I}_f \end{cases}$$

where α fixes the maximum time that θ will remain fixed at its maximum value.

4.4 Conclusions

In this chapter, the adaptive high-gain extended Kalman filter was completely defined. A sigmoid function allows us to take care of the influence of measurement noise on the final value of innovation. We proposed a clear methodology in order to efficiently tune the parameters of the observer. Its performance was compared to those of a pure high-gain and a non high-gain observer in simulation.

In the second part of this chapter, the implementation of the observer, in a hard real-time environment, on a simple process has been investigated in detail. The observer's behavior is as efficient as it is in simulations. The compliance to hard real-time constraints showed that the effective use of this algorithm is not a mathematician's mid-summer night's dream.

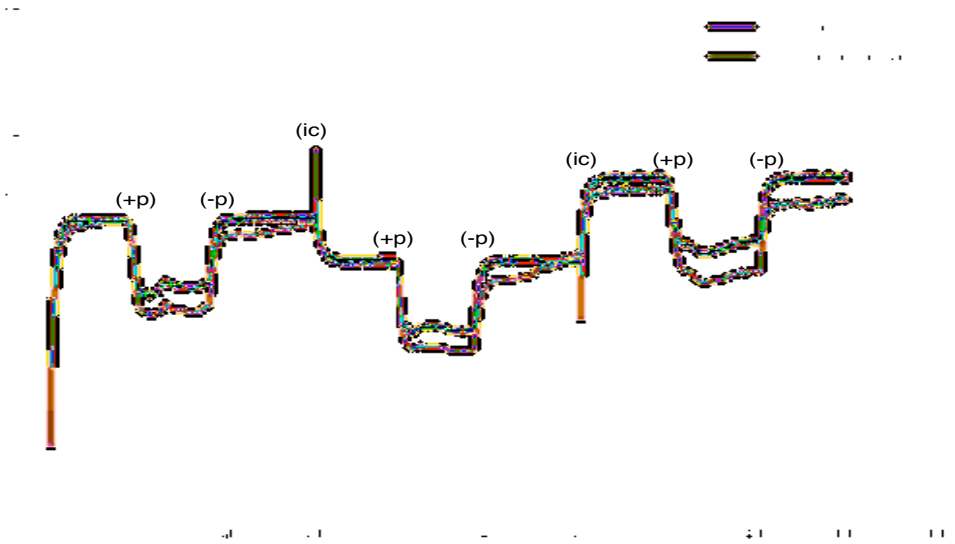

Figure 4.27: Speed estimation using a high-gain extended Luenberger observer.
(+p): begining of perturbation, (-p): end of perturabtion, (Ic): change of the input.

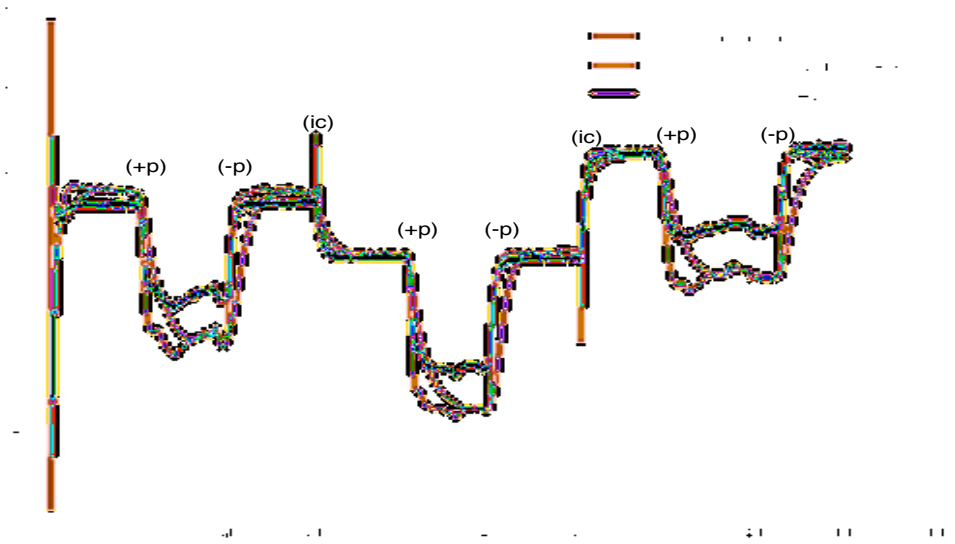

Figure 4.28: Speed estimation using a high-gain extended Kalman filter
(+p): begining of perturbation, (-p): end of perturabtion, (Ic): change of the input.

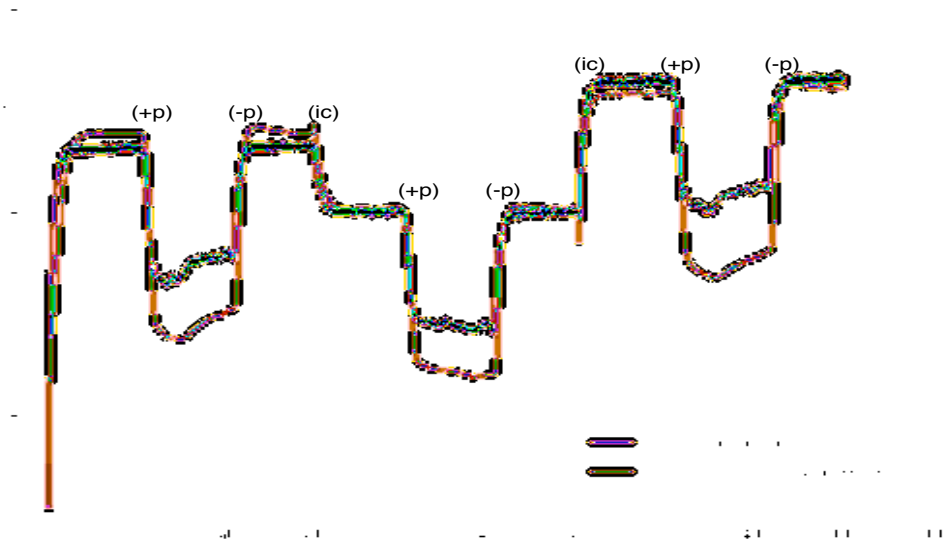

Figure 4.29: Speed estimation using the adaptive-gain extended Kalman filter (+p): begining of perturbation, (-p): end of perturabtion, (Ic): change of the input.

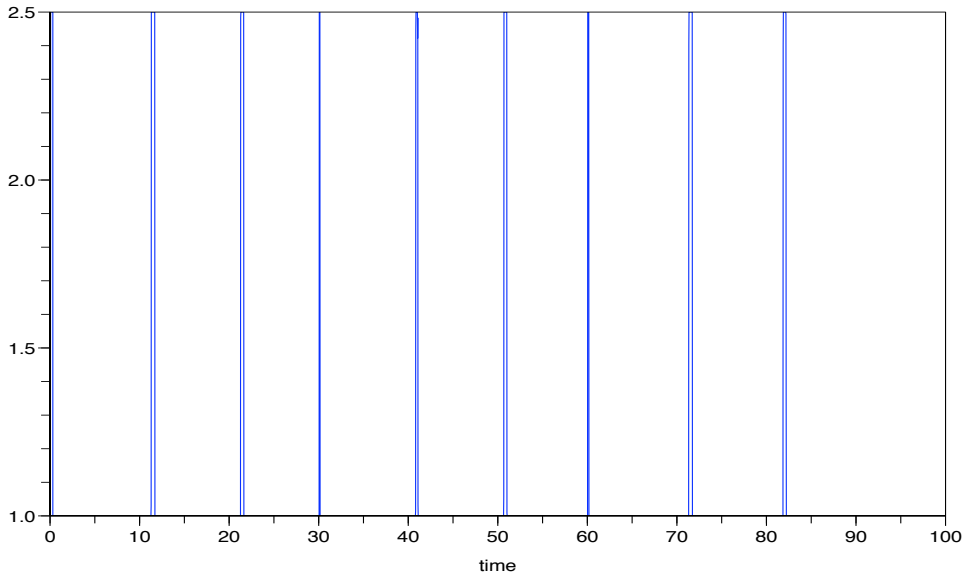

Figure 4.30: Evolution of theta.

Chapter 5

Complements

Contents

This chapter contains several additional and complementary considerations, which are related to adaptive high-gain observers. The first section provides some insight into the multiple outputs case. A generalization of the observer of Chapter 3 is also presented. In the second section, we develop an observer for a continuous-discrete system.

For these two cases, we define the requirements necessary to achieve exponential convergence.

5.1 Multiple Inputs, Multiple Outputs Case

From a theoretical point of view, the multiple outputs case is harder to handle than the case of a single output. Indeed, there is no unique observability form (see [19, 20, 27, 51, 57, 63, 84] and the references herein). The multiple outputs case is also more complex in practice since the various normal forms lead to different definitions of the observer. In the section below, we propose a generalization of the normal form (3.2) together with the definition of the corresponding observer. The proof of the convergence of this observer is given in Subsection 5.1.3.

Although in a more compact form than in Chapter 3, we keep the proof self contained, which implies that we will repeat ourselves to some extent. The modifications of the proof, which are specific to the Multiple Inputs/Multiple Outputs(MIMO) case, are denoted by a thin vertical line in the left margin. The single output case, which was presented before, is included in this generalized version.

5.1.1 System Under Consideration

We focus on a blockwise generalization of the multiple input, single output form (3.2) of Chapter 3. A similar form has been used in the Ph.D. thesis of F. Viel, [113] for a high-gain extended Kalman filter. Another choice has been made in [57] for an even dimension state vector.

- The state variable $x(t)$ resides, as before, within a compact subset $\chi \subset \mathbb{R}^n$,
- The input variable $u(t)$ resides within a subset $\mathcal{U}_{adm} \subset \mathbb{R}^{n_u}$,
- The output vector $y(t)$ is within $\mathbb{R}^{n_y}$ where $n_y \leq 1$.

The system is of the form:

$$\left\{ \begin{array}{rcl} \frac{dx}{dt} & = & A\left(u\right)x + b\left(x,u\right) \\ y & = & C\left(u\right)x. \end{array} \right. , \tag{5.1}$$

and the state variable is decomposed as

$$x(t) = \left(x_1(t), ..., x_{n_y}(t)\right)',$$

where for any i, $i \in \{1, ..., n_y\}$: $x_i \in \chi_i \subset \mathbb{R}^{n_i}$, χ_i compact. Therefore $n = \sum_{i=1}^{n_y} n_i$, and each element $x_i(t)$ is such that:

$$x_i(t) = \left(x_i^1(t), x_i^2(t), ..., x_i^{n_i}(t)\right)'.$$

The matrices $A(u)$ and $C(u)$ are given by:

$$A = \begin{pmatrix} A_1(u) & 0 & \dots & 0 \\ 0 & A_2(u) & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & A_{n_y}(u) \end{pmatrix}, \quad A_i(u) = \begin{pmatrix} 0 & \alpha_i^2 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \alpha_i^{n_i} \\ 0 & \dots & 0 & 0 \end{pmatrix}$$

and the matrix $C(u)$ is the generalization:

$$C = \begin{pmatrix} C_1(u) & 0 & ... & 0 \\ 0 & C_2(u) & ... & 0 \\ & & \ddots & \\ 0 & ... & & 0 \\ 0 & 0 & ... & C_{n_y}(u) \end{pmatrix}, \quad C_i = \begin{pmatrix} \alpha_i^1(u) & 0 & ... & 0 \end{pmatrix}.$$

Finally the vector field $b(x,u)$ is defined as:

$$b(x,u) = \begin{pmatrix} b_1(x,u) \\ b_2(x,u) \\ ... \\ b_{n_y}(x,u) \end{pmatrix}, \quad b_i(x,u) = \begin{pmatrix} b_i^1(x_i^1,u) \\ b_i^2(x_i^1,x_i^2,u) \\ ... \\ b_i^{n_i}(\mathbf{x},u) \end{pmatrix}.$$

Remark 47

The very last component of each element $b_i(.,.)$ of the vector field b is allowed to depend on the full state. As one can see, the linear part is a Brunovsky form of observability: this is clearly a generalization of system (3.2). Nevertheless not every observable system can be transformed into this form.

5.1.2 Definition of the Observer

In order to preserve the convergence result, apply a few modifications to the observer. There are mainly three points to consider

- the definition of innovation is adapted, rather trivially, to a multidimensional output space;
- the matrixΔ , generated with the parameter θ is not the generalization one would imagine initially;
- and the definition of the matrix R_θ must remain compatible with the matrixΔ.

Let us first recall the equations of the observer:

$$\begin{cases} \frac{dz}{dt} &= A(u)z + b(z,u) - S^{-1}C'R_\theta^{-1}(Cz - y(t)) \\ \frac{dS}{dt} &= -(A(u) + b^*(z,u))'S - S(A(u) + b^*(z,u)) + C'R_\theta^{-1}C - SQ_\theta S \\ \frac{d\theta}{dt} &= \mathcal{F}(\theta, \mathcal{I}_d(t)) \end{cases}$$

where Q and R are symmetric positive definite matrices of dimension $(n \times n)$ and $(n_y \times n_y)$ respectively. The innovation at time t is:

$$\mathcal{I}_d(t) = \int_{t-d}^{t} \| y(s) - y(t-d, z(t-d), s) \|^2_{\mathbb{R}^{n_y}} ds$$

where:

- $y\left(t-d, z(t-d), s\right)$ denotes the output of the system of Subsection 5.1 computed over the interval $s \in [t-d; t]$ with $z(t-d)$ as the initial state,
- y is the measured output of dimension n_y.

Let us now denote $n^* = \max\left(n_1, n_2, ..., n_{n_y}\right)$, the size of the largest block, and define the matrix:

$$\Delta_i = \begin{pmatrix} 1/\theta^{n^*-n_i} & 0 & ... & 0 \\ 0 & 1/\theta^{n^*-(n_i-1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1/\theta^{n^*-1} \end{pmatrix} \tag{5.2}$$

and Δ is given by $diag(\Delta_1, ..., \Delta_{n_y})$. The definition of Q_θ is the same as before:

$$Q_\theta = \theta\Delta^{-1}Q\Delta^{-1}.$$

We need to provide a new definition for the matrix R_θ. To do so, we begin with the matrix:

$$\delta_\theta = \begin{pmatrix} \theta^{n^*-n_1} & 0 & ... & 0 \\ 0 & \theta^{n^*-n_2} & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \theta^{n^*-n_{n_y}} \end{pmatrix},$$

and set:

$$R_\theta = \frac{1}{\theta}\delta_\theta R \delta_\theta.$$

The initial state of the observer is:

- $z(0) \in \chi \subset \mathbb{R}^n$,
- $S(0) \in S_n(+)$, the set of the symmetric positive definite matrices,
- $\theta(0) = 1$.

Remark 48

This definition can also be used for single output systems. Since $(n_y = 1) \Rightarrow (n = n_1$ *and* $n^* = n) \Rightarrow (n^* - n_1 = 0)$. *Therefore:*

- $\delta_\theta = 1$,
- Δ *is the same as in Chapter 3, equation (3.3).*

5.1.3 Convergence and Proof

The convergence theorems remain as presented in Chapter 3:

Theorem 49

For any time $T^* > 0$ *and any* $\varepsilon^* > 0$, *there exist* $0 < d < T^*$ *and a function* $\mathcal{F}\left(\theta, \mathcal{I}_d\right)$ *such that for any time* $t \geq T^*$:

$$\|x\left(t\right) - z\left(t\right)\|^2 \leq \varepsilon^* \mathrm{e}^{-a\left(t-T^*\right)}$$

where $a > 0$ *is a constant (independent from* ε^**).*

The proof of convergence is quite long. For the sake of simplicity, we divide the proof into several parts:

1. the innovation lemma,
2. a study of the properties of the Riccati matrix S,
3. the calculation of some preliminary inequalities (i.e. the preparation for the proof),
4. the three technical lemmas,
5. and the main body of the proof.

Points 2, 4 and 5 are essentially the same as in the single output case. Only points 1 and 3 require modifications for the multiple output case. The modifications are explained here.

5.1.3.1 Lemma on Innovation

Lemma 50 ***(Lemma on innovation, multiple output case)***
Let $x_1^0, x_2^0 \in \mathbb{R}^n$ and $u \in \mathcal{U}_{\text{adm}}$. Let us consider the outputs $y\left(0, x_1^0, \cdot\right)$ and $y\left(0, x_2^0, \cdot\right)$ of system (5.1) with initial conditions respectively x_1^0 and x_2^0. Then the following property (called persistent observability) holds:

$$\forall d > 0, \exists \lambda_d^0 > 0 \text{ such that } \forall u \in L_b^1(\mathcal{U}_{adm})$$

$$\|x_1^0 - x_2^0\|^2 \leq \frac{1}{\lambda_d^0} \int_0^d \|y\left(0, x_1^0, \tau\right) - y\left(0, x_2^0, \tau\right)\|_{\mathbb{R}^{n_y}}^2 d\tau \tag{5.3}$$

Proof.
Let $x_1(t) = x_{x_1^0,u}(t)$ and $x_2(t) = x_{x_2^0,u}(t)$ the solutions of (5.1) with $x_i(0) = x_i^0$, $i = 1, 2$. A few computations lead to

$$b(x_2, u) - b(x_1, u) = B(t)(x_2 - x_1)$$

where $B(t) = \int_0^1 \frac{\partial b}{\partial x}(\alpha x_2 + (1-\alpha)x_1, u) d\alpha$.
Set $\varepsilon = x_1 - x_2$. Let us consider the system:

$$\begin{cases} \dot{\varepsilon} &= [A(u) + B(t)]\,\varepsilon \\ y_\epsilon &= C(u)\,\varepsilon = a_1(u)\,\varepsilon_1. \end{cases}$$

It is uniformly observable[1] due to the structure of $B(t)$. We define the Gramm observability matrix, G_d, using $\Psi(t)$, the resolvent[2] of this system:

$$G_d = \int_0^d \Psi(v)' C' C \Psi(v)\, dv.$$

[1] See [57], for example. Alternatively, one could compute the observability matrix

$$\phi_{\mathcal{O}} = \left[C' | (CA)' | \ldots | (CA^n)'\right]',$$

and check the full rank condition for any input.

[2] Cf. Appendix A.1

Since $\|B(t)\| \leq L_b$, each $b_{i,j}(t)$ can be considered as a bounded element of $L^\infty_{[0,d]}(\mathbb{R})$.

We identify $\left(L^\infty_{[0,d]}(\mathbb{R})\right)^p$ with $L^\infty_{[0,d]}(\mathbb{R}^p)$ where[3]

$$p = \sum_{i=1}^{n_y} \frac{n_i(n_i-1)}{2} + n_y n.$$

We consider the function:

$$\begin{array}{rccc} \Lambda: & L^\infty_{[0,d]}(\mathbb{R}^p) \times L^\infty_{[0,d]}(\mathbb{R}^{n_u}) & \longrightarrow & \mathbb{R}^+ \\ & (b_{i,j})_{(j\leq i)\in\{1,..,n\}}, u^\varsigma & \to & \lambda_{\min}(G_d) \end{array}$$

where $\lambda_{\min}(G_d)$ is the smallest eigenvalue of G_d. Let us endow $L^\infty_{[0,d]}(\mathbb{R}^p) \times L^\infty_{[0,d]}(\mathbb{R}^{n_u})$ with the weak-* topology[4] and $\mathbb{R}$ has the topology induced by the uniform convergence. The weak-* topology on a bounded set implies uniform continuity of the resolvent, hence Λ is continuous[5].

Since control variables are supposed to be bounded,

$$\Omega_1 = \left\{ L^\infty_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right); \|B\| \leq L_b \right\}$$

and

$$\Omega_2 = \left\{ u \in L^\infty_{[0,d]}(\mathbb{R}^n); \|u\| \leq M_u \right\}$$

are compact subsets. Therefore $\Lambda(\Omega_1 \times \Omega_2)$ is a compact subset of $\mathbb{R}$ which does not contain 0 since the system is observable for any input. Thus G_d is never singular. Moreover, for M_u sufficiently large, $\left\{ u \in L^\infty_{[0,d]}(\mathbb{R}^n); \|u\| \leq M_u \right\}$ includes $L^\infty_{[0,d]}(\mathcal{U}_{\text{adm}})$.

Hence, there exists λ_d^0 such that $G_d \geq \lambda_d^0 Id$ for any u and any matrix $B(t)$ as above. We conclude that

$$\int_0^d \left\| y(0, x_1^0, \tau) - y(0, x_2^0, \tau) \right\|^2 d\tau \geq \lambda_d^0 \left\| x_1^0 - x_2^0 \right\|^2. \tag{5.4}$$

■

[3]The matrix $B(t)$ can be divided into n_y parts of dimensions $(n_i \times n)$, $i \in \{1, ..., n_y\}$. For each of those parts, the last line may be full (i.e. derivation of the vector field elements of the form $b_{i,n_i}(x,u)$ *w. r. t.* the state x). It gives a maximum of $n_y n$ elements.

For each one of the n_y parts, the lower triangular part of a square of dimension $(n_i - 1 \times n_i - 1)$ may contain non zero elements. That makes $n_i(n_i-1)/2$ elements.

Therefore the maximum number of non null elements of the matrix $B(t)$ is:

$$\sum_{i=1}^{n_y} \frac{n_i(n_i-1)}{2} + n_y n.$$

[4]The definition of the weak-* topology is given in Appendix A.

[5]This property is explained in Appendix A.

5.1.3.2 Preparation for the Proof

First, we remind the reader of the change of variables that we performed in Subsection 3.5.

- $\epsilon = x - z$ is the estimation error,
- $\tilde{x} = \Delta x$,
- $\tilde{b}(., u) = \Delta b(\Delta^{-1}., u)$,
- $\tilde{b}^*(., u) = \Delta b(\Delta^{-1}., u)\Delta^{-1}$.

The definition of the matrix Δ gives us the following property.

Lemma 51

1. *The vector field $\tilde{b}(\tilde{x}, u)$ has the same Lipschitz constant as $b(x, u)$.*
2. *The matrix $\tilde{b}^*(\tilde{x}, u)$ has the same bound as the Jacobian of $b(x, u)$*

Remark 52

This lemma is valid for both the definition of Chapter 3 and the definition of Section 5.1.2 provided above. The proof is given in [57], pg. 215*. We reproduce the proof for the multiple output case since it allows us to justify the definition of Δ.*

Proof.

Recall that $\theta(t) \geq 1$.

1. Consider a component of $\tilde{b}(., u)$ of the form $\tilde{b}_i^k(., u)$ with $i \in \{1, ..., n_y\}$ and $k \in \{1, ..., n_i - 1\}$. From the change of variables $\tilde{b}(., u) = \Delta b(\Delta^{-1}., u)$ we have:

$$\tilde{b}_i^k(x, u) = \frac{1}{n^* - n_i + k - 1} b\left(\theta^{n^*-n_i} x_i^1, \theta^{n^*-n_i+1} x_i^2, ..., \theta^{n^*-n_i+k-1} x_i^k, u\right).$$

We denote L_b as the Lipschitz constant of $b(., u)$ *w.r.t.* the variable x:

$$\begin{aligned}
&\left\|\tilde{b}_i^k(x, u) - \tilde{b}_i^k(z, u)\right\| \\
&\quad = \tfrac{1}{\theta^{n^*-n_i+k-1}} \left\| b\left(\theta^{n^*-n_i} x_i^1, ..., \theta^{n^*-n_i+k-1} x_i^k, u\right)\right. \\
&\qquad\qquad \left. - b\left(\theta^{n^*-n_i} z_i^1, ..., \theta^{n^*-n_i+k-1} z_i^k, u\right)\right\| \\
&\quad \leq \tfrac{L_b}{\theta^{n^*-n_i+k-1}} \left\| \left(\theta^{n^*-n_i} x_i^1, ..., \theta^{n^*-n_i+k-1} x_i^k, u\right)\right. \\
&\qquad\qquad \left. - \left(\theta^{n^*-n_i} z_i^1, ..., \theta^{n^*-n_i+k-1} z_i^k, u\right)\right\| \\
&\quad \leq \tfrac{L_b}{\theta^{n^*-n_i+k-1}} \theta^{n^*-n_i+k-1} \left\| \left(x_i^1, ..., x_i^k, u\right) - \left(z_i^1, ..., z_i^k, u\right)\right\| \\
&\quad = L_b \left\| \left(x_i^1, ..., x_i^k, u\right) - \left(z_i^1, ..., z_i^k, u\right)\right\|.
\end{aligned} \tag{5.5}$$

Therefore the Lipschitz constant of $b(., .)$ is the same in the two coordinate systems.

Consider now an element of the form $\tilde{b}_i^{n_i}(., u)$. Such an element can be a function of the full state. First of all we note that when $n^* = \max_{i \in \{1,...,n_y\}} n_i$ we have:

$$\| \left(\Delta^{-1} x - \Delta^{-1} z\right) \| \leq \theta^{n^*-1} \|x - z\|.$$

The matrix Δ has been defined such that, for all n_i, $i \in \{1, ..., n_y\}$:

$$\tilde{b}_i^{n_i}(., u) = \frac{1}{\theta^{n^*-1}} \tilde{b}_i^{n_i}(., u).$$

This implies that for all n_i, $i \in \{1, ..., n_y\}$:

$$\|\tilde{b}_{n_i}(\tilde{x}, u) - \tilde{b}_{n_i}(\tilde{z}, u)\| \leq L_b \|x - z\|,$$

proving the first part of the lemma.

