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

E. BUSVELLE, J.P.GAUTHIER

*Dedicated to Velimir Jurdjevic*

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

## 1. Introduction, systems under consideration

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

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

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

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

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

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

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

*Date*: December 2000.

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

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

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

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

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

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

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

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

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

**1. The extended Kalman Filter.**

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

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

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

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

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

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

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

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

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

## 2. The High Gain Extended Kalman Filter.

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

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

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

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

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

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

## 3. What we want to do in this paper.

The idea in this paper is the **very simple** following one: we give the parameter $\theta$ in the HGEKF an exponential decay from $\theta_0$ large, to 1.

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

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

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

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

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

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

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

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

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

## 2. Justification of the assumptions and observability

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

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

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

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

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

Basically, there are two cases.

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2. **Justification of the technical assumptions.** Let us consider successively the two technical assumptions we made in the section 1.1:

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

B. The functions $b_i$ are compactly supported.

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

Let us just notice the following.

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

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

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

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

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

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

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

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

## 3. Statement and proof of the theoretical result

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

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

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

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

**Comments:**

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

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

3. If $\theta_0$ is large, and if $\tau \leq T$, then, this equation is almost the equation of the high gain extended Kalman filter with gain $\theta(T)$. Hence, for $\tau \leq T$, setting $\varepsilon(\tau) = z(\tau) - x(\tau)$, ($\varepsilon$ is the **estimation error**), we can expect the following, for $\theta_0$ large enough in front of $T$:

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

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

In particular, this implies that the error $\varepsilon(t)$ **can be made arbitrarily small, in arbitrarily short time, increasing** $\theta_0$. For $\theta$ constant, this is the behavior of the "high gain extended Kalman filter. In that case ($\theta$ constant), this estimate follows from [10], [5]. We will prove it below for $\theta$ nonconstant.

Our main result herein will be the following:

**Theorem 1.** *1. For all* $0 \leq \lambda \leq \lambda_0$, *(*$\lambda_0 = \frac{Q_m \alpha}{4(n-2)}$, *where* $Q \geq Q_m Id$ *and* $\alpha$ *comes from Lemma 1 below), for all* $\theta_0$ *large enough, depending on* $\lambda$, *for all* $S_0 \geq c\ Id$, *for all* $K \subset \mathbb{R}^n$, $K$ *a compact subset, for all* $\varepsilon_0 = z_0 - x_0$, $\varepsilon_0 \in K$, *the following estimation holds, for all* $\tau \geq 0$ :

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

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

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

**Comments.**

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

b. This means that, provided that $\lambda$ is smaller than a certain constant $\lambda_0$, and $\theta_0$ is large in front of $\lambda$, the estimation error goes exponentially to zero, and can be made arbitrarily small in arbitrary short time.

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

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

## 3.2. **Proof of Theorem 1.**

### 3.2.1. *Preparation for the proof.*

Let us recall that:

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

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

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

The equations under consideration are:

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

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

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

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

An obvious computation gives:

$$\frac{d}{d\tau}(\tilde{\varepsilon}) = \theta[(A - \tilde{P}C'r^{-1}C)\tilde{\varepsilon} + \frac{1}{\theta}(\tilde{b}(\tilde{z}) - \tilde{b}(\tilde{x})) - \frac{\lambda(1-\theta)}{\theta^2}F\tilde{\varepsilon}], \tag{3.8}$$
$$\frac{d}{d\tau}(\tilde{S}) = \theta[-(A + \frac{1}{\theta}\tilde{b}^*(\tilde{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\theta)}{\theta^2})'\tilde{S}$$
$$- \tilde{S}(A + \frac{1}{\theta}\tilde{b}^*(\tilde{z}) - (\frac{Id}{2} + F)\frac{\lambda(1-\theta)}{\theta^2}) + C'r^{-1}C - \tilde{S}Q\tilde{S}], \tag{3.9}$$
$$\frac{d\theta}{d\tau} = \lambda(1-\theta).$$

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

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

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

to get the final set of equations:

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

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

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

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

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

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

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

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

Straightforward computations with (3.10) give:

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

where $Q \geq Q_m\ Id$.

