Université de Lyon
CNRS, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1

**Laboratoire Ampère**
Unité Mixte de Recherche du CNRS - UMR 5005
Génie Electrique, Automatique, Bio-ingénierie

# *Mémoire doctorant 1ère année*
# *2017 -2018*

| | |
|---|---|
| Nom - Prénom | **AYALA CUEVAS Jorge Ivan** |
| email | **jorge.ayala-cuevas@ec-lyon.fr** |
| Titre de la thèse | **Performance validation of MEMS sensors using nonlinear uncertain models.** |
| Directeur de thèse | **Gérard Scorletti** |
| Co- encadrants | **Anton Korniienko** |
| Dpt. de rattachement | **MIS** |
| Date début des travaux | **01/10/2017** |
| Type de financement | **Contrat doctoral Ecole Centrale de Lyon**<br>**Projet NEXT4MEMS – BPI France** |

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France
Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

# Performance validation of MEMS sensors using nonlinear uncertain models

Jorge Ivan AYALA CUEVAS
July 2, 2018

**Abstract**

*MEMS inertial sensors allow to measure acceleration and angular rate of an object in motion. These devices have interesting characteristics such as small volume, light weight, low power-consumption and reduced cost for mass production. However, their accuracy remains limited mainly because of manufacturing tolerances and and sensitivity to environmental changes. An important task on the design process of new approaches is the performance validation of the proposed solutions. The objective of this PhD is to develop new methods for performance evaluation of MEMS inertial sensors based on automatic control analysis approaches. The proposed methods must allow to evaluate if the system respects the desired performance specifications before the experimental stage. This report introduces the system operation and model, the performances criterion and review the literature of the main system imperfections. The main approaches that can be exploited for robustness analysis are studied. The proposed approaches and partial results of performance validation are presented. Finally, some perspectives of the future work are discussed.*

**Keywords:** MEMS inertial sensors, MEMS gyroscopes, robustness analysis, performance objectives, uncertainty, scale factor error, bias.

## Contents

# 1 Introduction

This report presents the work done during the first year of the doctoral program entitled *Performance validation of MEMS sensors using nonlinear uncertain models*. This PhD thesis is performed within the scope of the project *NEXT4MEMS*, the consortium consists of the french leaders in the inertial sensor and microelectronics industry and two academic laboratories including *Ampère* (http://www.ampere-lab.fr/spip.php?article885). This project is supported by *BPI France* (the French public investment bank) within the framework of the Global Innovation Cluster *Minalogic "PSPC"* (Structural Research and Development Project for Competitiveness). This PhD is supervised by Gérard Scorletti (Ampère) and co-supervised by Anton Korniienko (Ampère).
The objective of NEXT4MEMS project is the development of a new generation of MEMS (Micro-Electromechanical Systems) inertial sensors with higher performance, which is required for certain applications (e.g. aerospace industry). Ampère participates at this project to deal with control engineering aspects, this includes modelling, identification, control design and robustness analysis. To solve this different challenges, three other PhD projects will be carried out in Ampère:

- *Joint identification and control of MEMS sensors* by Kévin Colin.
- *Robust Control for MEMS Inertial Sensor* by Fabricio Saggin.
- *Towards a joint identification and control procedure tackling the static nonlinearities of MEMS sensors* by Federico Morelli.

The Micro-Electromechanical Systems Systems (MEMS) are systems that are integrated directly in Silicon electronic boards and they global length goes from $0,1\mu m$ to $1000\mu m$. These hybrid devices comprise a mechanical part and an electronic part which allow to obtain a lecture of a physical variable. NEXT4MEMS project focus specifically on the development of MEMS inertial sensors, which are composed of MEMS accelerometers and MEMS gyroscopes which allow to measure linear acceleration and angular rate respectively.
Nowadays, MEMS inertial sensors can be found in several applications such as mobile phones (step-counting applications, screen-rotation detection), gaming devices, air-bag deployment systems, etc. They are more attractive with respect to other types of inertial sensors (e.g. optical) because of their main features like small volume, light weight, low power-consumption and reduced cost for mass production. However, due to the scale, stability and performance of micro-machined gyroscopes can be easily affected by manufacturing tolerances, material sensibility and environmental factors. This lack of accuracy is not compatible with some of the potential applications (e.g. inertial navigation systems of planes and cars) demanding a precision level relatively high, which today is achieved by the sensors of *macro* size.
The objective of this PhD is to develop new methods for performance validation of MEMS inertial sensors based on Automatic Control analysis approaches. The proposed methods must allow to evaluate if the closed-loop controlled system will successfully respect the desired performance specifications before the experimental implementation stage. The solution of such a problem is crucial since it allow to save the time and resources necessaries for experimentation by integrating the performance evaluation stage into the control-design one. The methods for performance evaluation must give the maximal guarantee of performance requirements achievement even if unexpected scenarios arise, but also it is desired that this tool does not result on over-conservative results and unreasonable computation time.
This report is structured as follows:

- Section 2 presents the formulation of the problem, by presenting the system to be tested, the performance objectives, the phenomena that today limits the improvement of the performance levels, and finally a discussion of the necessity of improvement with respect to the current methods for performance evaluation.

- Section 3 presents the robustness analysis approaches that were studied and exploited. Then, the proposed methods for performance evaluation and the obtained results are introduced.
- Section 4 gives the conclusions and perspectives for the following of this PhD.

# 2 Problem formulation

## 2.1 Introduction

The micro-mechanical structure of MEMS inertial sensors can be considered as the core of the system. It uses resonance phenomena to give an image of the measured variable. Then, they can be classified in two types of operation modes with respect to the mechanical operation: *In resonance* and *out-of resonance* inertial sensors. Resonating operation mode is much more often used since it allow to considerably decrease the effect of noise on the measure. This project will focus mainly in MEMS resonating gyroscopes, since both accelerometers and gyroscopes are based on resonance principle, being the MEMS gyroscopes the systems with a more complex operation. MEMS gyroscopes are generally operated in a feedback loop in order to obtain an acceptable level of performance, in the literature we can find solutions based on electronics, automatic control, and/or mixed propositions to guarantee the operation of the system around the resonance frequency and, in some cases, to estimate the angular rate more precisely.
The role of Ampère in NEXT4MEMS is to propose solutions based on automatic control for the MEMS gyroscopes developed by our industrial partners in order to attain the level of precision required for the project. As automatic control engineers, our two general objectives can be listed as follows:

- **Control design**: To propose methods for the design of control architectures that allow to attain the specified performance levels. This solution has to be sufficiently general so it can be integrated in different models without big changes for the adaptation, simple enough so it can be exploited for engineers without a strong control research background, and implementable since all the solutions have to take into account the characteristics of the developed devices and being realistic with respect to this.
- **Performance validation**: To provide tools that allow us to guarantee the correct operation of the system when the proposed control solution would be implemented in the real system, this before the experimental tests.

This report will deal with the second objective: *performance validation*. The control design process, similarly to the the most of the systems-design tasks, is an iterative process. The system to be controlled is firstly studied to obtain a model sufficiently informative, this knowledge is then exploited when looking for a control solution which meet, at least qualitatively, the design requirements. The last step is to evaluate if the system will guarantee the performance specifications on the real system. If the system would not attain this requirements, then it is necessary to look back to the system identification or control design stages. Since this iterative process is generally repeated several times before to find the optimal solution, there exists a big interest in reducing the cost of time and resources in the different design stages.
Regarding the validation of performances, the first and most obvious approach is to test the control solution on the system itself, or in a low-cost version of the system for experimental tests. This is in many cases a long, costly and delicate process, and sometimes it is even impossible to carry out. In our particular case, the implementation of the control solution on the sensor can be relatively evaluated by experimental tests. Nevertheless, the technical requirements (presented in following sections) demand to achieve the performance specifications for different operation conditions (robustness property). For example we can consider the temperature (actually one of the most delicate aspects in

MEMS gyroscopes). The standard tests to evaluate the performances with respect to the changes of temperature demand some special equipment which is not available in all the facilities of the companies or laboratories. Therefore, this can easily increases the cost and slow-down the design process.

Another approach is based in developing complex and accurate mathematical models of the system that are used for simulation of the controlled system. In order to test the behaviour of the system with respect to changes in the environment (robustness evaluation), the closed-loop system is simulated by testing the model in different operating points, trying to consider mainly the extremal conditions of the system. The main drawbacks of this approach is the cost of time and computational resources for the simulation. Even more critical is the fact that this approach will not allow to completely guarantee the proper operation of the system in all the non-tested scenarios.

A third approach consists on the development of mathematical methods that allow to guarantee the performance specifications of the system, moreover, to guarantee this specifications in any possible scenario. This approach is considered relevant if it model properly the real closed-loop system and it offers the maximum of guarantees, but also it should avoid over-conservative results and unreasonable computation time. This last approach is the one that we are looking to develop for performance validation. From the automatic control point of view, this is known as the *system analysis problem.* For this purpose, three problems have to be solved:

- **The system representation:** The first point is to understand the operation of the system in order to obtain a representative model (section 2.2). Secondly, to guarantee a rich enough representation of the real system it is necessary to consider the possible unmodelled phenomena and the changes on the operation conditions, this means, to study what we know about the *unknown* (subsection 2.4) and to look for a suited representation. For this problem, two concepts will arise: the *set of models* and the *uncertainty representation* (section 3.2 and annexe A).
- **The performance criterion:** The question here is how can we consider if the system has a good performance. First, the performance objectives have to be established, and then we need to translate them into a mathematical criterion. In MEMS inertial sensors industry, the performance indicators and the tests to evaluate them have been already established and standardized (quantitative specifications) (section 2.3). However, this indicators were established for post-design experimental evaluation. The problem here is to translate this post-design analysis into a well-suited and exploitable mathematical criterion that can be tested before the implementation of the solution on the real system.
- **The analysis method:** We need to provide some mathematical method to test the stability and performance of the model (section 3). The analysis method has to be accurate enough and computationally reasonable.

## 2.2 Operation principle of MEMS gyroscope

The mechanical part of MEMS Gyroscope consists of a double vibrating mass ($m_x$ and $m_y$) poorly damped. The mass $m_x$ can only be displaced on the $\vec{x}$ direction through an applied force $F_x$, this is called the drive mode or primary mode. Similarly, the mass $m_y$ moves only on the $\vec{y}$ direction by applying a force $F_y$ (sense mode or secondary mode).

Both masses are linked to a fixed structure by the means of cantilever beams. This implies that there exist stiffness and damping coefficients between the masses and the fixed structure (Fig. 1). Therefore, the system can be ideally modelled as two independent mass-spring-damper systems as follows:

$$m_x\,\ddot{x}(t) + D_{xx}\,\dot{x}(t) + k_{xx}\,x(t) \;= F_x(t) \tag{1}$$

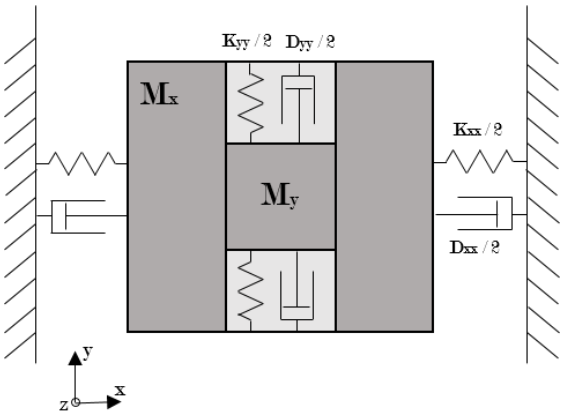



Figure 1: Mechanical resonator of gyroscope

$$m_y\,\ddot{y}(t) + D_{yy}\,\dot{y}(t) + k_{yy}\,y(t) \;= F_y(t) \tag{2}$$

where $k_{xx}$ and $k_{yy}$ are the stiffness constants of the drive and sense mode respectively, while $D_{xx}$ and $D_{yy}$ are he damping coefficients of drive and sense modes.
Consider now that the gyroscope is submitted to an angular velocity $\Omega_z$ around the $\vec{z}$ axis. Also, a force $F_x$ is applied in order to make vibrate the mass $m_x$ along the $\vec{x}$ direction (drive resonator). The Coriolis effect will affect the mass in movement with respect to the rotating frame producing an imaginary force. As a result, Coriolis force $F_{cory}$ will appear along the $y$ axis affecting the mass $m_y$. Therefore, there exists a transmission of energy from primary to secondary mode which is proportional to the angular velocity $\Omega_z$. The image of angular velocity $\Omega_z$ can then be obtained by exciting the primary mode and measuring the vibration amplitude of secondary mode.
Note that there exists as well a Coriolis force $F_{corx}$ due to the movement of the mass $m_y$ of the secondary resonator that will disturb the vibration of primary mode. Regarding the gyroscope considered in this project, the mass and vibrations of primary mode are considerably bigger than in secondary mode, so that the Coriolis effect from primary to secondary mode will be much more important than in the opposite direction. The Coriolis forces of both modes are then defined:

$$F_{cory}(t) \;= -2m_x\,\dot{x}(t)\,\Omega_z(t) \tag{3}$$

$$F_{corx}(t) \;= 2m_y\,\dot{y}(t)\,\Omega_z(t) \tag{4}$$

Including Coriolis forces expressions into 1 and 2, the complete equation of ideal gyroscope is:

$$\begin{bmatrix} m_x & 0 \\ 0 & m_y \end{bmatrix}\begin{bmatrix} \ddot{x}(t) \\ \ddot{y}(t) \end{bmatrix} + \begin{bmatrix} D_{xx} & 0 \\ 0 & D_{yy} \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} k_{xx} & 0 \\ 0 & k_{yy} \end{bmatrix}\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 0 & 2m_y\Omega_z(t) \\ -2m_x\Omega_z(t) & 0 \end{bmatrix}\begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} F_x(t) \\ F_y(t) \end{bmatrix} \tag{5}$$

Now, considering that Coriolis force affecting the primary mode is small enough and can be neglected, the transfer function of primary mode between the applied force $F_x$ and the displacement $x$ is:

$$\frac{X(s)}{F_x(s)} = \frac{1}{m_x s^2 + D_{xx} s + k_{xx}} \tag{6}$$

Making some changes of variables, we can transform this expression in terms of frequency response variables as follows:

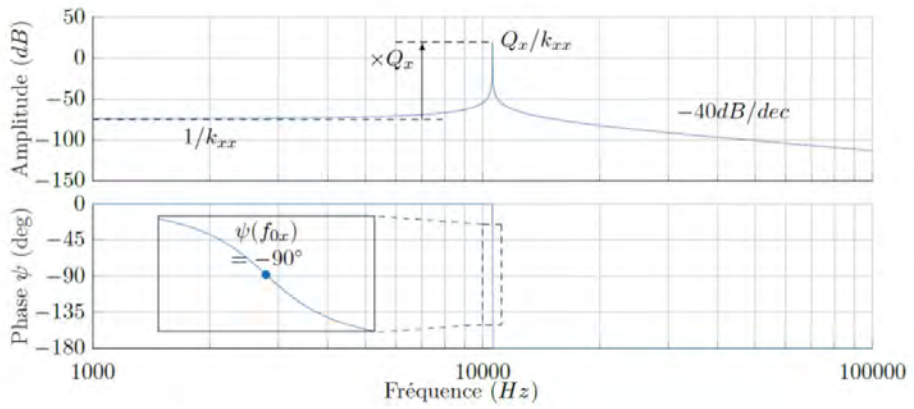



Figure 2: Frequency response of drive mode

$$\frac{X(s)}{F_x(s)} = \frac{1/m_x}{s^2 + \frac{\omega_{0x}}{Q_x}s + \omega_{0x}^2} \tag{7}$$

where $Q_x = \sqrt{m_x k_{xx}}/D_{xx}$ and $\omega_{0x}\sqrt{k_{xx}/m_x}$ are respectively the *Quality factor* and *resonance frequency* of drive mode. The mechanical architecture of MEMS gyroscopes is specified in terms of this parameters because they are more related to the frequency response of the system which is directly linked with the performance specifications of the drive mode.