2. The situation is simpler for the Jacobian matrix $\tilde{b}^*(., u)$. Consider any element denoted $\tilde{b}_{(i,j)}(\tilde{z})$. From the definition of the change of variables there exists $\tilde{i} \in \mathbb{N}$ and $\tilde{j} \in \mathbb{N}$ (i.e. they can be equal to one) such that:

$$\tilde{b}^*_{(i,j)}(\tilde{z}) = \frac{1}{\theta^{\tilde{i}}} b^*_{(i,j)}(\tilde{z}) \theta^{\tilde{j}}$$

with $\tilde{i} \geq \tilde{j}$ (otherwise the element $b^*_{(i,j)} = 0$ because of the structure of $b(x, u)$). Then

$$\|\tilde{b}^*_{(i,j)}(\tilde{z})\| \leq \theta^{\tilde{j}-\tilde{i}} \|b^*_{(i,j)}(\tilde{z})\| \leq \|b^*_{(i,j)}(\tilde{z})\| \leq L_b.$$

proving the lemma.

■

Remark 53

When Δ is defined in a different way, the Lipschitz constant of $\tilde{b}$ isn't L_b. The constant actually depends on the value of θ. This leads to a inconsistent proof since L_b is used in the definition of θ_1 in the proof of Theorem 49.

When an observability form distinct from (5.1) is used, Δ has to be redefined in such a way that the condition (5.5) is satisfied.

We have the following set of identities:

$$\begin{array}{llll} (a) & \Delta A = \theta A \Delta, & (b) & A' \Delta = \theta \Delta A', \\ (c) & A\Delta^{-1} = \theta \Delta^{-1} A, & (d) & \Delta^{-1} A' = \theta A' \Delta^{-1}, \\ (e) & \frac{d}{dt}(\Delta) = -\frac{\mathcal{F}(\theta,\mathcal{J})}{\theta} N\Delta, & (f) & \frac{d}{dt}\left(\Delta^{-1}\right) = \frac{\mathcal{F}(\theta,\mathcal{J})}{\theta} N\Delta^{-1}, \end{array} \tag{5.6}$$

and

$$\begin{array}{lrcl} (g) & \Delta^{-1} C' R_\theta^{-1} C \Delta^{-1} & = & \Delta^{-1} C' \left(\frac{1}{\theta} \delta_\theta R \delta_\theta\right)^{-1} C \Delta^{-1} \\ & & = & \theta C' R^{-1} C. \end{array} \tag{5.7}$$

The matrix N is defined by

$$N = \begin{pmatrix} N_1 & 0 & ... & 0 \\ 0 & N_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & N_{n_y} \end{pmatrix}, \quad N_i = \begin{pmatrix} n^* - n_i & 0 & ... & 0 \\ 0 & n^* - n_i + 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n^* - 1 \end{pmatrix}.$$

The dynamics of the error after the change of coordinates are given by:

$$\frac{d\tilde{\varepsilon}}{dt} = \theta\left[-\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}N\tilde{\varepsilon} + A\tilde{\varepsilon} - \tilde{S}^{-1}C'R^{-1}C\tilde{\varepsilon} + \frac{1}{\theta}\left(\tilde{b}\left(\tilde{z},u\right) - \tilde{b}\left(\tilde{x},u\right)\right)\right], \tag{5.8}$$

and the Riccati equation becomes

$$\frac{d\tilde{S}}{dt} = \theta\left[\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}\left(N\tilde{S}+\tilde{S}N\right) - \left(A'\tilde{S}+\tilde{S}A\right) + C'R^{-1}C - \tilde{S}Q\tilde{S} - \frac{1}{\theta}\tilde{S}\tilde{b}^*\left(\tilde{z},u\right) - \frac{1}{\theta}\tilde{b}^{*\prime}\left(\tilde{z},u\right)\tilde{S}\right]. \tag{5.9}$$

These two equations are used to compute the derivative of $\widetilde{\varepsilon}'\widetilde{S}\widetilde{\varepsilon}$:

$$\frac{d}{dt}\left(\widetilde{\varepsilon}'\widetilde{S}\widetilde{\varepsilon}\right) \leq -\theta q_m\widetilde{\varepsilon}'\widetilde{S}^2\widetilde{\varepsilon} + 2\widetilde{\varepsilon}'\widetilde{S}\left(\tilde{b}\left(\widetilde{z},u\right) - \tilde{b}\left(\widetilde{x},u\right) - \tilde{b}^*\left(\widetilde{z},u\right)\widetilde{\varepsilon}\right). \tag{5.10}$$

The proof of the theorem comes from the stability analysis of this last equation. This analysis requires the use of four lemmas, which we state below.

5.1.3.3 Intermediary Lemmas

The proofs of Lemmas 54 and 57 can be found in Sections 3.6 and 3.7 respectively. Lemmas 55 and 56 are proven in Appendix B.2.

Lemma 54 ***(Bounds for the Riccati equation)***

Let us consider the Riccati equation (5.8). We suppose that

- *the functions* $a_i\left(u\left(t\right)\right)$, $\left|\widetilde{b}^*_{i,j}\left(\overline{z},\overline{u}\right)\right|$,
- $\left|\frac{\mathcal{F}(\overline{\theta},\overline{\mathcal{I}})}{\overline{\theta}^2}\right|$ *are smaller than* $a_M > 0$ *and if* $a_i\left(u\left(t\right)\right) > a_m > 0$
- $S\left(0\right) = S_0$ *is symmetric definite positive, taken in a compact of the form* $aId \leq S_0 \leq bId$*, and*
- $\theta(0) = 1$

Then there exist two constants $0 < \alpha < \beta$ *such that, for all* $t \geq 0$,

$$\alpha\, Id \leq \widetilde{S}\left(t\right) \leq \beta\, Id.$$

Lemma 55 ***(Technical lemma one)***

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

$$\frac{dx(t)}{dt} \leq -k_1 x + k_2 x\sqrt{x},$$

for almost all $t > 0$*, for* k_1, $k_2 > 0$*. Then, if* $x\left(0\right) < \frac{k_1^2}{4k_2^2}$, $x(t) \leq 4\,x\left(0\right)e^{-k_1 t}$.

Lemma 56 ***(Technical lemma two)***

Consider $\widetilde{b}\left(\tilde{z}\right) - \widetilde{b}\left(\tilde{x}\right) - \widetilde{b}^*\left(\tilde{z}\right)\tilde{\varepsilon}$ *as in the inequality (3.12) (omitting to write* u *in* $\tilde{b}$*) and suppose* $\theta \geq 1$*. Then* $\left\|\widetilde{b}\left(\tilde{z}\right) - \widetilde{b}\left(\tilde{x}\right) - \widetilde{b}^*\left(\tilde{z}\right)\tilde{\varepsilon}\right\| \leq K\theta^{n-1}\left\|\tilde{\varepsilon}\right\|^2$*, for some* $K > 0$.

Lemma 57 ***(adaptation function)***

For any $\Delta T > 0$, there exists a positive constant $M(\Delta T)$ such that:

- *for any $\theta_1 > 1$,and*
- *any $\gamma_1 > \gamma_0 > 0$,*

there is a function $\mathcal{F}(\theta, \mathcal{I})$ such that the equation

$$\dot{\theta} = \mathcal{F}(\theta, \mathcal{I}(t)), \tag{5.11}$$

for any initial value $1 \leq \theta(0) < 2\theta_1$, and any measurable positive function $\mathcal{I}(t)$, has the properties:

1. *(5.11) has a unique solution $\theta(t)$ defined for all $t \geq 0$, and this solution satisfies $1 \leq \theta(t) < 2\theta_1$,*
2. *$\left|\frac{\mathcal{F}(\theta,\mathcal{I})}{\theta^2}\right| \leq M$,*
3. *if $\mathcal{I}(t) \geq \gamma_1$ for $t \in [\tau, \tau + \Delta T]$ then $\theta(\tau + \Delta T) \geq \theta_1$,*
4. *while $\mathcal{I}(t) \leq \gamma_0$, $\theta(t)$ decreases to 1.*

5.1.3.4 Proof of the Theorem

First of all let us choose a time horizon d (in $\mathcal{I}_d(t)$) and a time T such that $0 < d < T < T^*$. Let $\mathcal{F}$ be a function as in Lemma 57 with$\Delta\, T = T - d$, and M such that *fact* 2 of Lemma 57 is true. Let α and β be the bounds from Lemma 54.

From the preparation of the proof, inequality (5.10) can be written, using Lemma 54 (i.e. using $\tilde{S} \geq \alpha\, Id$)

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \theta \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + 2\tilde{\varepsilon}'\tilde{S}\left(\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right). \tag{5.12}$$

From (5.12) we can deduce two inequalities: the first one, global, will be used mainly when $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)$ is not in a neighborhood of 0 and θ is large. The second one, local, will be used when $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)$ is small, whatever the value of θ.
Using

$$\left\|\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right\| \leq 2L_b \|\tilde{\varepsilon}\|,$$

together with $\alpha\, Id \leq \tilde{S} \leq \beta\, Id$ (Lemma 54), (5.12) becomes the "global inequality"

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq \left(-\alpha q_m \theta + 4\frac{\beta}{\alpha}L_b\right)\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t). \tag{5.13}$$

Thanks to Lemma 56 we get the"local inequality" as follows:

$$\left\|\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right\| \leq K\theta^{n-1}\|\tilde{\varepsilon}\|^2.$$

Since $1 \leq \theta \leq 2\theta_1$ inequality (5.12) implies

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + 2K(2\theta_1)^{n-1}\left\|\tilde{S}\right\|\|\tilde{\varepsilon}\|^3.$$

Since $\|\tilde{\varepsilon}\|^3 = \left(\|\tilde{\varepsilon}\|^2\right)^{\frac{3}{2}} \leq \left(\frac{1}{\alpha}\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)\right)^{\frac{3}{2}}$, it becomes

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \leq -\alpha q_m \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + \frac{2K(2\theta_1)^{n-1}\beta}{\alpha^{\frac{3}{2}}}\left(\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)\right)^{\frac{3}{2}}. \tag{5.14}$$

Let us apply[6] Lemma 55, which states that if

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) \leq \frac{\alpha^5 q_m^2}{16K^2(2\theta_1)^{2n-2}\beta^2},$$

then, for any $t \geq \tau$,

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \leq 4\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau)\mathrm{e}^{-\alpha q_m(t-\tau)}.$$

Provided there exists a real γ, such that

$$\gamma \leq \frac{1}{(2\theta_1)^{2n-2}}\min\left(\frac{\alpha\varepsilon^*}{4}, \frac{\alpha^5 q_m^2}{16K^2\beta^2}\right), \tag{5.15}$$

then $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) \leq \gamma$ implies, for any $t \geq \tau$,

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \leq \frac{\alpha\varepsilon^*}{(2\theta_1)^{2n-2}}\mathrm{e}^{-\alpha q_m(t-\tau)}. \tag{5.16}$$

From (5.13)

$$\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(T) \leq \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(0)\mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha}L_b\right)T},$$

and if we suppose $\theta \geq \theta_1$ for $t \in [T, T^*]$, $T^* > T$, using (5.13) again:

$$\begin{aligned}\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(T^*) &\leq \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(0)\mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha}L_b\right)T}\mathrm{e}^{\left(-\alpha q_m\theta_1 + 4\frac{\beta}{\alpha}L_b\right)(T^*-T)} \\ &\leq M_0\mathrm{e}^{-\alpha q_m T}\mathrm{e}^{4\frac{\beta}{\alpha}L_b T^*}\mathrm{e}^{-\alpha q_m\theta_1(T^*-T)},\end{aligned} \tag{5.17}$$

where

$$M_0 = \sup_{x,z\in X}\varepsilon' S\varepsilon(0). \tag{5.18}$$

Now, we choose θ_1 and γ for

$$M_0\mathrm{e}^{-\alpha q_m T}\mathrm{e}^{4\frac{\beta}{\alpha}L_b T^*}\mathrm{e}^{-\alpha q_m\theta_1(T^*-T)} \leq \gamma \tag{5.19}$$

and (5.15) to be satisfied simultaneously, which is possible since $\mathrm{e}^{-\mathrm{cte}\times\theta_1} < \frac{\mathrm{cte}}{\theta_1^{2n-2}}$ for θ_1 sufficiently large. Let us choose a function $\mathcal{F}$ as in Lemma 57 with$\Delta\ T = T - d$ and $\gamma_1 = \frac{\lambda_d^0\gamma}{\beta}$.

[6]This lemma cannot be applied if we use Q_θ and R instead of Q_θ and R_θ in the definition of the observer as in [38]. This is due to the presence of a $\frac{\mathcal{F}}{\theta}$ term that prevents parameters k_1 and k_2 to be positive for all times.

We claim that there exists $\tau \leq T^*$ such that $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) \leq \gamma$. Indeed, if $\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) > \gamma$ for all $\tau \leq T^*$ then because of Lemma 50:

$$\gamma < \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(\tau) \leq \beta \|\tilde{\varepsilon}(\tau)\|^2 \leq \beta \|\varepsilon(\tau)\|^2 \leq \frac{\beta}{\lambda_d^0}\mathfrak{I}_d(\tau + d).$$

Therefore, $\mathfrak{I}_d(\tau + d) \geq \gamma_1$ for $\tau \in [0, T^*]$ and hence $\mathfrak{I}_d(\tau) \geq \gamma_1$ for $\tau \in [d, T^*]$, so we have $\theta(t) \geq \theta_1$ for $t \in [T, T^*]$, which results in a contradiction ($\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(T^*) \leq \gamma$) because of (5.17) and (5.19).

Finally, for $t \geq \tau$, using (5.16)

$$\begin{array}{rcl} \|\varepsilon(t)\|^2 & \leq & (2\theta_1)^{2n-2}\|\tilde{\varepsilon}(t)\|^2 \\ & \leq & \frac{(2\theta_1)^{2n-2}}{\alpha}\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) \\ & \leq & \varepsilon^* e^{-\alpha q_m (t-\tau)}, \end{array} \tag{5.20}$$

and the theorem is proven.

Remark 58

Just as in the single output case, an alternative result can be derived. Refer to Remark 44, Chapter 3.

5.2 Continuous-discrete Framework

In the present section we develop an adaptation of the observer to continuous discrete systems. In such systems, the evolution of the state variables is described by a continuous process, and the measurement component of the system by a discrete function. When we cannot use the quasi-continuity assumption of measurements, as we did before, this version is the one to use.

In engineering sciences, when pure discrete filters are used, the discrete model needed is sometimes obtained from a continuous formulation. The model

$$\dot{x} = f(x(t), u(t), t)$$

is transformed into:

$$x(t^* + \delta_t) = x(t^*) + f(x(t^*), u(t^*), t^*)\,\delta_t,$$

where δ_t represents the sampling time of the process. This equation represents nothing more than Euler's numerical integration method. In other words, this modeling technique is a special case of the continuous-discrete framework. The mechanization equation of inertial navigation systems[7] is an example of such modeling, details of which can be found in [10, 114] for example.

Although we directly consider a single output continuous-discrete normal form, we want to draw the reader's attention to the problem of *the preservation of observability under sampling*. In [15], the authors prove that for a continuously observable[8] system, observability is

[7] The mechanization process merges the data coming from 3-axis accelerometers to those coming from a gyroscope.

[8] Uniformly observable for a class of input functions, and uniformly infinitesimally observable in the sense of the definitions given in Chapter 2.

preserved provided that the sample time δ_t is small enough. (Both examples and counterexamples can be found in this article (see also [14] on the same topic)).

Let us state the main theorem of [15].

Theorem 59

Assume that a nonlinear system is observable for every input $u(.)$ and uniformly infinitesimally observable[9]*, then for all $M > 0$, there exists a $\delta_0 > 0$ such that the associated $\delta-$sampled system is observable for all $\delta \leq \delta_0$ and all M, $D-$bounded input u^δ.*

5.2.1 System Definition

Let us consider the continuous-discrete version of the multiple input, single output system of Chapter 3 (Equation 3.2):

$$\left\{ \begin{array}{rcl} \frac{dx}{dt} & = & A(u(t))x + b(x(t), u(t)) \\ y_k & = & Cx_k \end{array} \right. \tag{5.21}$$

where

- δ_t is the constant sampling time of the measurement procedure,
- $x(t) \in \chi \subset \mathbb{R}^n$, χ compact, and $x_k = x(k\delta_t)$, $k \in \mathbb{N}$,
- $u(t) \in \mathcal{U}_{\text{adm}} \subset \mathbb{R}^{n_u}$ bounded, and $u_k = u(k\delta_t)$, $k \in \mathbb{N}$,
- $y(t) \in \mathbb{R}$, and $y_k = y(k\delta_t)$, $k \in \mathbb{N}$.

The matrices $A(u)$ and $C(u)$ are defined by:

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

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

with $0 < a_m \leq a_i(u) \leq a_M$ for any u in $\mathcal{U}_{\text{adm}}$. The $\mathcal{C}^1$ vector field $b(x,u)$ is assumed to be compactly supported and to have the following triangular structure:

$$b(x,u) = \begin{pmatrix} b_1(x_1, u) \\ b_2(x_1, x_2, u) \\ \vdots \\ b_n(x_1, \ldots, x_n, u) \end{pmatrix}.$$

We denote L_b as the bound of the Jacobian matrix $b^*(x,u)$ of $b(x,u)$ *w.r.t.* x (i.e. $\|b^*(x,u)\| \leq L_b$). Since $b(x,u)$ is compactly supported and u is bounded, b is Lipschitz uniformly in x: $\|b(x_1,u) - b(x_2,u)\| \leq L_b \|x_1 - x_2\|$.

[9] See Chapter 2.

Remark 60

Notice that the matrix C is defined differently than in the previous systems. Details lie in Appendix B, Remark 94 in particular.

5.2.2 Observer Definition

In the continuous-discrete setting the observer is defined by:

1. a set of prediction equations for $t \in [(k-1)\delta_t, k\delta_t[$,
2. a set of correction equations at times $t = k\delta_t$.

In the following we use the notations:

- $z(t)$ is the estimated state for all $t \in](k-1)\delta_t, k\delta_t[$,
- $z_k(-)$ is the estimated state at the end of a prediction period,
- $z_k(+)$ is the estimated state after a correction step (i.e. at the beginning of a new prediction period).

The prediction equations for $t \in [k\delta_t, (k+1)\delta_t[$, with initial values $z_{k-1}(+)$, $S_{k-1}(+)$ are

$$\left\{ \begin{array}{rcl} \dot{z} & = & A(u)z + b(z,u) \\ \dot{S} & = & -\left(A(u) + b^*(z,u)\right)' S - S\left(A(u) + b^*(z,u)\right) - SQ_\theta S \end{array} \right. \tag{5.22}$$

where S_0 is a symmetric definite positive matrix taken inside a compact subset of the form $aId \leq S_0 \leq bId$.

The correction equations are:

$$\left\{ \begin{array}{rcl} z_k(+) & = & z_k(-) - S_k(+)^{-1} C' r_\theta^{-1} \delta_t (C z_k(-) - y) \\ S_k(+) & = & S_k(-) + C' r_\theta^{-1} C \delta_t \\ \mathcal{I}_{k,d} & = & \sum_{i=0}^{i=d} \|y_{k-i} - \hat{y}_{k-i}\|^2 \\ \theta_k & = & \mathcal{F}\left(\theta_{k-1}, \mathcal{I}_{k,d}\right) \end{array} \right. \tag{5.23}$$

where

- x_0 and z_0 belongs to χ, a compact subset of $\mathbb{R}^n$,
- $\theta(0) = \theta_0 = 1$,
- r and Q are symmetric definite positive matrices[10]:
  - $Q_\theta = \theta \Delta^{-1} Q \Delta^{-1}$,
  - $r_\theta = \frac{1}{\theta} r$,

[10] r is written in capital letters to emphasize the fact that the system is single output.

with $\Delta = diag\left(\{1, \frac{1}{\theta}, ..., \frac{1}{\theta^{n-1}}\}\right)$,

- the innovation, $\mathcal{I}_{k,d}$ ($d \in \mathbb{N}^*$), is computed over the time window $[(k-d)\delta_t; k\delta_t]$ with:
  - y_{k-i} denotes the measurement at epoch $k-i$, and
  - $\hat{y}_{k-i}$ denotes the output of system (5.21) at epoch $k-i$ with $z\left((k-d)\delta_t\right)$ as initial value at epoch $k-d$.

Remark 61

In [24] we presented a different version of this observer. In this paper, the adaptation of the high-gain parameter was determined during the prediction steps via a differential equation. We changed our strategy with respect to the peculiar manner in which the continuous discrete version of the observer works.

Recall that the estimation is a sequence of continuous prediction periods followed by discrete correction steps when a new measurement/observation is available. Consequently, since innovation is based on the measurements available, it is computed at the correction steps.

Suppose now that θ is adapted via a differential equation. It starts to change during the prediction step following the computation of a large innovation value, and reaches θ_1 after some time. If alternatively θ is adapted at the end of the correction step, directly after the computation of innovation, then it reaches θ_1 much faster.
We opt for the strategy in which θ is adapted at the end of the correction step[11].

An advantage brought about by this approach is that now we may remove one of the assumptions on the adaptation function: "there exists M such that $|\frac{\mathcal{F}}{\theta^2}| < M$".

5.2.3 Convergence Result

Theorem 62

For any time $T^* > 0$ and any $\epsilon^* > 0$, there exist

- *two real constants μ and θ_1,*
- *$d \geq n-1 \in \mathbb{N}^*$, and*
- *an adaptation function $\mathcal{F}(\theta, \mathcal{I})$,*

such that for all δ_t sufficiently small (i.e. $2\theta_1\delta_t < \mu$ and $0 < d < \frac{T^*}{\delta_t}$), any time $t \geq T^*$, and any $(z(0), x(0)) \in \chi^2$ then:

$$\|z(t) - x(t)\|^2 \leq \epsilon^* e^{-a(t-T^*)},$$

where $a > 0$ does not depend on θ.

The proof of convergence in the continuous-discrete case is developed in the subsections to come. The strategy is again to:

[11] Getting rid of the differential equation in the *continuous case*, creates a technical problem. Indeed the computation of the derivative of θ would require the computation of the derivative of innovation.

– let θ increase when innovation is large,

– have θ decrease toward 1 when innovation is small.

In the continuous-discrete setting, an extra hypothesis, related to the sampling time, appears (see Lemma 63 below). As a consequence the beginning of the proof of Subsection 5.2.6 is performed independently from δ_t.

5.2.4 Innovation

The lemma for innovation in the continuous-discrete case is:

Lemma 63

Let $x^0, \xi^0 \in \mathbb{R}^n$ and $u \in \mathcal{U}_{\text{adm}}$. Let us consider the outputs $y_j\left(0, x^0\right)$ and $y_j\left(0, \xi^0\right)$ of system (5.21) with initial conditions x^0 and ξ^0 respectively. Then the following condition (called persistent observability) holds:

$$\forall d \in \mathbb{N}^* \text{ large enough (i.e. } d \geq n-1\text{)}, \exists \lambda_d^0 > 0 \text{ such that } \forall u \in L_b^1(\mathcal{U}_{adm})$$

$$\|x^0 - \xi^0\|^2 \leq \tfrac{1}{\lambda_d^0} \sum_{i=0}^{i=d} \|y_i\left(0, x^0\right) - y_i\left(0, \xi^0\right)\|^2.$$

Proof.

Let $x(t)$ and $\xi(t)$ be the solutions of the first equation of system (5.21) with x^0 and ξ^0 as initial values. We denote the controls by $u(t)$.
As in the proof of Lemma 33 we have:

$$b(\xi, u) - b(x, u) = B(t)(\xi - x) \tag{5.24}$$

where $B(t) = (b_{i,j})_{(i,j)\in\{1,..,n\}}$ is a lower triangular matrix since $b(x, u)$ is a lower triangular vector field. Set $\varepsilon = x - \xi$. Then

$$\dot{\varepsilon} \quad = \quad [A(u) + B(t)]\,\varepsilon. \tag{5.25}$$

We consider the system formed by equation (5.25) and $C(u_k)\,\varepsilon_k = a_1(u_k)\,\varepsilon_{1,k}$ as output. Let us consider $\Psi(t)$, the resolvent (5.25), and the Gramm observability matrix

$$G_d = \sum_{i=0}^{i=d} \Psi(i\delta_t)' C_i' C_i \Psi(i\delta_t).$$

From the lower triangular structure of $B(t)$, the upper triangular structure of $A(u)$ and the form of the matrix $C_i = C(u(i\delta_t))$, we can deduce that G_d is invertible[12] when $d \geq n-1$. It

[12] $\psi_0(.)$ is the resolvent of system (5.25) (refer to Appendix A.1). The Gramm observability matrix can be written:

$$G_d = \begin{pmatrix} C_0 \\ C_1\Psi_0(k\delta_t) \\ ... \\ C_d\Psi_0(d\delta_t) \end{pmatrix}' \begin{pmatrix} C_0 \\ C_1\Psi_0(k\delta_t) \\ ... \\ C_d\Psi_0(d\delta_t) \end{pmatrix} = \phi'\phi.$$

Therefore G_d is invertible provided it is of rank n, which can be achieved for $d \geq n-1$.

is, therefore, also symmetric positive definite. We consider the same function as before (Cf. Lemma 33):

$$\Lambda : \; L^{\infty}_{[0,d]}\left(\mathbb{R}^{\frac{n(n+1)}{2}}\right) \times L^{\infty}_{[0,d]}\left(\mathbb{R}^{n_u}\right) \;\longrightarrow\; \mathbb{R}^{+} \; .$$

With the same reasoning as in Lemma 33, we deduce the existence of a scalar $\lambda_d^0 > 0$ such that $G_d \geq \lambda_d^0 \, Id$ for any u and any matrix $B(t)$ having the structure specified above. Therefore:

$$\begin{aligned}\sum_{i=0}^{i=d} \left\| y_i\left(0,x^0\right) - y_i\left(0,\xi^0\right)\right\|^2 &= \left(x^0-\xi^0\right)' G_d \left(x^0-\xi^0\right) \\ &\geq \lambda_d^0 \left\|x^0-\xi^0\right\|^2 .\end{aligned}$$

■

5.2.5 Preparation for the Proof

In order to establish the preliminary inequalities needed in the proof, we first recall the *matrix inversion lemma*:

Lemma 64 ***(matrix inversion lemma [67], Section*** 0.7.4***)***

If M is symmetric positive definite, and $\lambda > 0$ then

$$(M + \lambda MC'CM)^{-1} = M^{-1} - C'(\lambda^{-1} + CMC')^{-1}C.$$

The estimation error is denoted $\epsilon(t) = z(t) - x(t)$, and we consider the change of variables:

- $\tilde{x} = \Delta x$, $\tilde{z} = \Delta z$ and $\tilde{\epsilon} = \Delta\epsilon$,
- $\tilde{b}(.,u) = \Delta b(\Delta^{-1}.,u)$, and $\tilde{S} = \Delta^{-1} S \Delta^{-1}$,
- $\tilde{b}^*(.,u) = \Delta b^*(\Delta^{-1}.,u)\Delta^{-1}$.