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

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

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

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

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

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

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

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

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

**Step 2,** $t \leq T_0$.

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

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

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

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

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

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

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

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

Gronwall's inequality implies that:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

for another positive constant $k_2'$.

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

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

as soon as

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

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

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

If :

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

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

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

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

Then, if:

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

for $\theta_0$ large enough, for $||\varepsilon_0||$ bounded, $q$ is arbitrarily small.

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

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

$$\begin{aligned}\bar{\varepsilon}(t)'\bar{S}(t)\bar{\varepsilon}(t) &\leq 4e^{-k_1(t-T^*)}\bar{\varepsilon}(T^*)'\bar{S}(T^*)\bar{\varepsilon}(T^*),\\ &\leq 4e^{-k_1t}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1}{\lambda}}\bar{\varepsilon}(T^*)'\bar{S}(T^*)\bar{\varepsilon}(T^*).\end{aligned}$$

This implies, with (3.18):

$$\begin{aligned}||\bar{\varepsilon}(t)||^2 &\leq 4\frac{\beta}{\alpha}e^{-k_1t}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1}{\lambda}}||\bar{\varepsilon}(T^*)||^2,\\ &\leq 4\frac{\beta}{\alpha}H(T_0,c)e^{L'T^*}e^{-k_1t}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1}{\lambda}}||\varepsilon_0||^2,\\ &\leq 4\frac{\beta}{\alpha}H(T_0,c)e^{-k_1t}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,\end{aligned}$$

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

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

$$\begin{aligned}||\bar{\varepsilon}(t)||^2 &\leq H(T_0,c)e^{-k_1t}e^{L't}||\varepsilon_0||^2,\\ &\leq H(T_0,c)e^{-k_1t}4\frac{\beta}{\alpha}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,\end{aligned}$$

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

Therefore, for all $t \geq 0$ :

$$\begin{aligned}||\bar{\varepsilon}(t)||^2 &\leq 4\frac{\beta}{\alpha}H(T_0,c)e^{-k_1t}\Big(\frac{\theta_0-1}{\frac{2L'}{Q_m\alpha}-1}\Big)^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,\\ &\leq \tilde{H}(T_0,c,\lambda)e^{-k_1t}{\theta_0}^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,\end{aligned}$$

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

$$||\tilde{\varepsilon}(\tau)||^2 \leq \tilde{H}(T_0,c,\lambda)e^{-k_1t}{\theta_0}^{\frac{k_1+L'}{\lambda}}||\varepsilon_0||^2,$$

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

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

Finally,

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

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

This is the long term estimation. It holds as soon as $\lambda$ satisfies (3.23), and for $\theta_0$ large, depending on $\lambda$.

3.3. **Practical implementation: making the observer "recursive".** We consider a one parameter family $\{O_\tau, \tau \geq 0\}$ of observers of type (3.1), indexed by the time, each of them starting from $S_0$, $\theta_0$, at the current time $\tau$. In fact, in practice, it will be sufficient to consider, at time $\tau$, a slipping window of time, $[\tau - T, \tau[$, and a finite set of observers $\{O_{t_i}, \tau - T \leq t_i \leq \tau\}$, with $t_i = \tau - i\frac{T}{N}$, $i = 1, ..., N$.

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

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

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

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

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

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

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

c. the error can be made arbitrarily small in arbitrary short time, provided that $\theta_0$ is large enough.

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

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

## 4. Application: observation of a binary distillation column

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

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

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

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

The equations of this model are:
Total condenser:

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

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

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

Feed tray:

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

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

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

Bottom of the column:

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

The parameters have the following physical meaning:

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

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

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

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

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

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

There are also two disturbances to be counteracted:
a. changes in the feed rate $F = FL + FV$. In general this is a "measured disturbance", (a flowrate measurement),
b. the in-feed composition $Z_F$. In general, it is unknown, and it is practically very expensive to "observe it". Moreover, it may change brutally. We will consider this feed composition $Z_F$ as an unknown (constant) state variable. When $Z_F$ changes, the consequence is **a jump of the state of the system.**

The qualitative properties of this model are very nice (see [10], [18], [17]):
a. For positive control variables $L$ and $V$, (negative doesn't physically makes sense), the "physical" domain $D = [O, 1]^n$ is positively invariant under the dynamics. This means that all the state variables $x_j$ remain between 0 and 1.
b. In the domain $D$, all other variables (than the $x_i$'s and the $y_i$'s) being constant, **there is a single equilibrium, which is globally asymptotically stable.**
c. It has very nice observability properties, as will be discussed later on.