Let take the example of a MEMS gyroscope with $Q_x = 60000$ and $\omega_{0x} = 10200$ Hz. The Bode diagram of transfer function (7) for this gyroscope is presented in figure 2. We can observe that the gain of the system at the resonance frequency with respect to its static gain is equal to $Q_x$. This important amplification allows to obtain an important signal to noise ratio (SNR), and by consequence, to get a more precise measure of the angular velocity $\Omega_z$. From where the importance of guaranteeing the operation of the drive mode at the resonance frequency.

### Synchronous demodulation

Once the displacement of masses $m_x$ or $m_y$ has been transformed into a voltage signal, it is possible to obtain amplitude and phase information that can be used for the control loop in both drive and sense mode, it can also allow to extract the image of the angular rate $\Omega_z$ on the sense mode by demodulating the Coriolis force of sense mode $F_{cory}$. In order to show the operation of synchronous demodulation, let take the example of drive mode displacement measurement $x$: assuming that the drive mode is excited at the frequency $\omega_{exc}$. Therefore, at the output of the resonator (including the position to voltage converter) there is a signal $V_{dem}(t)$ at the frequency $\omega_{exc}$:

$$V_{dem}(t) = a\, sin(\omega_{exc}t + \phi). \tag{8}$$

with $a$ the amplitude of the signal and $\phi$ the phase with respect to the excitation signal $V_{exc}$. Now, the synchronous demodulation is done by multiplying $V_{dem}$ by signals at the same frequency: $sin(\omega_{exc})$ and $cos(\omega_{exc})$ (figure 3). The multiplication gives the intermediate signals $i(t)$ and $q(t)$ which pass through low-pass filters to finally obtain the phase signal $I(t)$ and the quadrature signal $Q(t)$:

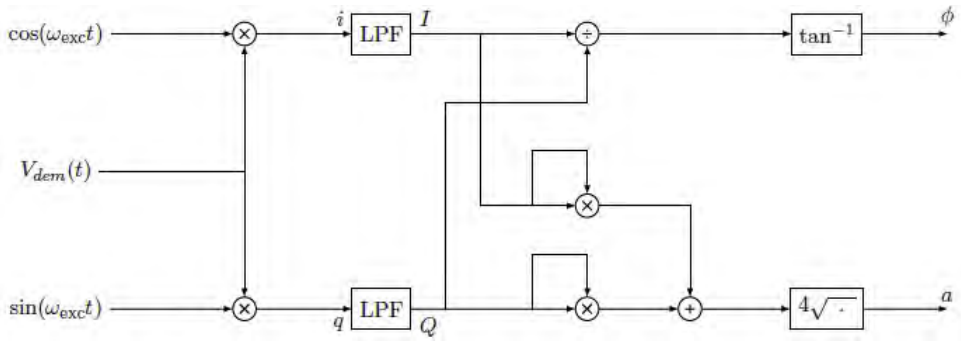



Figure 3: Synchronous demodulation

$$i(t) = a\frac{1}{2}\left(sin(\phi) + sin(2\omega_{exc} + \phi)\right) \tag{9}$$

$$q(t) = a\frac{1}{2}\left(cos(\phi) - cos(2\omega_{exc} + \phi)\right) \tag{10}$$

$$I(t) = a\frac{1}{2}sin(\phi) \tag{11}$$

$$Q(t) = a\frac{1}{2}cos(\phi) \tag{12}$$

Finally, it is possible to obtain amplitude and phase just by using arithmetic calculations:

$$a = 2\sqrt{I^2 + Q^2} \tag{13}$$

$$\phi = tan^{-1}\left(\frac{I}{Q}\right) \tag{14}$$

## 2.3 Performance specifications

Within the framework of NEXT4MEMS project, we aim to develop new approaches that allow to improve the global performance of MEMS inertial sensors. Therefore, it is necessary to specify what are the main indicators that are used to evaluate and compare the achieved performances using different methods. We can consider two main classes of requirements: the operation range specifications and the maximal errors tolerances.

### Operation range specifications

Operation range specifications indicate the expected interval of operation conditions in which the gyroscope has to work properly, this main requirements are:

- Input range: The minimum range of angular rate than the gyroscope has to be capable of measure respecting the maximal error tolerances.
- Angular acceleration: The gyroscope has to be capable of work under a minimal angular acceleration.
- Minimal bandwidth.
- Rotation Latency: Time interval between the implementation of a rate signal on the input and the availability of the corresponding data on the output.

- Full performance start up time: Time interval between power up and the delivering of nominal data of the sensor operating within full performance.

This information will impose some performance constraints in the control design stage.

### Maximal tolerated errors

The second class of MEMS gyroscopes specifications is related to the different types of errors tolerances that are used to evaluate the precision of the sensor. This specifications will be exploited in the performance analysis stage in order to provide mathematical criterion that can be tested using control-based performance validation methods.
The error tolerances specifications are defined from the relationship between the real angular rate at the input $\Omega_{input}$ and the measured lecture at the output $\Omega_{output}$:

$$\Omega_{output} = (SF\,\Omega_{input}) + B \tag{15}$$

where $SF$ is the scale factor and $B$ the bias (figure 4). Therefore, the errors are grouped into bias and scale factor errors and subdivisions therein. These errors and the methods to analyse them will be presented in following sections.

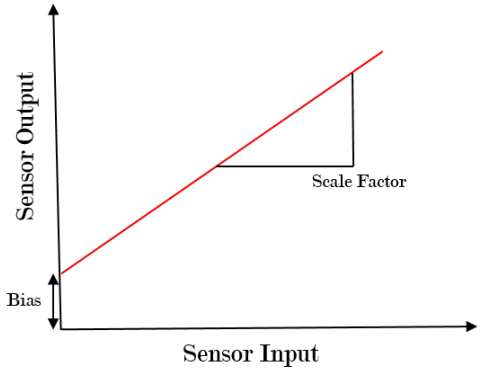



Figure 4: Angular rate input-output relation.

### Bias

Bias is the measured output when the gyroscope is not under rotation, this means, when there is not angular rate input. The bias is classified in two parts depending in the nature of the cause: the deterministic bias or *bias offset* and the stochastic bias or *bias drift*. Bias offset is caused by non-ideal coupling phenomena (either mechanical or electrical) between drive and sense modes of the gyroscope, and it can be also influenced by time-delays on the electronic system. In practice, this Bias offset will be surrounded by some noise, so it is determined by computing the average of the output during a certain period $T$:

$$\Omega_{offset} = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \Omega(t)dt \tag{16}$$

Once the "nominal" bias is known, the gyroscope is calibrated in order to have no bias in nominal operating conditions, actually, since this error can be easily compensated, there is no specification for the bias offset. The main interest is to maintain this bias offset at minimal levels when the operation conditions change, this specifications are:

- Run to run bias repeatability: Standard deviation of several repeated bias measurements (7 according to the norm) under the same operating conditions and with specific non-operating periods between two measurements.
- Bias thermal sensitivity: Standard deviation of bias measurement over the specified range of temperatures.

In the other side, bias drift is caused by phenomena which are random in nature or stochastic processes . There are mainly three specifications for the impact of noise in the output:

- RMS noise: The RMS noise level at the output of the sensor in a specific band at nominal conditions has to be lower than specified values.
- Bias instability: It is the random variation in bias as computed over finite sample time and averaging time intervals. It is a non stationary process characterized by 1/f power spectral density. It is evaluated by observing the lowest point on the named *Allan Deviation plot.*
- Angular Random Walk: It is a zero-mean Gaussian stochastic process with stationary independent increments and with standard deviation that grows as the square root of time. It is evaluated by observing the -1/2 slope on the Allan Deviation plot.

Since Allan Variance is the standard method used to evaluate the impact of different noises into the bias of the MEMS gyroscope, in following sections we will present this method in the classical form, and to propose a reinterpretation that can be exploitable with control analysis approaches.

**Scale Factor**

Scale factor is the ratio between the measured angular rate at the output and the real angular rate at the input. It can be seen as the slope where the output is a function of the input expressed in units of another nature, in the ideal case it is expected to be a linear relationship. Scale Factors errors are considered as the deviations of this slope usually caused by changes in the environment conditions or system imperfections. Whatever the cause of the slope deviation is, this error is a measure of the relative error between the output and input angular rate:

$$\varepsilon_{SF} = \frac{|\Omega_{out} - \Omega_{in}|}{|\Omega_{in}|} \tag{17}$$

where $\varepsilon_{SF}$ is the scale factor error, similarly to the bias offset, the scale factor error is attempted to remain at minimum levels when operating conditions change. The specifications for the scale factor are then:

- Run to run scale factor repeatability: Scale factor error after several repeated scale factor measurements under the same operating conditions and with non-operating periods between two measurements.
- Scale Factor thermal sensitivity: Scale factor error over the specified range of temperature.
- Scale factor non-linearity: Variation of the linear relationship between the input angular rate and the measure at the output over the complete range of angular rate operation.

The technical specifications with respect to the maximal tolerated errors can be grouped as combinations of the potential cause of the error and the type of error tested at the output:

- **Causes:** Temperature, fatigue or noise.

- **Errors:** Bias or scale factor error.

This will allow us to finely define the analysis tests by adjusting the different mathematical criterion. In one side, the cause will provide the necessary information about the type and range of uncertainty on the model, and, in the other side, the type of error will allow to define the adequate approach for robustness analysis.

## 2.4 Non-ideal behaviour sources

The knowledge of the operation principle of MEMS gyroscopes and the performance specifications can be sufficient to test the *nominal stability* and *nominal performance* of the system, this means, to guarantee stability and performance requirements of a specific model. However, one of the biggest challenges in MEMS gyroscopes performance improvement is their sensibility to fabrication tolerances, modelling limitations and changes in the environment conditions. Then, the main objective for us is to develop methods that allow to guarantee the *robust stability* and *robust performance*, this is, to guarantee the same properties as in the nominal case, but for every possible model that can be included in the real system. The development of such a method demands prior knowledge about the possible sources of *uncertainty* on the system and their impact on system behaviour, in the following lines, we will visit some of the studies done in literature with respect to all this system imperfections and other aspects that increases the error between the model and the real system.

### System imperfections

In practice, achieving the optimal operation of MEMS gyroscopes becomes quite difficult due to system imperfections caused by manufacturing tolerances of the current microfabrication techniques, as well as changes on environmental conditions such as temperature variations and the ageing of components. This non-ideal conditions may affect the stability and global performance of the gyroscope.
In ideal MEMS gyroscope model, the system is seen as two independent mass-damper-spring systems that are only coupled by Coriolis force. Indeed, these two modes can be coupled by other phenomena caused either by natural parasite effects inherent to the system or by manufacturing dispesion. Non proportional damping has its origin in two main phenomena: [You11] *squeeze-film damping*, which occurs as a result of the massive movement of the air between the plates of interdigital capacitors used for resonator excitation, which is resisted by the viscosity of the fluid; and *thermoelastic damping* which results from the irreversible heat flow generated by the compression and decompression of the oscillating structure. Both phenomena cause the apparition of forces that are orthogonal to the displacement direction and proportional to the velocity of the displacement direction. This can be modelled by the inclusion of non-diagonal parameters $D_{xy}$ and $D_{yx}$. Anisoelasticity is mainly caused by the non-perfect alignment between the mechanic system and the measurement axis [SSP+06] (figure 5). Similarly, this can be translated by non-diagonal stiffness constants $k_{xy}$ and $k_{yx}$. Including this non-diagonal coupling terms, the equation (18) becomes:

$$\begin{bmatrix} m_x & 0 \\ 0 & m_y \end{bmatrix} \begin{bmatrix} \ddot{x}(t) \\ \ddot{y}(t) \end{bmatrix} + \begin{bmatrix} D_{xx} & D_{xy} \\ Dyx & D_{yy} \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} k_{xx} & k_{xy} \\ k_{yx} & k_{yy} \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} 0 & 2m_y\Omega_z(t) \\ -2m_x\Omega_z(t) & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} + \begin{bmatrix} F_x(t) \\ F_y(t) \end{bmatrix} \tag{18}$$

The spring-softening (or spring-hardening) is an effect where a parasite electrostatic force caused by the DC component of the excitation signal behaves as the addition of a stiffness constant $k_es$ on the primary mode as follows:

$$k_{eff} = k_{xx} + k_{es} \tag{19}$$

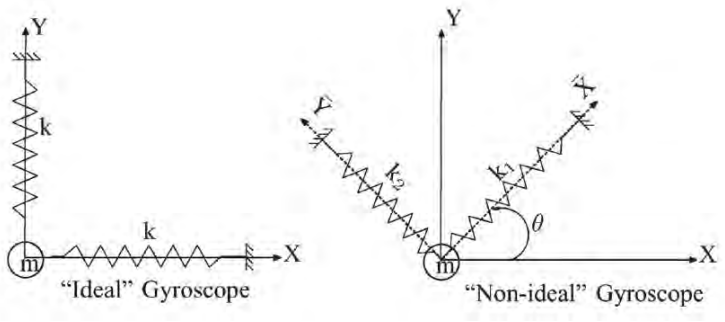



Figure 5: Angular mismatch of gyroscope [SSP+06]

This has an impact on the variation of the resonance frequency since:

$$\omega_0 = \sqrt{\frac{k_{eff}}{m}} \tag{20}$$

In spring softening, the resonance frequency decreases as the amplitude of the of the DC component increases, whereas in the case of spring hardening, the resonance frequency of the structure increases as the amplitude of oscillation increases. Both phenomena are present simultaneously, but one of them is dominant [EKT+11].

### Noise

When MEMS gyroscopes are incorporated in navigation systems, the main interest is to integrate the angular rate $\Omega(t)$ to obtain the orientation angle $\theta(t)$. In the ideal case, if there is not movement, the measure angular rate is zero and also the angle obtained by integration. Noise is the main phenomena that causes deviation on the obtained angle:

$$\theta_{meas}(t) = \int_0^T (\Omega_{meas}(t) + \Omega_{noise}(t))dt \tag{21}$$

This is known as the angular random walk, this one of the main sources of error in MEMS gyroscopes and there exists a big interest in minimizing its effect. In general, two types of noise are analysed when study the performance of MEMS inertial sensors (both accelerometers and gyroscopes): the white noise and the Flicker noise ($1/f$ noise).

As every dissipative system, resonator of the MEMS gyroscope is affected by mechanical-thermal noise, this is caused by the random collisions of gas molecules around the resonator which are excited when the temperature increases, this type of noise is found directly at the output of the resonators of both drive and sense modes [Lel05]. Its power spectral density (PSD) is a linear function of temperature given by:

$$S_{MTN}(f) = 4K_BTD \tag{22}$$

where $K_B$ is the Boltzmann constant, $T$ the absolute temperature and $D$ the damping coefficient [**?** ]. Mechanical-thermal noise is the main cause of white noise measured at the output of the gyroscope [Lel05].

Flicker noise is a non stationary low-frequency noise [Kes82], the origin of this type of noise comes mainly from the electronics, more specifically is believed to be a result of fluctuations in conductivity in a semiconductor device [MYM12]. However, its origin is not exactly known [Vos79] and from

our knowledge, the most of the studies are focused in its calibration at the output rather that in attenuating its effect from the origin [KJCT12]. The PSD of a Flicker noise is modelled as:

$$S_{Flicker}(f) = \frac{B^2_{Flicker}(K,R,I)}{f} \quad \forall\ f \leq f_0 \tag{23}$$

with $B_{Flicker}$ the Flicker noise coefficient, which is a function of the device constant $K$, the electrical resistance $R$ and the current $I$, and $f_0$ is the cut off frequency, at higher frequencies than $f_0$, the noise is considered to be drawn into the white noise.