As before, the Lipschitz constant of the vector field remains the same in the new system of coordinates (Cf. Lemma 51).
The error dynamics are given by

$$\dot{\epsilon} = A(u)\epsilon + (b(z,u) - b(x,u)) \tag{5.26}$$

in the continuous case, and

$$\epsilon_k(+) = \epsilon_k(-) - \delta_t S^{-1}(+) C' r_\theta^{-1} C \epsilon_k(-) \tag{5.27}$$

in the discrete case.
We highlight the relations below as they are useful for the following computations:

$$\begin{aligned}\Delta A(u) &= \theta A(u) \Delta, \\ \Delta^{-1} A'(u) &= \theta A'(u) \Delta^{-1},\end{aligned} \tag{5.28}$$

where $N = diag\left(\{0,1,\ldots,n-1\}\right)$.

Remark 65

Notice that on intervals of the form $[k\delta_t, (k+1)\delta_t[$, the derivative of θ is equal to zero.

For $t \in [k\delta_t; (k+1)\delta_t[$**,**

$$\frac{d\tilde{\epsilon}}{dt} = \theta\left[A(u)\tilde{\epsilon} + \frac{1}{\theta}\left(\tilde{b}(\tilde{z},u) - b(\tilde{x},u)\right)\right], \tag{5.29}$$

and

$$\frac{d\tilde{S}}{dt} = \theta\left[-\left(A(u) + \tfrac{1}{\theta}\tilde{b}^*(z,u)\right)'\tilde{S} - \tilde{S}\left(A(u) + \tfrac{1}{\theta}\tilde{b}^*(z,u)\right) - \tilde{S}Q\tilde{S}\right]. \tag{5.30}$$

We now consider the Lyapunov function $\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}$ and use identities (5.29, 5.30) to obtain the equality below:

$$\frac{d\left(\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}\right)}{dt} = \theta\left[\tfrac{2}{\theta}\tilde{\epsilon}'\tilde{S}\left(\tilde{b}(\tilde{z},u) - \tilde{b}(\tilde{x},u) - \tilde{b}^*(\tilde{z},u)\tilde{\epsilon}\right) - \tilde{\epsilon}'\tilde{S}Q\tilde{S}\tilde{\epsilon}\right]. \tag{5.31}$$

Similarly, at time $k\delta_t$**,**

$$\tilde{\epsilon}_k(+) = \left(Id - \theta\delta_t\tilde{S}_k^{-1}(+)C'r^{-1}C\right)\tilde{\epsilon}_k(-), \tag{5.32}$$

and,

$$\tilde{S}_k(+) = \tilde{S}_k(-) + \theta\delta_t C'r^{-1}C. \tag{5.33}$$

As we did for the differential equations, we use (5.32) and (5.33) to compute the Lyapunov function at time $k\delta_t$:

$$\begin{aligned}\left(\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}\right)_k(+) &= \tilde{\epsilon}_k'(-)\left(Id - \theta\delta_t\tilde{S}_k^{-1}(+)C'r^{-1}C\right)'\tilde{S}_k(+) \\ &\qquad \times\left(Id - \theta\delta_t\tilde{S}_k^{-1}(+)C'r^{-1}C\right)\tilde{\epsilon}_k(-) \\ &= \tilde{\epsilon}_k'(-)\Big[\tilde{S}_k(+) - 2\theta\delta_t C'r^{-1}C \\ &\qquad +(\theta\delta_t)^2C'r^{-1}C\tilde{S}_k(+)^{-1}C'r^{-1}C\Big]\tilde{\epsilon}_k(-).\end{aligned} \tag{5.34}$$

From (5.33), we replace $\left(\theta\delta_t C'r^{-1}C\right)$ with $\left(\tilde{S}_k(+) - \tilde{S}_k(-)\right)$:

$$\begin{aligned}\left(\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}\right)_k(+) &= \tilde{\epsilon}_k'(-)\left[\tilde{S}_k(-)\tilde{S}_k(+)^{-1}\tilde{S}_k(-)\right]\tilde{\epsilon}_k(-) \\ &= \tilde{\epsilon}_k'(-)\left[\tilde{S}_k(-)^{-1}\tilde{S}_k(+)\tilde{S}_k(-)^{-1}\right]^{-1}\tilde{\epsilon}_k(-).\end{aligned} \tag{5.35}$$

From equation (5.33) we write:

$$S_k^{-1}(-)S_k(+)S_k^{-1}(-) = S_k^{-1}(-) + \frac{\theta\delta_t}{r}S_k^{-1}(-)C'CS_k^{-1}(-), \tag{5.36}$$

and compute $[S_k^{-1}(-)S_k(+)S_k^{-1}(-)]^{-1}$ by using Lemma 64 with $\lambda = \frac{\theta\delta_t}{r}$ and $M = \tilde{S}_k^{-1}(-)$. This results in:

$$\begin{aligned}\left(\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}\right)_k(+) &= \tilde{\epsilon}'\left[\tilde{S}_k(-) - C'(\tfrac{r}{\theta\delta_t} + C\tilde{S}_k^{-1}(-)C')^{-1}C\right]\tilde{\epsilon} \\ &= \left(\tilde{\epsilon}'\tilde{S}\tilde{\epsilon}\right)_k(-) - \tilde{\epsilon}_k'(-)\left[C'(\tfrac{r}{\theta\delta_t} + C\tilde{S}_k^{-1}(-)C')^{-1}C\right]\tilde{\epsilon}_k(-).\end{aligned} \tag{5.37}$$

As in Section 3.5, we need to write the prediction-correction Riccati equations in a different time scale (τ), so that we can bound the Riccati matrix independently from θ. We consider $d\tau = \theta(t)dt$ or equivalently $\tau = \int_0^t \theta(v)dv$ and keep the notation $\bar{x}(\tau) = \tilde{x}(t)$.

$$\left\{ \begin{array}{rcl} \frac{d\bar{S}}{d\tau} & = & -\left(A(\overline{u}) + \frac{\tilde{b}^*(\overline{z},\overline{u})}{\bar{\theta}}\right)' \overline{S} - \overline{S}\left(A(\overline{u}) + \frac{\tilde{b}^*(\overline{z},\overline{u})}{\bar{\theta}}\right) - \overline{S}Q\overline{S} \\ \bar{S}_k(+) & = & \bar{S}_k(-) + \theta\delta_t C' r^{-1} C. \end{array} \right. \tag{5.38}$$

Since $\theta(t)$ varies within an interval of the form $[1, \theta_{max}]$, the instants $t_k = k\delta_t$, $k \in \mathbb{N}$ are difficult to track in the τ time scale. For convenience, $t_k = k\delta_t$ is denoted τ_k in the τ time scale.

With the help of this representation we are able to derive the following lemma:

Lemma 66

Let us consider the prediction correction Riccati equations (5.38), and the assumptions:

- *the functions* $a_i(u(t))$, $\left|\tilde{b}^*_{i,j}(\overline{z},\overline{u})\right|$, *are smaller than* $a_M > 0$,
- $a_i(u(t)) \geq a_m > 0$, $i = 2, ..., n$,
- $\theta(0) = 1$, *and*
- $S(0)$ *is a symmetric positive definite matrix taken from a compact subset of the form* $aId \leq S(0) \leq bId$.

Then, there exists a constant μ, *and two scalars* $0 < \alpha < \beta$, *such that, if* $\theta(t)\delta_t \leq \mu$ *for all* $t \geq 0$,

$$\alpha Id \leq \bar{S}(\tau) \leq \beta Id$$

for all $k \in \mathbb{N}$, *for all* $\tau \in [\tau_k, \tau_{k+1}]$ *(this notations includes* $\bar{S}$ *before and / or after the correction step).*
Here, α *and* β *are independent from* δ_t *and* $\theta(t)$.
Since this relation is valid for all time, τ, *is is also true in the time scale* t.

Proof.

The proof of this lemma is quite long and technical. In order to facilitate the reading of this section, we detailed the proof in Appendix B.1.

■

5.2.6 Proof of the Theorem

First of all, consider $T^* > 0$, and ε^* as in Theorem 62.
Let us now set a time T such that $0 < T < T^*$.
Let α and β be the bounds of Lemma 66.
For $t \in [k\delta_t; (k+1)\delta_t[$, inequality (5.31) can be written (i.e. using $\alpha Id \leq \tilde{S}$),

$$\frac{d\tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \theta \tilde{\varepsilon}'\tilde{S}\tilde{\varepsilon}(t) + 2\tilde{\varepsilon}'\tilde{S}\left(\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z})\tilde{\varepsilon}\right) \tag{5.39}$$

with $q_m > 0$ such that $q_m Id < Q$ (and omitting to write the control variable u).

From (5.39) we can deduce two bounds: the first bound, the local bound, will be useful when $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t)$ is small independent of the value of θ. The second bound, the global bound, will be useful mainly when $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t)$ is not in the neighborhood of 0.

Global bound: Starting from:

$$\left\| \tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z}) \tilde{\varepsilon} \right\| \leq 2L_b \|\tilde{\varepsilon}\|,$$

together with $\alpha Id \leq \tilde{S} \leq \beta Id$ (Lemma 66), (5.39) becomes

$$\frac{d\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t)}{dt} \leq \left(-\alpha q_m \theta + 4 \frac{\beta}{\alpha} L_b \right) \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t). \tag{5.40}$$

Local bound: Using Lemma 56

$$\left\| \tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x}) - \tilde{b}^*(\tilde{z}) \tilde{\varepsilon} \right\| \leq K \theta^{n-1} \|\tilde{\varepsilon}\|^2,$$

which because $1 \leq \theta \leq 2\theta_1$, implies that

$$\frac{d\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t)}{dt} \leq -\alpha q_m \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) + 2K (2\theta_1)^{n-1} \left\| \tilde{S} \right\| \|\tilde{\varepsilon}\|^3.$$

The fact that $\|\tilde{\varepsilon}\|^3 = \left(\|\tilde{\varepsilon}\|^2 \right)^{\frac{3}{2}} \leq \left(\frac{1}{\alpha} \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) \right)^{\frac{3}{2}}$, allows us to write

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) \leq -\alpha q_m \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) + \frac{2K (2\theta_1)^{n-1} \beta}{\alpha^{\frac{3}{2}}} \left(\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) \right)^{\frac{3}{2}}. \tag{5.41}$$

Let us apply Lemma 55: If there exists ξ such that

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(\xi) \leq \frac{\alpha^5 q_m^2}{16 K^2 (2\theta_1)^{2n-2} \beta^2},$$

then for any $k\delta_t \leq \xi \leq t \leq (k+1)\delta_t$

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) \leq 4\tilde{\epsilon}' \tilde{S} \tilde{\varepsilon}(\xi) \, \mathrm{e}^{-\alpha q_m (t-\xi)}.$$

If $\gamma \in \mathbb{R}$ such that

$$\gamma \leq \frac{1}{(2\theta_1)^{2n-2}} \min \left(\frac{\alpha \varepsilon^*}{4\beta}, \frac{\alpha^5 q_m^2}{16 K^2 \beta^2} \right), \tag{5.42}$$

then $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(\xi) \leq \gamma$ implies

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(t) \leq \frac{\alpha \varepsilon^*}{\beta (2\theta_1)^{2n-2}} \mathrm{e}^{-\alpha q_m (t-\xi)}. \tag{5.43}$$

Given any arbitrary value of δ_t, there exists $k_T \in \mathbb{N}$ such that $T \in [k_T \delta_t; (k_T+1)\delta_t[$. From the global bound (5.40), with $\theta_k \geq 1$, for all $k \in \mathbb{N}$:

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(T) \leq \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(k_T \delta_t) \, \mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha} L_b\right)(T - k_T \delta_t)}.$$

When we consider $t \in [k\delta_t, (k+1)\delta_t[$, this requirement means that $\left(\tilde{\epsilon}' \tilde{S} \tilde{\epsilon}\right)(k\delta_t) = \left(\tilde{\epsilon}' \tilde{S} \tilde{\epsilon}\right)_k (+)$. We know from (5.37) that in full generality

$$\left(\tilde{\epsilon}' \tilde{S} \tilde{\epsilon}\right)_k (+) \leq \left(\tilde{\epsilon}' \tilde{S} \tilde{\epsilon}\right)_k (-). \tag{5.44}$$

Thus

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(T) \leq \left(\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\right)_{k_T} (-) \mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha} L_b\right)(T - k_T \delta_t)},$$

and since $\left(\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\right)_{k_T} (-)$ is the end value of the equation (5.31) for $t \in [(k_T - 1)\delta_t; k_T \delta_t[$, then:

$$\left(\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\right)_{k_T} (-) \leq \left(\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\right)_{k_T - 1} (+) \mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha} L_b\right)\delta_t}.$$

We can therefore, iteratively, independently from δ_t, obtain the inequality:

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(T) \leq \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(0)\, \mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha} L_b\right)T}. \tag{5.45}$$

We now suppose that $\theta \geq \theta_1$ for $t \in [T, T^*]$, $T^* \in [\tilde{k}\delta_t, (\tilde{k}+1)\delta_t[$ and use (5.40):

$$\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(T^*) \leq \tilde{\epsilon}' \tilde{S} \tilde{\varepsilon} \left(\tilde{k}\delta_t\right) \mathrm{e}^{\left(-\alpha q_m \theta_1 + 4\frac{\beta}{\alpha} L_b\right)(T^* - \tilde{k}\delta_t)}. \tag{5.46}$$

As before, independent of δ_t, we obtain:

$$\begin{aligned} \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(T^*) &\leq \tilde{\epsilon}' \tilde{S} \tilde{\varepsilon}\,(T)\, \mathrm{e}^{\left(-\alpha q_m \theta_1 + 4\frac{\beta}{\alpha} L_b\right)(T^* - T)}, \\ &\leq \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}\,(0)\, \mathrm{e}^{\left(-\alpha q_m + 4\frac{\beta}{\alpha} L_b\right)T} \mathrm{e}^{\left(-\alpha q_m \theta_1 + 4\frac{\beta}{\alpha} L_b\right)(T^* - T)}, \\ &\leq M_0 \mathrm{e}^{-\alpha q_m T} \mathrm{e}^{4\frac{\beta}{\alpha} L_b T^*} \mathrm{e}^{-\alpha q_m \theta_1 (T^* - T)}, \end{aligned} \tag{5.47}$$

where $M_0 = \sup_{x,z \in \chi} \varepsilon' S \varepsilon\,(0)$.

Now, we choose θ_1 and γ such that

$$M_0 \mathrm{e}^{-\alpha q_m T} \mathrm{e}^{4\frac{\beta}{\alpha} L_b T^*} \mathrm{e}^{-\alpha q_m \theta_1 (T^* - T)} \leq \gamma \tag{5.48}$$

and (5.42) are satisfied simultaneously, which results because $\mathrm{e}^{-\mathrm{cte} \times \theta_1} < \frac{\mathrm{cte}}{\theta_1^{2n-2}}$ for θ_1 sufficiently large.
We check that the condition[13] $2\theta_1 \delta_t < \mu$ is satisfied, and if necessary δ_t shortened (up to now, all the parameters we use now do not depend on δ_t).
We set $d \in \mathbb{N}^*$ in order to satisfy the conditions $0 < d\delta_t < T < T^*$, and $d \geq n - 1$. We still have the freedom to shorten δ_t again if necessary.
Because innovation is defined, then so is the parameter λ_d^0 of Lemma 63.

We now need to design an adaptation function $\mathcal{F}$. Since the sample time δ_t is now fixed, we can compute and set k_T such that $T \in [k_T \delta_t, (k_T + 1)\delta_t[$. In order to have $\theta(T) > \theta_1$ we must have $\theta_{k_T} > \theta_1$. The adaptation function must have the following features:

[13] We arbitrarily set $\theta_{max} = 2\theta_1$.

- θ is such that $1 \leq \theta_k \leq 2\theta_1$, for all $k \in \mathbb{N}$,
- Set $\gamma_1 = \frac{\lambda_d^0 \gamma}{\beta}$, and $0 \leq \gamma_0 \leq \gamma_1$:
  - suppose, for any arbitrary $j \in \mathbb{N}$, $I_{d,k} > \frac{\lambda_d^0 \gamma}{\beta}$ in the time interval $[j\delta_t, j\delta_t + (T - d\delta_t)]$ then $\theta\left(j\delta_t + (k_T - d)\delta_t\right)) > \theta_1$.
  - when $\mathcal{I}_d, k <\gamma_0$, then θ_k decreases to 1.

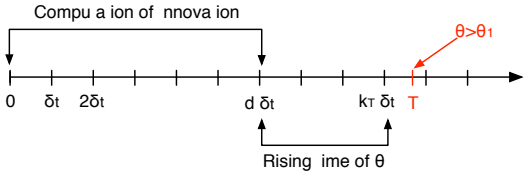

Figure 5.1: It can happen that $k_T = d$. The rising time is then zero.

Remark 67

1. *Figure 5.1 displays a clear representation of the situation with regards to the influence of the subdivision on the rising time.*
2. *The most basic choice for such a function is a switch*
   - *from* 1 *to* $2\theta_1$ *when* $\mathcal{I}_d, k \geq \gamma_1$, *and*
   - *from* $2\theta_1$ *to* 1 *when* $\mathcal{I}_d, k \leq \gamma_0 = \gamma_1$.

We claim that there exists $\xi \leq T^*$ such that $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(\tau) \leq \gamma$.

We recall the previous notation $T^* \in [\tilde{k}\delta_t, (\tilde{k}+1)\delta_t[$. If $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(\xi) > \gamma$ for all $\xi \leq T^*$ then $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(k\delta_t) > \gamma$ for all $k \in \{0, ..., \tilde{k}\}$, therefore

$$\gamma < \tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(k\delta_t) \leq \beta \left\|\tilde{\varepsilon}(k\delta_t)\right\|^2 \leq \beta \left\|\varepsilon(k\delta_t)\right\|^2 \leq \frac{\beta}{\lambda_d^0} \mathcal{I}_d(k\delta_t + d\delta_t),$$

because of Lemma 63.
This means that, $\mathcal{I}_{k+d,d} \geq \gamma_1$ for all $k \in \{0, ..., \tilde{k}\}$, hence $\mathcal{I}_{k,d} \geq \gamma_1$ for all $k \in \{d, ..., \tilde{k}\}$. Therefore, $\theta_k \geq \theta_1$ for $t \in [T, T^*]$, since the function $\mathcal{F}$ has been designed for that purpose. We obtain the contradiction: $\tilde{\varepsilon}' \tilde{S} \tilde{\varepsilon}(T^*) \leq \gamma$, because (5.47) and (5.48).

Finally, for $t \geq \xi$, using (5.43) and (5.42):

$$\begin{aligned} \|\epsilon(t)\|^2 &\leq (2\theta_1)^{2n-2} \|\tilde{\epsilon}(t)\|^2 \\ &\leq \frac{(2\theta_1)^{2n-2}}{\alpha} \tilde{\epsilon}' \tilde{S} \tilde{\epsilon}(t) \\ &\leq 4\frac{(2\theta_1)^{2n-2}}{\alpha} \tilde{\epsilon}' \tilde{S} \tilde{\epsilon}(\xi) e^{-\alpha q_m (t-\xi)} \\ &\leq \epsilon^* e^{-\alpha q_m (t-\xi)} \\ &\leq \epsilon^* e^{-\alpha q_m (t-T^*)}. \end{aligned} \tag{5.49}$$

Remark 68

As for the continuous case, the convergence proof can be extended to the multiple output case following the instructions of Section 5.1. The only difference lies in the usage of Lemma 64 with equation (5.36), where we cannot write $\frac{\theta\delta_t}{r}$ *since* R *is no longer a scalar. Since* R *is definite positive, we can assume the existence of a scalar* $\tilde{r}$ *such that* $R > \tilde{r}Id$*. Then (5.36) can be written:*

$$S_k^{-1}(-)S_k(+)S_k^{-1}(-) \leq S_k^{-1}(-) + \frac{\theta\delta_t}{\tilde{r}} S_k^{-1}(-)C'CS_k^{-1}(-),$$

and Lemma 64 can be used.

Chapter 6

Conclusion and Perspectives

The work described in this thesis deals with the design of an observer of the Kalman type for nonlinear systems.

More precisely, we considered the high-gain formalism and proposed an improvement of the *high-gain extended Kalman filter* in the form of an adaptive scheme for the parameter at the heart of the method. Indeed, although the high-gain approach allows us to analytically prove the convergence of the algorithm in the deterministic setting, it comes with an increased sensitivity to measurement noise. We propose to let the observer evolve between two end point configurations, one that rejects noise and one that makes the estimate converge toward the real trajectory. The strategy we developed here allowed us to analytically prove this convergence.

Observability theory constitutes the framework of the present study. Thus, we began this thesis by providing a review and some insight into the main results of the theory of [57]. We also provided a review of similar adaptive strategies. In this introduction and background review, we stated that the main concern of the thesis would be theoretically proving that the observer is convergent.

The observer has been described in Chapters 3 and 5. It was initially described in the continuous setting, and afterwards extended to the continuous-discrete setting. The adaptive strategy was also explained in those chapters. This strategy is composed of two elements:

1. a measurement of the quality of the estimation, and
2. an adaptation equation.

The quality measurement is called innovation or *innovation for an horizon of length d*. It is slightly different than the usual concept of innovation. The major improvement provided by our definition is a proof that shows that innovation places an upper bound on the past estimation error (the delay equals the parameter d above mentioned). This fact is a corner stone of the overall convergence proof.

The second element of the strategy is the adaptation equation that drives the high-gain parameter. A differential equation was used in the continuous setting, and a function in the continuous-discrete setting (i.e. the adaptation is performed at the end of the update procedure). The sets of requirements for those two applications have been proposed such that several adaptation functions can be conceived. The set of possible applications isn't void, as we demonstrated by actually displaying an eligible function.

The full proof of convergence has been developed in the single output continuous setting, and afterwards extended to the multiple output and continuous-discrete settings.

A second major concern of this work, was the applicability of the observer. We therefore extensively described its implementation on a single output system: a series-connected DC machine. The time constraints where investigated via experiments performed using a real motor in a hard real-time environment. The testbed was described in Chapter 4 and the compatibility with real-time constraints assessed.

We conclude this work with some ideas for future investigations.

The Luenberger Case

It is much more direct to prove the convergence of a high-gain Luenberger observer. This is because of the absence of the Riccati equation. However, it prevents us from providing a local result of convergence when $\theta = 1$. Therefore the adaptation strategy has to be different from that used here (Cf. [11]), or performed for a specific class of nonlinear systems (see for example [19]).

Automatic code generation

The implementation procedure for this algorithm is now well known. It can be roughly classified into two parts: 1) coding specific to the model, and 2) coding related to the observer mechanisms. It would be interesting to create a utility that automatically generates the code of the observer once the model has been provided. We could save development time, and implementation errors cause by typos.

Dynamic output stabilization

Dynamic output stabilization is considered in the second part of [57]. An extension of this work to a closed loop containing an adaptive observer is a natural development. This may be accomplished because the observer presented here is an exponential observer. Since θ is allowed to increase when convergence is not achieved, we can expect to deliver a good estimate to the control algorithm. The ability to quickly switch between modes will be important.

Cascaded systems

Let us consider an observable nonlinear cascaded system of the form:

$$\left\{ \begin{array}{rcl} \dot{x} & = & f(x,u), \\ \dot{\xi} & = & g(x,\xi), \\ y & = & h(x,\xi). \end{array} \right.$$

One could imagine a situation where the state variable x is well known or estimated, but not the variable ξ. Does the θ parameter of the observer really need to be high for the part of the estimation that corresponds to x? We could consider a high-gain observer with two varying high-gain parameters.