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

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

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

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

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

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

or:

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

where,

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

and,

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

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

The observations are then given by

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

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

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

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

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

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

Due to these observability properties, we will be able to apply the observer of the previous section 3.3. In fact, it will be an adaptation of the results of section 3, Theorem 1, to this multi-output case.

**We leave the reader to check (this is really straightforward) that all the reasoning in the proof of Theorem 1 can be strictly repeated, and that the statements of this theorem are valid for the distillation column.**

Of course, in practice, we didn't compute the theoretical bounds $\lambda_0$ and $\theta_0(\lambda)$. We have just got some values for them by experimentation. Also, the number $N$ of "parallel" observers, and the "sampling times" $t_i$ of section 3.3 have been chosen experimentally.

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

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

where $u = (L, V)$.

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

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

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

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

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

We have implemented our observer as described in the previous section. Since the state has dimension 6, each observer requires to solve 28 ordinary differential equations (for the state, the Riccati matrix, and the very simple equation for $\theta$).

Finally, our observer is a set of 140 ODE's. We have solved it in conjunction with the model (6 equations) using LSODAR from ODEPACK ([14]), without taking into account the possibility of decoupling these equations (which are indeed equivalent to five systems of 34 equations, including the model into each system). A simulation of 3 hours of real time takes about 40 seconds on a Pentium III machine.

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

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

We have applied the following scenario:

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

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

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

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

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

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

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

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

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

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

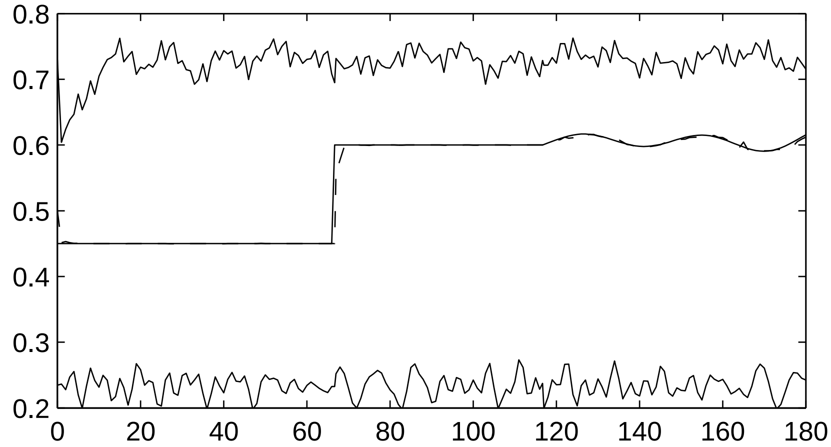

FIGURE 1. Measured output and estimation of the feed quality.

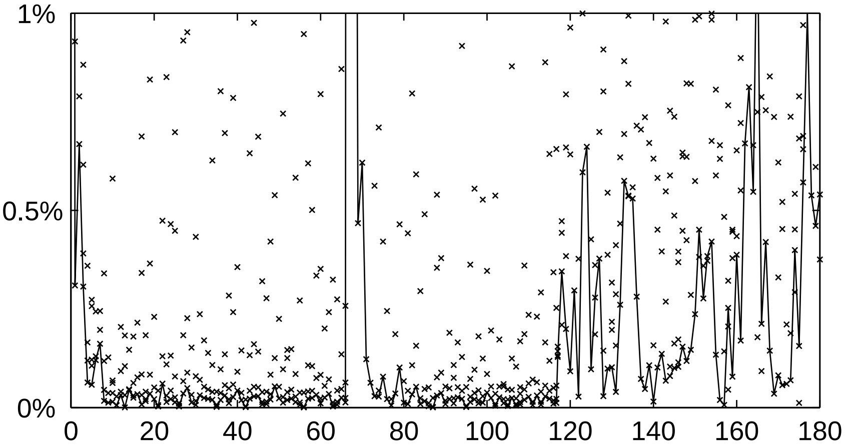

FIGURE 2. Relative error between the actual feed quality and its estimation by the selected observer (continuous line) and the others.