### Temperature variation

The properties of materials that compose the MEMS gyroscopes are affected by the temperature variation, for example, Young module of monocrystalline silicon which leads to a variation on stiffness. This phenomen is identified by observing the changes of *resonance frequency* and *Quality factor* as a function of temperature. While resonance frequency is a linear function of temperature [FKP+05], [GHL+15], [FLW11], Quality factor has been observed to have an approximate relation of $1/T^3$ [KHC+08], [PTS12] . This relation is shown in figure 6.
This results can be mapped as functions of the variation of damping coefficient and stiffness with respect to temperature. This will allow to determinate which is the uncertainty interval of system parameters due to temperature variations.

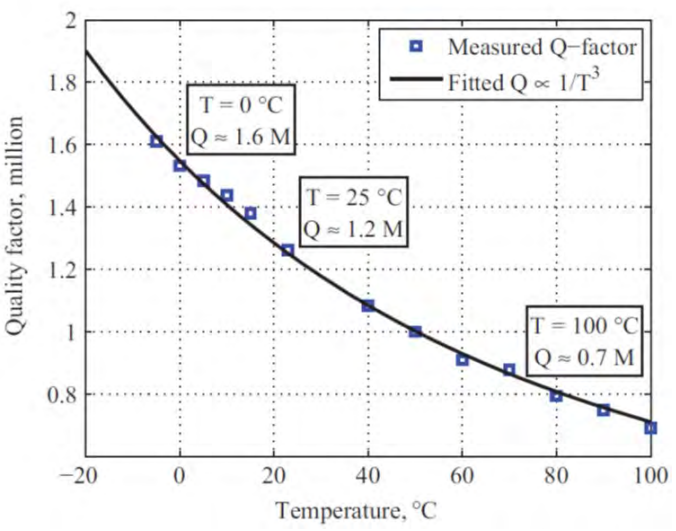



Figure 6: Variation of Q-factor with temperature [PTS12]

### Modelling errors

In general, the determination of MEMS gyroscopes model is made by using identification methods. This allows to estimate the approximative values of system parameters at nominal operation conditions and eventually, to determinate which of this parameters are uncertain and which is the level of uncertainty.
The identification process is in general designed to obtain a relatively simple model that can be exploited for control design purposes. Even if this is not the case, obtain a model that describes perfectly a real system is highly complicated (in fact, impossible). In general, the interest is focus on deriving an accurate enough model for the frequency interval in which the system will normally operate, and all the dynamics on other frequencies are usually neglected.

Either by ignorance or by practical issues, there exist an error between the model and the real system, this is translated as an uncertainty on the dynamic response of the system. The representation of this type of uncertainty and the way to analyse how this can affect or not the stability and performance of the controlled system will be presented in Annexe A.

### 2.5 Current methods for performance evaluation of MEMS sensors

It is well known that MEMS inertial sensors are a very promising option for several applications in the future, but they have still critical limitations with respect to their sensibility to changes in the operating conditions and to the noise. This does not allow to keep the desired performance levels permanently. The literature has mainly focused in the design aspect that allow to improve this performances. However, from the author knowledge, performance validation of the proposed solutions is made in many cases directly by experimentation [LLL08], [ABSS09].

The general and almost only used pre-experimental approach for performance validation of MEMS inertial sensors is based on Montecarlo simulations ( [GLLG03], [SH06], [ZTSL10]), this methods demand an accurate and complex model, which consumes time and computational resources. Moreover, when looking for robust performance assessment, this approach can not give a complete guarantee of reliable results.

The only attempt of using control analysis methods is found on [DZG07], nevertheless, the analysis is based on testing the classical stability margins when the temperature is at the maximal and minimal values of the temperature operation range, which does not give a guarantee of robustness. For this reason, it is obvious that there exist a need of developing new approaches that allow to test the stability and performance of MEMS gyroscopes in a reliable and efficient way.

Next section will introduce some of the system analysis approaches born from the control theory, and the first performance validation methods developed during this PhD will be introduced, and finally the results and perspectives will be presented.

## 3 Robustness analysis

### 3.1 Introduction

When the controller is implemented on the real system, it is crucial to ensure that the actual behaviour is close to the expected one. For practical reasons, the controller is designed using a reduced and simplified model usually named the *nominal model*. Therefore, when physical tests are done, several physical imperfections, non nominal environmental conditions and unmodeled dynamics may cause an unexpected behaviour, a degradation of system performances and, in the worst case, the system may become unstable. Here lies the importance of developing methods that allow to foresee such scenarios and to assess if the system works well or not when facing all these uncertainties. Typically, the robust performance of MEMS sensors is evaluated based-on Monte Carlo methods which use complex and highly accurate models to test if the performance specifications remain respected considering different combinations of parameter values. In spite of its simplicity and universality, this approach is costly and time-consuming. Moreover, the main drawback of Monte Carlo method is that the system is investigated only for randomly chosen parameter combinations, therefore, it does not offer a rigorous guarantee of robust stability and robust performance. Actually, the "worst-case" parameter combination can not be considered when sampling over the parameter-values range. Considering all these reasons, it is obvious the need to use and adapt new techniques for robustness validation in order to obtain the maximal guarantee of the proper operation of MEMS inertial sensors.

The *set of models* representing the *the uncertain system* is shown in figure 7. Monte-Carlo approaches will consider only a certain number of scenarios, which will never be sufficient for ensuring robustness since it does not consider the whole set of models. The *worst-case* approaches will allow to guarantee

the stability and a certain level of performance for all the models included in the set. If all the possible operation conditions of the real system are included in the uncertain model set, then some properties can be ensured when testing the real system in any scenario.
In following sections, we will introduce the main approaches for the representation of uncertain systems, focusing mainly on Linear Fractional Representations (LFR), which is the representation exploited for the input-output robustness analysis methods that will be addressed later on this document.

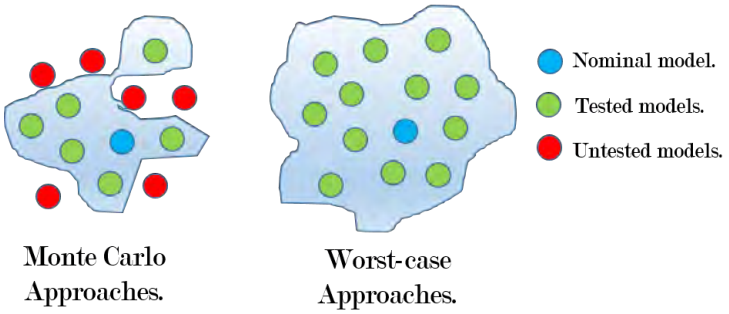



Figure 7: Set of Models.

## 3.2 Lineal Fractional Representation

The first step before applying any approach for robustness analysis consists in finding a way to model the uncertain system. As said before, the design of a controller for the system is done by using a model which represents the main features of the real system to be controlled, this model is a simplified approximation of the reality obtained from physical laws and/or using identification methods. For classical linear methods for control design, this model is a Linear and Time-Invariant (LTI) system. Then, the controller will ensure stability and some performance specifications for this specific model. Indeed, a certain tolerance will be associated to the identified parameters, in other cases the values of the parameters may be affected because of changes in the environmental or internal conditions. In addition, the simplification of the model is done neglecting intentionally some high-frequency dynamics and/or non-linearities which are present in the original system. Therefore, this real and relatively "unknown" system can not be studied using the classical control theory methods for stability (e.g. gain and phase margins) and performance. All the sources of uncertainty mentioned above, can be classified in two main groups:

- **Dynamic uncertainties:** Caused by missing dynamics in the model, these dynamics could be neglected for the sake of simplicity, or unknown, as usually occurs for high-frequency dynamics.
- **Parametric uncertainties:** The structure of the model is known, but the values of the model parameters are uncertain.

Nevertheless, some information about the uncertainties is known, the intervals to which the parameters and uncertain dynamics belong to. This, means, the uncertainties (and sometimes their rates of variation) are assumed to be bounded. Then, a dynamic uncertainty can be defined as follows:
The parametric uncertainties can be still be grouped in different classes with respect to their rate of variation:

- **Time-Invariant parameters.**
- **Rate-bounded Time-Varying parameters.**

- **Arbitrary-fast Time-Varying parameters.**

The most general framework for representing uncertainties is to represent the system in the form of a lineal fractional transformation (LFT). The LFT representation proposes to represent the uncertain model as the interconnection of a certain LTI model representing the "nominal model" and an operator representing the uncertainties and/or non-linearities of the system. Quite often, the system depends rationally on the uncertain parameters, even in those cases, it is always possible to find an LFR of such a system; which is not possible with other types of representations such as polytopic representation. The LFR in figure A, is denoted $\Delta \star M$ , where $\star$ defines the star product of Redheffer. This product allows to describe the interconnection of several systems. In the robustness analysis of uncertain LTI systems framework, we consider the interconnection of a LTI stable system (bloc $M_d$ in figure **??**) representing the nominal model, and a bounded bloc operator $\Delta$ containing all the uncertainties of the real system.

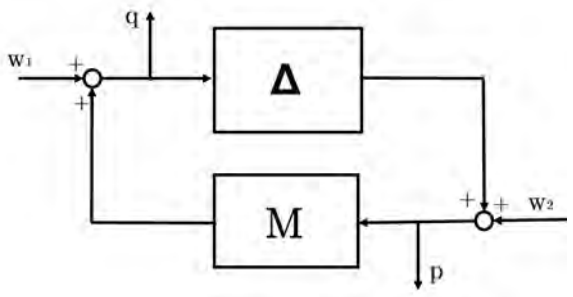



Figure 8: Lineal Fractional Representation.

## 3.3 Input-output approaches for Robustness Analysis

### Small-Gain Theorem

The small-gain theorem was introduced by George Zames in 1966 [Zam66]. It derives the conditions for the stability of the system $(M, \Delta)$ represented by the feedback interconnection in figure **??** of two stable operators. Considering the bounded exogenous signals $w_1$ and $w_2$ with respect to some norm, the stability implies that all the internal signals in the interconnection will remain bounded. This is guaranteed if the system is internally stable.

**Internal Stability:** Consider the closed-loop system $(M, \Delta)$ represented in figure **??**. This system is said to be internally stable if the transfer function matrix defined by:

$$\begin{bmatrix} q(s) \\ p(s) \end{bmatrix} = \begin{bmatrix} (I - M(s)\Delta(s))^{-1} & (I - M(s)\Delta(s))^{-1}M(s) \\ (I - \Delta(s)M(s))^{-1}\Delta(s) & (I - \Delta(s)M(s))^{-1} \end{bmatrix} \begin{bmatrix} w_1(s) \\ w_2(s) \end{bmatrix} \tag{24}$$

is stable. In addition, a family of systems is stable if all its members are stable.

**Theorem 3.1** (Small-Gain theorem)**.** *Assuming that there exists a causal inverse of* $(I - M(s)\Delta(s))$*, the family of systems* $(M, \Delta)$ *represented in* **??** *is stable for every stable* $\Delta$ *such that* $\|\Delta\|_\infty \leq \beta$ $(\|\Delta\|_\infty < \beta)$ *if* $\|M\|_\infty < \beta$ $(\|\Delta\|_\infty \leq \beta)$*.*

The power of the Small-Gain Theorem is that it guarantees robust stability as long as the subsystems of the interconnection can be characterized for some induced norm. This means that the bloc operator $\Delta$ (and even $M$) can be either a LTI system, uncertain real parameters, a time-varying or non-linear operator. The two main drawbacks of this theorem are related to the conservatism:

- It does not consider any additional information about the structure of the uncertainty bloc than its boundedness to guarantee robustness.
- The obtained result is the same if the uncertain bloc is either the uncertainty is a time-invariant uncertain parameter or a non-linearity, which actually will not have the same impact on the system stability.

In order to reduce the conservatism of the robustness analysis methods,, it will be necessary to take into account some additional information about the structure and nature of the uncertain bloc $\Delta$.

### $\mu$-Analysis

When dealing with a system including several uncertainties, the small-gain theorem provides only sufficient condition to guarantee robust stability since there is no information about the structure of the uncertain bloc $\Delta$. John Doyle proposed in 1982 [Doy82] a systematic mechanism for exploiting information about the structure of a perturbation. It was particularly important the formulation of the block-diagonal perturbation problem. This problem is quite general since any norm-bounded perturbation problem, regardless of structure, can be trivially rewritten as a block-diagonal perturbation problem.
Now $\Delta$ is a structured stable LTI perturbation such that $\Delta(s)$ belongs to a set $\underline{\Delta}$ of bloc diagonal uncertain matrices that can include complex uncertainties $\Delta(j\omega)$ because the uncertainty is defined by a constraint on its frequency response which is a complex number, and real uncertainties in the case where the uncertainty is a physical parameter, its frequency response being real. Then, we define the set of complex matrices having the structure of the uncertainty considered:

$$\underline{\Delta} = \left\{ \Delta \in \mathbb{C}^{k\times k} \,|\, \Delta = \begin{bmatrix} \Delta_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \Delta_q & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & \delta_1 I_{r_1} & 0 & \vdots \\ \vdots & \ddots & 0 & 0 & \ddots & 0 \\ \vdots & \ddots & 0 & 0 & \ddots & \delta_r I_{r_r} \end{bmatrix} \; with \left\{ \begin{array}{r} \Delta_i \in \mathbb{C}^{k_i\times k_i},\ i \in \{1,...,q\} \\ \delta_j \in \mathbb{R},\ j \in \{1,...,r\} \\ k \;=\; \sum_{i=1}^{q} k_i \;+\; \sum_{j=1}^{r} r_j \end{array} \right\} \right\} \tag{25}$$

Now, in order to establish a similar condition to that of small-gain theorem, we define a new indicator for complex matrices, the Structured Singular Value is a generalization of the singular value $\overline{\sigma}$ used in the small-gain theorem.
**Structured Singular Value (SSV):** Let be the matrix $A \in \mathbb{C}^{k\times k}$. Then the Structured Singular Value of $A$ with respect to the structure $\underline{\Delta}$, denoted $\mu_{\underline{\Delta}}(A)$, is defined by:

$$\mu_{\underline{\Delta}}(A) \;=\; \frac{1}{\underset{\Delta\in\underline{\Delta}}{inf}\ (\overline{\sigma}(\Delta) \mid det(I - \Delta A) = 0)} \tag{26}$$

Here, the indicator $\mu$ depends not only on the system $M(s)$ but also on the structure of $\Delta$. The mathematical expression can be defined as the research of the smallest structured $\Delta$ which makes the term $I - \Delta A$ to be singular, then we get the inverse of the obtained value. If there not exist any structured $\Delta \in \underline{\Delta}$ which makes the determinant of $I - \Delta A$ equals to zero, then $\mu_{\underline{\Delta}}(A) = 0$.
With the previous definitions, now we dispose of all the necessary elements to introduce an extension of the small-gain theorem for the case of systems families with diagonal structured uncertainties.

**Theorem 3.2** (Structured Small-Gain theorem)**.** *Let be the family of closed-loop systems $(M, \Delta)$, with $M(s)$ a stable transfer functions matrix, and $\Delta(s)$ a stable transfer function such that $\Delta(s) \in \underline{\Delta}$. The family of closed-loop systems $(M, \Delta)$ is stable for all $\Delta$, such that $\|\Delta\|_\infty < \beta$ if and only if:*

$$\forall \omega \in [0, +\infty], \quad \mu_{\underline{\Delta}}(M) \leq \frac{1}{\beta} \tag{27}$$

Indeed, there exist a strong relationship between the structured small gain theorem and the classical Nyquist criterion for the MIMO case. This approach can be seen ten not as a replacement of classical stability margins, but as an extension that allow to develop more efficient but still intuitive toolboxes for robustness analysis.

In spite of he simplicity of the concept of the indicator $\mu$, it has been established that its exact computation is a non-polynomial (NP-hard) problem, this means that the computational complexity growth with the number of parameters involved even for purely complex perturbations [TO95]. However, practical algorithms for computing efficiently upper bounds of the SSV for cases with complex, real or mixed real/complex perturbations are available [FTD91], usually based on optimization problems under LMI constraints.