Unscented Kalman filter

The unscented Kalman filter is a derivative free, nonlinear observer that has received a lot of attention recently [71]. This observer is based on the *unscented transformation*:

a method to calculate the statistics of a random variable which undergoes a nonlinear transformation [112]. Continuous and continuous-discrete versions of this observer have been proposed in [108]. The idea is to study, the extent to which the high-gain formalism, and consequently adaptive high-gain mechanisms can be embedded into this observer.

Appendices

Appendix A

Mathematics Reminder

Contents

A.1 Resolvent of a System

In this subsection, we recall basic concepts from the theory of linear differential equations. Details can be found in [106].

A first order system of linear differential equations is given by:

$$\frac{dx}{dt} = A(t)x(t) + b(t) \tag{A.1}$$

where

1. $t \in \mathbb{I}$, an interval of $\mathbb{R}$,
2. $x \in \mathbb{R}^n$,
3. $A(t)$ is a t dependent matrix of dimension $(n \times n)$,
4. $b(t)$ is a t dependent vector field of dimension n.

First, remind, that provided that the applications $A(t)$ and $b(t)$ are continuous then for all $s \in \mathbb{I}$ and for all $x_0 \in \mathbb{R}^n$, this equation has a unique solution on $\mathbb{I}$ such that $x(s) = x_0$.

We now consider the associated homogenous equation:

$$\frac{dx}{dt} = A(t)x(t). \tag{A.2}$$

Let us denote by $x(t, s, x_0)$ the solution of (A.2) at time t with initial condition $x(s, s, x_0) = x_0$. We define the application:

$$\xi \mapsto x(t, s, \xi), \tag{A.3}$$

that associates to any element ξ of $\mathbb{R}^n$ the solution of (A.2) starting from ξ.
Let c_1, c_2 be two positive scalars and ξ_1, ξ_2 two elements of $\mathbb{R}^n$. Consider the trajectory $c_1 x(t, s, \xi_1) + c_2 x(t, s, \xi_2)$. For $t = s$, we have:

$$c_1 x(s, s, \xi_1) + c_2 x(s, s, \xi_2) = c_1\xi_1 + c_2\xi_2 = x(s, s, c_1\xi_1 + c_2\xi_2).$$

From the unicity of solutions, we conclude that

$$c_1 x(t, s, \xi_1) + c_2 x(t, s, \xi_2) = x(t, s, c_1\xi_1 + c_2\xi_2).$$

That is to say that for all t and s in $\mathbb{I}$, (A.3) is linear. Therefore there is a t and s dependent matrix such that:

$$x(t, s, x_0) = \phi(t, s)x_0,$$

with $\phi(s, s) = Id$. This matrix is called the resolvent[1] of (A.2). The resolvent has the following properties:

Theorem 69

1. *$\phi(t, s)$ is linear w.r.t. the variables s and t,*

[1] $\phi(s, s) = Id$ is also said to be the resolvent of the equation (A.1).

2. $\phi(s,s) = Id$,
3. *for all* $\tau \in \mathbb{I}$, $\phi(t,\tau)\phi(\tau,s) = \phi(t,s)$,
4. $\phi^{-1}(t,s) = \phi(s,t)$,
5. $\frac{\partial}{dt}\phi(t,s) = A(t)\phi(t,s)$ *and* $\frac{\partial}{ds}\phi(t,s) = -\phi(t,s)A(s)$,
6. *in particular:* $\frac{\partial}{dt}\phi^{-1}(t,s)' = -A'(s)\phi^{-1}(t,s)'$,
7. *the solution of (A.2) at time* t *with,* $x(s) = x_0$ *is given by:* $x(t) = \phi(t,s)x_0$,
8. *the solution of (A.1) at time* t *with,* $x(s) = x_0$ *is given by:*

$$x(t) = \phi(t,s)x_0 + \int_s^t \phi(t,v)b(v)dv.$$

A.2 Weak-∗ Topology

Let E be a Banach space[2] and denote its norm by $\|.\|_E$. E^* is the dual vector space[3] of E, and E^{**} is the dual vector space of E^*: the bi-dual space of E. A dual space is embedded with the dual norm defined as:

$$\|f\|_{E^*} = \sup_{x\in E, \|x\|\leq 1} |f(x)|.$$

Therefore E^* has the topology induced by the dual norm. The object of this section is to give the definition of two additional topologies that can be built on the space E^*: the *weak* and the *weak-∗* topologies. The topology described above, induced by the dual norm, is also called *strong topology*.

For $f \in E^*$ and $x \in E$ we denote $< f, x >$ instead of $f(x)$. It is called the *dual scalar product*.

Topology Associated to a Family of Applications

We consider a set X and a family of topological spaces $(Y_i)_{i\in\mathbf{I}}$. For all $i \in \mathbf{I}$, let φ_i denote an application from X to Y_i. We want to embed X with the coarser topology that makes all the applications φ_i continuous. In the following this topology is denoted $\mathcal{T}$.

Let i be an element of $\mathbf{I}$. Let ω_i be an open subset of Y_i. If the application φ_i is continuous then $\varphi_i^{-1}(\omega_i)$ is an open subset of X. Consequently $\mathcal{T}$ shall contain the family of subsets of X defined by $\varphi_i^{-1}(\omega_i)$ when ω_i scans all the open subsets of Y_i. This remark applies to all $i \in \mathbf{I}$. We denote $(\mathcal{O}_\lambda)_{\lambda\in\Lambda}$ the family composed by all the subsets of the form $\varphi_i^{-1}(\omega_i)$, where ω_i is an open subset of Y_i, and $i \in \mathbf{I}$.

[2] i.e. A normed vector space, complete with respect to the topology of its norm.

[3] The space of all the continuous linear forms on E.

The problem *"find the coarser*[4] *topology that makes the applications φ_i continuous"* is transformed into: *"find the coarser family of subsets of X that contains $(\mathcal{O}_\lambda)_{\lambda\in\Lambda}$ and that is stable under any finite intersection and any infinite union of its elements"*[5].

Consider first the family of all the subsets of X obtained as the intersection of a finite number of elements of the form $\mathcal{O}_\lambda$, $\lambda \in \Lambda$. It is a family of subsets of X stable under finite intersections.

Secondly, consider all the infinite unions of elements of this latter family[6]. We finally end up with the family

$$\tilde{\mathcal{T}} = \left\{ \bigcup_{infinite} \left[\bigcap_{finite} \mathcal{O}_\lambda \right], \lambda \in \Lambda \right\}.$$

$\tilde{\mathcal{T}}$ is a topology: stability under infinite union is obvious, and the proof of the stability under finite intersection is left as an exercise of sets theory.

Consider now any topology $\hat{\mathcal{T}}$ of X that makes all the applications $\varphi_i : X \to Y_i$, $i \in \mathbf{I}$ continuous. Then all the subsets of the form $\varphi_i^{-1}(\omega_i)$, ω_i being an open subset of Y_i, $i \in \mathbf{I}$, are contained in $\hat{\mathcal{T}}$. And since $\hat{\mathcal{T}}$ is a topology, it contains all the elements of $\tilde{\mathcal{T}}$. Thus any topology that makes all $(\varphi_i)_{i\in\mathbf{I}}$ continuous contains $\tilde{\mathcal{T}}$: it is the coarser topology we are searching for.

Definition of the Weak Topology

For a fixed $f \in E^*$, we define the application $\varphi_f : E \to \mathbb{R}$ by $\varphi_f(x) =< f, x >$. When f describes the set E^*, we have a family of applications $(\varphi_f)_{f\in E^*}$.

Definition 70

*The **weak topology** is the coarser topology that renders all the applications $(\varphi_f)_{f\in E^*}$ continuous.*

A weak topology is built on the dual space E^* by considering the applications $(\varphi_\xi)_{\xi\in E^{**}}$.

Definition of the Weak-$*$ Topology

We define a canonical injection from E to E^{**}:

- let $x \in E$ be fixed
- $\begin{array}{ccc} E^* & \to & \mathbb{R} \\ f & \mapsto & < f, x > \end{array}$ is a continuous linear form on E^*, that is to say an element of E^{**}, denoted Jx.
- we have:
$$< Jx, f >_{(E^{**},E^*)} =< f, x >_{(E^*,E)}, \forall x \in E,\ \forall f \in E^*.$$
The application $x \mapsto Jx$ is a linear injection.

[4] The coarser topology is the one that contains "the least" number of open sets.

[5] Obviously $\emptyset$ and X are contained in $(\mathcal{O}_\lambda)_{\lambda\in\Lambda}$. Therefore the stability under infinite unions, and the stability under finite intersections define a topology.

[6] If we perform the infinite union first and then the finite intersection, we obtain a family that is not stable under infinite union anymore.

- Let x and y be two elements of E, $J(x+y)$ is the associated element of E^{**}. For all $f \in E^*$: $< J(x+y), f >=< x+y, f >=< x, f > + < y, f >$. Thus $J(x+y) \equiv Jx + Jy$.
- If Jx is an identically null application, then $< f, x >= 0 \ \forall f \in E^*$, which means that $x = 0$. The kernel is composed of the null element.

It may happen that J is not a surjective application: E is therefore identified to a subspace of E^{**}.

We now consider a fixed $x \in E$ and the application $\varphi_x : E^* \to \mathbb{R}$ defined as $f \mapsto \varphi_x(f) =< f, x >$. Since x can be any elements of E, we obtain a family of applications: $(\varphi_x)_{x \in E}$.

Definition 71

*The **weak-* topology** is the coarser topology that renders all the applications $(\varphi_x)_{x \in E}$ continuous.*

The weak-$*$ topology makes continuous the family of applications $\mathcal{F}_1 = (\varphi_x)_{x \in E}$ and the weak topology makes continuous the family of applications $\mathcal{F}_2 = (\varphi_\xi)_{\xi \in E^{**}}$. Since E can be seen as a subset of E^{**}, the family $\mathcal{F}_1$ can be seen as a subfamily of $\mathcal{F}_2$. Therefore the topology constructed using the family $\mathcal{F}_1$ is coarser than the one obtained from the family $\mathcal{F}_2$.

More details concerning those topologies can be found in [33] (in Moliere's mother tongue) or [107] (in Shakespear's mother tongue).

A.3 Uniform Continuity of the Resolvant

In the proof of some lemmas[7] we use the fact that *"the weak-* topology on a bounded set implies uniform continuity of the resolvent"*. We explain how it works in the present subsection. We consider a control affine system of the form:

$$(\Sigma) \begin{cases} \dot{x} &= f(x) + \sum_{i=1}^{i=p} g_i(x) u_i, \\ x(0) &= x_0. \end{cases}$$

Let $P_\Sigma : Dom\left(P_\Sigma\right) \subset L^\infty\left([0;T], \mathbb{R}^{n_u}\right) \to C^0\left([0;T], X\right)$, be the *"input to state"* mapping ofΣ , i.e., the mapping that, to any $u(.)$, associates the corresponding state trajectory $(t \in [0;T] \mapsto x(t))$.
We endow $L^\infty\left([0;T], \mathbb{R}^{n_u}\right)$ with the weak-* topology, and $C^0\left([0;T], X\right)$ with the topology of uniform convergence. The following lemma is proven in [57].

Lemma 72

P_Σ has open domain and is continuous on bounded sets (w.r.t. the topologies above).

The proof of this lemma is reproduced below (refer to [57] for the complete statement). In order to ease the understanding of this proof we first recall a few facts concerning the properties of the weak-* topology.

[7] Lemmas 33 and 38, and their equivalent in Chapter 5.

The weak-* topology on L^∞ sets

According to what is said in Section A.2, we can embed L^∞ with either the weak or the weak-$*$ topology. We choose the weak-$*$ topology, the coarser one[8]. The following theorem, stated with the same notations as before, gives us an important property of this topology:

Theorem 73 ***([33], III-25, Pg 48)***

Let E be a separable Banach space, then $B_{E^*} = \{f \in E^*; \|f\| \leq 1\}$ is metrizable for the weak-$*$ topology[9]*.*

Conversely, if $B_{E^*} = \{f \in E^*; \|f\| \leq 1\}$ is metrizable for the weak-$*$ topology then E is separable.

Now, recall that $(L^1)^* = L^\infty$ and $L^\infty \subset (L^1)^*$ [33, 107] and consider the three theorems below.

Theorem 74 ***([33])***

I - IV-7, Pg 57: *L^p is a normed vector space for $1 \leq p \leq \infty$.*

II - (Fischer-Riesz) - IV-8, Pg 57: *L^p is a Banach space for $1 \leq p \leq \infty$.*

III - IV-13, Pg 62: *L^p is separable*[10] *for $1 \leq p < \infty$.*

Therefore the closed unit ball (with respect to the strong metric) of L^∞ is metrizable for the weak-$*$ topology.

Let us consider any bounded subset of L^∞, denotedΩ $_1$. There exists $r < \infty$ such thatΩ is contained in a closed ball of radius r. This latter ball has a weak-* metric induced by the one of the unit ball B_{E^*}. Consequently, in order to prove Lemma 72, we only need to prove that the "input to state" mapping is sequentially continuous[11].

Proof of the lemma

We can assume that (1) the manifold X equals $\mathbb{R}^n$, and that (2) f and g_i's in Σ are compactly supported vector fields.

Let us fix u and consider a sequence (u_n) converging $*$-weakly to u. The corresponding trajectories of Σ are denoted x and x_n respectively. The proof is done by considering the case $n = 1$ for clarity of the computations only.

Lemma 75

There is a $k > 0$ such that, $\forall n$, $\forall t \in [0; T]$, we have

$$\|x_n(t) - x(t)\| \leq k \sup_{\theta \in [0;T]} \left\| \int_0^\theta (u_n(s) - u(s))\, G(x(s)) ds \right\|.$$

[8] The interest is that the coarser topology has the "largest amount" of compacts sets.

[9] This means that there exist a metric on B_{E^*} which open sets are the same than those induced by the weak-$*$ topology.

[10] Actually, L^∞ isn't separable.

[11] i) For an application between two metrizable spaces, continuity is equivalent to sequential continuity. ii) (Heine's theorem) Any continuous application between metric spaces is uniformly continuous on compact subsets.

Here $G(x(s))$ denotes a column vector field whose elements are the g_i's of Σ, and u denote the line vector composed of elements u_i.

Proof.

First of all, let us write:

$$\begin{aligned} x(t) &= x_0 + \int_0^t \left[f(x(s)) + \sum_{i=1}^{i=n_u} g_i(x(s))u_i(s) \right] ds \\ &= x_0 + \int_0^t \left[f(x(s)) + u(s)G(x(s)) \right] ds. \end{aligned}$$

Then

$$\begin{aligned} \|x_n(t) - x(t)\| &\leq \left\| \int_0^t \left[f(x_n(s)) + u_n(s)G(x_n(s)) - f(x(s)) - u(s)G(x(s)) \right] ds \right\| \\ &\leq \left\| \int_0^t \big[f(x_n(s)) + u_n(s)G(x_n(s)) + u_n(s)G(x_(s)) \right. \\ &\qquad \left. - u_n(s)G(x_(s)) - f(x(s)) - u(s)G(x(s)) \big] ds \right\| \\ &\leq \left\| \int_0^t \big[f(x_n(s)) + u_n(s)G(x_n(s)) - f(x(s)) - u_n(s)G(x_(s)) \big] ds \right\| \\ &\qquad + \left\| \int_0^t \big[(u_n(s) - u(s))G(x_(s)) \big] ds \right\| = A + B. \end{aligned}$$

The terms A and B are such that:

$$\begin{aligned} B &\leq \sup_{\theta \in [0;T]} \left\| \int_0^\theta (u_n(s) - u(s))G(x(s))ds \right\| \\ A &\leq \int_0^t \|f(x_n(s)) - f(x(s))\| \, ds + \int_0^t \|(G(x_n) - G(x)\| \, |u_n(s)| ds. \end{aligned}$$

Because u_n converges $*-$weakly, the sequence $(\|u_n\|_\infty : n \in \mathbb{N})$ is bounded. The Lipschitz properties of f and G give $A \leq m \int_0^t \|x_n(s) - x(s)\| \, ds$.

Therefore

$$\|x_n(t) - x(t)\| \leq \sup_{\theta \in [0;T]} \left\| \int_0^\theta (u_n(s) - u(s))G(x(s))ds \right\| + m \int_0^t \|x_n(s) - x(s)\| \, ds.$$

The result is given by Gronwall's lemma.

■

For all $\theta \in [0, T]$, for all $\delta > 0$, let us consider a subdivision $\{t_j\}$ of $[0, T]$, such that $\theta \in [t_i, t_{i+1}]$ and $t_{j+1} - t_j < \delta$ for all j. We have,

$$\left\| \int_0^\theta (u_n(s) - u(s))G(x(s))ds \right\| \leq \left\| \int_0^{t_i} (u_n(s) - u(s))G(x(s))ds \right\| + \left\| \int_{t_i}^\theta (u_n(s) - u(s))G(x(s))ds \right\|.$$

With $\varepsilon > 0$ being given, define $\varepsilon^* = \frac{\varepsilon}{k}$, where k comes from the preceding lemma. We take $\delta \leq \frac{\varepsilon^*}{2\gamma m}$, where $\gamma = \sup_{x \in \mathbb{R}^n} \|G(x)\|$, and $m = \sup_n \|u_n - u\|_\infty$. We get, for all $\theta \in [0, T]$:

$$\left\| \int_{t_i}^{\theta} (u_n(s) - u(s)) G(x(s)) ds \right\| \leq m\gamma\delta \leq \frac{\varepsilon^*}{2}$$

By the weak-* convergence of (u_n), there exists $N \in \mathbb{N}^*$ such that for all $n > N$,

$$\left\| \int_{0}^{t_i} (u_n(s) - u(s)) G(x(s)) ds \right\| \leq \frac{\varepsilon^*}{2}.$$

By the lemma, there exists $N \in \mathbb{N}^*$ such that for all $n > N$, for all $t \in [0, T]$, $\|x_n(t) - x(t)\| \leq \varepsilon$, which proves the sequential continuity.

A.4 Bounds on a Gramm Matrix

This section's material is taken from [57], Chapter 6, Section 2.4.2. We present a lemma useful to investigate the properties of the Riccati matrix of extended Kalman filters, both in the continuous and continuous discrete settings. This lemma is used in Appendix B.

Let Σ be a system[12] on $\mathbb{R}^n$, of the form

$$(\Sigma) \begin{cases} \frac{dx}{d\tau} &= A(u)x + \sum\limits_{\substack{k,l=1 \\ l \leq k}}^{n} u_{k,l} e_{k,l} x \\ y &= C(u)x, \end{cases} \tag{A.4}$$

where

- $A(u)$ is a an anti-shift matrix whose elements never equals zero, and are bounded,
- $C(u) = \begin{pmatrix} \alpha(u) & 0 & \dots & 0 \end{pmatrix}$, where α never equals zero and is bounded,
- $e_{i,j}$ is such that $e_{i,j} x_k = \delta_{jk} v_i$, where $\{v_k\}$ denotes the canonical basis of $\mathbb{R}^n$.

The term on the right of the " + " sign is a lower triangular $(n \times n)$ matrix. For $(u_{i,j})$, a measurable bounded control function, defined on $[0, T]$, we denote $\psi_u(t, s)$ the associated resolvent matrix (see Section A.1 above). We define the Gramm observability matrix of Σ by

$$G_u = \int_0^T \psi_u'(v, T) C' C \psi_u(v, T) dv.$$

The matrix G_u is symmetric and positive semi-definite. The system (A.4) is observable for all $u(.)$ measurable and bounded.

[12] In order to make this system easier to understand, we only describe it as single output. The result of the lemma is though valid for multi outputs systems.

Lemma 76

If a bound B is given on the controls $u_{i,j}$, then there exist two positive scalars $0 < \alpha < \beta$ depending on B and T only, such that

$$\alpha Id \leq G_u \leq \beta Id.$$

Proof.

The complete proof is decomposed into two parts:

1. Lemma 72 is used to prove the continuity of the map $u(.) \rightarrow G_u$,
2. The precompactness of the weak-* topology, and the observability of Σ are used to prove the lemma.

■

Appendix B

Proof of Lemmas

Contents

B.1 Bounds on the Riccati Equation

This section deals with the properties of the Riccati matrix S. We consider the continuous discrete framework. The proof follows the ideas of [57] where those properties are investigated in details for continuous time systems[1].

Lemma 77

Let us consider the prediction correction equations:

$$\left\{\begin{array}{rcl} \frac{d\bar{S}}{d\tau} & = & -\left(A(\bar{u})+\frac{\bar{b}(\bar{z},\bar{u})}{\theta}\right)' \bar{S} - \bar{S}\left(A(\bar{u})+\frac{\bar{b}(\bar{z},\bar{u})}{\theta}\right) - \bar{S}Q\bar{S} \\ \bar{S}_k(+) & = & \bar{S}_k(-) + \theta\delta_t C' R^{-1} C \end{array}\right. \tag{B.1}$$

with the notations of Section 5.2.1, and the set of assumptions:

- *Q and R are fixed symmetric positive definite matrices,*
- *the functions $a_i(u(t))$, $\left|\tilde{b}^*_{i,j}(z,u)\right|$, are smaller than $a_M > 0$,*
- *$a_i(u(t)) \geq a_m > 0$,*
- *$\theta(0) = 1$, and*
- *$S(0)$ is a symmetric positive definite matrix taken in a compact of the form $aId \leq S(0) \leq bId$.*

Then, there exist a constant μ, and two scalars $0 < \alpha < \beta$, such that, if $\theta_k\delta_t \leq \mu$,

$$\alpha Id \leq S(\tau) \leq \beta Id$$

for all $k \in \mathbb{N}$, for all $\tau \in [\tau_k, \tau_{k+1}]$.
Here, α and β are independent from δ_t and $\theta(t)$.
Since this relation is valid for all times τ, it is also true in the time scale t.

We divide the proof of this lemma into two Subsections:

1. in the first one we prove the existence of the upper bound β,
2. the second one is dedicated to the lower bound α.

Remember that the τ time scale is defined by $\delta_\tau = \theta\delta_t$. Since θ is constant on intervals of the form $[k\delta_t, (k+1)\delta_t[$, the length of an interval between two correction steps equals $\theta_k\delta_t$. This duration depends on values of θ that cannot be predicted.
Moreover, the maximum value of θ that has to be reached for the convergence of the observer to take place, is still unknown.
Therefore, the lemma above has to be proven independently from the time subdivision used to define correction steps.

Notations

[1] In Chapter 6, Part 2.4.2.

- S_n is the set of $(n \times n)$ symmetric matrices having their values in $\mathbb{R}$,
- $S_n(+)$ is the set of positive definite matrices of S_n,
- $T_r(S)$ denotes the trace of the matrix S,
- for any matrix S, $|S| = \sqrt{T_r(S'S)}$ is the Frobenius norm, when $S \in S_n$, $|S| = \sqrt{T_r(S^2)}$,
- for any matrix S, $\|S\|_2 = \sup_{\|x\|_2=1} \|Sx\|_2$, is the norm induced by the second euclidean norm, we also write it $\|S\|$ by omission,
- we keep the τ time scale notation, but we use S instead of $\bar{S}$ to ease the reading of equations,
- $\tau_k = \int_0^{k\delta_t} \theta(v)dv$. Since θ is fixed during prediction periods, the time elapsed between two correction steps is $(\tau_k - \tau_{k-1}) = \theta_{k-1}\delta_t$.
- $\mathcal{A}$ stands for the matrix $\left(A(\bar{u}) + \frac{\bar{b}(\bar{z},\bar{u})}{\theta}\right)$, we omit to write the dependencies to u, z and θ.

Matrix facts

Notice that according to equation (B.1), if $S(0)$ is symmetric then $S(t)$ is symmetric.

1. On the Frobenius norm (Cf. [67], Section 5.6):
   (a) If U and V are orthogonal matrices then $|UAV| = |A|$,
   (b) if A is symmetric semi positive then $|A| = |D_A|$, where D_A is the diagonal form of A,
   (c) if A is as in (b), $|A| = \left(\sum \lambda_i^2\right)^{\frac{1}{2}}$, where $\lambda_i \geq 0$ denotes the eigenvalues of A,
   (d) $\|A\|_2 \leq |A| \leq \sqrt{(n)}\|A\|_2$,
   (e) A is symmetric, $A \leq \|A\|_2 Id \leq |A| Id$.
2. On the trace of square matrices (Cf. [67]):
   (a) If A, B are $(n \times n)$ matrices, then $|T_r(AB)| \leq \sqrt{T_r(A'A)}\sqrt{T_r(B'B)}$,
   (b) if S is $(n \times n)$, and symmetric semi positive, then $T_r(S^2) \leq T_r(S)^2$,
   (c) if S is as in (b) then $T_r(S^2) \geq \frac{1}{n} T_r(S)^2$,
   (d) if S is as in (b) then, $T_r(SQS) \geq \frac{q}{n} tr(S)^2$, with $q = \min_{\|x\|=1} x'Qx$, and Q symmetric definite positive,
   (e) as a consequence of (b) and (c), if S is as in (b), and $\|.\|$ denotes any norm on $(n \times n)$ matrices, then there exist $l, n > 0$ such that
   $$l\|S\| \leq T_r(S) \leq m\|S\|.$$
3. On inequalities between semi positive matrices (Cf. [67], Sections 7.7 and 7.8), A and B are symmetric in this paragraph:

(a) if $A \geq B \geq 0$, and U is orthogonal, then $U^{'}AU \geq U^{'}BU$,

(b) A, and B as in (a), and A is invertible, then $\rho(BA^{-1}) \leq 1$, where ρ denotes the spectral radius,

(c) A, and B as in (a), then $\lambda_i(A) \geq \lambda_i(B)$, where $\lambda_i(M)$ denote the i^{th} eigenvalue of the matrix M sorted in ascending order,

(d) A, and B as in (a), and both invertible then $B^{-1} \geq A^{-1} \geq 0$,

(e) A, and B as in (a), then $det(A) \geq det(B)$, and $Tr(A) \geq Tr(B)$,

(f) if $A_1 \geq B_1 \geq 0$, and if $A_2 \geq B_2 \geq 0$ then $A_1 + A_2 \geq B_1 + B_2 \geq 0$.