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

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

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

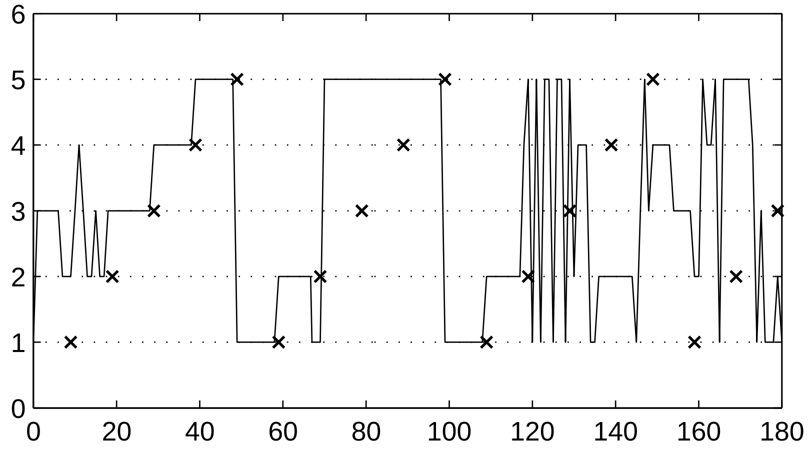

Figure 3. The 5 observers. Time of reinitialization of each observer (×), and the best one (continuous line).

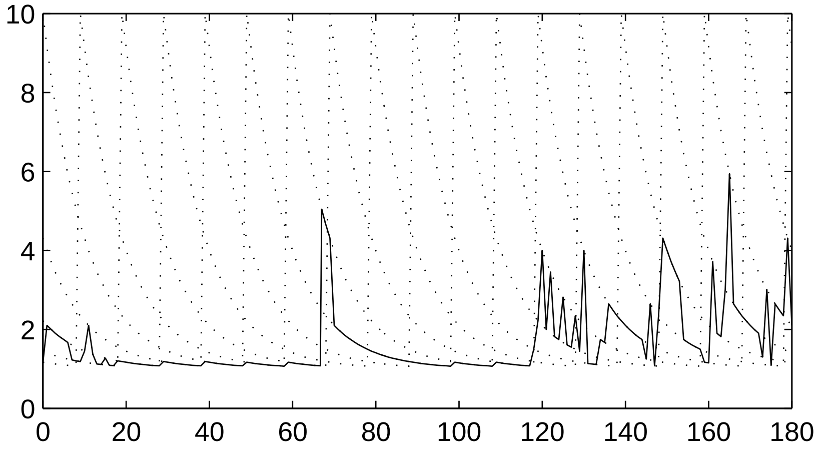

Figure 4. Various values of $\theta$ versus time (dotted lines), and best observer (continuous line).

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

Our conclusion, from these simulations, is that even if the use of several observers in parallel requires the introduction of new tuning parameters ($\theta_0$, $\lambda$, $N$ and $T$), the choice of these new parameters is very easy, due to their very clear effect on the results.

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

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

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

## 5. Appendix. Technical lemmas

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

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

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

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

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

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

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

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

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

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

$$z(t)\geq e^{\frac{\lambda}{2}t}(\frac{1}{\sqrt{x(0)}}-\frac{k}{\lambda})+\frac{k}{\lambda},$$

$$y(t)\leq\frac{1}{e^{\frac{\lambda}{2}t}(\frac{1}{\sqrt{x(0)}}-\frac{k}{\lambda})+\frac{k}{\lambda}},$$

$$x(t)\leq\frac{x(0)e^{-\lambda t}}{(1-\sqrt{x(0)}\frac{k}{\lambda})^{2}}.$$

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

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

□

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

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

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

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

We have, for $t>0$:

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

and:

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

Hence,

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

Since $f$ is compactly supported, we get:

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

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

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

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

This gives the result. □

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Université de Bourgogne, Département de Mathématiques, Laboratoire d'Analyse Appliquée et Optimisation, BP 47870, 21078, Dijon cedex, France.

*E-mail address*: busvelle@u-bourgogne.fr