It has to be noted that the $\mu$-analysis allows to address the stability of interconnections of LTI operators. So that the main drawback of this approach is that no time-varying systems can be analysed with this theorem. Nevertheless, $\mu$-analysis is in general a powerful, efficient and relatively not conservative tool for robustness analysis when dealing with time-invariant systems or with systems that can be assumed as not varying with time (e.g. slowly-varying systems).

### Integral Quadratic Constraints

The IQC-based robustness analysis method has been introduced by Megretski and Rantzer [MR97] as a unifying approach between the absolute stability theory, input-output theory and robust control theory. IQC approach is attractive since it is compatible with LFR representation, but in this case, the uncertain block can also consider time-varying parameters and certain types of non-linearities.

For this purpose let consider the system of the figure 9, we try to capture the main characteristics of the uncertain bloc using an input-output approach, this means, we observe if the inputs and outputs of the uncertain bloc verify an integral quadratic constraint of the form:

$$\sigma(\omega) \int_{-\infty}^{\infty} \begin{bmatrix} p(j\omega) \\ q(j\omega) \end{bmatrix}^T \Pi(j\omega) \begin{bmatrix} p(j\omega) \\ q(j\omega) \end{bmatrix} d\omega \geq 0 \tag{28}$$

The matrix $Pi(j\omega)$ is the so-called multiplier, it allow to characterise the type of considered uncertainty. A collection of different multipliers sets describing different types of uncertainties can be found in [MR97], [VSK16]. Another advantage of IQC approach is its modularity, this means, it is possible to consider different types of uncertainties/non-linearities for the analysis quite easily.

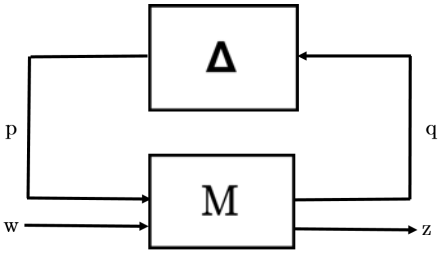



Figure 9: Lineal Fractional Representation.

In order to evaluate the robust performance of the system, we will consider the system which considers the uncertain channel $p \to q$ and a performance channel $w \to z$ as represented in figure 9, the system is ow defined through the linear fractional representation:

$$\begin{bmatrix} q(j\omega) \\ z(j\omega) \end{bmatrix} = \begin{bmatrix} M_{qp} & M_{qw} \\ M_{zp} & M_{zw} \end{bmatrix} \begin{bmatrix} p(j\omega) \\ w(j\omega) \end{bmatrix}, \quad p(j\omega) = \Delta q(j\omega) \tag{29}$$

where $M$ si a stable LTI system and $\Delta$ the bloc containing all the uncertainties. Now, we would like to impose a performance criterion through an IQC as follows:

$$\sigma(\omega) \int_{-\infty}^{\infty} \begin{bmatrix} z(j\omega) \\ w(j\omega) \end{bmatrix}^T \Pi_{perf} \begin{bmatrix} z(j\omega) \\ w(j\omega) \end{bmatrix} d\omega < 0 \tag{30}$$

Then, the robust performance of the system with respect to the specified performance criterion can be evaluated using the following result:

**Theorem 3.3** (Robust performance with respect to )**.** *Let be the . Assume that for all $\Delta \in \underline{\Delta}$ and for all $\Pi \in \underline{\Pi}$ the IQC (30) is satisfied. Then the interconnection (29) is robustly stable and robust performance with respect to $\Pi_{perf}$ on the channel $w \to z$ is guaranteed if there exist $Pi \in \underline{\Pi}$ satisfying the frequency domain inequality:*

$$\begin{bmatrix} M_{qp}(j\omega) & M_{qw}(j\omega) \\ I & 0 \\ M_{z2p}(j\omega) & M_{z2w}(j\omega) \\ M_{z1p}(j\omega) & M_{z1w}(j\omega) \end{bmatrix}^* \begin{bmatrix} \Pi(j\omega) & 0 \\ 0 & \Pi_{perf} \end{bmatrix} \begin{bmatrix} M_{qp}(j\omega) & M_{qw}(j\omega) \\ I & 0 \\ M_{z2p}(j\omega) & M_{z2w}(j\omega) \\ M_{z1p}(j\omega) & M_{z1w}(j\omega) \end{bmatrix} < 0 \tag{31}$$

### 3.4 Analysis of Scale Factor error in MEMS gyroscope

The objective is to develop an approach that allow to determine which is the minimal scale factor error of the controlled closed-loop system (using the controllers obtained by Fabricio Saggin on the project PhD thesis *Robust Control for MEMS Inertial Sensor*) that can be obtained considering all the uncertainties of the model. For this purpose, we used a model provided by one of our industrial partners, it includes information about the variation of the parameters with respect to the change of temperature, and the aniso-elasticity. It is important to note that this linear model is being refined thanks to the work done on the PhD project *Joint identification and control of MEMS sensors* of Kévin Colin, so this results will be complemented in the future using those new models. Our task here is the development of a general method for scale factor error evaluation that can be tested with any model ad any controller.

Let consider a given angular rate input $\Omega_{in}$, the scale factor error will is in fact, a relative error given by:

$$\varepsilon_{SF} = \frac{|\Omega_{out} - \Omega_{in}|}{|\Omega_{in}|} \tag{32}$$

This definitions is the same for all the scale factors errors, the only changes are on the uncertainty levels and accepted values of $\varepsilon_{SF}$. Let note the maximum accepted scale factor error $\varepsilon_{SFmax}$. Then the objective will be:

$$\frac{|\Omega_{out}(t) - \Omega_{in}|}{|\Omega_{in}|} \leq \varepsilon_{SFmax} \tag{33}$$

which can be rewritten as:

$$|\Omega_{out} - \Omega_{in}| \leq \varepsilon_{SFmax}|\Omega_{in}| \tag{34}$$

As seen in previous section, the angular rate is estimated by using the Coriolis force that couples the drive mode to the sense mode. Therefore, equation (34) can be redefined in terms of real and estimated Coriolis forces as follows:

$$\frac{|F_{CorEst}(t) - F_{CorReal}(t)|}{2m_x|\dot{x}(t)|} \leq \frac{\varepsilon_{SFmax}\,|F_{CorReal}(t)|}{2m_x|\dot{x}(t)|} \tag{35}$$

Thus, the scale factor error constraint can be written as:

$$|F_{CorEst}(t) - F_{CorReal}(t)| \leq \varepsilon_{SFmax}\,|F_{CorReal}(t)| \tag{36}$$

This constraint will be also respected if we consider the quadratic inequality:

$$(F_{CorEst}(t) - F_{CorReal}(t))^2 \leq \varepsilon_{SFmax}^2\,(F_{CorReal}(t))^2 \tag{37}$$

Since the angular rate input $\Omega_{in}$ is considered to remain constant or to vary considerably slowly with respect to the dynamics of the system, we can consider the inequality of the average during a certain period $T$:

$$\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}(F_{CorEst}(t) - F_{CorReal}(t))^2dt \leq \varepsilon_{SFmax}^2\,\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}(F_{CorReal}(t))^2dt \tag{38}$$

If we consider that $T \to \infty$, it s possible to use the Parseval's identity for the energy of a signal expressed on the frequency domain:

$$\int_{-\infty}^{\infty}(\hat{F}_{CorEst}(j\omega) - \hat{F}_{CorReal}(j\omega))^2d\omega \leq \varepsilon_{SFmax}^2\int_{-\infty}^{\infty}(\hat{F}_{CorReal}(j\omega))^2d\omega \tag{39}$$

For this performance objective, the gyroscope is a MIMO sytem which consist of the reference signal $x_{ref}(t)$ as the only input and two outputs: the real Coriolis force and the estimated Coriolis force. Moreover, we know that in the frequency domain, the system will operate only at frequencies close to the resonance frequency, then we define:

$$x_{ref}(j\omega) = H_w(j\omega)w(j\omega) \tag{40}$$

where $H_w(j\omega)$ is a generator function whoch constraints the set of system input signals $w(j\omega)$ to the set of signals around the resonance frequency. In the other hand, the output signals $z_1(j\omega)$ and $z_2(j\omega)$ : are defined as:

$$z_1(j\omega) = \hat{F}_{CorReal}(j\omega) = H_1(j\omega)\,w(j\omega) \;= H_R(j\omega)H_w(j\omega)\,w(j\omega) \tag{41}$$

$$z_2(j\omega) = \hat{F}_{CorEst}(j\omega) - \hat{F}_{CorReal}(j\omega) = H_2(j\omega)\,w(j\omega) \;= H_E(j\omega)H_w(j\omega)\,w(j\omega) \tag{42}$$

where $H_R(j\omega)$ is the transfer function between the reference and the real Coriolis force, and $H_E(j\omega)$ the transfer function between the reference and the difference of real and estimated Coriolis forces. Then, inequality 39 can be rewritten as:

$$\int_{-\infty}^{\infty} z_2(j\omega)^*z_2(j\omega)^* - w(j\omega)^*H_1^*(j\omega)\varepsilon_{SFmax}^2H_1(j\omega)w(j\omega) \leq 0 \tag{43}$$

From where it is possible to derive the following integral quadratic inequality:

$$\int_{-\infty}^{\infty} \begin{bmatrix} z_2(j\omega) \\ w(j\omega) \end{bmatrix}^* \begin{bmatrix} 1 & 0 \\ 0 & -H_1^*(j\omega)\varepsilon_{SFmax}^2 H_1(j\omega) \end{bmatrix} \begin{bmatrix} z_2(j\omega) \\ w(j\omega) \end{bmatrix} d\omega \leq 0 \tag{44}$$

This integral will be negative for all input signal $w(j\omega)$ if:

$$\begin{bmatrix} H_2(j\omega) \\ H_1(j\omega) \end{bmatrix}^* \begin{bmatrix} 1 & 0 \\ 0 & -\varepsilon_{SFmax}^2 \end{bmatrix} \begin{bmatrix} H_2(j\omega) \\ H_1(j\omega) \end{bmatrix} \leq 0 \tag{45}$$

which establish a quadratic inequality that describes the performance objective for the analysis.
The question now is to use a suitable method to perform the robust performance analysis. It is possible to use $\mu$-analysis for robust performance analysis by introducing a false uncertainty and a weighting function in the LFR. This approach is in general an adequate method when we are facing time-invariant uncertainties, which is the case here, so this was the first approach explored to solve this problem. However, $\mu$-analysis allow only to test robust performance when the performance objective is strictly an input-output property between the inputs $w(j\omega)$ and the output $z(j\omega)$. When analysing the scale factor error, we are evaluating an output-output property since both the real Coriolis force and the error signal $z_2(j\omega)$ are both outputs of the system (figure 10 ). So, this approach is limited not only for the type of uncertainties to be considered, but also for the type of performance objective to be evaluated.

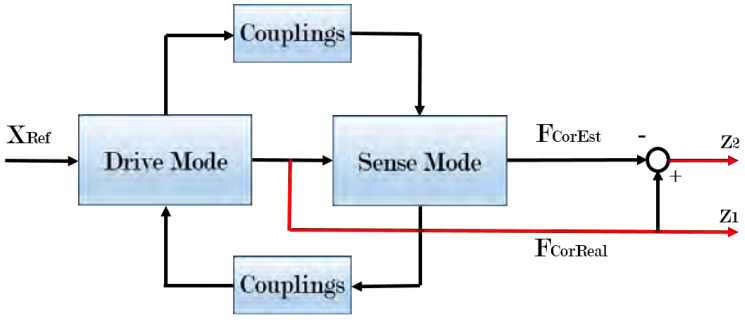



Figure 10: Scale Factor error, system schema.

Therefore, IQC approach was explored and it has shown to offer a solution by introducing a slightly modified version of theorem (3.3) as follows:

**Theorem 3.4.** *Let be the family of closed-loop systems $(M, \Delta)$, with $M(s)$ a stable transfer functions matrix, and $\Delta(s)$ a stable transfer function such that $\Delta(s) \in \underline{\Delta}$. Given a frequency $\omega_0$, the The family of closed-loop systems $(M, \Delta)$ is stable for all $\Delta$, such that $\|\Delta\|_\infty < \beta$ and the scale factor error will be lower than $\varepsilon_{SFmax}^2$ if there exists $\Pi \in \underline{\Pi}$ such that the following constraint is verified :*

$$\begin{bmatrix} M_{qp}(j\omega_0) & M_{qw}(j\omega_0) \\ I & 0 \\ M_{z2p}(j\omega_0) & M_{z2w}(j\omega_0) \\ M_{z1p}(j\omega_0) & M_{z1w}(j\omega_0) \end{bmatrix}^* \begin{bmatrix} \Pi & 0 \\ 0 & \Pi_{perf} \end{bmatrix} \begin{bmatrix} M_{qp}(j\omega_0) & M_{qw}(j\omega_0) \\ I & 0 \\ M_{z2p}(j\omega_0) & M_{z2w}(j\omega_0) \\ M_{z1p}(j\omega_0) & M_{z1w}(j\omega_0) \end{bmatrix} < 0 \tag{46}$$

From this result, an optimisation algorithm under LMI constraints was developed for the frequency analysis of the scale factor error. In order to validate the results, they were compared with the scale factor error in nominal conditions, and with several random-generated models using different combinations of parameters, the results are shown in figure 11.
Assuming general conclusions about the performance of the MEMS gyroscope using this results can be pessimistic, since during the last stages of *Kévin Colin* work, it could be seen that the used model in this first stages is not rigorous enough and consider some assumptions (mainly the case

of aniso-elasticity) that can lead to to over conservative results. Nevertheless, it is worthy to note that the proposed control architecture allow to achieve the performance specifications in the *nominal conditions*. Also, we noticed, as expected, that the change of temperature is still the main contribution for the degradation of the scale factor error, which tells us about te complexity of achieving the robust performance specifications with the used linear control techniques in spite of the good nominal results.

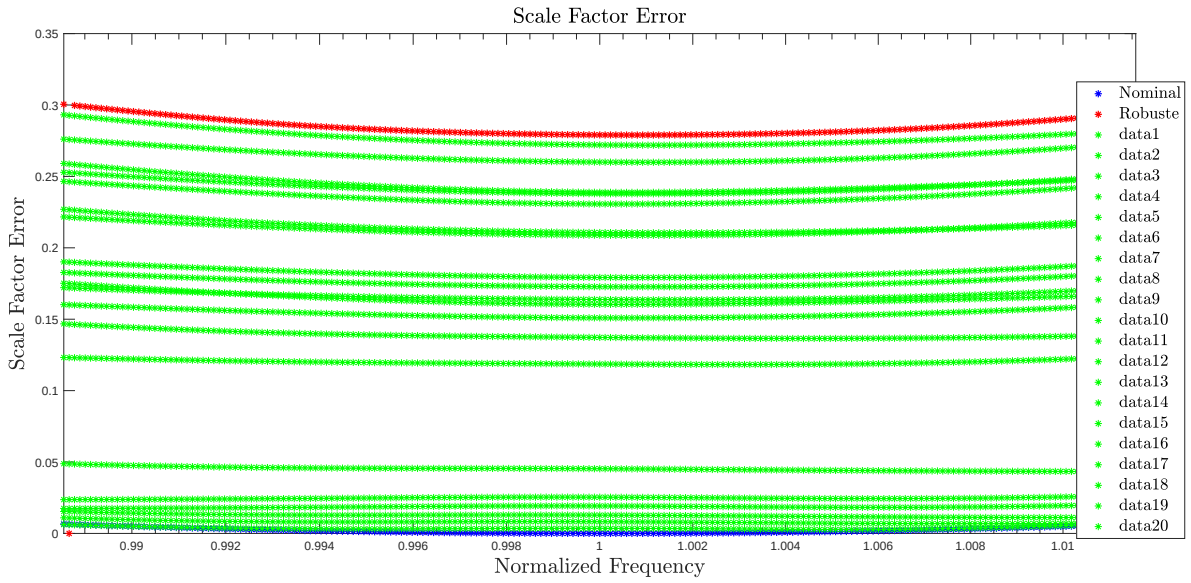



Figure 11: Scale Factor error analysis results.