4. From *facts* $1.(c)$ and $3.(f)$ we deduce that:

$$(A \geq B \geq 0) \Rightarrow (|A| \geq |B| \geq 0)\,.$$

B.1.1 Part One: the Upper Bound

This first part of the proof is decomposed into three steps. Let T^* be a fixed, positive scalar.

1. We prove that there is β_1 such that for all $\tau_k \leq T^*$, $k \in \mathbb{N}$, $S_k(+) \leq \beta_1 Id$,
2. we show that there exists β_2 such that for all $T^* \leq \tau_k$, $k \in \mathbb{N}$, $S_k(+) \leq \beta_2 Id$,
3. we deduce the result for all times.

The first fact can be directly proven as follows.

Lemma 78

Consider equation (B.1) and the assumptions of Lemma 77. Let $T^* > 0$ be fixed. There exists $\beta_1 > 0$ such that

$$S_k(+) \leq \beta_1 Id,$$

for all $\tau_k \leq T^*$, $k \in \mathbb{N}$, independently from the subdivision $\{\tau_i\}_{i\in\mathbb{N}}$.

Proof.

For all $\tau \in [\tau_{k-1}; \tau_k]$, equation (B.1) gives:

$$\begin{aligned} S(\tau) &= S_{k-1}(+) + \int_{\tau_{k-1}}^{\tau} \frac{dS(v)}{d\tau} dv, \\ &= S_{k-1}(+) \\ &\quad + \int_{\tau_{k-1}}^{\tau} \left[-\left(A(u) + \frac{\widetilde{b}^*(z,u)}{\theta} \right)^{'} S - S\left(A(u) + \frac{\widetilde{b}^*(z,u)}{\theta} \right) - SQS \right] dv. \end{aligned}$$

Since SQS is symmetric definite positive, we can write

$$S(\tau) \leq S_{k-1}(+) + \int_{\tau_{k-1}}^{\tau} \left[-\left(A(u) + \frac{\widetilde{b}^*(z,u)}{\theta} \right)^{'} S - S\left(A(u) + \frac{\widetilde{b}^*(z,u)}{\theta} \right) \right] dv,$$

and *fact* 4 gives us

$$|S(\tau)| \quad \leq | \quad S_{k-1}(+)| + \int_{\tau_{k-1}}^{\tau} 2s|S| dv, \tag{B.2}$$

where $A_M = \sup_{[0;T^*]} (|A(u(\tau)|)$, $\left|\widetilde{b}^* (z,u)\right| \leq L_b$ and $s = A_M + L_b$.
Gronwall's lemma gives

$$|S(\tau)| \leq| S_{k-1}(+)|e^{2s(\tau-\tau_{k-1})}. \tag{B.3}$$

Therefore, with $c = |C^{'} RC|$,

$$|S_k(+)| \leq| S_{k-1}(+)|e^{2s(\tau_k-\tau_{k-1})} + c\,(\tau_k - \tau_{k-1})\,. \tag{B.4}$$

Consider a subdivision $\{\tau_k\}_{k\in\mathbb{N}}$ such that $\tau_0 = 0$. From equation (B.4):

$$|S_1(+)| \leq| S_0|e^{2s\tau_1} + c\tau_1.$$

Since we set $\theta(0) = 1$ then $\bar{S}_0 = S_0$ and there is no ambiguity in the notation. We iterate to obtain

$$|S_2(+)| \leq| S_0|e^{2s\tau_2} + c\tau_1 e^{2s(\tau_2-\tau_1)} + c(\tau_2 - \tau_1),$$

and for all $k \in \mathbb{N}$

$$\begin{aligned} |S_k(+)| &\leq| S_0|e^{2s\tau_k} + \sum_{i=1}^{i=k} c\,(\tau_i - \tau_{i-1})\,e^{2s(\tau_k-\tau_i)} \\ &\leq |S_0|e^{2s\tau_k} + ce^{2s\tau_k} \sum_{i=1}^{i=k} (\tau_i - \tau_{i-1})\,e^{-2s\tau_i}. \end{aligned}$$

Since $e^{-2s\tau}$ is a decreasing function of τ, the sum on the right hand side of the inequality is smaller than the integral of $e^{-2s\tau}$ over the interval $[0, \tau_k]$. Therefore

$$\begin{aligned} |S_k(+)| &\leq| S_0|e^{2s\tau_k} + ce^{2s\tau_k} \int_0^{\tau_k} e^{-2s\tau} d\tau \\ &\leq |S_0|e^{2s\tau_k} + \tfrac{c}{2s}\left(e^{2s\tau_k} - 1\right). \end{aligned}$$

From this last equation we conclude that for any subdivision and any $k \in \mathbb{N}$ such that $\tau_k \leq T^*$:

$$|S_k(+)| \leq \left(|S_0| + \frac{c}{2s}\right) e^{2sT^*} = \beta_1.$$

■

In order to prove the result for times greater than T^*, consider a symmetric semi positive matrix S. We have $S \leq |S|Id = \sqrt{Tr(S^2)}Id$, and according to *fact* 2.(*b*): $S \leq Tr(S)Id$. Therefore, investigations on the upper bound are done in the form of investigations on the trace of S.

Lemma 79

If $S : [0, T[\to S_n$ is a solution to $\frac{dS}{d\tau} = -\mathcal{A}^{'} S - S\mathcal{A} - SQS$ then for almost all $\tau \in [0, T[$,

$$\frac{d}{d\tau} Tr(S) \leq -a\,(T_r\,(S(\tau)))^2 + 2bT_r\,(S(\tau))\,,$$

where:

$$\begin{aligned} a &= \tfrac{\lambda_{min}(Q)}{n} \\ b &= \sup_\tau T_r(\mathcal{A}^{'}(\tau)\mathcal{A}(\tau))^{\frac{1}{2}}. \end{aligned}$$

Proof.

$$\begin{aligned} \frac{d}{d\tau}T_r(S(\tau)) &= -T_r(SQS) - T_r(\mathcal{A}'S) - T_r(S\mathcal{A}) \\ &\leq -T_r(SQS) + 2|T_r(\mathcal{A}'S)|. \end{aligned}$$

From the matrix facts: $T_r(SQS) > \frac{q}{n}\left(T_r(S)\right)^2$, with $q = \min_{\|x\|=1} x'Qx$, and

$$\begin{aligned} |T_r(\mathcal{A}'S)| &\leq \sqrt{T_r(\mathcal{A}'\mathcal{A})}\sqrt{T_r(S'S)} \\ &\leq T_r(\mathcal{A}'\mathcal{A})^{\frac{1}{2}}\left[T_r(S)^2\right]^{\frac{1}{2}} \\ &\leq \sup_t\left(T_r(\mathcal{A}'\mathcal{A})^{\frac{1}{2}}\right)T_r(S). \end{aligned}$$

Therefore

$$\frac{d}{d\tau}T_r(S) \leq -aT_r(S)^2 + 2bT_r(S)$$

where a and b are as in the lemma.

■

Useful bounds for $Tr(S(\tau))$ can be derived from Lemma 79. In the following, x stands for $Tr(S)$.

Lemma 80

Let a, b be two positive constants. Let $x : [0, T[\to \mathbb{R}^+$ (possibly $T = +\infty$) be an absolutely continuous function satisfying for almost all $0 < \tau < T$ the inequality:

$$\dot{x}(\tau) \leq -ax^2(\tau) + 2bx(\tau).$$

The roots of $-aX^2 + 2bX$ are $\frac{2b}{a}$ and 0. The solution $x(\tau)$ is such that:

$$x(\tau) \leq \max\left\{(x(0), \tfrac{2b}{a})\right\} \text{ for all } \tau \in [0, T[.$$

In addition if $x(0) > \frac{2b}{a}$ then for all $\tau > 0 \in [0, T[$ we have the two inequalities:

$$\begin{cases} x(\tau) \leq \dfrac{2b}{a} + \dfrac{2b}{a}\dfrac{1}{e^{2b\tau} - 1}, \\ x(\tau) \leq \dfrac{2bx_0e^{2b\tau}}{ax_0\left(e^{2b\tau} - 1\right) + 2b}. \end{cases} \tag{B.5}$$

Proof.

- If $x(\tau) \leq \frac{2b}{a}$ for all $\tau \in [0, T[$, trivially we have $x(\tau) \leq \max\left(x(0), \frac{2b}{a}\right)$.
- Otherwise, we define

  $$E = \left\{\tau \in [0, T[: \; x(\tau) > \frac{2b}{a}\right\}$$

  which is a non empty set. Take any connected component $]\alpha,\beta\,[\subset E$ such that $\alpha > 0$. There are two situations:

  1. $\beta = T$, and $x(\beta) \geq \frac{2b}{a}$, $x(\alpha) = \frac{2b}{a}$,

2. $\beta < T$, and $x(\beta) = \frac{2b}{a} = x(\alpha)$.

In both cases, $x(\alpha) \leq x(\beta)$ ★.
Let us define $F = \{\tau \in]\alpha,\beta[: \dot{x}(\tau) > 0\}$: it has positive measure.

> *Suppose it is not the case: therefore x decreases from time α to time β. This implies that $x(\alpha) > x(\beta)$ which is in contradiction with ★. Therefore the measure of F is strictly positive.*

However for any $\tau \in]\alpha,\beta[\subset E$ we have $x(\tau) > \frac{2b}{a}$: $-aX^2 + 2bX < 0$ since $x(\tau) > \frac{2b}{a}$. Then $\dot{x}(\tau) < 0$ and $\tau \notin F$, which means that F is not of positive measure. This gives us a contradiction: either $\alpha = 0$ or $E = \emptyset$ (first item at the beginning of the proof).

Consider $\alpha = 0$: $E = [0, \tau_1[$, $\tau_1 > 0$ and $x(t) \leq \max(x(0), \frac{2b}{a})$, for all $\tau \geq \tau_1$. For $\tau \in [0, \tau_1[$:

$$\dot{x}(\tau) \leq a\left(\frac{2b}{a} - x(\tau)\right) x(\tau) < 0,$$

which means that $x(\tau)$ is decreasing on $[0, \tau_1[$ ($x(\tau) < x_0$). Therefore, for all $\tau \in [0, T[$, $x(\tau) \leq \max(x(0), \frac{2b}{a})$ in full generality.

Let us suppose that $x_0 > \frac{2b}{a}$, then $x(\tau) > \frac{2b}{a}$ for $\tau \in [0, \tau_1[$ as above. It means that:

$$\dot{x} \leq a\left(\frac{2b}{a} - x(\tau)\right) x(\tau).$$

We rewrite it:

$$\frac{-\dot{x}(\tau)}{a\left(\frac{2b}{a} - x\right) x} \geq 1.$$

Consider

$$d\left[ln\left(\frac{x}{x - \frac{2b}{a}}\right)\right] = \frac{-\frac{2b}{a}dx}{x(x - \frac{2b}{a})}$$

or in other terms

$$\frac{a}{2b}\frac{d}{d\tau}\left[ln\left(\frac{x}{x - \frac{2b}{a}}\right)\right] = \frac{-\dot{x}}{(x - \frac{2b}{a})x} \geq a.$$

Integration with respect to time gives:

$$ln\left(\frac{x}{x - \frac{2b}{a}}\right) \geq a\frac{2b}{a}\tau + ln\left(\frac{x_0}{x_0 - \frac{2b}{a}}\right).$$

Therefore, since $x_0 > \frac{2b}{a}$, and $x(\tau) > \frac{2b}{a}$ for $\tau < \tau_1$,

$$\frac{x}{(x - \frac{2b}{a})} \geq \frac{x_0}{(x_0 - \frac{2b}{a})} e^{2b\tau}. \tag{B.6}$$

Since $x(\tau) > \frac{2b}{a}$, equation (B.6) implies:

$$\frac{x}{(x - \frac{2b}{a})} \geq e^{2b\tau}.$$

Which gives the first inequality:

$$x \geq \left(x - \frac{2b}{a}\right) e^{2b\tau}$$

therefore

$$x \leq \frac{2b}{a} + \frac{2b}{a} \frac{1}{e^{2b\tau} - 1}.$$

We obtain the second inequality simply from (B.6):

$$x(\tau) \leq \frac{2bx_0 e^{2b\tau}}{ax_0 \left(e^{2b\tau} - 1\right) + 2b}.$$

■

Remark 81

1. *The first inequality of (B.5) can be used for any initial value $S(0)$, since it tends toward $+\infty$ when $\tau \to 0$,*
2. *the second one requires some knowledge on $S(0)$ in order to be useful,*
3. *the two bounds obtained are higher than $\frac{2b}{a}$, therefore they are true for any $\tau \in [0, T]$.*

Let us denote $r = \sup\left(Tr(C' R^{-1} C)\right)$. According to equation (B.1), the problem turns into proving that $x_k(+)$, the solution of:

$$\begin{cases} \frac{dx}{d\tau} & = & -ax^2 + 2bx \\ x_k(+) & \leq & x_k(-) + (\tau_k - \tau_{k-1})\, r, \end{cases} \tag{B.7}$$

is upper bounded for all $T^* \leq \tau_k$, $k \in \mathbb{N}$, independently from the subdivision $\{\tau_i\}, i \in \mathbb{N}$.

The bounds of Lemma 80 are valid on intervals of the form[2] $]\tau_{i-1}, \tau_i]$. Let us find an expression that we can use in order to upper bound x at any time.

Lemma 82

The solution of (B.7) is such that:

$$x(\tau) \leq \frac{2b}{a} + \frac{2b}{a} \frac{1}{e^{2b\tau} - 1} + r\tau,$$

for any $\tau > 0$, before or after an update.

[2] Or of the form $[\tau_{i-1}, \tau_i]$, it depends on which bound one considers.

Proof.

The first bound of (B.5) gives:

$$x_1(+) \leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1,$$

and the second is rewritten:

$$x_2(-) \leq \frac{2b}{a} + \frac{2b}{a}\frac{x_1(+) - \frac{2b}{a}}{x_1(+)\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a}}.$$

We want to replace $x_1(+)$ by the upper bound found above.
Let us define the function

$$h(x) = \frac{2bxe^{2b\tau}}{ax\left(e^{2b\tau}-1\right)+2b}.$$

Its derivative *w.r.t.* x is

$$\begin{aligned} h^{'}(x) &= \frac{e^{2b\tau}2b\left[ax(e^{2b\tau}-1)+2b\right] - a(e^{2b\tau}-1)xe^{2b\tau}2b}{\left[ax(e^{2b\tau}-1)+2b\right]^2} \\ &= \frac{e^{2b\tau}(2b)^2}{\left[ax(e^{2b\tau}-1)+2b\right]^2}. \end{aligned}$$

It is positive for all $\tau > 0$, and we can replace $x_1(+)$ by its upper bound:

$$\begin{aligned} x_2(-) &\leq \frac{2b}{a} + \frac{2b}{a}\frac{\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1\right] - \frac{2b}{a}}{\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a}} \\ &\leq \frac{2b}{a} + \frac{2b}{a}\frac{2b}{a}\frac{1}{e^{2b\tau_1}-1}\frac{1}{\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a}} \\ &\quad + \frac{2b}{a}\frac{r\tau_1}{\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a}}. \end{aligned}$$

We upper bounds the denominator of the last term with:

$$\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + r\tau_1\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a} \geq \frac{2b}{a},$$

and the denominator of the second term with:

$$\left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1} + \mathbf{r\tau_1}\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a} \geq \left[\frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1}\right]\left(e^{2b(\tau_2-\tau_1)}-1\right) + \frac{2b}{a}.$$

We also simplify $(2b/a)$ in those two terms:

$$\begin{aligned} x_2(-) &\leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_1}-1}\frac{1}{\left(\left[1+\frac{1}{e^{2b\tau_1}-1}\right]\left(e^{2b(\tau_2-\tau_1)}-1\right)+1\right)} + r\tau_1, \\ &\leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_2} - e^{2b\tau_1} + e^{2b\tau_1} - 1} + r\tau_1. \end{aligned}$$

Thus we have:

$$x_2(+) \leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_2}-1} + r\tau_2.$$

This last inequality is of the same form as the first bound of (B.5), and is independent from the values of τ_1 and τ_2. We can therefore generalize it to any $i \in \mathbb{N}^*$:

$$x_i(+) \leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau_i}-1} + r\tau_i.$$

Moreover, we can generalize this inequality to any $\tau > 0$, before or after the update (i.e. the correction step):

$$x(\tau) \leq \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau}-1} + r\tau.$$

■

Remark 83

This last bound cannot be used to obtain an upper bound for x for all times $\tau \geq T^*$ because it is not a decreasing function of time (this is proven later on). In other words, suppose we have two times $0 < \xi_1 < \xi_2$, then there exist two positive scalars, β_1 and β_2 such that $x(\xi_1) < \beta_1$ and $x(\xi_2) < \beta_2$. But we don't know if β_2 is higher or not than β_1.

In the following we'll see that such a relation can be derived provided that the length between two samples is bounded.

Lemma 84

Let us define the functions

$$\begin{aligned} \phi(\tau) &= \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2b\tau}-1} + r\tau, \\ \psi_{x_0}(\tau) &= \frac{2bx_0e^{2b\tau}}{ax_0\left(e^{2b\tau}-1\right)+2b} + r\tau. \end{aligned}$$

There exists $\mu_\phi > 0$, and $\mu_\psi(x_0) > 0$ such that $\phi(\tau)$, respectively $\psi_{x_0}(\tau)$, is a decreasing function for $\tau \in]0, \mu_\phi]$, respectively for $\tau \in [0, \mu_\psi(x_0)]$.
Moreover $\mu_\psi(x_0)$ is an increasing function of x_0.

Proof.

The proof basically consists in the computation of the variation tables of the two functions.

1. $$\phi'(\tau) = \frac{ra(e^{2b\tau})^2 - (4b^2+2ar)e^{2b\tau} + ra}{a(e^{2b\tau}-1)^2}.$$

   Let us consider the polynomial

   $$P(X) = raX^2 - (4b^2+2ar)X + ra.$$

   Its discriminant $\left(\left((2b)^2+2ar\right)^2 - 4a^2r^2\right)$ is positive. Therefore $P(X)$ has two real roots whose product and sum are both positive. Thus both roots are positive.

   We denote them $0 < X^* < X^\mu$, and remark that $P(X)$ is a non positive function on the interval $[X^*, X^\mu]$.

Those two roots are:

$$\begin{aligned} X^* &= \frac{(4b^2+2ar)-\sqrt{(4b^2+2a)^2-4a^2r^2}}{2ar} \\ &\leq \frac{\sqrt{(4b^2+2ar)^2-4a^2r^2}+\sqrt{4a^2r^2}-\sqrt{(4b^2+2a)^2-4a^2r^2}}{2ar}, \\ &\leq 1, \end{aligned}$$

and

$$X^\mu = \frac{(4b^2+2ar)+\sqrt{(4b^2+2a)^2-4a^2r^2}}{2ar} > 1.$$

We conclude that there exists $\mu_\phi > 0$ such that $\phi(\tau)$ is a decreasing function over the interval $[0, \mu_\phi]$.

2. Let us remark first that $\psi_{x_0}(0) = x_0$. A few computations give us:

$$\psi'_{x_0}(\tau) = \frac{ra^2x_0^2(e^{2b\tau})^2 - x_0(ax_0-2b)(4b^2+2ra)e^{2b\tau} + r(ax_0-2b)^2}{[ax_0(e^{2b\tau}-1)+2b]^2}.$$

As before we consider the polynomial

$$P(X) = ra^2x_0^2X^2 - x_0(ax_0-2b)(4b^2+2ra)X + r(ax_0-2b)^2.$$

Its discriminant is

$$x_0^2(ax_0-2b)^2((2b)^2+2ra)^2 - 4r^2a^2x_0^2(ax_0-2b)^2.$$

It is positive. Again both roots are positive: $0 < X^* < X^\mu$.

We show[3] that for b is high enough then $X^* \leq 1$, and $X^\mu > 1$. From the definition of b in Lemma 79 we can consider that it is the case from the very beginning (b comes the upper bounds of the matrices $A(u)$ and b^* of the Riccati equation). We conclude that there exists $\mu_\psi > 0$, a function of x_0, such that $\psi(\tau)$ is a decreasing function over the interval $]0, \mu_\psi(x_0)]$.

In order to check that $\mu_\psi(x_0)$ increases with x_0, we show that X^μ is an increasing function of x_0.

$$\begin{aligned} X^\mu &= \frac{x_0(ax_0-2b)(4b^2+2ar)}{2ra^2x_0^2} \\ &\qquad + \frac{\sqrt{(ax_0-2b)^2(4b^2+2ar)^2x_0^2 - 4r^2a^2x_0^2(ax_0-2b)^2}}{2ra^2x_0^2} \\ &= \frac{(a-\frac{2b}{x_0})(4b^2+2ar)}{2ra^2} \\ &\qquad + \frac{\sqrt{(a-\frac{2b}{x_0})^2(4b^2+2ar)^2 - 4r^2a^2(a-\frac{2b}{x_0})^2}}{2ra^2} \\ &= \frac{A+\sqrt{B}}{D}. \end{aligned}$$

[3] Since the coefficient of the second order term of $P(X)$ is positive, then the polynomial has negative sign for $X \in]X^*, X^\mu[$. We only to check that $P(1) \leq 0$.

- from the definition of r, $D > 0$,
- $(a - \frac{2b}{x_0})$ is an increasing function of x_0, so is A,
- B can be rewritten $(a - \frac{2b}{x_0})^2 \left[(4b^2 + 2ar)^2 - 4r^2a^2\right]$, the right part is positive, and the left one is an increasing function of x_0, so is B.

$X^\mu(x_0)$ is therefore an increasing function of x_0, so is $\mu_\psi(x_0)$ (i.e. $P(X)$ was obtained from the change of variables $X = e^{2b\tau}$).

■

Remark 85

1. *Notice that since $\psi(0) = x_0$, therefore for all $\tau \in [0, \mu_\psi(x_0)]$, $\psi(\tau) \leq \psi(0) = x_0$.*

2. *Let $x_0^* > 0$ be fixed, and denote $\psi_{x_0^*}(\tau)$ the associated ψ function. It is a decreasing function over $[0, \mu_\psi(x_0^*)]$. The fact that $\mu_\psi(x_0^*)$ is an increasing function of x_0 tells us that for any $x_0 > x_0^*$, $\psi_{x_0}(\tau)$ is a decreasing function of τ over the interval $[0, \mu_\psi(x_0^*)] \subset [0, \mu_\psi(x_0)]$.*
   In other words $\psi_{x_0}(\mu_\psi(x_0^*)) < x_0$, for all $x_0 \geq x_0^*$.

We now have all the building blocks we need to prove the other half of this subsection's result.

Lemma 86

Consider equation (B.1) and the assumptions of Lemma 77. Let $T^* > 0$ be fixed. There exist two scalars $\beta_2 > 0$ and $\mu > 0$ such that

$$S_k(+) \leq \beta_2 Id,$$

for all $T^* \leq \tau_k$, $k \in \mathbb{N}$, for all subdivision $\{\tau_i\}_{i\in\mathbb{N}}$, $\tau_i - \tau_{i-1} < \mu$.

Proof.

Let $T^* > 0$ be arbitrarily fixed. We define $B(T^*) = \frac{2b}{a} + \frac{2b}{a}\frac{1}{e^{2bT^*}-1} + rT^*$. From Lemma 82, $x(T^*) \leq B(T^*)$ independently from the subdivision $\{\tau_i\}_{i\in\mathbb{N}}$.
Let us define μ as the maximum value such that the functions $\phi(\tau)$, and $\psi_{B(T^*)}(\tau)$ of Lemma 84 are both decreasing functions over the interval $[0, \mu]$. Notice that μ depends on $B(T^*)$.

We claim that $\beta_2 = B(T^*) + r\mu$ solves the problem.

Let us consider a time subdivision $\{\tau_i\}_{i\in\mathbb{N}}$ such that $\tau_i - \tau_{i-1} \leq \mu$, for all $i \in \mathbb{N}$.
We show first that if $x_k(+) \leq \beta_2$, then $x_{k+1}(+) \leq \beta_2$.

1. If $x_k(+) \leq \frac{2b}{a}$ then, from Lemma 80, $x_{k+1}(-) \leq \max\left(\frac{2b}{a}, x_k(+)\right) = \frac{2b}{a}$. Therefore $x_{k+1}(+) \leq \frac{2b}{a} + r\mu \leq \beta_2$,

2. if $x_k(+) \geq \frac{2b}{a}$, and $x_k(+) \leq B(T^*)$, we use Lemma 80 again to deduce $x_{k+1}(-) \leq \max\left(\frac{2b}{a}, x_k(+)\right) \leq B(T^*)$. Finally we have $x_{k+1}(+) \leq B(T^*) + r\mu = \beta_2$,

3. if $x_k(+) \geq \frac{2b}{a}$, and $x_k(+) \geq B(T^*)$ then , Lemma 80 tells us that $x_{k+1} \leq \psi_{x_k}(\tau_{k+1} - \tau_k)$. From Lemma 84, we know that it is a decreasing function, and according to Remark 85 we have $x_{k+1}(+) \leq \psi_{x_k}(0) \leq \beta_2$.