The previous result confirm the fact that our worst-case approach will allow to obtain more guarantees than with the MonteCarlo validation tests. It is important to note that, since IQC approach is also exploitable for time-varying parameters and certain non-linearities, this is a promising result that can be extended for more complex and complete models of the MEMS gyroscope.

## 3.5 Analysis of stochastic errors

In some applications, like Inertial Navigation Systems (INS), the main interest of using gyroscopes is not to provide a measure of the angular rate, but the angular position with respect to a reference frame. The position of the device is typically obtained by integrating the output of the gyroscope, which is usually distorted by several types of errors such as it was presented in previous chapters. The integration operation will cause an accumulation of the error and then, the computed position will diverge quickly from the real angular position value.

One of this sources of error are the noises presented in the gyroscope. This noises can be of different natures such as white noise, Flicker noise, quantification noise, etc. It is therefore necessary to identify and discriminate the different types of noise in order to evaluate them and to develop methods to reduce its impact on system performance.

Several tools have been introduced for the study and characterization of noises in atomic oscillators, mainly in the time domain using variances. The variance of Allan was then adopted for the study of measurement noise of MEMS accelerometers and gyroscopes as it allows to characterize the main types of noises that are present in MEMS devices.

This section first introduces the Allan variance method as found in the official standard [Iee98]. Then, another interpretation will be proposed in order to find a method that can translate the classical statistical procedure into an optimization problem.

### Allan Variance

Allan variance has become the standard method for evaluating the performance of MEMS gyroscopes, more specifically it is used to characterise the bias drift caused by stochastic processes of several natures. It was first introduced by David Allan in 1966 in order to study the frequency stability of oscillators. Thanks to its simplicity for computation and understanding, it has been adapted and standardized for the study of bias drifts on MEMS gyroscopes and accelerometers [Iee98]. this time-domain analysis method is presented in the following

**Standard computation of Allan Variance** Let consider $\Omega(t)$ the value of the sensor output at the instant $t$ when it remains static (no input) and the bias offset (of deterministic nature) has been already calibrated. Then $\Omega(t)$ will be the result of a random realization with mean value equals to zero.
An averaging time $T$ is set. Then the time history of the signal is divided into clusters of duration $T$. The average of a given cluster is given by:

$$\bar{\Omega}_k(T) = \frac{1}{T}\int_{t_k}^{t_k+T} \Omega(t)dt \tag{47}$$

where $\bar{\Omega}_k(T)$ is the cluster average of the gyroscope output in the interval $t_k$ and $t_k + T$. So the average of the subsequent cluster is given by:

$$\bar{\Omega}_{k+m}(T) = \frac{1}{T}\int_{t_{k+m}}^{t_{k+m}+T} \Omega(t)dt \tag{48}$$

The Allan variance of averaging time $T$ is defined as the variance between two averaged subsequent clusters:

$$\sigma^2(T) = \frac{1}{2}\left\langle \left[\bar{\Omega}_{k+m}(T) - \bar{\Omega}_k(T)\right]^2 \right\rangle \tag{49}$$

where $\langle\rangle$ denote the averaging operation. Finally, the Allan deviation is computed by obtaining the square root of equation **??**. This result is plot as a function of the averaging time $T$, from where it is possible to characterize the noise in the MEMS gyroscope. The different random processes appear in the form of different slopes on the Allan deviation plot as we can see in figure 12.

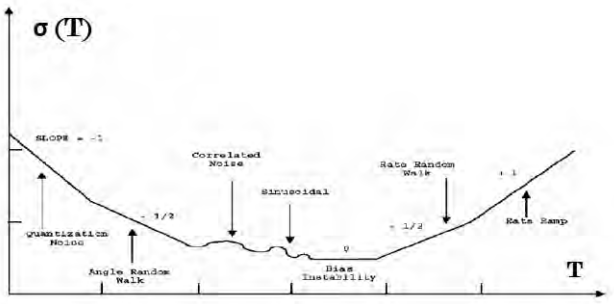



Figure 12: Allan Deviation Plot

### Reinterpretation of Allan Variance

Indeed, 47 can be seen as an LTI convolution system with input $\Omega(t)$ and output $\bar{\Omega}_k(t)$ with an impulse response $h(t)$ given by the rectangle function:

$$h(t) = \frac{1}{T} rect\left(\frac{t + T/2}{T}\right) \tag{50}$$

Then, by definition, the frequency response $H(\nu)$ of this system will be given by the Fourier transform of its impulse response, leading to:

$$H(\nu) = e^{2\pi j T\nu} sinc(T\nu) \tag{51}$$

From where, it is possible to express the Power Spectral Density (PSD) of the averaged output $\bar{\Omega}_k(t)$ with respect to the DSP of the input $\Omega(t)$:

$$\forall\ \nu \in \mathbb{R} \quad S_{\bar{\Omega}}(\nu) = sinc^2(T\nu) S_{\Omega}(\nu) \tag{52}$$

The second step is to determine the variation of the averaged signal $\bar{\Omega}_k(T)$. This can be interpreted in a first time, as the discretization of $\bar{\Omega}_k(T)$ at a sample rate $T$, then we can define the signal $d(T)$ given by:

$$\forall\ k,\ \ \forall\ t \in [kT\ \ (k+1)T] \quad d(T) = \frac{1}{\sqrt{2}} \bar{\Omega}((k+1)T) - \bar{\Omega}(kT) \tag{53}$$

Which is the weighted difference between the signal at current instant and the same signal delayed of $kT$. Then, Allan variance with respect to averaging time $T$ is equivalent to the power of the signal $d(T)$:

$$\sigma^2(T) = P_d = \int_{-\infty}^{\infty} d^2(t) dt \tag{54}$$

An interesting aspect is the fact that considering the stationarity and ergodicity of $\bar{\Omega}_k(t)$, it can be shown that the power $P_d$ is given by:

$$P_d = R_{\bar{\Omega}}(0) - R_{\bar{\Omega}}(T) \tag{55}$$

From where, it is possible to determine (ANNEXE) the frequency relationship between the output of the gyroscope $\Omega(t)$ and the Allan variance $\sigma^2(T)$:

$$\sigma^2(T) = 4 \int_0^{\infty} S_{\Omega}(\nu) \frac{sin^4(\pi\nu T)}{(\pi\nu T)^2} d\nu \tag{56}$$

Which can be seen as the power of the signal $\Omega(t)$ which passes through a filter with frequency response $\frac{sin^4(\pi f T)}{(\pi f T)^2}$. This filter allow us to isolate the signal in a certain range of frequency where a certain type of noise is predominating. hen, depending on the averaging time $T$ we will evaluate the impact of a given predominating type of noise. Figure 13 it shown the frequency response of the "filter" that links the DSP of the angular rate $\Omega$ to the DSP of signal $d(T)$ for different values of averaging time $T$.

Let consider now the following example. Let assume that at the output of the sensor $\Omega(t)$, we obtain exclusively a white noise signal with DSP $S_{\Omega} = Q_0^2$ as represented in figure 14. Then:

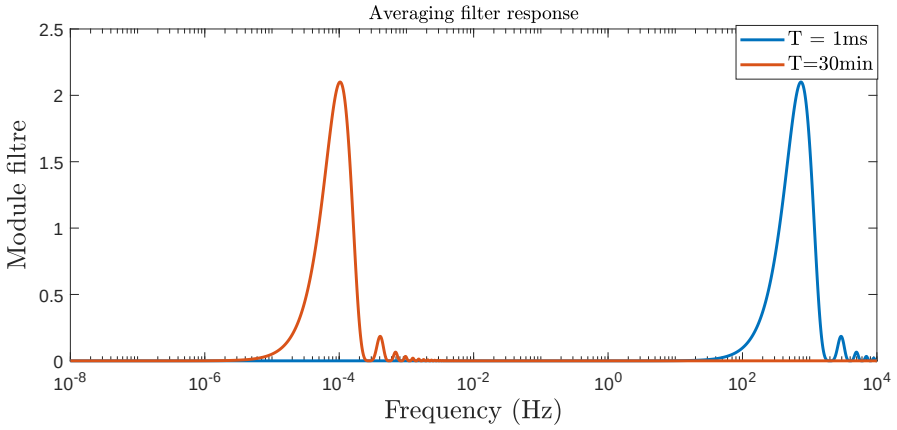



Figure 13: Frequency response of averaging filters at different $T$.

$$\begin{aligned}\sigma(T) &= 4\int_0^\infty S_\Omega(f)\frac{sin^4(\pi\nu T)}{(\pi\nu T)^2}d\nu \\ &= 4Q_0^2\int_0^\infty \frac{sin^4(\pi\nu T)}{(\pi\nu T)^2}d\nu \\ &= \frac{4Q_0^2}{\pi T}\int_0^\infty \frac{sin^4(\pi\nu T)}{(\pi\nu T)^2}d\nu \\ &= \frac{Q_0^2}{T}\end{aligned}$$

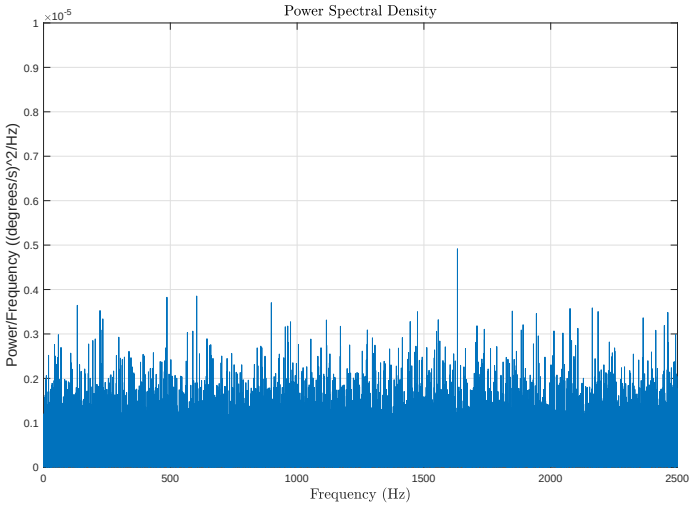



Figure 14: White Noise signal to be tested.

Since the Allan Deviation plot is represented using the log-log plot of the square root of Allan variance (or Allan deviation) with respect to the averaging time $T$, the white noise signal is represented in the Allan deviation plot as a linear function with a slope of -1/2. The computation of the Allan deviation with respect to the averaging time $T$ for the white noise of figure 14 was computed and as expected, the figure 15 represents the obtained result.

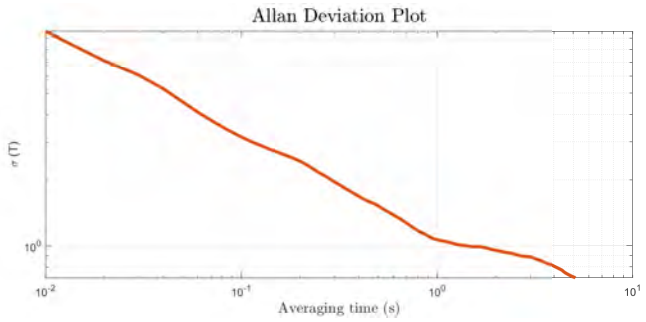



Figure 15: Allan deviation plot for a white noise signal

It is interesting to note that since the value of the Allan deviation for a given averaging time $T$ can be actually seen as the value of the $L_2$ norm of the previously defined signal $d(t)$

$$\sigma(T) = \|d\|_2 = \sqrt{\int_{-\infty}^{\infty} d^2(T)dt} \tag{57}$$

In conclusion, it could be possible to transform the Allan Variance post-design analysis into an optimization problem, where the objective would be the minimization of the norm $H_2$ of the system in figure 16 as long as an equivalent LTI system of the global system could be determined. The work to be done in this line of research will be presented in the perspectives of the report.

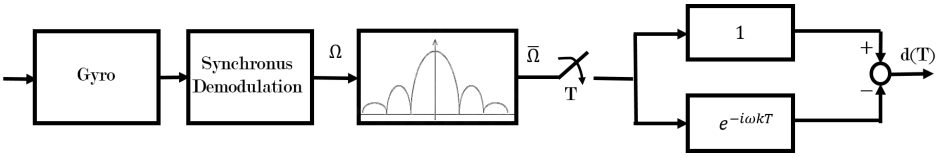



Figure 16: Equivalent system of Allan Variance

# 4 Conclusion and thesis Roadmap

This report presented the subject and challenges of this PhD. First a bibliographic review about the main imperfections and phenomena that degrade the performance of MEMS gyroscopes, and about the current methods used nowadays for performance validation, which put on evidence the need of developing new pre-experimentation methods for performance validation.

During this year, a first available model was studied and exploited to take into account the possible uncertainties of the system, from where an uncertain model of the system was developed and implemented on *matlab* for the different robustness tests. After that, studying the performance objectives, it was found that there were two main problems to be solved: to develop a tool that allow to estimate the scale factor error considering all the possible uncertainties. Secondly, the study of the noise impact on the system. For solving the scale factor error aspect, several classical results where explored and tried to be mastered, and using a slightly modified version of an IQC-based robust performance theorem, it was possible to obtain a tool for robust performance validation of the scale factor error. The noise aspect was faced by studying the Allan variance method considered in the standards, and to look for a reinterpretation in order to derive a new method to predict the noise impact on the closed-loop system.

Tackling this problems and extending the partial results presented in this report, we aim to propose a general and systematic method to guarantee the robust performance validation of MEMS sensors before experimentation.

The future work to be achieved in the following includes:

1. Analysis of noise impact on MEMS measurement:
    - Identification of the noise at the output of the gyroscope in order to establish the range of frequency where they are predominant.
    - Determine the origin of the different noises, this means, in which part of the closed-loop system they appear.
    - To look for a linear approximation of the system when the synchronous demodulation is included.
    - Establish the general optimization problem for the norm $H_2$ of the equivalent system of figure 16.
2. Analysis of scale factor error:
    - Inclusion of the phenomenons identified by Kévin Colin (see the report *Joint identification and control of MEMS sensors*), a remarkable result to consider is the identification of the electrical coupling.
    - Extension of the approach to the case of time-varying uncertainties and non-linearities.
3. Improvement of the uncertain model:
    - The future results of identification with respect to the change of temperature and angular rate will be exploited to obtain a more finely derived model of the parametric uncertainties.
    - Possible reduction of the uncertain parameters, since the temperature could be modelled as the main uncertainty, and the variation of the on the parameters can be defined as the function of temperature on the new model.

## References

[ABSS09] Al-Hussein Albarbar, Abdellatef Badri, Jyoti K Sinha, and A Starr. Performance evaluation of mems accelerometers. *Measurement*, 42(5):790–795, 2009.

[Bay12] Benoît Bayon. *Estimation robuste pour les systèmes incertains.* PhD thesis, Ecole Centrale de Lyon, 2012.

[Cha13] Julien Chaudenson. *Robustness analysis with integral quadratic constraints, application to space launchers.* PhD thesis, Supélec, 2013.

[Doy82] John Doyle. Analysis of feedback systems with structured uncertainties. In *IEE Proceedings D-Control Theory and Applications*, volume 129, pages 242–250. IET, 1982.

[DZG07] Lili Dong, Qing Zheng, and Zhiqiang Gao. A novel oscillation controller for vibrational mems gyroscopes. In *American Control Conference, 2007. ACC'07*, pages 3204–3209. IEEE, 2007.

[EKT+11] Amro M. Elshurafa, Kareem Khirallah, Hani H. Tawfik, Ahmed Emira, Ahmed K.S. Abdel Aziz, and Sherif M. Sedky. Nonlinear dynamics of spring softening and hardening in folded-mems comb drive resonators. *Journal of Microelectromechanical Systems*, 20(4):943–958, 2011.

[FKP+05] Michael I. Ferguson, Didier Keymeulen, Chris Peay, Karl Yee, and Daliang Leon Li. Effect of temperature on MEMS vibratory rate gyroscope. *IEEE Aerospace Conference Proceedings*, 2005, 2005.

[FLW11] Yongzhen Fan, Bing Luo, and Ancheng Wang. Analysis of temperature adaptability for frequency control loop for silicon micromechanical gyroscope. *Proceedings - IEEE 2011 10th International Conference on Electronic Measurement and Instruments, ICEMI 2011*, 4(4):346–349, 2011.

[FTD91] Michael KH Fan, Andre L Tits, and John C Doyle. Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics. *IEEE Transactions on Automatic Control*, 36(1):25–38, 1991.

[GHL+15] Rui Guan, Chunhua He, Dachuan Liu, Qiancheng Zhao, Zhenchuan Yang, and Guizhen Yan. A temperature control system used for improving resonant frequency drift of MEMS gyroscopes. *2015 IEEE 10th International Conference on Nano/Micro Engineered and Molecular Systems, NEMS 2015*, pages 397–400, 2015.

[GLLG03] Richard Giroux, René Jr Landry, Barrie Leach, and Richard Gourdeau. Validation and performance evaluation of a simulink inertial navigation system simulator. *Canadian aeronautics and space journal*, 49(4):149–161, 2003.

[Iee98] Ieee. *IEEE Standard Specification Format Guide and Test Procedure for Single –Axis Interferometric Fiber Optic Gyros*, volume 1997. 1998.

[Kes82] M.S. Keshner. 1/F Noise. *Proceedings of the IEEE*, 70(3):212–218, 1982.

[KHC+08] Bongsang Kim, Matthew A. Hopcroft, Rob N. Candler, Chandra Mohan Jha, Manu Agarwal, Renata Melamud, Saurabh A. Chandorkar, Gary Yama, and Thomas W. Kenny. Temperature dependence of quality factor in MEMS resonators. *Journal of Microelectromechanical Systems*, 17(3):755–766, 2008.

[KJCT12] Martti Kirkko-Jaakkola, Jussi Collin, and Jarmo Takala. Bias prediction for mems gyroscopes. *IEEE Sensors Journal*, 12(6):2157–2163, 2012.

[Lel05] Robert P. Leland. Mechanical-thermal noise in MEMS gyroscopes. *IEEE Sensors Journal*, 5(3):493–500, 2005.

[LLL08] Di Li, Rene Landry, and Philippe Lavoie. Low-cost mems sensor-based attitude determination system by integration of magnetometers and gps: A real-data test and performance evaluation. In *Position, Location and Navigation Symposium, 2008 IEEE/ION*, pages 1190–1198. IEEE, 2008.

[MR97] Alexandre Megretski and Anders Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, 1997.

[MYM12] A. Mohammadi, M. R. Yuce, and S. O. R. Moheimani. A Low-Flicker-Noise MEMS Electrothermal Displacement Sensing Technique. *Journal of Microelectromechanical Systems*, 21(6):1279–1281, dec 2012.

[PTS12] I P Prikhodko, a a Trusov, and a M Shkel. Achieving Long-Term Bias Stability in High-Q Inertial Mems By Temperature Self-Sensing With a 0.5 Millicelcius Precision. *Solid-State Sensors, Actuators, and Micosystems Workshop, Hilton Head*, 2012.

[Sco97] Gérard Scorletti. *Approche Unifiée de l'Analyse et de la Commande des Systèmes par Optimisation LMI*. PhD thesis, Université Paris Sud-Paris XI, 1997.

[SH06] Isaac Skog and Peter Händel. Calibration of a mems inertial measurement unit. In *XVII IMEKO World Congress*, pages 1–6, 2006.

[SP05] S. Skogestad and I. Postlethwaite. Multivariable feedback control: analysis and design. *International Journal of Robust and Nonlinear Control*, 8(14):575, 2005.

[SP07] Sigurd Skogestad and Ian Postlethwaite. *Multivariable feedback control: analysis and design*, volume 2. Wiley New York, 2007.

[SSP+06] A. Srikantha Phani, Ashwin A. Seshia, Moorthi Palaniapan, Roger T. Howe, and John A. Yasaitis. Modal coupling in micromechanical vibratory rate gyroscopes. *IEEE Sensors Journal*, 6(5):1144–1152, 2006.

[TO95] Onur Toker and Hitay Ozbay. On the np-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback. In *American Control Conference, Proceedings of the 1995*, volume 4, pages 2525–2526. IEEE, 1995.

[Vos79] Richard F Voss. 1/f (flicker) noise: A brief review. In *33rd Annual Symposium on Frequency Control. 1979*, pages 40–46. IEEE, 1979.

[VSK16] Joost Veenman, Carsten W Scherer, and Hakan Köroğlu. Robust stability and performance analysis based on integral quadratic constraints. *European Journal of Control*, 31:1–32, 2016.

[You11] Mohammad I. Younis. *MEMS Linear and Nonlinear Statics and Dynamics*, volume 20 of *Microsystems*. Springer US, Boston, MA, 2011.

[Zam66] George Zames. On the input-output stability of time-varying nonlinear feedback systems part one: Conditions derived using concepts of loop gain, conicity, and positivity. *IEEE transactions on automatic control*, 11(2):228–238, 1966.

[ZTSL10] Hao Zhou, Hailin Tang, Wei Su, and Xianxue Liu. Robust design of a mems gyroscope considering the worst-case tolerance. In *Nano/Micro Engineered and Molecular Systems (NEMS), 2010 5th IEEE International Conference on*, pages 1012–1016. IEEE, 2010.

# A Uncertainty representation

As seen in previous sections, there are several phenomena that affect the ideal operation of MEMS gyroscopes. In general, these "imperfections" are not consider for the model used in control design. However, it is desired to obtain the maximum of guarantees that the controlled system will satisfy the design requirements despite the differences between the design model and the actual system.
The first step is to find a mathematical representation of the model uncertainties. This demands to exploit the maximal available information about the uncertain part of the system.
Whatever the source of model uncertainty is, it may be possible to represent it by one of the two main classes:

- Parametric uncertainties: Uncertainty on a specific model parameter.
- Dynamic uncertainties: It represents directly the error caused by the missing dynamics (unknown or intentionally neglected).

## Parametric uncertainties

### Representation of parametric uncertainties

If some parameter is uncertain, we say that there exist a lack of knowledge about which value the parameter can take at a given time, however, we may know at least which can be the interval of its variation. Let consider the uncertain parameter $\theta$, we can define a set of all its possible values as follows:

$$\theta \in [\theta_{min}, \theta_{max}] \tag{58}$$

The uncertain parameter can be introduced in its "normalized" representation [SP05], it means that the uncertainty is centered at zero and has a size equals to one:

$$\theta = \theta_0 + \theta_g \, \delta_\theta \tag{59}$$

Where $\theta_0 = \frac{\theta_{min}+\theta_{max}}{2}$ is the nominal value of the parameter, $\theta_g = \frac{\theta_{min}-\theta_{max}}{2}$ is the maximal variation of $\theta$ such that $\theta_{max}\,(\theta_{min}) = \theta_0 + (-)\theta_g$, and $\delta_\theta$ is the normalized uncertainty such that $|\delta_\theta| \leq 1$.

### Uncertainties on MEMS Gyroscope model

Some information about the interval of the uncertain parameters can be obtained from the GYPRO3300 model:

- Linear relationship between the resonance frequency and temperature, also we now that the temperature range of operation for the gyroscopes is between $-40$ and $60°$, then it is possible to translate this as the uncertainty interval for the resonance frequency.
- Linear approximation of the relationship between the Q-factor and temperature, then it is possible to determine the uncertainty interval of the Q-factor.
- Using the maximal quadrature error measured during the identification, it is possible to define the maximum value of the non-diagonal stiffness constant $k_{xy} = k_{yx}$ by using equation **??**.
- For robustness analysis, it is possible to consider the angular rate $\Omega_z$ as an uncertain parameter whose uncertainty interval is given by the design specifications.

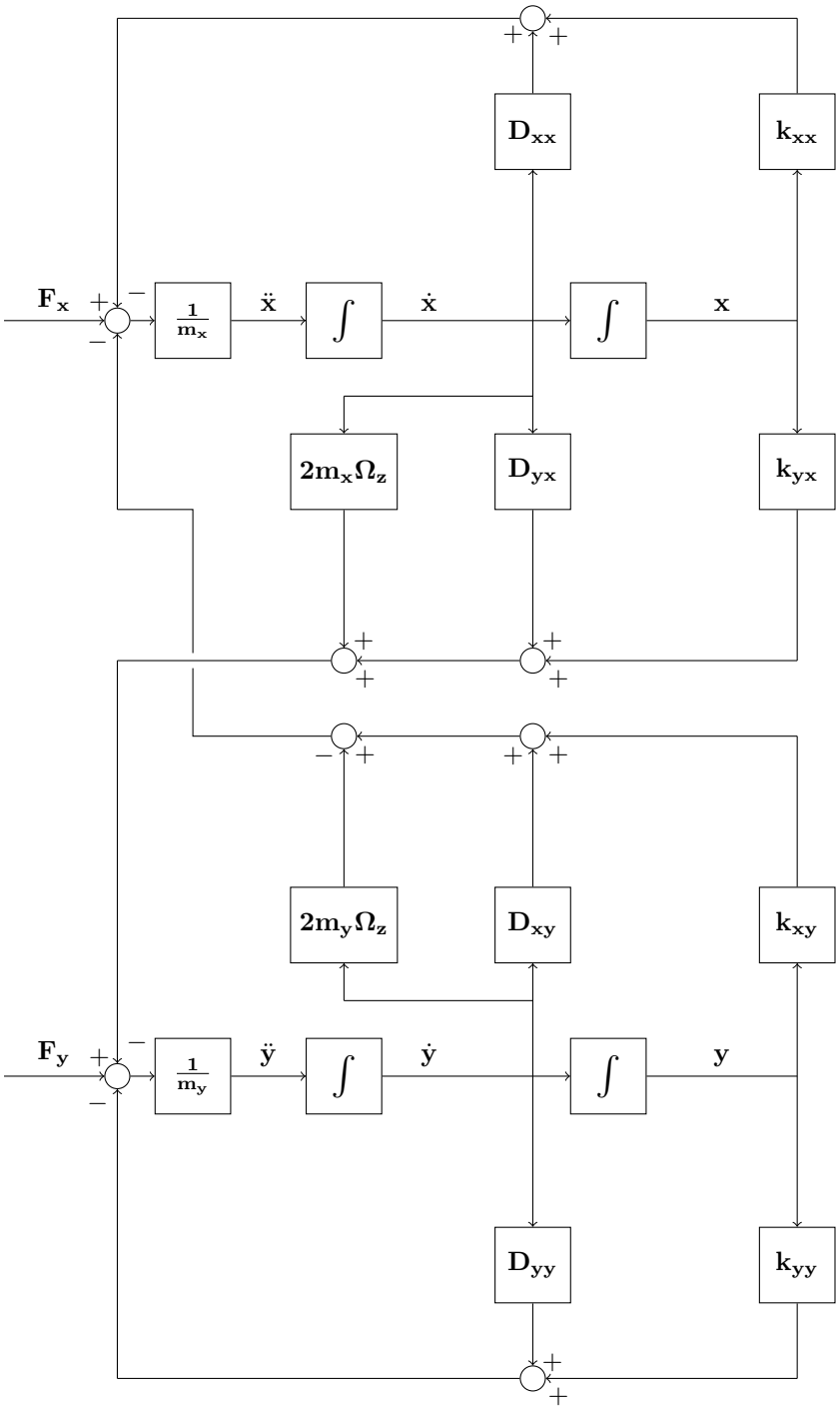



Figure 17: Block diagram of Non-ideal MEMS Gyroscope

Since the masses $m_x$ and $m_y$ can not change, it is possible to translate the uncertainties on the resonance frequency and Q-factor to the space of parameters by using equations **??** and **??**. The parameter space representation of uncertainty was chosen because the number of uncertain parameters is reduced with respect to "frequency" response parameters (resonance frequencies and Q-factors).
The block diagram if the MEMS gyroscope model is shown on figure 17. The next step is to observe if it is possible to find a minimal representation for the uncertain model.

### Dynamic uncertainties

The objective of considering a dynamic uncertainty is to evaluate if the controlled system designed from an identified model can remain stable and achieve the specifications when operates on the real system. A real mechanical can include an infinite number of flexible modes which appear at high frequencies. However, it is only possible to identify the system until a certain frequency $\omega_{max}$. Thus, all the dynamics beyond this frequency are unknown. This lack of knowledge can

#### Representation of dynamic uncertainty

To evaluate the uncertainty, it is possible to evaluate the absolute error or the relative error between the real system and the model. The relative error can be more convenient in this case, since it will express the fact that the system is well known in the interval of frequencies where the system was identified and outside this interval, there exists some neglected dynamics. The question now is to look for a dynamic uncertainty which can represent this relative error.

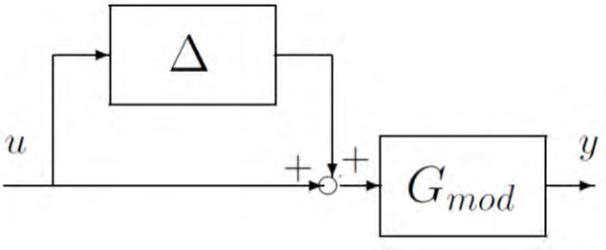



Figure 18: Direct multiplicative uncertainty

We denote the model $G_{mod}$ and the real system $G_{real}$. Let consider the input multiplicative uncertainty $\Delta_{dir}$ as shown in figure A . The real system is written as follows:

$$G_{reel} = G_{mod}(I + \Delta_{dir}) \tag{60}$$

And then, the direct relative error is:

$$\Delta_{dir} = G_{mod}^{-1}(G_{reel} - G_{mod}) \tag{61}$$

Which is actually a relative error, the main motivation of using multiplicative uncertainty is to avoid the mismatch of possible zeros when several flexible modes can be neglected in the model. Now, we are considering a frequency dependent uncertainty $\Delta(j\omega$, to represent this aspect, it is necessary to consider some stable transfer functions $W_1(j\omega)$ and $W_2(j\omega$ called "weighting functions" such that $\forall\ \omega\ |\Delta(j\omega)| \leq |W_1(j\omega)W_2(j\omega)|$ (figure A). Normalizing the dynamic uncertainty (as it was done with parametric uncertainties), we can define the set of $\Delta(j\omega) = W_1(j\omega)\hat{\Delta}W_2(j\omega)\ \forall\omega$ with $|\hat{\Delta}| \leq 1$ (or $\|\hat{\Delta}\| \leq 1$ for the MIMO case).
The question now, is to determine how to choose the weighting functions $W_1(j\omega)$ and $W_2(j\omega)$. Since there is not any information about the model errors, we will simplify the representation by choosing $W_2(j\omega == I$, and then we will define the level and ranges of frquency of the dynamic uncertainty by defining $W_1(j\omega)$. Also the input multiplicative uncertainty will represent the relative error with respect to all the transfer functions. Therefore, $\hat{\Delta}$ will be a $2 \times 2$ full-block uncertain matrix:

$$\hat{\Delta} = \begin{bmatrix} \hat{\Delta}_{drive \to drive} & \hat{\Delta}_{drive \to sense} \\ \hat{\Delta}_{sense \to drive} & \hat{\Delta}_{sense \to sense} \end{bmatrix} \tag{62}$$

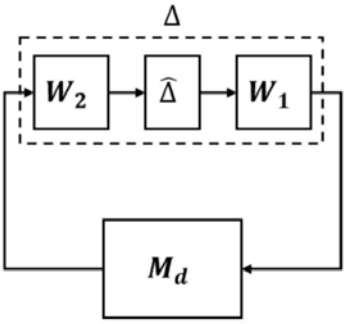



Figure 19: Weighted Dynamic Uncertainty

Then, it is necessary to consider two weighting filters $W_1^1(j\omega)$, and $W_1^2(j\omega)$. In this case, we will consider the same amount of uncertainty for all the transfer functions, so the weighting filters will be the same. Now, to define the weighting functions, we assume that inside the range of frequencies in which the system was identified, the relative error is very small, and, beyond the maximal frequency of identification ($\omega > 17000Hz$), the relative error can be considerable. So, the weighting functions were chosen as show in figure A.