Let us now define k such that $\tau_{k-1} < T^* \leq \tau_k$. We claim that $x_k(+) \leq \beta_2$.

1. If $T^* \in]0, \tau_1]$, since μ is such that the function $\phi(\tau)$ is decreasing over $[0, \tau_1] \subset [0, \mu]$, and from Lemma 82:
$$x_1(+) \leq \phi(\tau_1) \leq \phi(T^*) \leq \beta_2,$$
2. if $\tau_{k-1} > 0$, we have $x(T^*) \leq B(T^*)$ by definition. From equation (B.7) and Lemma 80:
   - if $x(T^*) \leq \frac{2b}{a}$ therefore $x_k(-) \leq \frac{2b}{a} \leq B(T^*)$,
   - if $x(T^*) \geq \frac{2b}{a}$, then $\frac{dx}{d\tau}$ is negative according to equation (B.7), x is decreasing and $x_k(-) \leq B(T^*)$.

   In both cases $x_k(+) \leq B(T^*) + \mu r = \beta_2$

We finally conclude that, there are two positive constants β_2 and μ such that $S_k(+) \leq \beta_2 Id$ for all $\tau_k \geq T^*$, $k \in \mathbb{N}$. It is true for all subdivision $\{\tau_i\}_{i\in\mathbb{N}}$, such that $\tau_i - \tau_{i-1} \leq \mu$, $\forall i$.

■

Remark 87

The constraint $\tau_i - \tau_{i-1} \leq \mu$, interpreted in the original time scale, gives $\theta_i \delta_t \leq \mu$, $\forall i \in \mathbb{N}$.

We conclude this subsection giving the upper bound property for all times τ.

Lemma 88

Consider the prediction correction equation (B.1) and the hypothesis of Lemma 77. There exist two positive constants $\mu > 0$ and β such that $S(\tau) \leq \beta Id$, for all $k \in \mathbb{N}$, and all $\tau \in [\tau_k, \tau_{k+1}[$, for all subdivision $\{\tau_i\}_{i\in\mathbb{N}}$ such that $(\tau_i - \tau_{i-1}) \leq \mu$.
This inequality is then also true in the t time scale.

Proof.

Let us define $\hat{\beta} = \max(\beta_2, \beta_1)$. Lemmas 86 and 78 give us $S_k(+) \leq \beta_2 Id$, for all $k \in \mathbb{N}$. Then Lemmas 79 and 80 implies the existence of $\hat{\beta}^*$ such that $S(\tau) \leq \hat{\beta} Id$ for any $\tau \in [\tau_k, \tau_{k+1}[$, and any $k \in \mathbb{N}$.
We define β as the maximum value between $\hat{\beta}^*$ and $\hat{\beta}$.

■

B.1.2 Part Two: the Lower Bound

As in the previous subsection, the proof is separated into three parts. Let T_* be a fixed positive scalar, possibly distinct from T^* (Cf. previous subsection).

1. We prove that there is α_1 such that for all $k \in \mathbb{N}$, such that $\tau_k \leq T_*$, $\alpha_1 Id \leq S_k(+)$,

2. we show that there exists α_2 such that for all $k \in \mathbb{N}$, such that $T_* \leq \tau_k$, $\alpha_2 Id \leq S_k(+)$,

3. we give the result for all times.

As before, the first fact is proven rather directly.

Lemma 89

Consider equation (B.1) and the assumptions of Lemma 77. Let $T_* > 0$ be fixed. There exists $\alpha_1 > 0$ such that

$$\alpha_1 Id \leq S_k(+),$$

for all $\tau_k \leq T^*$, $k \in \mathbb{N}$, independently from the subdivision $\{\tau_i\}_{i\in\mathbb{N}}$.

Proof.

We denote $P = S^{-1}$. For all $\tau \in [\tau_{k-1}, \tau_k[$, equation (B.1) gives:

$$\begin{aligned} P(\tau) &= P_{k-1}(+) + \int_{\tau_{k-1}}^{\tau} \frac{dP(v)}{d\tau} dv, \\ P(\tau) &= P_{k-1}(+) \\ &\quad + \int_{\tau_{k-1}}^{\tau} \left[P\left(A(u) + \frac{\tilde{b}^*(z,u)}{\theta}\right)' + \left(A(u) + \frac{\tilde{b}^*(z,u)}{\theta}\right) P + Q \right] dv. \end{aligned} \tag{B.8}$$

Computations performed as in Lemma 78 lead to

$$|P_k(-)| \leq \left(|P_{k-1}(+)| + |Q|\,(\tau_k - \tau_{k-1})\right) e^{2s(\tau_k - \tau_{k-1})} \tag{B.9}$$

where $s > 0$ is as before. From (B.1) and Lemma 64

$$\begin{aligned} P_k(+) &= \left[S_k(-) + \theta\delta_t C' r^{-1} C\right]^{-1} \\ &= S_k(-)^{-1}\left[S_k(-)^{-1} + \theta\delta_t S_k(-)^{-1} C' r^{-1} C S_k(-)^{-1}\right]^{-1} S_k(-)^{-1} \\ &= S_k(-)^{-1}\left[S_k(-) - C'\left(\tfrac{r}{\theta\delta_t} + C S_k(-)^{-1} C'\right)^{-1} C\right] S_k(-)^{-1} \\ &\leq S_k(-)^{-1} S_k(-) S_k(-)^{-1} \\ &\leq P_k(-). \end{aligned} \tag{B.10}$$

We consider a subdivision of $\{\tau_k\}_{k\in\mathbb{N}}$ such that $\tau_0 = 0$. Since $\bar{P}_0 = P_0$:

$$|P_1(+)| \leq \left(|P_0| + |Q|\,\tau_1\right) e^{2s\tau_1}.$$

We iterate to obtain

$$|P_2(+)| \leq |P_0|\, e^{2s\tau_2} + |Q|\left(\tau_1 e^{2s\tau_2} + (\tau_2 - \tau_1)\, e^{2s(\tau_2 - \tau_1)}\right),$$

and for all $k \in \mathbb{N}$

$$|P_k(+)| \leq |P_0|\, e^{2s\tau_k} + |Q| \sum_{i=1}^{i=k} (\tau_i - \tau_{i-1})\, e^{2s(\tau_k - \tau_i)}.$$

In the same manner as in Lemma 78 we conclude that for all $\tau_k \leq T_*$, $k \in \mathbb{N}$:

$$|P_k(+)| \leq \left(|P_0| + \frac{|Q|}{2s} \right) e^{2sT_*} = \frac{1}{\alpha_1}.$$

Therefore $P_k(+) \leq \frac{1}{\alpha_1} Id$, for all $\tau_k \leq T_*$, $k \in \mathbb{N}$, for all subdivisions $\{\tau_i\}_{i \in \mathbb{N}}$.
Equivalently, from matrix fact 3.(c): $\alpha_1 Id \leq S_k(+)$.

■

We now proceed with the proof of the second part of this subsection. We need a different set of tools than the one used in the preceding section. It is a series of lemmas adapted from [57]. As we will see, in addition to the proof of the existence of a lower bound for S, we also prove that it is a symmetric positive definite matrix for all times.

Lemma 90

For any $\lambda \in \mathbb{R}^*$, any solution $S : [0, T[\to S_m$ (Possibly, $T = +\infty$) of

$$\frac{dS}{d\tau} = -\mathcal{A}^{'}(\tau)S - S\mathcal{A}(\tau) - SQS,$$

we have for all $\tau \in [0, T[$:

$$\begin{aligned} S(\tau) = e^{-\lambda\tau}\phi_u(\tau,0)S_0\phi_u^{'}(\tau,0) \\ +\lambda \int_0^\tau e^{-\lambda(\tau-v)}\phi_u(\tau,v)\left(S(v) - \frac{S(v)QS(v)}{\lambda}\right)\phi_u^{'}(\tau,v)dv \end{aligned} \tag{B.11}$$

where $\phi_u(\tau, s)$ is such that:

$$\left\{ \begin{array}{rcl} \frac{d\phi_u(\tau,s)}{d\tau} & = & -\mathcal{A}^{'}(\tau)\phi_u(\tau,s), \\ \phi_u(s,s) & = & Id. \end{array} \right.$$

Remark 91

Notice that $S(\tau)$ is not λ-dependent. This latter scalar is used only to provide us with a convenient way to express $S(\tau)$.

Proof.

Let us consider an equation of the form

$$\frac{d}{d\tau}\Lambda(\tau) = -\mathcal{A}^{'}\Lambda(\tau) - \Lambda(\tau)\mathcal{A} + F(\tau). \tag{B.12}$$

$\phi_u(\tau, s)$ denotes the resolvent of the system $\frac{dx}{d\tau} = -\mathcal{A}^{'}x$:

$$\left\{ \begin{array}{rcl} \frac{d\phi_u(\tau,s)}{d\tau} & = & -\mathcal{A}^{'}\phi_u(\tau,s), \\ \phi_u(s,s) & = & Id. \end{array} \right.$$

We search for a solution of the form$\Lambda(\ \tau) = \phi_u(\tau, s)h(\tau)\phi_u^{'}(\tau, s)$.

$$\begin{aligned} \frac{d\Lambda}{d\tau} &= \left(\tfrac{d}{d\tau}\phi_u(\tau,s)\right)h(\tau)\phi_u^{'}(\tau,s) + \phi_u(\tau,s)h(\tau)\left(\tfrac{d}{d\tau}\phi_u^{'}(\tau,s)\right) \\ &\qquad +\phi_u(\tau,s)\left(\tfrac{d}{d\tau}h(\tau)\right)\phi_u^{'}(\tau,s) \\ &= -\mathcal{A}^{'}\phi_u(\tau,s)h(\tau)\phi_u^{'}(\tau,s) - \phi_u(\tau,s)h(\tau)\phi_u^{'}(\tau,s)\mathcal{A}(\tau) \\ &\qquad +\phi_u(t,s)\left(\tfrac{d}{d\tau}h(\tau)\right)\phi_u^{'}(\tau,s) \\ &= -\mathcal{A}^{'}\Lambda(\tau) - \Lambda(\tau)\mathcal{A} + \phi_u(\tau,s)\left(\tfrac{d}{d\tau}h(\tau)\right)\phi_u^{'}(\tau,s), \end{aligned}$$

therefore $\phi_u(\tau,s)\left(\frac{d}{d\tau}h(\tau)\right)\phi_u^{'}(\tau,s)=F(\tau)$, and

$$h(\tau)=h(0)+\int_0^\tau \phi_u(s,v)F(v)\phi_u^{'}(s,v)dv.$$

Since$\Lambda(\ \tau)=\phi_u(\tau,s)h(\tau)\phi_u^{'}(\tau,s)$, thus $\phi_u(s,0)\Lambda(0)\phi_u^{'}(s,0)=h(0)$. Therefore:

$$\begin{aligned}\Lambda(\tau) &= \phi_u(\tau,s)h(0)\phi_u^{'}(\tau,s)+\phi_u(\tau,s)\left(\int_0^\tau \phi_u(s,v)F(v)\phi_u^{'}(s,v)dv\right)\phi_u^{'}(t,s)\\ &= \phi_u(\tau,s)\phi_u(s,0)\Lambda(0)\phi_u^{'}(s,0)\phi_u^{'}(\tau,s)+\int_0^\tau \phi_u(\tau,v)F(v)\phi_u^{'}(\tau,v)dv\\ &= \phi_u(\tau,0)\Lambda_0\phi_u^{'}(\tau,0)+\int_0^\tau \phi_u(\tau,v)F(v)\phi_u^{'}(\tau,v)dv.\end{aligned}$$

Let us take $\lambda\in\mathbb{R}^*$, and define $\hat{S}=e^{\lambda\tau}S$. The differential equation associated to $\hat{S}$ is

$$\begin{aligned}\tfrac{d}{d\tau}\hat{S}(\tau) &= \lambda e^{\lambda\tau}S+e^{\lambda\tau}\tfrac{dS}{d\tau}\\ &= \lambda\hat{S}(\tau)-\mathcal{A}^{'}\hat{S}(\tau)-\hat{S}(\tau)\mathcal{A}-e^{-\lambda\tau}\hat{S}(\tau)Q\hat{S}(\tau).\end{aligned}$$

This equation is of the form (B.12), with $F(\tau)=\lambda\hat{S}(\tau)-e^{-\lambda\tau}\hat{S}Q\hat{S}$. According to the computation above, we have:

$$\hat{S}(\tau)=\phi_u(\tau,0)\hat{S}(0)\phi_u^{'}(\tau,0)+\int_0^\tau \phi_u(\tau,v)\left[\lambda\hat{S}(v)-e^{-\lambda v}\hat{S}Q\hat{S}\right]\phi_u^{'}(\tau,v)dv,$$

and consequently

$$S(\tau)=\phi_u(\tau,0)S_0\phi_u^{'}(\tau,0)e^{-\lambda\tau}+\lambda\int_0^\tau e^{-\lambda(\tau-v)}\phi_u(\tau,v)\left[S-\frac{SQS}{\lambda}\right]\phi_u(\tau,v)dv.$$

■

Lemma 92

Let $S:[0;e(S)[\rightarrow S_m$ be a maximal semi solution of

$$\frac{d}{d\tau}S=-\mathcal{A}^{'}S-S\mathcal{A}-SQS.$$

If $S(0)=S_0$ is positive definite then

$$e(S)=+\infty \text{ and } S(\tau) \text{ is positive definite for all } \tau\geq 0.$$

Thus, for any arbitrary time subdivision $\{\tau_i\}_{i\in\mathbb{N}^*}$ the solution to the continuous discrete Riccati equation (B.1) is positive definite for all times provided that S_0 is positive definite.

Proof.

Assume that S is not always positive definite. Let us define

$$\theta=\inf\left(\tau|S(\tau)\notin S_m(+)\right).$$

In other words $S(\tau)\in S_m(+)$ for all $\tau\in[0;\theta[$. From Lemma 79:

$$\frac{d}{d\tau}T_r(S)\leq -aT_r(S)^2+2bT_r(S)$$

which, in combination with Lemma 80, gives $T_r(S) \leq \max\left(T_r(S_0), \frac{2b}{a}\right)$ and

$$|S| = \sqrt{T_r(S^2)} \leq \sqrt{T_r(S)^2} = T_r(S) \leq \max\left(T_r(S_0), \frac{2b}{a}\right).$$

Choose $\lambda > |Q| \max\left(T_r(S_0), \frac{2b}{a}\right)$ and apply Lemma 90:

$$S(\tau) = e^{-\lambda\tau}\phi_u(\tau,0)S_0\phi_u^{'}(\tau,0) + \int_0^\tau e^{-\lambda(\tau-v)}\phi_u(\tau,v)\left(S(\tau) - \frac{S(\tau)QS(\tau)}{\lambda}\right)\phi_u^{'}(\tau,v)dv.$$

Let τ be equal to θ then $S(\theta) = (I) + (II)$ with

$$\begin{aligned}(I) &= e^{-\lambda\theta}\phi_u(\theta,0)S_0\phi_u^{'}(\theta,0),\\ (II) &= \int_0^\theta e^{-\lambda(\theta-v)}\phi_u(\theta,v)\left(S - \frac{SQS}{\lambda}\right)\phi_u^{'}(\theta,v)dv.\end{aligned}$$

(I) is definite posittive
(II) depends on $\left(S - \frac{SQS}{\lambda}\right)$, which we can rewrite $\sqrt{S}\left(Id - \frac{\sqrt{S}Q\sqrt{S}}{\lambda}\right)\sqrt{S}$ since $S(\tau)$ is positive definite for $0 < \tau < \theta$. Therefore the positiveness of (II) depends on $\left(Id - \frac{\sqrt{S}Q\sqrt{S}}{\lambda}\right)$. From the definition of λ we have

$$\frac{\sqrt{S}Q\sqrt{S}}{\lambda} < Id.$$

Therefore (II) is definite positive and so is $S(\theta)$. This is in contradiction with the definition of θ (i.e. such that $S(\theta)$ is not positive definite). Therefore $S(t)$ is always positive definite provided that S_0 is.

Let us consider an arbitrary time subdivision $\{\tau_i\}_{i\in\mathbb{N}^*}$, and $j \in \mathbb{N}$. We assume that $S_j(+)$ is symmetric positive definite. Then from the demonstration above, $S_{j+1}(-)$ exists and is symmetric definite positive. Thus $S_{j+1}(+)$ is symmetric positive definite, which recursively proves the result.

■

We now give a lemma that allows us to use the properties of the continuous Gramm matrix given in Appendix A.4.

Lemma 93

Let $m(t)$, $t \in [0,T]$, be a $(n \times n)$ symmetric matrix, at least differentiable once.

Let μ be a positive constant, and $\{\tau_i\}_{i\in\{0,1,\dots,k\}}$ an arbitrary subdivision of $[0,T]$, with $\tau_0 = 0$, $\tau_k = T$ such that $\tau_i - \tau_{i-1} \leq \mu$, for all i.

We suppose that all the coefficients of m have their derivative bounded over time. Then

$$\int_0^T m(v)dv - \sum_{i=1}^{k} m(\tau_i)\left(\tau_i - \tau_{i-1}\right) \leq (\mu K T)\, Id,$$

where $K = \frac{n}{2}\max_{k,l,\tau}\left(\left|m^{'}(\tau)\right|\right)$, and Id is the identity matrix.

Proof.

Let $M(t)$ be a primitive matrix of $m(t)$, that is to say a matrix whose elements are the primitives of the elements of $m(t)$. We have the identity

$$\int_0^T m(v)dv = M(T) - M(0) = \sum_{i=1}^{k} \left[M(\tau_i) - M(\tau_{i-1})\right].$$

We can apply the Taylor-Lagrange expansion on all elements M_{kl}:

$$M_{kl}(\tau_{i-1}) = M_{kl}(\tau_i) + (\tau_{i-1} - \tau_i)\, m_{kl}(\tau_i) + \frac{(\tau_{i-1} - \tau_i)^2}{2} m'_{kl}(\xi_{kl,i})$$

where $\xi_{kl,i} \in [\tau_{i-1}, \tau_i]$. We have thus, the relation

$$\sum_{i=1}^{k} M(\tau_{i-1}) = \sum_{i=1}^{k} M(\tau_i) + \sum_{i=1}^{k} m(\tau_i)\left((\tau_{i-1} - \tau_i)\right) + \sum_{i=1}^{k} \left(\frac{(\tau_{i-1} - \tau_i)^2}{2} R_i\right)$$

where $(R_{kl})_i = m'_{kl}(\xi_{kl,i})$, with $\xi_{kl,i} \in [\tau_i, \tau_{i-1}]$. Therefore

$$\begin{aligned}
\int_0^T m(v)dv - \sum_{i=1}^{k} m(\tau_i)(\tau_i - \tau_{i-1}) &= \sum_{i=1}^{k} \left[M(\tau_i) - M(\tau_{i-1})\right] - \sum_{i=1}^{k} \left((\tau_i - \tau_{i-1}) m(\tau_i)\right) \\
&= \textstyle\sum_{i=1}^{k} \left(\frac{(\tau_{i-1} - \tau_i)^2}{2} R_i\right).
\end{aligned}$$

We now use the definition of the matrix inequality. Let x be a non zero element of $\mathbb{R}^n$,

$$\begin{aligned}
x' \left[\sum_{i=1}^{k} \left(\frac{(\tau_{i-1} - \tau_i)^2}{2} R_i\right)\right] x &= \frac{1}{2} \sum_{i=1}^{k} \left((\tau_{i-1} - \tau_i)^2\, x' R_i x\right) \\
&\leq \frac{1}{2} \sum_{i=1}^{k} \left((\tau_{i-1} - \tau_i)^2 \sum_{k,l} |x_k|\, |R_{k,l}|_i\, |x_l|\right) \\
&\leq \frac{1}{2} \mu \max_{k,l,i} \left(|R_{k,l}|_i\right) \left(\sum_{i=1}^{k} \tau_{i-1} - \tau_i\right) \left(\sum_{k,l} |x_k|\, |x_l|\right) \\
&\leq \frac{1}{2} \mu \max_{k,l,i} \left(|R_{k,l}|_i\right) \left(\sum_{i=1}^{k} \tau_{i-1} - \tau_i\right) \frac{1}{2} \left(\sum_{k,l} |x_k|^2 + |x_l|^2\right) \\
&\leq \frac{1}{2} \mu \max_{k,l,i} \left(|R_{k,l}|_i\right) T \frac{1}{2} \left(2n \|x\|^2\right).
\end{aligned}$$

Thus giving the result.

■

Remark 94

Suppose that $m = \psi'(v,T) C' C \psi(v,T)$ where $\psi(v,T)$ is a resolvent matrix as in Lemma 90. From the assumptions we put on our continuous discrete system, and the fact that the

observation matrix is fixed (i.e. not u-dependent), the derivative of such a m matrix is bounded.

If we want to consider C matrices defined with a function α_1 as in the continuous case, the derivative of this function must have a bounded time derivative. The consequence is that bang-bang like control inputs become inadmissible.

We now have gathered all the elements necessary to prove the second half of this subsection's result.

Lemma 95

Consider equation (B.1) and the assumptions of Lemma 77. Let $T_* > 0$ be fixed. There exist $\alpha_2 > 0$ and $\mu > 0$ such that for all subdivision $\{\tau_i\}_{i\in\mathbb{N}}$ with $\tau_i - \tau_{i-1} < \mu$, and for all τ_k, $k \in \mathbb{N}$, $T_* \leq \tau_k$:

$$\alpha_2 Id \leq S_k(+).$$

Proof.

From Lemma 88, there are two scalars β and μ such that, for all subdivisions $\{\tau_i\}_{i\in\mathbb{N}}$, such that $(\tau_i - \tau_{i-1}) \leq \mu$, $S(\tau) \leq \beta Id$ for all $\tau \in [\tau_{i-1}, \tau_i]$, $k \in \mathbb{N}$.

Let us define[4] $T_* \geq n\mu$, and consider a subdivision of time $\{\tau_i\}_{i\in\mathbb{N}}$ with $\tau_i - \tau_{i-1} \leq \mu$, $\forall i$. Apply equation (B.11) with $\lambda > \beta\ |Q|$:

$$\begin{aligned} S_1(+) = e^{-\lambda\tau_1}\phi_u(\tau_1,0)S_0\phi_u^{'}(\tau_1,0) \\ +\lambda\int_0^{\tau_1} e^{-\lambda(\tau_1-v)}\phi_u(\tau_1,v)\left(S(v) - \frac{S(v)QS(v)}{\lambda}\right)\phi_u^{'}(\tau_1,v)dv + C^{'}R^{-1}C\tau_1. \end{aligned}$$

In the same manner we obtain at time τ_2:

$$\begin{aligned} S_2(+) = e^{-\lambda(\tau_2-\tau_1)}\phi_u(\tau_2,\tau_1)S_1(+)\phi_u^{'}(\tau_2,\tau_1) \\ +\lambda\int_{\tau_1}^{\tau_2} e^{-\lambda(\tau_2-v)}\phi_u(\tau_2,v)\left(S(v) - \frac{S(v)QS(v)}{\lambda}\right)\phi_u^{'}(\tau_2,v)dv + C^{'}R^{-1}C(\tau_2-\tau_1). \end{aligned}$$

A few computations (see Appendix A.1 for the resolvent's properties) lead to:

$$\begin{aligned} S_2(+) = e^{-\lambda\tau_2}\phi_u(\tau_2,0)S_0\phi_u^{'}(\tau_2,0) \\ +\lambda\int_0^{\tau_2} e^{-\lambda(\tau_2-v)}\phi_u(\tau_2,v)\left(S(v) - \frac{S(v)QS(v)}{\lambda}\right)\phi_u^{'}(\tau_2,v)dv \\ +e^{-\lambda(\tau_2-\tau_1)}\phi_u(\tau_2,\tau_1)C^{'}R^{-1}C\phi_u^{'}(\tau_2,\tau_1)\tau_1 + C^{'}R^{-1}C(\tau_2-\tau_1). \end{aligned}$$

[4]The matrix inequality we want to achieve implies, in particular, the invertibility of the matrices $S_k(+)$, $k \in \mathbb{N}$. It implies the invertibility of the sum that appears in (B.13). This matrix cannot be inverted if there are less than n summation terms.
As is explained at the end of the proof, the result is achieved provided that the step size of the subdivision used in the summation is small enough. Since we have the freedom to make μ as small as we want, the summation above mentioned can be performed with sufficiently many points whatever the value of T^*. It makes the requirement $T_* \geq n\mu$ redundant.
Recall now that μ represents the maximum sampling that allows us to use the observer. Lemma 89 provides us with a first value for μ, and it is pretty much system dependent. By removing the assumption, we probably have to shorten μ for the sake of mathematical elegance only.
We think that the proof is better that way.