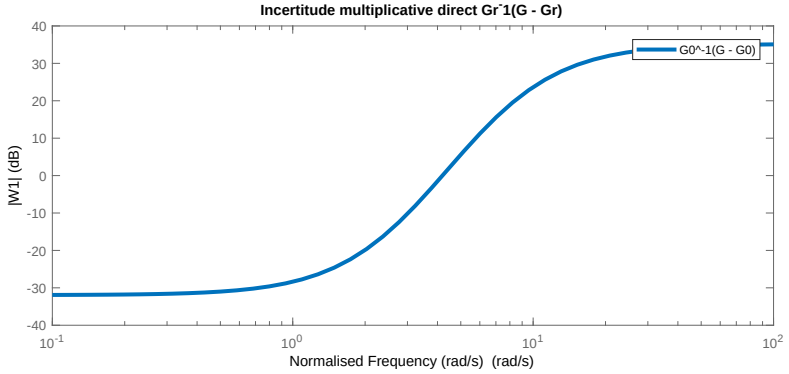



Figure 20: Dynamic Uncertainty Weighted Function

$$M_{d_{dir}} = (I + L(j\omega)^{-1}L(j\omega) = T[j\omega) \tag{63}$$

where $L(j\omega)$ is the open loop transfer function $K(j\omega)G(j\omega)$ and $T(j\omega)$ is the complementary sensitivity function. Let denote $\Omega$ the operation frequency interval, in this range of frequencies, the open loop system has a high gain

### Linear fractional representation of uncertain system

The linear fractional representation allows to separate the nominal part of a dynamic system (represented by the nominal LTI system), and the uncertain part of the system (represented by an uncertain structured block), the Linear Fractional representation is shown in figure A.
where $M_d$ is the nominal LTI system, whose state-space representation is given by:

$$\begin{bmatrix} \dot{x} \\ q \\ z \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} x \\ p \\ u \end{bmatrix} \tag{64}$$

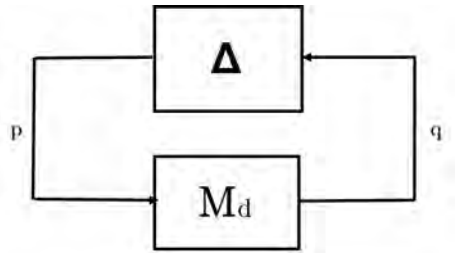



Figure 21: Linear Factional Representation

and the uncertain block is:

$$\Delta = \begin{bmatrix} \delta_{kxx} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \delta_{Dxx} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_{kyy} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_{Dyy} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_{kyx} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta_{Dyx} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta_{kxy} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta_{Dyx} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \Delta_{\Omega_z} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \hat{\Delta} \end{bmatrix} \quad with \quad \Delta_{\Omega_z} = \begin{bmatrix} \delta_{\Omega_z} & 0 \\ 0 & \delta_{\Omega_z} \end{bmatrix} \tag{65}$$

As all the normalized parametric uncertainties and dynamic uncertainties are bounded by 1, and considering the block diagonal structure of the uncertainty block. Therefore, all $\Delta =\in \Delta$ is bounded by $\|\Delta\|_\infty \leq 1$.

Université de Lyon
CNRS, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1

**Laboratoire Ampère**
Unité Mixte de Recherche du CNRS - UMR 5005
Génie Electrique, Automatique, Bio-ingénierie

# *Mémoire doctorant 1ère année*
# *2017 -2018*

| | |
|---|---|
| Nom - Prénom | **COLIN Kévin** |
| email | **kevin.colin@ec-lyon.fr** |
| Titre de la thèse | **Joint identification and control of MEMS sensors.** |
| Directeur de thèse | **BOMBOIS Xavier** |
| Co- encadrants | **KORNIIENKO Anton** |
| Dpt. de rattachement | **MIS** |
| Date début des travaux | **01/10/17** |
| Type de financement | **Contrat de recherche ECL, dans le cadre du projet NEXT4MEMS financé par la BPI** |

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France
Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

**Abstract**: This report presents the work done during the first year of the PhD thesis entitled "Joint identification and control of MEMS sensors". It is about the modeling of a particular MEMS sensor, the gyroscope GYPRO3300, implemented on the electronic card manufacturer of this consortium, called Mobyle. A new way to model linearly the phenomena involved in a MEMS gyroscope is presented in this report. We use a black-box approach with prediction-error minimization identification. This model is more accurate than the ones proposed in the literature survey. Some perspectives for its improvement are given in this report.

**Key-words**: MEMS gyroscope, model, system modeling, system identification, black-box modeling, grey-box modeling prediction-error minimization, uncertainties, LTI models, LPV models, parallel modeling, Experiment Design



## Contents

## List of Figures



## List of Tables

## List of Acronyms

| | |
|---|---|
| BPI | Banque Publique d'Investissement |
| CNRS | Centre National de la Recherche Scientifique |
| DSP | Density Spectral of Power |
| ECL | École Centrale de Lyon |
| FFT | Fast Fourier Transform |
| LTI | Linear Time Invariant |
| LPV | Linear Parameter Varying |
| MEMS | Micro Electro-Mechanical Structure |
| MIMO | Multiple Inputs Multiples Outputs |
| MIS | Méthodes pour l'Ingénierie des Systèmes |
| OE | Output Error |
| PEM | Prediction-Error Method |
| SISO | Single Input Single Output |
| SNR | Signal to Noise Ratio |

## List of Notations

| | |
|---|---|
| $\Omega$ | Real angular rate |
| $\hat{\Omega}$ | Deduced angular rate |
| $O$ | Center of the mass of the gyroscope |
| $x$ | Displacement along the drive mode |
| $y$ | Displacement along the sense mode |
| $F_x$ | External force on the drive mode |
| $F_y$ | External force on the sense mode |
| $V_{in_x}$ | Input voltage on the drive mode |
| $V_{in_y}$ | Input voltage on the sense mode |
| $V_{in}$ | Input voltage vector $V_{in} = (V_{in_x}\ V_{in_y})^T$ |
| $V_{in}^2$ | Square input voltage vector $V_{in}^2 = (V_{in_x}^2\ V_{in_y}^2)^T$ |
| $V_{out_x}$ | Output voltage on the drive mode |
| $V_{out_y}$ | Output voltage on the sense mode |
| $V_{out}$ | Output voltage vector $V_{out} = (V_{out_x}\ V_{out_y})^T$ |
| $C^{act}$ | Actuation capacitance |
| $C^{mea}$ | Measurement capacitance |
| $t$ | time variable |
| $s$ | Laplace variable |
| $z$ | $\mathcal{Z}$ transform variable, shift operator |
| $E$ | $2 \times 2$ transfer function matrix for parasite electrical effect of the gyroscope |
| $H$ | $2 \times 2$ transfer function matrix for measurement noise of the gyroscope |
| $G$ | $2 \times 2$ transfer function matrix for mechanical dynamics of the gyroscope |
| $T$ | Temperature of the gyroscope |
| $m_x$ | Mass of the drive mode |
| $m_y$ | Mass of the sense mode |
| $d_{xx}$ | Damping coefficient on the drive mode |
| $d_{yy}$ | Damping coefficient on the sense mode |
| $d_{xy}$ | Cross-damping coefficient from sense mode to drive mode |
| $d_{yx}$ | Cross-damping coefficient from drive mode to sense mode |
| $k_{xx}$ | Stiffness coefficient on the drive mode |
| $k_{yy}$ | Stiffness coefficient on the sense mode |
| $k_{xy}$ | Cross-stiffness coefficient from sense mode to drive mode |
| $k_{yx}$ | Cross-stiffness coefficient from drive mode to sense mode |
| $f_{0_x}$ | Resonance frequency of the drive mode (in Hz) |
| $f_{0_y}$ | Resonance frequency of the sense mode (in Hz) |
| $\omega_{0_x}$ | Resonance frequency of the drive mode (in rad/s) |
| $\omega_{0_y}$ | Resonance frequency of the sense mode (in rad/s) |
| $\epsilon$ | Prediction-error signal |
| $\theta$ | Vector of the model parameters |
| $\theta_0$ | Vector of the true parameters |
| $N$ | Number of consecutive data |
| $\hat{\theta}_N$ | Vector of the estimated parameters with $N$ consecutive data |
| $P(\theta_0)$ | Covariance matrix of the parameter vector $\theta_0$ |
| $\Phi_u$ | Input power spectrum |

# 1 Introduction

This report presents the work done during the first year of the PhD work "Joint identification and control of MEMS sensors". This PhD thesis takes place within the MIS department (Méthodes pour l'Ingénierie des Systèmes), one of the departments of the laboratory Ampère at the ECL (Ecole Centrale de Lyon) site. It is supervised by Xavier Bombois (CNRS) and Anton Korniienko (Ampère). This thesis is founded by the BPI (Banque Publique d'Investissement) and is attached to a research project called NEXT4MEMS. It focuses on micro electromechanical structure (MEMS) motion sensors. Two other PhD students are working on that project: Fabrício Saggin working on the control of the MEMS sensors and Jorge Ayala working on the robustness analysis of the MEMS. Our three PhD subjects are linked and force us to work as a team which is a strong advantage of that project. This project deals with motion sensors, encountered in several applications (smartphones, planes, etc). The consortium of this project is composed by two laboratories including Ampère [1]. The aim of the NEXT4MEMS project is

> "the development of a novel industrial sector aiming at the production of a new generation of MEMS inertial sensors with higher performance (as e.g. required by the aerospace industry). To cover the multiple facets of this ambitious project, the project consortium consists of the French leaders in the inertial sensor [...] industry and two academic laboratories that will be in charge of the related fundamental research challenges [...] Ampère for the control engineering aspects" [One, 2018]

The attraction for this type of motion sensors is due to the fact that they are less expensive, less cumbersome and less power consuming compared to other technologies like the optical ones. They are also less accurate justifying the use of control techniques to improve their accuracy. To have an efficient control we need an accurate model of the behaviors of those sensors. We will focus on the MEMS gyroscope, measuring angular rate by using Coriolis effect.

The MEMS gyroscope using Coriolis effect is composed by a structure with a mass oscillating in two directions linked to the structure by dampers and springs in each direction [Saukoski, 2008]. The structure is fixed to the object whose angular rate needs to be measured. Each direction is acting as a mechanical resonant mode. The Coriolis effect couples both modes. In each mode there is an external force which can be applied. The Coriolis effect is also an external force for one mode which is function of the angular rate multiplied by the velocity of the other mode. We can actuate and measure the displacements with a capacitive instrumentation [Saukoski, 2008]. The common strategy for the deduction of the angular rate is to keep a constant motion in one mode (drive mode) by tracking a reference signal with a controller and demodulate the output of the other mode (sense mode) (as developed for instance in [Chen et al., 2003] and [Egretzberger et al., 2012]). Then an accurate model is needed for the controller design and the demodulation of the sense displacement. The research question explored in this report and currently followed in this PhD thesis is the following one:

**How to compute an accurate model for the gyroscope in order to minimize the angular rate deduction error when this model is used for the deduction of the angular rate?**

In a first step we identify the dynamics without angular rate. The first approach to model a system is to use physics equations (white box-modeling). In the literature survey the most simple model is described by two resonance equations with the Coriolis force effect that couples both modes. External forces are used for the actuation of both modes. In this model the outputs are the displacements and the inputs are the external forces as presented in [Fei and Yang, 2011]. However due to manufacturing nonidealities, such as non proportional damping and anisoelasticity effects, the modes are coupled

1. We will not give he names of the members of the consortium in this document. One can report to the document on the site http://www.ampere-lab.fr/spip.php?article885 to get more informations.

mechanically even without Coriolis effect. This has been considered in the paper [Fei and Yang, 2011]. Moreover this model does not take account of the instrumentation of the modes (the way to actuate and measure the displacements of both modes). In the MEMS gyroscope of this study, the GYPRO330, we use capacitive actuations and measurements like in [Saukoski, 2008]. The model must take account of this instrumentation and the new considered inputs and outputs for the model are voltages. In our case the force is proportional to the square of the input voltage and the displacement is proportional to the output voltage. But due to electrical field interactions there is a parasite effect from the input voltage to the output voltage that is not modeled by the equations. The input of this phenomene is directly the input voltage. The model in [Saukoski, 2008] is not accurate anymore in our case. Then the approaches using models from physics equations are not sufficient for the modeling of GYPRO3300.

The second possibility is to use a data-based approach called System Identification, i.e. developing a model from experimental data. Indeed unmodeled dynamics with physics models can be captured by using this second approach. In our case we are able to model the mechanical dynamics and the parasite electrical effect. It has been done in [Chen et al., 2005] and [M'Closkey et al., 1999]. But here the inputs of both phenomena are the same as the author considered the electrical field as input and output. In our case the mechanical dynamics are driven by the square of the input voltage and the parasite electrical coupling by the input voltage. Therefore we cannot use the method in [M'Closkey et al., 1999].

In this report we present our contribution on the modeling of MEMS gyroscope when the input is a voltage for capacitive instrumentation. The idea is to separate both effects in a parallel model structure. We can directly do it in MIMO (Multiple Inputs Multiple Outputs) configuration but due to the complexity of the algorithm we can face some local minimum issues in the computation of the model. Input correlation can also make the estimation wrong. Then our approach is to do it with a SISO-by-SISO approach (by estimating SISO models separately where SISO means Simple Input Simple Output) by preventing the effect of parasite electrical coupling on the model of the mechanical dynamic in a first step and then the contrary in a second step. Therefore we had to be careful for the chosen experiment such that both effects can be separated. We chose to excite with multisine such that the effect of the mechanical excitation is at one frequency range and the parasite electrical effect is at another frequency range. We use PEM (Prediction-Error Method) identification ([Zhu Yu-Cal, 2007] for MIMO, [Ljung, 1999] for SISO) for the computations of the SISO models which are LTI (Linear Time Invariant) models. The computed model has been verified with a controller design. However some effects are not modeled yet. Future perspectives are presented for the improvement of this model.

First with the computed SISO models we can use them as initial point for the MIMO algorithm, more appropriate for a MIMO model. To model like that we need MIMO experimental data, i.e. all inputs and outputs are simultaneously excited. But with input correlation can lead to a wrong model. This is the first perspective of this PhD thesis.

We did not consider the angular rate effect on the dynamics for the modeling in a first consideration. It will be our second perspective: we need to verify the dependency of the model with the angular rate (verification of the Coriolis expression). If it is not satisfying we can model the angular rate dynamics with a LPV (Linear Parameter Varying) approach by interpolating LTI models at different angular rates (local approach presented in [Ghosh et al., 2018]).

The temperature affects the mechanical and electrical properties of the MEMS gyroscope. The computed controller is not robust to these variations. We need an adaptive controller and so we need to compute a LPV model by interpolating LTI models at different temperatures (local approach), third perspective of this PhD thesis.

Finally all computed models have uncertainties due to the presence of the noise in the data. Thanks to the statistical properties of the PEM estimator [Ljung, 1999] we can estimate the modeling error. This error will lead to error in the deduction of the angular rate. First we can try to quantify these uncertainties on the angular rate deduction in the time-domain for different scenarii of angular rate. Then, if they are too important, thanks to Experiment Design, illustrated for instance

in [Bombois and Scorletti, 2012] we can minimize them. The complexity of the problem needs a second order approximation already studied in [Forgione et al., 2015] and [Hjalmarsson, 2009]. This is another perspective of this work.