For any k, we compute $S_k(+)$ iteratively:

$$\begin{aligned} S_k(+) &= e^{-\lambda\tau_k}\phi_u(\tau_k,0)S_0\phi_u^{'}(\tau_k,0) \\ &\quad +\lambda\int_0^{\tau_k} e^{-\lambda(\tau_k-v)}\phi_u(\tau_k,v)\left(S(v)-\frac{S(v)QS(v)}{\lambda}\right)\phi_u^{'}(\tau_k,v)dv \\ &\quad +\sum_{i=1}^{k} e^{-\lambda(\tau_k-\tau_i)}\phi_u(\tau_k,\tau_i)C^{'}R^{-1}C\phi_u^{'}(\tau_k,\tau_i)\left(\tau_i-\tau_{i-1}\right). \end{aligned} \tag{B.13}$$

We consider this last equation for any $k \in \mathbb{N}^*$ such that $\tau_k \geq T_*$. It is of the form $S_k(+) = (I) + (II) + (III)$, in the same order as before.

- Since S_0 is positive definite, (I) is positive definite,
- from the definition of λ, $\left(S(v)-\frac{S(v)QS(v)}{\lambda}\right)$ is positive definite, and the same goes for (II),
- let us define $l < k$ as the maximal element of $\mathbb{N}$ such that $\tau_k - \tau_l \geq T_*$. Since we consider (B.13) for times $\tau_k \geq T_*$ such an element always exists[5]. Note that because the subdivisions we use have a step less than μ, $\tau_k - \tau_l \leq T_* + \mu$. All the terms of the sum (III) are symmetric definite positive matrices and thus

$$(III) \geq \sum_{i=l+1}^{k} e^{-\lambda(\tau_k-\tau_i)}\phi_u(\tau_k,\tau_i)C^{'}R^{-1}C\phi_u^{'}(\tau_k,\tau_i)\left(\tau_i-\tau_{i-1}\right).$$

  From the properties of the resolvent, the right hand side expression can be rewritten, with $\bar{u}(\tau) = u(\tau+\tau_l)$:

$$\sum_{i=l+1}^{k} e^{-\lambda(\tau_k-\tau_i)}\phi_{\bar{u}}(\tau_k-\tau_l,\tau_i-\tau_l)C^{'}R^{-1}C\phi_{\bar{u}}^{'}(\tau_k-\tau_l,\tau_i-\tau_l)\left(\tau_i-\tau_{i-1}\right). \tag{B.14}$$

  For all $i = \{l+1,..,k\}$ we have $e^{-\lambda(\tau_k-\tau_i)} \geq e^{-\lambda(T_*+\mu)}$, and R^{-1} is a fixed symmetric positive definite matrix. Therefore in order to lower bound (B.14) we have to find a lower bound for

$$\sum_{i=l+1}^{k} \phi_{\bar{u}}(\tau_k-\tau_l,\tau_i-\tau_l)C^{'}C\phi_{\bar{u}}^{'}(\tau_k-\tau_l,\tau_i-\tau_l)\left(\tau_i-\tau_{i-1}\right). \tag{B.15}$$

  Note that this sum is now computed for a subdivision of the interval $[0,\tau_k-\tau_l]$ that has the very specific property $T_* \leq \tau_k - \tau_l \leq T_* + \mu$. Let us redefine this subdivision as follows:

  - let k_* be equal to $k-l$, the subdivision has k_*+1 elements,
  - we denote $\tilde{\tau}_i = \tau_{i+l} - \tau_l$ –i.e. $\tilde{\tau}_0 = 0$ and $\tilde{\tau}_{k_*} = \tau_k - \tau_l$.

[5] It can happen that $k = 0$. Recall that $\tau_0 = 0$ for all subdivisions.

We can prove that (III) has a lower bound if we can prove that

$$G_{\bar{u},d} = \sum_{i=1}^{k_*} \phi_{\bar{u}}(\tilde{\tau}_{k_*}, \tilde{\tau}_i) C^{'} C \phi_{\bar{u}}^{'}(\tilde{\tau}_{k_*}, \tilde{\tau}_i) \, (\tilde{\tau}_i - \tilde{\tau}_{i-1})$$

is lower bounded, for all subdivisions $\{\tilde{\tau}_i\}_{i\in\{0,\dots,k_*\}}$ such that $\tilde{\tau}_{i+1} - \tilde{\tau}_i \leq \mu$ and that[6] $T_* \leq \tilde{\tau}_{k_*} \leq T_* + \mu$.

Let us define $\psi_{\bar{u}}(t,s) = (\phi_{\bar{u}}^{-1}(t,s))^{'}$. Since $\phi_{\bar{u}}$ is the resolvent of $\frac{dx}{d\tau} = -\mathcal{A}^{'}x$, therefore $\psi_{\bar{u}}$ is the resolvent of $\frac{dx}{d\tau} = \mathcal{A}x$, (Cf. Appendix A.1).
Actually[7] $G_{\bar{u},d}$ is the Gramm observability matrix of the continuous discrete version of a system of the form (A.4).

Let us denote $G_{\bar{u}}(T)$ the grammian of (A.4) when the integral is computed from 0 to T. Let us apply Lemma 76 on $G_{\bar{u}}(T_*)$. There is a $a > 0$, independent from $\bar{u}$, such that $aId \leq G_{\bar{u}}(T_*) \leq G_{\bar{u}}(\tilde{\tau}_{k_*})$.

From Lemma 93 and Remark 94 there is a $K > 0$ such that

$$\begin{aligned} aId \quad \leq \quad G_{\bar{u}}(\tilde{\tau}_{k_*}) - G_{\bar{u},d} + G_{\bar{u},d} \quad &\leq \quad G_{\bar{u},d} + \mu K \tilde{\tau}_{k^*} Id \\ &\leq \quad G_{\bar{u},d} + \mu K (T_* + \mu) Id. \end{aligned}$$

Therefore

$$\left[a - \mu K (T_* + \mu)\right] Id \leq G_{\bar{u},d},$$

and μ can be shortened such that $(a - \mu K(T_* + \mu)) > 0$, independently from the subdivision $\{\tilde{\tau}_i\}$.

We conclude that there exists a $\alpha_2 > 0$ such that $\alpha_2 Id \leq (III)$.

Consequenltly $\alpha_2 Id \leq S_k(+)$, for all k such that $\tau_k \geq T_*$.

■

We finally state the equivalent of Lemma 88.

Lemma 96

Consider the prediction correction equation (B.1) and the hypothesis of Lemma 77. There exist two positive constants $\mu > 0$ and α such that $\alpha Id \leq S(\tau)$, for all $k \in \mathbb{N}$, and all $\tau \in [\tau_k, \tau_{k+1}[$, for all subdivisions $\{\tau_i\}_{i\in\mathbb{N}}$ such that $(\tau_i - \tau_{i-1}) \leq \mu$.
This inequality is then also true in the t time scale.

Lemma 77 is the sum of Lemmas 88 and 96.

[6] This second hypothesis implies that $k_* + 1$, the number of elements can vary from one subdivision to the other.

[7] Notice that with the $\psi_{\bar{u}}$ notation

$$G_{\bar{u},d} = \sum_{i=1}^{k_*} \psi_{\bar{u}}^{'}(\tilde{\tau}_i, \tilde{\tau}_{k_*}) C^{'} C \psi_{\bar{u}}(\tilde{\tau}_i, \tilde{\tau}_{k_*}) \, (\tilde{\tau}_i - \tilde{\tau}_{i-1}) \, .$$

B.2 Proofs of the Technical Lemmas

Lemma 97 *([38]) Let $\{x(t) > 0, t \geq 0\} \subset \mathbb{R}^n$ be absolutely continuous, and satisfying:*

$$\frac{dx(t)}{dt} \leq -k_1 x + k_2 x\sqrt{x},$$

for almost all $t > 0$, for k_1, $k_2 > 0$. Then, if $x(0) < \frac{k_1^2}{4k_2^2}$, $x(t) \leq 4\,x(0)\,e^{-k_1 t}$.

Proof.

We make three successive change of variables: $y = \sqrt{x}$, $z = \frac{1}{y}$ and $w(t) = e^{-\frac{k_1}{2}t} z(t)$. Then all the quantities $y(t)$ $z(t)$, $w(t)$ are positive and absolutely continuous on any finite time interval $[0, T]$. We have

$$\begin{array}{rclcl} \dot{y} & = & \frac{1}{2\sqrt{x}}\dot{x} & \leq & -\frac{k_1}{2}y + \frac{k_2}{2}y^2 \\ \dot{z} & = & -\frac{1}{y^2}\dot{y} & \geq & \frac{k_1}{2}z - \frac{k_2}{2} \\ \dot{w} & = & -\frac{k_1}{2}e^{-\frac{k_1}{2}t}z(t) + e^{-\frac{k_1}{2}t}\dot{z} & \geq & -\frac{k_2}{2}e^{-\frac{k_1}{2}t} \end{array}$$

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

$$w(t) \geq \frac{1}{\sqrt{x(0)}} - \frac{k_2}{k_1} + \frac{k_2}{k_1}e^{-\frac{k_1}{2}t}$$

If $\frac{1}{\sqrt{x(0)}} - \frac{k_2}{k_1} > 0$, then $w(t) > 0$ and we can go backwards in the previous inequalities:

$$\begin{array}{rcl} w(t) & \geq & \frac{1}{\sqrt{x(0)}} - \frac{k_2}{k_1}\left(1 - e^{-\frac{k_1}{2}t}\right) \\ z(t) & \geq & e^{\frac{k_1}{2}t}\left(\frac{1}{\sqrt{x(0)}} - \frac{k_2}{k_1}\right) + \frac{k_1}{2} \\ y(t) & \leq & \frac{1}{e^{\frac{k_1}{2}t}\left(\frac{1}{\sqrt{x(0)}} - \frac{k_2}{k_1}\right) + \frac{k_1}{2}} \\ x(t) & \leq & \frac{x(0)e^{-k_1 t}}{\left(1 - \sqrt{x(0)}\frac{k_2}{k_1}\right)^2} \end{array} .$$

Hence, if $x(0) \leq \frac{k_1^2}{4k_2^2}$, or $1 - \sqrt{x(0)}\frac{k_2}{k_1} \geq \frac{1}{2}$, then:

$$x(t) \leq 4x(0)e^{-k_1 t}$$

■

Lemma 98 *([38])*

Consider $\widetilde{b}(\tilde{z}) - \widetilde{b}(\tilde{x}) - \widetilde{b}^*(\tilde{z})\,\tilde{\varepsilon}$ that appears in the inequality (3.12) (omitting to write u in $\tilde{b}$) and suppose $\theta \geq 1$. Then $\left|\widetilde{b}(\tilde{z}) - \widetilde{b}(\tilde{x}) - \widetilde{b}^*(\tilde{z})\,\tilde{\varepsilon}\right| \leq K\theta^{n-1}\,|\tilde{\varepsilon}|^2$, for some $K > 0$.

Proof.

Let us denote $\varepsilon = z - x$ and consider a smooth expression $E(z, x)$ of the form:

$$E(z, x) = f(z) - f(x) - df(z)\varepsilon,$$

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

We have, for $t > 0$:

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

and:

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

Hence,

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

Since f is compactly supported, we get :

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

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

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

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

This gives the result.

■

Appendix C

Source Code for Realtime Implementation

Contents

C.1 Replacement Code for the File: *rtai4_comedi_datain.sci*

```
function [x,y,typ] = rtai4_comedi_datain(job,arg1,arg2)
  x=[];y=[];typ=[];
  select job
  case 'plot' then
    exprs=arg1.graphics.exprs;
    ch=exprs(1)
    name=exprs(2)
    standard_draw(arg1)
  case 'getinputs' then
    [x,y,typ]=standard_inputs(arg1)
  case 'getoutputs' then
    [x,y,typ]=standard_outputs(arg1)
  case 'getorigin' then
    [x,y]=standard_origin(arg1)
  case 'set' then
    x=arg1
    model=arg1.model;graphics=arg1.graphics;
    exprs=graphics.exprs;
    while %t do
      [ok,ch,name,range,aref,exprs]=getvalue('Set RTAI-..
          COMEDI DATA block parameters',['Channel:';'Device:..
          ';'Range:';'Aref:'],list('vec',-1,'str',1,'vec'..
          ,-1,'vec',-1),exprs)
      if ~ok then break,end
      if exists('outport') then out=ones(outport,1), in=[],
                                else out=1, in=[], end
      [model,graphics,ok]=check_io(model,graphics,in,out..
          ,1,[])
      if ok then
        graphics.exprs=exprs;
        model.ipar=[ch;
                    range;
                    aref;
                    length(name);
                    ascii(name)'];
        model.rpar=[];
        model.dstate=[1];
        x.graphics=graphics;x.model=model
        break
      end
    end
  case 'define' then
    ch=0
    name='comedi0'
    range=0
    aref=0
    model=scicos_model()
    model.sim=list('rt_comedi_datain',4)
```

```
    if exists('outport') then model.out=ones(outport,1), ..
        model.in=[],
                          else model.out=1, model.in=[], end
    model.evtin=1
    model.rpar=[]
    model.ipar=[ch;
                range;
                aref;
                length(name);
                ascii(name)']
    model.dstate=[1];
    model.blocktype='d'
    model.dep_ut=[%t %f]
    exprs=[sci2exp(ch),name,sci2exp(range),sci2exp(aref)]
    gr_i=['xstringb(orig(1),orig(2),[''COMEDI A/D'';name+'' ..
        CH-''+string(ch)],sz(1),sz(2),''fill'');']
    x=standard_define([3 2],model,exprs,gr_i)
  end
endfunction
```

C.2 Computational Function for the Simulation of the DC Machine

This first code are given only to provide a complete picture with regard to the implementation. This simulation presents no difficulties. Let us remark that an alternative solution to the definition of the parameter in the code itself is to pass them through the *real parameter vector* entry of the interfacing function, see [41].

function name	DCmachine	number of zero crossing surfaces	0
implicit	n	initial discrete state	[]
input port size	2	real parameter vector	[]
output port size	2	integer parameter vector	[]
input event port size	[]	initial firing vector	[]
output event port size	[]	direct feed through	y
initial continuous state	⊛	time dependence	y

Table C.1: Arguments of the DC Simulation.
$\circledast = [I(0), \omega_r(0)]$.

```
#include <math.h>
#include <stdlib.h>
#include <scicos/scicos_block.h>
void DCmachine(scicos_block *block,int flag)
{ /* this is the list of the fields we have to use
```

```
  int block->nevprt;
  int block->nz;
  double* block->z;
  int block->nx;
  double* block->x;
  double* block->xd;
  double* block->res;
  int block->nin;
  int *block->insz;
  double **block->inptr;
  int block->nout;
  int *block->outsz;
  double **block->outptr;
  int block->nevout;
  int block->nrpar;
  double *block->rpar;
  int block->nipar;
  int *block->ipar;
  int block->ng;
  double *block->g;
  int *block->jroot;
  char block->label[41];
  */
  if (flag == 4) { /* initialization */
   DCmachine_bloc_init(block,flag);
  } else if(flag == 1) { /* output computation*/
   set_block_error(DCmachine_bloc_outputs(block,flag));
  } else if(flag == 0) { /* derivative or residual ..
      computation*/
   set_block_error(DCmachine_bloc_deriv(block,flag));
  } else  if (flag == 5) { /* ending */
     set_block_error(DCmachine_bloc_ending(block,flag));
  }
}

int DCmachine_bloc_init(scicos_block *block,int flag)
{return 0;}

int DCmachine_bloc_outputs(scicos_block *block,int flag)
{block->outptr[0][0]=block->x[0];
block->outptr[0][1]=block->x[1];
return 0;}

int DCmachine_bloc_deriv(scicos_block *block,int flag)
{double L=1.22, Res=5.4183, Laf=0.068;
double J=0.0044, B=0.0026, p=0;
double I=block->x[0], wr=block->x[1];
double V=block->inptr[0][0], Tl=block->inptr[0][1];
block->xd[0]=(V-Res*I-Laf*wr*I)/L;
block->xd[1]=(Laf*I*I-B*wr-p*pow(wr,2.08)-Tl)/J;
return 0;}
```

```
int DCmachine_bloc_ending(scicos_block *block,int flag)
{return 0;}
```

C.3 AEKF Computational Functions C Code

In the following code, we used the notation:

- the array $Ab[.]$ is composed of the elements of the matrix $A(u) + b^*(x, u)$,
- the variables of the form Pij are the elements of the matrix P,
- the tuning matrices of the EKF are set to $r = 1$ and $Q = diag(\{q_1, q_2, q_3\})$.

Those two matrices are organized as follows:

$$\begin{pmatrix} Ab[0] & Ab[3] & 0 \\ Ab[1] & Ab[4] & Ab[7] \\ Ab[2] & Ab[5] & Ab[8] \end{pmatrix} \quad and \quad \begin{pmatrix} P11 & P21 & P31 \\ P21 & P22 & P23 \\ P31 & P23 & P33 \end{pmatrix}$$

function name	AEKF	number of zero crossing surfaces	0
implicit	n	initial discrete state	[]
input port size	3	real parameter vector	[]
output port size	4	integer parameter vector	[]
input event port size	[]	initial firing vector	[]
output event port size	[]	direct feed through	y
initial continuous state	⊛	time dependence	y

Table C.2: Arguments of the Main Function.
⊛ = $[I(0); \omega_r(0); T_l(0); P(0); \theta(0)]$, initial state is a dim 10 vector.

```
#include <scicos/scicosblock.h>
#include <math.h>
#include <stdlib.h>
/*  input = [V;Y;I(t)] */
void AEKF(scicos_block *xblock,int flag)
{
int n=3;
int N=n*(n+1)/2;
if (flag==4)
{/* INITIALISATION */
/* change of variables from original coordinates -> normal ..
    form coordinates */
```

```
block->x[1]=block->x[1]*block->x[0];
block->x[2]=block->x[2]*block->x[0];
}
else if (flag==1)
{/* OUTPUT */
/* change of variables, normal form coordinates -> original ..
    coordinates */
block->outptr[0][0]=block->x[0];
block->outptr[0][1]=block->x[1]/block->x[0];
/* Estimated torque */
block->outptr[0][2]=block->x[2]/block->x[0];
/* High-gain parameter */
block->outptr[0][3]=block->x[n+N];
}
else if (flag==0)
{/* DERIVATIVE */
double p=0,K2=0.068, J=0.0044, B=0.0026;
double L=1.22, Res=5.4183, K1=0.068;
double DT=0.01,theta1=1.25,lambda=100,Beta=2000;
double m1=0.005,m2=0.004,m=m1+m2;
double V=block->inptr[0][0];
double Y=block->inptr[0][1];
double Inn=block->inptr[0][2];
double SI,F0;
double z1=block->x[0],z2=block->x[1];
double z3=block->x[2],theta=block->x[n+N];
double erreur=Y-z1;
double Ab[9];
double P11=block->x[3],P21=block->x[4],P31=block->x[5],P22=..
    block->x[6],P32=block->x[7],P33=block->x[8];
double q1=1, q2=0.1, q3=0.01;

if(Y<0.5){Y=0.5}; /* avoids ending with a value of I too ..
    close to 0 */

/* d(z)/dt=b(z,u)+A(u)*z+PC'R^-1(Y-Cz) */
block->xd[0]=(V-Res*z1-K1*z2)/L+theta*block->x[n+0]*erreur;..
    //dz1/dt=...+P11*erreur
block->xd[1]=(V*z2/z1-Res*z2-K1*z2*z2/z1)/L+(K2*z1*z1*z1-B*..
    z2-z3-p*pow(z2,2.08)/pow(z1,1.08))/J+theta*block->x[n..
    +1]*erreur;//dz2/dt=...+P21*erreur
block->xd[2]=(V*z3/z1-Res*z3-K1*z2*z3/z1)/L+theta*block->x[n..
    +2]*erreur;//dz3/dt=...+P31*erreur

/* Definition of the matrix Ab=(A(u)+bstar) */
Ab[0]=-Res/L;//b(1,1)
Ab[1]=(K1*z2*z2-V*z2)/(L*z1*z1)+(3*K2*z1*z1+1.08*p*pow(z2/z1..
    ,2.08))/J;//b(2,1)
Ab[2]=(K1*z2*z3-V*z3)/(L*z1*z1);//b(3,1)
Ab[3]=-K1/L;
```

```
Ab[4]=((V-2*K1*z2)/z1-Res)/L-(B+2.08*p*pow(z2/z1,1.08))/J;//..
    b(2,2)
Ab[5]=-K1*z3/(L*z1);//b(3,2)
//Ab[6]=0;
Ab[7]=-1/J;
Ab[8]=((V-K1*z2)/z1-Res)/L;//b(3,3)

/* Computation of the Riccati equation */
block->xd[3]=2*(Ab[0]*P11+Ab[3]*P21)-theta*P11*P11+theta*q1;
block->xd[4]=Ab[1]*P11+Ab[4]*P21+Ab[7]*P31+Ab[0]*P21+Ab[3]*..
    P22-theta*P11*P21;
block->xd[5]=Ab[2]*P11+Ab[5]*P21+Ab[8]*P31+Ab[0]*P31+Ab[3]*..
    P32-theta*P11*P31;
block->xd[6]=2*(Ab[1]*P21+Ab[4]*P22+Ab[7]*P32)-theta*P21*P21..
    +pow(theta,3)*q2;
block->xd[7]=Ab[2]*P21+Ab[5]*P22+Ab[8]*P32+Ab[1]*P31+Ab[4]*..
    P32+Ab[7]*P33-theta*P31*P21;
block->xd[8]=2*(Ab[2]*P31+Ab[5]*P32+Ab[8]*P33)-theta*P31*P31..
    +pow(theta,5)*q3;

/* Computation of the adaptation function */
SI=1/(1+exp(-Beta*(Inn-m)));
F0=(theta<=theta1)?(1/delT*theta*theta):(1/delT*pow(theta-2*..
    theta1,2));

/* NOTE a?b:c means if a then b else c */

block->xd[9]=SI*F0+lambda*(1-SI)*(1-theta);}
elseif (flag==5)
{/* ENDING */}
}
```

C.4 Innovation Computational Functions C Code

In the code below, the vector containing the system output past values and, the system input past values are not updated at the same place.

Let us suppose that we want to compute $\mathcal{I}_d(t)$. The computation is done via a discrete process with sample time δ. Therefore we need $\frac{d}{\delta}+1$ values for the output signal (i.e. $y(t-d)$, $y(t-d+\delta)$),...,$y(t)$), and $\frac{d}{\delta}$ values for the input signal[1] (i.e. $u(t-d)$,...,$u(t-\delta)$) in order to compute a trajectory[2].
Therefore the computation of *innovation* at time t requires:

- to update the vector of past values of y from $[y(t-d-\delta),...,y(t-\delta)]$ to $[y(t-d),...,y(t)]$,
- to compute a prediction of the state trajectory with the help of the vectors $[y(t-d),...,y(t)]$ and $[u(t-d),...,u(t-\delta)]$,

[1] Knowing $u(t)$ is useless here since we don't want to predict any trajectory for times higher than t.

[2] Remember that the initial point of this prediction is the estimated state at time $t-d$. It is fed to the function via a delay block.

– to update the vector of past values of u from $[u(t-d),...,u(t-\delta)]$ to $[u(t-d+\delta),...,u(t)]$.

function name	innovation	number of zero crossing surfaces	0
implicit	n	initial discrete state	⊛
input port size	5	real parameter vector	[]
output port size	1	integer parameter vector	[]
input event port size	[1]	initial firing vector	[]
output event port size	[]	direct feed through	y
initial continuous state	[]	time dependence	n

Table C.3: Arguments of the Innovation Function.
⊛ = a null vector of length 2*(d/Dt)+2, (i.e. 22).

```
#include <math.h>
#include <stdlib.h>
#include <scicos/scicos_block.h>

/*  input =  [Y;V;z(t-d)] */

void function(double *,double *);

void innovation(scicos_block *block,int flag)
{
double d=1,Dt=0.1;
int n=10;/*(d/Dt);*/  /* Auto conversion double -> int */
int nz,i,j;
/* Remark that there is NO TEST:Dt has to divide d!! */
nz=2*n+2;
if (flag==4)
{/* INITIALISATION */
/* Already done. It is impossible to change the dimension of..
     the state space from here */
}

else if(flag==1|flag==6)
{/* OUTPUT */
block->outptr[0][0]=block->z[nz-1];}

else if(flag==2)
{/* DERIVATIVE */
/* delayed input (index 2,3,4) */
double I,wr=block->inptr[0][3],Tl=block->inptr[0][4];
if (block->inptr[0][2]>0.5){I=block->inptr[0][2];}
else{I=0.5;}
double Y=block->inptr[0][0], V=block->inptr[0][1];
```

```
double x[3]={I,I*wr,Tl*I},xh[4]={0,0,0,0};
/* those ZEROs will be redefined later on */

double preInn[2]={pow((block->z[0]-x[0]),2),0};
double Inn=0; /* this is the INNOVATION */
double hh=Dt*0.5,h6=Dt/6;
double dx1[3],dx2[3],dx3[3],dx4[3];
int neq=3;/* number of ODEs to solve */

/* Update of the output signal stack */
for(i=1;i<=n;i++){block->z[i-1]=block->z[i];}
block->z[n]=Y;

for(j=1;j<=n;j++){
    V=block->z[n+j];
    xh[3]=V;
/* BEGIN RK4 */
    for(i=0;i<neq;i++){xh[i]=x[i];}
    function(&xh[0],&dx1[0]);
    for(i=0;i<neq;i++){xh[i]=x[i]+hh*dx1[i];}
    function(&xh[0],&dx2[0]);
    for(i=0;i<neq;i++){xh[i]=x[i]+hh*dx2[i];}
    function(&xh[0],&dx3[0]);
    for(i=0;i<neq;i++){xh[i]=x[i]+Dt*dx3[i];}
    function(&xh[0],&dx4[0]);
    for(i=0;i<neq;i++){x[i]=x[i]+h6*(dx1[i]+dx4[i]+2*dx2[i..
        ]+2*dx3[i]);}
/* END RK4 */
    preInn[1]=pow(block->z[j]-x[0],2);
    Inn=Inn+(preInn[0]+preInn[1])*Dt/2;
    preInn[0]=preInn[1];
}
/* Update of the input signal stack */
for(i=1;i<=n-1;i++){block->z[n+i]=block->z[n+i+1];}
block->z[n+n]=V;
block->z[nz-1]=Inn;
}
else if(flag==5)
{/* ENDING */}
} /* end of <innovation>  */

void function(double *xc,double *xcdot)
{double z1,z2,z3,V;
/* This function is called by <innovation>
The model parameters are defined at the beginning of
<innovation>: there is no need to redefine them.
The corresponding variable are declared globally, outside of
the function <innovation> */
double L=1.22, Res=5.4183, K1=0.068;
double K2=0.068, J=0.0044, B=0.0026, p=0;
z1=*(xc);
```

```
z2=*(xc+1);
z3=*(xc+2);
V=*(xc+3);
*(xcdot)=(V-Res*z1-K1*z2)/L;
*(xcdot+1)=(V*z2/z1-Res*z2-K1*z2*z2/z1)/L+(K2*z1*z1*z1-B*z2-..
    z3-p*pow(z2,2.08)/pow(z1,1.08))/J;
*(xcdot+2)=(V*z3/z1-Res*z3-K1*z2*z3/z1)/L;
return;}
```

C.5 Ornstein-Ulhenbeck Process

The facts compiled in this section are mainly taken from the book [28] and the article [59]. A note from S. Finch ([1]) and the book [100] were also convenient sources of information.

Introductory books on probability and stochastic processes can also be useful (e.g. [29, 49, 81])[3].

A stochastic process represents the state of a system that depends both on time and on random events. It is represented as a collection of random variables indexed by the time. We denote it $\{X_t : t \geq 0\}$.

Definition 99

- *A **Brownian motion or Wiener process with variance parameter** σ^2, starting at 0, is a stochastic process $\{B_t : t \geq 0\}$ taking values in $\mathbb{R}$, and having the properties:*
  1. $B_0 = 0$,
  2. *for any $t_1 < t_2 < ... < t_n$, the variables B_{t_1}, $B_{t_2} - B_{t_1}$, ..., $B_{t_n} - B_{t_{n-1}}$ are independents,*
  3. *for any $s < t$, the random variable $B_t - B_s$ has a normal distribution with mean 0 and variance $(t-s)\sigma^2$,*
  4. *the function $t \mapsto B_t$ is continuous.*
- *A **Ornstein-Ulhenbeck process** is a stochastic process that is solution of a stochastic differential equation of the form[4]:*

$$dX_t = -\rho X_t dt + \alpha dB_t \tag{C.1}$$

where B_t is a Brownian motion, and ρ and α are positive constants[5].

[3] [28] and [29] are in French.

[4] This equation is called the Langevin equation, see [82], the english translation of the original article from 1908.

[5] A more general definition of a Ornstein-Ulhenbeck, or stationary Gauss-Markov process is:

- stationary,
- Gaussian,
- Markovian,
- continuous in probability.

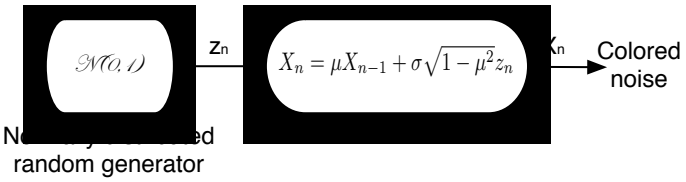

Figure C.1: Simulation of colored noise via block wise programing.

The law of the random variable at time 0 *is supposed the invariant probability associated to (C.1) such that* X_t *is a stationary process.*

Equation (C.1) can be rewritten:

$$d\left[e^{\rho t}X_t\right] = \alpha e^{\rho t} dB_t,$$

and then

$$X_t = e^{-\rho t}\left[X_0 + \int_0^t \alpha e^{\rho s} dB_s\right].$$

The stochastic process X_t is such that [28]:

1. If X_0 is a gaussian variable with zero mean and variance equals to $\frac{\alpha^2}{2\rho}$, then $X(t)$ is gaussian and stationary of covariance:

$$\mathbb{E}[X(t)X(s+t)] = \frac{\alpha^2}{2\rho} e^{-r\rho|s|}.$$

2. $X(t)$ is a markovian process.

3. When $X(0) = c$, the stochastic law of $X(t)$ is a normal law with mean $e^{-\rho t}c$, and variance $\frac{\alpha^2}{2\rho}\left(1 - e^{2\rho t}\right)$.

As explained in [59], Part 3, equation (3.15), for a sample time$\Delta\ t$, the exact updating formulas of the Ornstein-Uhlenbeck process is given by:

$$X(t + \Delta t) = X(t)\mu + \gamma z_n \tag{C.2}$$

where:

- $\mu = e^{-\rho\Delta t}$,
- $\gamma^2 = \left(1 - e^{-2\rho\Delta t}\right)\left(\frac{\alpha^2}{2\mu}\right)$,
- z_n is the realization of a standard gaussian distribution (i.e. $\mathcal{N}(0,1)$).

Let us now rewrite equation (C.1) as in [100], with the help of the positive constant parameters β and σ:

$$dX_t = -\beta X_t dt + \sigma\sqrt{2\beta}dB_t.$$

Therfore, with $X(0) = 0$, the mean value of X_t is 0 and the term $\frac{\alpha^2}{2\rho}$ of the variance, becomes σ^2.

In order to make a simulation of the Ornstein-Uhlenbeck process, we consider equation (C.2) and replace:

- α by $\sigma\sqrt{2\beta}$,
- ρ by β.

The update equation is (with $\mu = e^{-\beta\Delta t}$)

$$X(t + \Delta t) = X(t)\mu + \sigma(1 - \mu^2)z_n.$$

The actual simulation is done according to the diagram of Figure C.1, with $X(0) = 0$, $\mu \in]0; 1[$ and $\sigma > 0$. Remark that σ is the asymptotic standard deviation of the variables $X(t), t > 0$.

An important theorem due to J. L. Doob [48] ensures that such a process necessarily satisfies a linear stochastic differential equation identical to the one used in the definition above.

References

[1] *http://algo.inria.fr/bsolve/*, 2010. 162

[2] *http://maxima.sourceforge.net/*, 2010. 87

[3] *http://www.lar.deis.unibo.it/people/gpalli*, 2010. 78

[4] *www.comedi.org*, 2010. 77

[5] *www.rtai.org*, 2010. 77

[6] *www.scicoslab.org*, 2010. 77

[7] *www.scicos.org*, 2010.

[8] *www.scilab.org*, 2010. 77

[9] R. Abraham and J. Robbin. *Transversal Mappings and Flows*. W. A. Benjamin, inc., 1967. 13

[10] I. Abuhadrous. *Onboard real time system for 3D localization and modelling by multi-sensor data fusion.* PhD thesis, Mines ParisTech (ENSMP), 2005. 103

[11] J. H. Ahrens and H. K. Khalil. High-gain observers in the presence of measurement noise: A switched-gain approach. *Automatica*, 45:936–943, 2009. 27, 29, 116

[12] M. Alamir. *Nonlinear Observers and Applications*, volume 363 of *LNCIS*, chapter Nonlinear Moving Horizon Observers: Theory and Real-Time Implementation. Springer, 2007. 3, 40

[13] H. Souley Ali, M. Zasadzinski, H. Rafaralahy, and M. Darouach. Robust $\mathcal{H}_\infty$ reduced order filtering for unertain bilinear systems. *Automatica*, 42:405–415, 2005. 32

[14] S. Ammar. Observability and observateur under sampling. *International Journal of Control*, 9:1039–1045, 2006. 104

[15] S. Ammar and J-C. Vivalda. On the preservation of observability under sampling. *Systems and control letters*, 52:7–15, 2003. 103, 104

[16] V. Andrieu, L. Praly, and A. Astolfi. High-gain observers with updated gain and homogeneous correction terms. *Automatica*, 45(2), 2009. 25, 26

[17] J. S. Baras, A. Bensoussan, and M. R. James. Dynamic observers as asymptotic limits of recursive filters: Special cases. *SIAM Journal of Applied Mathematics*, 48(5):1147–1158, 1988. 38

[18] Y. Becis-Aubry, M. Boutayeb, and M. Darouach. State estimation in the presence of bounded disturbances. *Automatica*, 44:1867–1873, 2008. 33

[19] G. Besançon. *Nonlinear observers and Applications*, volume 363 of *LNCIS*, chapter An Overview on Observer Tools for Nonlinear Systems. Springer, 2007. 13, 93, 116

[20] G. Besançon, G. Bornard, and H. Hammouri. Observer synthesis for a class of nonlinear control systems. *European Journal of Control*, 2:176–192, 2996. 93

[21] N. Boizot and E. Busvelle. *Nonlinear Observers and Applications*, volume 363 of *LNCIS*, chapter Adaptive-gain Observers and Applications. Springer, 2007. xx, 2, 4, 7, 15, 71

[22] N. Boizot, E. Busvelle, and J-P. Gauthier. An adaptive high-gain observer for nonlinear systems. *Automatica (Accepted for publication)*. xx, 44, 71

[23] N. Boizot, E. Busvelle, and J-P. Gauthier. Adaptive-gain in extended kalman filtering (oral presentation). In *International conference on Differential equations and Topology, MSU, Moscow, Fed. of Russia*, 2008. xx, 44

[24] N. Boizot, E. Busvelle, and J-P. Gauthier. Adaptive-gain extended kalman filter: Extension to the continuous-discrete case. In *Proceedings of the 10th European Control Conference (ECC09)*, 2009. 106

[25] N. Boizot, E. Busvelle, J-P. Gauthier, and J. Sachau. Adaptive-gain extended kalman filter: Application to a series connected dc motor. In *Conference on Systems and Control, Marrakech, Morocco*, May 16-18 2007. xx, 76

[26] N. Boizot, E. Busvelle, and J. Sachau. High-gain observers and kalman filtering in hard realtime. In *9th Real-Time Linux Workshop, Institute for Measurement Technology, Johannes Kepler University of Linz*, November 2-4 2007. xx

[27] G. Bornard. Observability and high gain observer for multi-output systems. In 28^{th} *Summer School in Automatics, Gipsa Lab-Grenoble, France (www.gipsa-lab.inpg.fr/summerschool/session/s_28_uk.php), Lecture* 3, 2007. 93

[28] N. Bouleau. *Processus Stochastiques et Applications (Nouvelle Edition)*. Hermann, 2000. 162, 163

[29] N. Bouleau. *Probabilités de l'ingénieur, Variables Aléatoires et simulation (Nouvelle Edition)*. Hermann, 2002. 162

[30] M. Boutayeb and D. Aubry. A strong tracking extended kalman observer for nonlinear discrete-time systems. *IEEE Transactions on Automatic Control*, 44(8):1550–1556, 1999. 32, 33, 34

[31] M. Boutayeb, H. Rafaralahy, and M. Darouach. Convergence analysis of the extended kalman filter as an observer for nonlinear discrete-time systems. In *Proceedings of the* 34^{th} *Conference on Decision and Control*, 1995. 32, 33

[32] K. J. Bradshaw, I. D. Reid, and D. W. Murray. The active recovery of 3d motion trajectories and their use in prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):219–233, 1997. 32

[33] H. Brezis. *Analyse Fonctionnelle: Théorie et applications*. Dunod, 1999. 125, 126

[34] R. Bucher, S. Mannori, and T. Netter. *RTAI-Lab tutorial: Scilab, Comedi, and real-time control*, 2008. 77, 78, 79, 84

[35] E. Bullinger and F. Allgőwer. An adaptive high-gain observer for nonlinear systems. In *Proceedings of the* 36^{th} *Conference on Decision and Control*, 1997. 24, 26

[36] E. Bullinger and F. Allgőwer. Adaptive λ-tracking for nonlinear higher relative degree systems. *Automatica*, 41:1191–1200, 2005. 24

[37] F. Busse and J. How. Demonstration of adaptive extended kalman filter for low earth orbit formation estimation using cdgps. In *Institute of Navigation GPS Meeting*, 2002. 22, 32

[38] E. Busvelle and J-P. Gauthier. *High-gain and Non High-gain Observer for Nonlinear Systems*. World Scientific, 2002. 11, 17, 18, 22, 35, 36, 38, 40, 49, 50, 52, 63, 65, 67, 102, 151

[39] E. Busvelle and J-P. Gauthier. On determining unknown functions in differential systems, with an application to biological reactors. *COCV*, 9:509, 2003.

[40] E. Busvelle and J-P. Gauthier. Observation and identification tools for nonlinear systems. application to a fluid catalytic cracker. *International Journal of Control*, 78(3), 2005. xi, 11, 14, 15, 36

[41] S. L. Campbell, J-P. Chancelier, and R. Nikoukhah. *Modeling and Simulation in Scilab/Scicos*. Springer, 2006. 77, 86, 155

[42] H. Carvalho, P. Del Moral, A. Monin, and G. Salut. Optimal nonlinear filtering in gps/ins integration. *IEEE Transactions on Aerospace and Electronic Systems*, 33(3), 1997. 7

[43] C. K. Chui and G. Chen. *Kalman Filtering with Real-Time Applications,* 3^{rd} *Edition*. Springer, 1998. x, 5, 7, 30, 61

[44] P. Connolly. Hyperion prop talk. Technical report, http://aircraft-world.com/prod_datasheets/hp/emeter/hp-proptalk.htm. 79, 81

[45] B. d'Andrea Novel and M. Cohen de Lara. *Cours d'automatique : commande linéaire des systèmes dynamiques*. Masson, 2000. 5, 7

[46] F. Deza, E. Busvelle, and J-P. Gauthier. Exponentially converging observers for distillation columns and internal stability of the dynamic output feedback. *Chemical Engineering Science*, 47(15/16), 1992. 20

[47] F. Deza, E. Busvelle, J-P. Gauthier, and D. Rakotopara. High-gain estimation for nonlinear systems. *Systems and control letters*, 18:295–299, 1992. x, 11, 20, 21, 22, 38, 40

[48] J. L. Doob. The brownian movement and stochastic equations. *Annals of Math.*, 43:351–369, 1942. 164

[49] R. Durett. *Stochastic Calculus, A practical introduction*. CRC Press, 1996. 162

[50] F. Esfandiari and H. K. Khalil. Output feedback stabilization of fully linearizable systems. *International Journal of Control*, 56:1007–1037, 1992. 20, 28

[51] M. Farza, M. M'Saad, and L. Rossignol. Observer design for a class of mimo nonlinear systems. *Automatica*, 40:135–143, 2004. 93

[52] J-P. Gauthier and G. Bornard. Observability for any $u(t)$ of a class of nonlinear systems. *IEEE Transactions on Automatic Control*, 26(4):922–926, 1981. 11

[53] J-P. Gauthier, H. Hammouri, and I. Kupka. Observers for nonlinear systems. In *IEEE CDC conference*, pages 1483–1489, 1991. 11

[54] J-P. Gauthier, H. Hammouri, and S. Othman. A simple observer for nonlinear systems applications to bioreactors. *IEEE Transactions on Automatic Control*, 37(6):875–880, 1992. x, 11, 20

[55] J-P. Gauthier and I. Kupka. Observability and observers for nonlinear systems. *SIAM Journal on Control*, 32(4):975–994, 1994.

[56] J-P. Gauthier and I. Kupka. Observability with more outputs than inputs. *Mathematische Zeitschrift*, 223:47–78, 1996. 11

[57] J-P. Gauthier and I. Kupka. *Deterministic Observation Theory and Applications*. Cambridge University Press, 2001. x, xi, 4, 11, 14, 15, 17, 21, 40, 42, 47, 93, 96, 98, 115, 116, 125, 128, 131, 144

[58] A. Gelb, editor. *Applied Optimal Estimation*. The MIT Press, 1974. x, xvii, 3, 5, 6, 20, 30

[59] D. T. Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical Review E*, 54(2):2084–2091, 1996. 162, 163

[60] M. S. Grewal, L. Weill, and A. P. Andrews. *Global Positioning Systems, Inertial Navigation and Integration*. John Wiley and Sons, 2007. 20, 22, 30

[61] L. Z. Guo and Q. M. Zhu. A fast convergent extended kalman observer for nonlinear discrete-time systems. *International of Systems Science*, 33(13):1051–1058, 2002. 34

[62] T. Hagglund. *New Estimation Techniques for Adaptive Control*. PhD thesis, Lund Institue of Technology, 1985. 31

[63] H. Hammouri. *Nonlinear Observers and Applications*, volume 363 of *LNCIS*, chapter Uniform Observability and Observer Synthesis. Springer, 2007. 93

[64] S. Haugwitz, P. Hagander, and T. Norén. modeling and control of a novel heat exchange reactor, the open plate reactor. *Control Engineering Practices*, 15:779–792, 2007. xi, 22

[65] R. Hermann and A. J. Krener. Nonlinear controlability and observability. *IEEE Transactions on Automatic Control*, AC-22:728–740, 1977. 11

[66] M. W. Hirsch. *Differential Topology*. Springer-Verlag, 1972. 12

[67] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985. 108, 132

[68] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Publications, INC, 1970. x

[69] L. Jetto and S. Longhi. Development and experimental validation of an adaptive extended kalman filter for the localization of mobile robots. *IEEE Transactions on Robotics and Automation*, 15(2):219–229, 1999. 32

[70] 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. 11

[71] S. Julier and J. K. Ulhmann. A new extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI; Proceedings of the Conference, Orlando*, pages 182–193, 1997. 116

[72] D. J. Jwo and F. Chang. *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, chapter A Fuzzy Adaptive Fading Kalman Filter for GPS Navigation, pages 820–831. LNCIS. Springer, 2007. 23

[73] Radhakrishnan K. and Hindmarsh A. C. Description and use of lsode, the livermore solver for ordinary differential equations. Technical Report UCRL-ID-113855, LLNL, 1993. 87

[74] T. Kailath. *Linear Systems*, volume xxi. Prentice-Hall, New Jersey, 1980. 3

[75] R. E. Kalman. A new approach to linear filtering. *Transactions of the ASME - Journal of Basic Engineering*, 82 (Series D):35–45, 1960. 5

[76] R. E. Kalman and B. S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83:95–108, 1961. 5

[77] P. C. Krause, O. Wasynczuk, and S. D. Sudhoff. *Analysis of Electric Machinery and Drive Systems, 2nd Edition*. Wiley-interscience, 2002. 57

[78] P. Krishnamurthy and F. Khorrami. Dynamic high-gain scaling: State and output feedback with application to systems with iss appended dynamics driven by all states. *IEEE Transactions on Automatic Control*, 49(12):2219–2239, 2004. 25

[79] P. Krishnamurthy, F. Khorrami, and R. S. Chandra. Global high-gain based observer and backstepping controler for generalized output-feedback canonical form. *IEEE Transactions on Automatic Control*, 48(12), 2003. 25

[80] H. J. Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control AC*, 12(5), 1967. 7

[81] G. F. Lawler. *Introduction to Stochastic Processes, Second Edition*. Chapman and Hall/CRC, 2006. 162

[82] D. S. Lemons and A. Gythiel. Paul langevin's 1908 paper "on the theory of brownian motion". *American Journal of Physics*, 65(11), 1997. 162

[83] V. Lippiello, B. Siciliano, and L. Villani. Adaptive extended kalman filtering for visual motion estimation of 3d objects. *Control Engineering Practices*, 15:123–134, 2007. 32

[84] F. L. Liu, M. Farza, M. M'Saad, and H. Hammouri. High gain observer based on coupled structures. In *Conference on Systems and Control, Marrakech, Morocco*, 2007. 93

[85] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification*. The MIT Press, 1983. 83

[86] D. G. Luenberger. Observers for multivariable systems. *IEEE Transactions on Automatic Control*, 11:190–197, 1966. 5

[87] P. Martin and P. Rouchon. Two remarks on induction motors. In *Symposium on Control, Optimization and Supervision*, 1996. 57

[88] The Mathworks, www.mathworks.com. *Optimization Toolbox for Use with Matlab, User's Guide*. 83

[89] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, 1979. 30

[90] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 2. Academic Press, 1982. 30, 32

[91] O. Mazenc and B. Olivier. Interval observers for planar systems with complex poles. In *European Control Conference*, 2009. 4

[92] R. K. Mehra. Approaches to adaptive filtering. *IEEE Transactions on Automatic Control*, 17(5):693– 698, 1972. 30, 32

[93] S. Mehta and J. Chiasson. Nonlinear control of a series dc motor: Theory and experiment. *IEEE Transactions on Industrial Electronics*, 45(1), 1998. 57, 58

[94] C. Melchiori and G. Palli. A realtime simulation environment for rapid prototyping of digital control systems and education. *Automazione e Strumentazione*, Febbraio 2007. 78

[95] H. Michalska and D. Q. Mayne. Moving horizon observers and observer based control. *IEEE Transactions on Automatic Control*, 40:995–1006, 1995. 40

[96] A. H. Mohamed and K. P. Schwarz. Adaptive kalman filtering for INS/GPS. *Journal of Geodesy*, 73:193–203, 1999. 22, 30

[97] L. La Moyne, L. L. Porter, and K. M. Passino. Genetic adaptive observers. *Engineering Applications of Artificial intelligence*, 8(3):261–269, 1995. 23

[98] K. S. Narendra and A. M. Annaswamy. *Stable Adaptive Systems*. Prentice-Hall, New Jersey, 1989. 3

[99] M. G. Pappas and J. E. Doss. Design of a parallel kalman filter with variable forgetting factors. In *American Control Conference*, pages 2368–2372, 1988. 30

[100] E. Pardoux. *Ecole d'été de Probabilités de Saint-Flour XIX*, volume 1464 of *LNM*, chapter Filtrage non linéaire et Equations aux Dérivées Partielles Stochastiques Associées. Springer-Verlag, 1989. 29, 162, 164

[101] J. Picard. Efficiency of the extended kalman filter for nonlinear systems with small noise. *SIAM Journal of Applied Mathematics*, 51(3):843–885, 1991. x, 7, 29, 38

[102] S. H. Pourtakdoust and H. Ghanbarpour. An adaptive unscented kalman filter for quaternion-based orientation estimation in low-cost ahrs. *International Journal of Aircraft Engineering and Aerospace Technology*, 79(5):485–493, 2007. 32

[103] L. Praly. Asymptotic stabilization via output feedback for lower triangular systems with output dependent incermental rate. *IEEE Transactions on Automatic Control*, 48(6), 2003. 25

[104] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 2007. 83, 87

[105] A. Rapaport and D.Dochain. Interval observers for biochemical processes with uncertain kinetics and inputs. *Mathematical Biosciences*, 193(2):235–253, 2005. 4

[106] H. Reinhard. *Equations Différentielles. Fondements et Applications*. Gauthier-Villars, 1982. 122

[107] W. Rudin. *Functional Analysis (2^{nde} edition)*. McGraw-Hill Science Engineering, 1991. 125, 126

[108] S. Sarkka. On unscented kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 52(9):1631–1641, 2007. 7, 117

[109] S. Stubberud, R. Lobbia, and M. Owen. An adaptive extended kalman filter using artificial neural networks. *The international journal on smart systems design*, 1:207–221, 1998. 23

[110] H. J. Sussmann. Single-input observability of continuous-tiime systems. *Mathematical systems Theory*, 12:371–393, 1979. 11

[111] A. Tornambè. High-gain observers for nonlinear systems. *International Journal of Systems Science*, 13(4):1475–1489, 1992. 20, 23, 24

[112] J. K. Ulhmann. Simultaneous map building and localization for real-time applications. Technical report, University of Oxford, 1994. 117

[113] F. Viel. *Stabilité des systèmes non linéaires contrôlés par retour d'état estimé. Application aux réacteurs de polymérisation et aux colonnes à distiller*. PhD thesis, Université de Rouen, Mont-Saint-Aignan, FRANCE, 1994. 4, 22, 93

[114] B. Vik, A. Shiriaev, and Thor I. Fossen. *New Directions in Nonlinear Observer Design*, volume 244 of *LNICS*, chapter Nonlinear Observer Design for Integration of DGPS and INS. Springer, 1999. 103

[115] K. C. yu, N. R. Watson, and J. Arrillaga. An adaptive kalman filter for dynamic harmonic state estimation and harmonic injection tracking. *Transactions on Power Delivery*, 20(2), 2005. 22, 23, 30

[116] A. Zemouche and M. Boutayeb. Observer synthesis method for a class of nonlinear discrete-time systems with extension to observer-based control. In *Proceedings of the* 17^{th} *World Congress of the International Federation of Automatic Control*, 2008. 32

[117] A. Zemouche, M. Boutayeb, and G. Iulia Bara. Observer design for nonlinear systems: An approach based on the differential mean value theorem. In *Proceedings of the* 44^{th} *Conference on Decision and Control and the European Control Conference*, 2005. 32