The report is composed like this

- Section 2 : Presentation of the MEMS gyroscope, and the models from the ideal case to the modeling problem formulation.
- Section 3 : Modeling of GYPRO3300 with a parallel structure identified with PEM.
- Section 4 : Presentation of the perspectives of this PhD thesis linked to the limitations of the computed models. Roadmap of the PhD thesis.
- Section 5 : Conclusion.

# 2 Description of the MEMS gyroscope, literature survey of its modeling and problem formulation

A gyroscope is a motion sensor capable of measuring the angular rate of an object. There are different types: optical gyroscope, mechanical gyroscope with gyroscopic effect and MEMS gyroscope. The last type is the less expensive, less cumbersome, less consuming energy but also less accurate. That is why in the NEXT4MEMS project we focus on this kind of gyroscope and try to improve the accuracy of the angular rate measurement. Here we focus on the gyroscope GYPRO3300 manufactured by one company of the consortium, implemented on the electronic card (for the deduction of the angular rate), called Mobyle.

In general all MEMS gyroscopes use the Coriolis effect, described in the Appendix A. There are other effects which can be neglected and will not be described in this report.

## 2.1 Description of the MEMS gyroscope

The MEMS gyroscope is composed of a mass which vibrates in two directions in the plane $(O, x, y)$: $(O, x)$ and $(O, y)$ with $O$ the mass center of the gyroscope. The angular rate that is measured, denoted $\Omega$, is on the $z-$axis, orthonormal to $(O, x, y)$ and such that $(O, x, y, z)$ forms a direct coordinate system. This reference frame is linked to the object whose angular rate needs to be measured: the reference frame seems to not move from "the point of view of" the mass. This mass is attached to a structure (attached to the object whose angular rate is measured) with microscopic silicon cantilever beams described in [Yazdi et al., 1998] acting like springs (due to the deformations of the beam) and dampers (due to viscous effects caused by the air for instance) for both directions $(O, x)$ and $(O, y)$. The Figure 2.1 gives a scheme of the MEMS gyroscope.

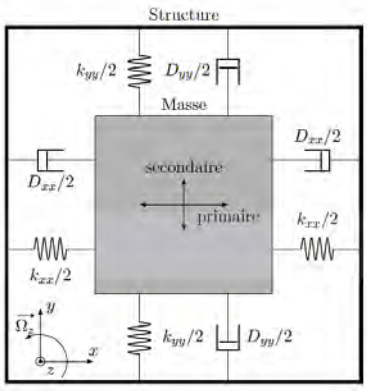



Figure 2.1 – General scheme of the MEMS gyroscope [Korniienko et al., 2017].

In the GYPRO3300 there is a difference from that general scheme. There are two masses: $m_x$ oscillating along $x-$axis and $m_y$ oscillating along $y-$axis. By applying the mechanical equations (Newton's second law) on the masses we see that both motions will act as a lightly damped oscillator. Therefore for both directions we have a resonant mode. We will call drive mode or primary mode the resonant mode along the $x-$axis and sense or secondary mode the resonant mode along the $y-$axis. The names drive and sense come from the working of each mode, explained later in this report. In the next paragraph we present the principle of the measurement of the angular rate.

## 2.2 Principle of angular rate measurement in the ideal case

In this paragraph we present briefly the way to deduce the angular rate and the importance of having an accurate model. In the sequel the term *measurement* will be used for the measurement of the displacements and the term *deduction* will be linked to the angular rate measurement.

We will first do some assumptions to explain the ideal principle of deduction: the stiffness and damping coefficients ($k_{xx}$, $k_{yy}$, $d_{xx}$ and $d_{yy}$) are known and there is no mechanical coupling between the two modes, i.e. when the angular rate $\Omega = 0$, a motion in the drive mode does not create a motion in the sense mode and also in the reciprocal way. In that case, by denoting $F_x(t)$ an external force used to excite the gyroscope in the $x-$axis (drive) and $F_y(t)$ the one of the sense mode, we can put the system in equations (equation (2.1) for the drive mode and equation (2.2) for the sense mode) thanks to the Newton's second law and by taking account of the Coriolis effect (see Appendix A):

$$m_x\ddot{x}(t) + d_{xx}\dot{x}(t) + k_{xx}x(t) = F_x(t) + 2m_y\Omega(t)\dot{y}(t) \quad (2.1)$$

$$m_y\ddot{y}(t) + d_{yy}\dot{y}(t) + k_{yy}y(t) = F_y(t) - 2m_x\Omega(t)\dot{x}(t) \quad (2.2)$$

The idea of the deduction of the angular rate is to exploit the Coriolis term on the secondary mode $2m_x\dot{x}\Omega$. If we excite the primary mode then we have the sense equation (2.3) with $F_y(t) = 0$:

$$m_y\ddot{y}(t) + d_{yy}\dot{y}(t) + k_{yy}y(t) = -2m_x\Omega(t)\dot{x}(t) \quad (2.3)$$

Then the secondary mode displacement is an image of the Coriolis force. If we measure $\dot{x}(t)$ we can deduce $\Omega$ but in reality we cannot measure $\dot{x}(t)$. Moreover in this configuration, as the sense mass will also move, it will induce a Coriolis term on the drive mode. To reject this disturbance (i.e. the term $2m_y\Omega(t)\dot{y}(t)$) on the drive mode we generally put it in closed loop by designing a controller. This justifies the names drive for the primary mode and sense for the secondary mode.

Generally the drive mode is excited with a sinusoidal signal, i.e. $F_x$ is sinusoidal, at its resonance frequency[2] $\omega_{0_x} \approx \sqrt{k_{xx}/m_x}$ (so that we minimize the power consumption of the signal $F_x$). Then the controller on the drive will allow the drive output $x(t)$ to track a reference signal $x_{ref}(t) = A_{ref}\cos(\omega_{0_x}t)$ at the resonance frequency and reject all disturbances like the Coriolis one. Let suppose that the tracking is done, i.e. $x(t) = A_{ref}\cos(\omega_{0_x}t)$ and $\dot{x}(t) = -A_{ref}\omega_{0_x}\sin(\omega_{0_x}t)$. Then we can estimate $\dot{x}$ very simply (derivative of $x_{ref}$): it is also a sinusoidal signal at $\omega_{0_x}$. As the sense output is an image of the Coriolis force on this mode (when $F_y(t) = 0$), the angular rate will be modulated by this sinusoidal signal. Then we see that from the equation (2.3) the angular rate $\Omega$ is modulated by this sinusoidal signal. Then we can demodulate the Coriolis force estimation to retrieve $\Omega$.

The demodulation is done by using a synchronous demodulation technique whose application on the MEMS gyroscope is described in [Saukoski, 2008]. It is used when we want to estimate the amplitude

2. For a resonant mode, the resonance frequency can be considered equal to the natural frequency of the mode when this mode is lightly damped. We will do this assumption in this report for both modes, drive and sense ones.

and the phase shift of a sinusoidal signal at a frequency $\omega$, by multiplying it by $\cos(\omega t)$ and $\sin(\omega t)$ (at the same frequency $\omega$). We do that because every sinusoidal signal can be put in the form $A\cos(\omega t + \phi) = A\cos(\phi)\cos(\omega t) + A\sin(\phi)\sin(\omega t) = P\cos(\omega t) + Q\sin(\omega t)$ and we can isolate $P$ by multiplying by $\cos(\omega t)$ and filtering the obtained signal with a lowpass filter which cuts the frequency $\omega$. It is the same for $Q$ but we multiply by $\sin(\omega t)$. A scheme of the synchronous demodulation is presented in the Figure 2.2. This is applied on the MEMS gyroscope sense output to deduce $\hat{\Omega}$.

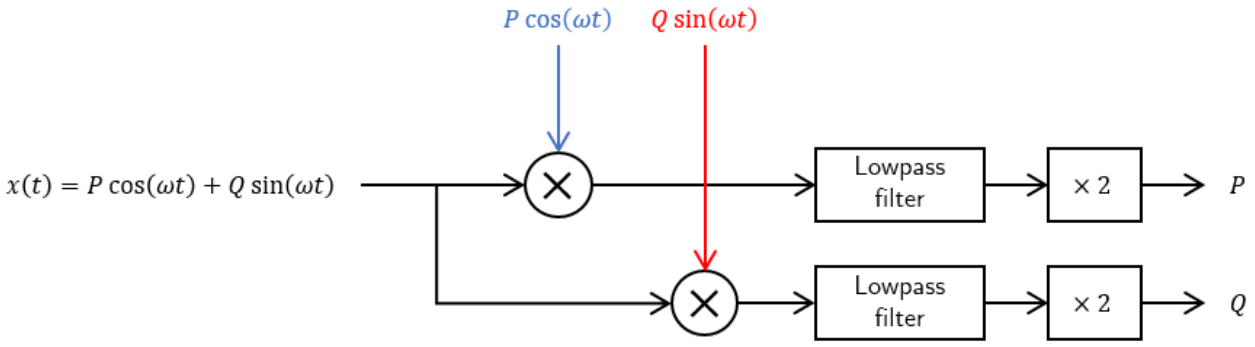



Figure 2.2 – General scheme of the synchronous demodulation.

To illustrate the importance of the model we give an example of angular rate scenario. In the case when the angular rate is constant the sense output $y(t)$ is given by (2.4)

$$y(t) = A_y A_{ref} \Omega \cos(\omega_{0_x} t + \phi_y) \tag{2.4}$$

where $A_y$ and $\phi_y$ are function of the parameters in (2.1) and (2.2). Then we can demodulate (2.4) to retrieve the angular rate $\Omega$. The deduced angular rate will be denoted $\hat{\Omega}(t)$ in the sequel.

What is important to observe is that we have to know all the parameters of both equations (2.1) and (2.2). Indeed to excite at $\omega_{0_x}$ we need to know accurately $k_{xx}$ and $m_x$. For the controller design it is important to know all the drive parameters and for the demodulation all the sense parameters. This is the main motivation of this work: we need an accurate estimate of these parameters as they have an impact on the accuracy of the deduced $\hat{\Omega}$ !

Now that we have explained the measurement principle in the ideal case and the importance of an accurate model we can investigate the modeling of real gyroscopes. **To simplify the modeling we will first consider the case when there is no angular rate, i.e $\Omega(t) = 0$ and this for the sequel of this report.** We will also try to find a linear model, simpler to obtain and to use for controller design. In the sequel we will present only zero-angular rate models. But first we need to explain the different representations of a system used in linear automatic control, based on the ideal equations without angular rate, i.e. the equations presented in (2.5) and (2.6).

$$m_x \ddot{x}(t) + d_{xx}\dot{x}(t) + k_{xx}x(t) = F_x(t) \tag{2.5}$$
$$m_y \ddot{y}(t) + d_{yy}\dot{y}(t) + k_{yy}y(t) = F_y(t) \tag{2.6}$$

## 2.3 Different types of linear representation of the zero-angular rate models for ideal MEMS gyroscope

In automatic control we often use another representation for the linear systems. The first representation is the state-space one. By denoting $q(t) = \begin{pmatrix} x(t) & \dot{x}(t) & y(t) & \dot{y}(t) \end{pmatrix}^T$ and $p(t) = \begin{pmatrix} x(t) & y(t) \end{pmatrix}^T$, the

state-space representation of the linear equations in (2.1) and (2.2), when $\Omega = 0$, is the form displayed in the equation (2.7):

$$\begin{cases} \dot{q}(t) &= Aq(t) + Bu(t) \\ p(t) &= Cq(t) + Du(t) \end{cases} \tag{2.7}$$

with $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_{xx}}{m_x} & -\frac{d_{xx}}{m_x} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -\frac{k_{yy}}{m_y} & -\frac{d_{yy}}{m_y} \end{pmatrix}$, $B = \begin{pmatrix} 0 & 0 \\ \frac{1}{m_x} & 0 \\ 0 & 0 \\ 0 & \frac{1}{m_y} \end{pmatrix}$, $C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$.

The other representation is the transfer function. It is obtained by taking the Laplace transform [3] of these equations and searching the ratio between output and input [4]. Then for $\Omega = 0$ we simply have a transfer function for each mode expressed by

$$G_{xx}^{ideal}(s) = \frac{X(s)}{F_x(s)} = \frac{1}{m_x s^2 + d_{xx} s + k_{xx}} \qquad G_{yy}^{ideal}(s) = \frac{Y(s)}{F_y(s)} = \frac{1}{m_y s^2 + d_{yy} s + k_{yy}}$$

where $X(s)$, $Y(s)$, $F_x(s)$ and $F_y(s)$ are respectively the Laplace transform of $x(t)$, $y(t)$, $F_x(t)$ and $F_y(t)$.

These two representation consider only the ideal case of the MEMS gyroscope. In reality it is more complex due to manufacturing issues, instrumentation of the modes, etc. These nonidealities are presented in the next subsection.

## 2.4 Nonidealities of the MEMS gyroscope and problem formulation

► **Mechanical cross-coupling:** Due to manufacturing nonidealities there are cross coupling terms between both modes: $k_{xy}$, $k_{yx}$, $d_{xy}$ and $d_{yx}$. Then both mode equations (2.1) and (2.2) become the ones presented in (2.8) and (2.9):

$$m_x \ddot{x}(t) + d_{xx}\dot{x}(t) + d_{xy}\dot{y}(t) + k_{xx}x(t) + k_{xy}y(t) = F_x(t) \tag{2.8}$$
$$m_y \ddot{y}(t) + d_{yy}\dot{y}(t) + d_{yx}\dot{x}(t) + k_{yy}y(t) + k_{yx}x(t) = F_y(t) \tag{2.9}$$

Indeed both modes are coupled mechanically due to two effects: anisoelasticity and non-proportional damping. The anisoelasticity effect explains the presence of the terms $k_{xy}$ and $k_{yx}$ and the non-proportional damping the presence of $d_{xy}$ and $d_{yx}$, more detailed in [Phani et al., 2006]. These terms can be accurately identified by following the methods proposed in [Painter and Shkel, 2002] and [Phani and Seshia, 2004].

If we consider only the mechanical part of the gyroscope, these terms have an effect on the systemic models (state-space representation and transfer function). Indeed we have now cross effects between modes. Then we cannot anymore represent the dynamics with two direct transfer functions. Indeed we have to model also $G_{xy}$ (corresponding of the transfer from the input of the sense mode to the output of the drive mode) and $G_{yx}$ (corresponding of the transfer from the input of the drive mode to

3. The Laplace transform of a time-domain signal $f$, denoted $\mathcal{L}[f]$ is defined by the following relation:

$$\mathcal{L}[f](s) = F(s) = \int_0^{+\infty} f(t)e^{-st}dt$$

where the complex $s$ is called the Laplace variable.

4. Indeed the transfer function of a system with an input $u$ and an output $y$ is obtained by using that Laplace transform and then calculating the ratio $\mathcal{L}[y](s)/\mathcal{L}[u](s) = Y(s)/U(s)$.

the output of the sense mode). The transfer function model becomes a matrix and is in that case the one presented in (2.10)

$$\begin{pmatrix} X(s) \\ Y(s) \end{pmatrix} = \underbrace{\begin{pmatrix} G_{xx}(s) & G_{xy}(s) \\ G_{yx}(s) & G_{yy}(s) \end{pmatrix}}_{G(s)} \begin{pmatrix} F_x(s) \\ F_y(s) \end{pmatrix} \tag{2.10}$$

Note that here $G_{xx}(s)$ and $G_{yy}(s)$ are different from $G_{xx}^{ideal}(s)$ and $G_{yy}^{ideal}(s)$ respectively. These terms lead to error in the angular rate deduction obtained in the synchronous demodulation as explained in [Saukoski, 2008]. The above equations have been the basis of the model chosen in [Fei and Yang, 2012] and [Fei et al., 2010].

The company which manufactures this gyroscope, uses also this kind of model for GYPRO3300, adding the mechanical cross-coupling terms. However they only consider the terms $k_{yx}$ and $d_{yx}$ ($k_{xy} = 0$ and $d_{xy} = 0$). The magnitude of the frequency response of this model is presented in the Figure 2.3.

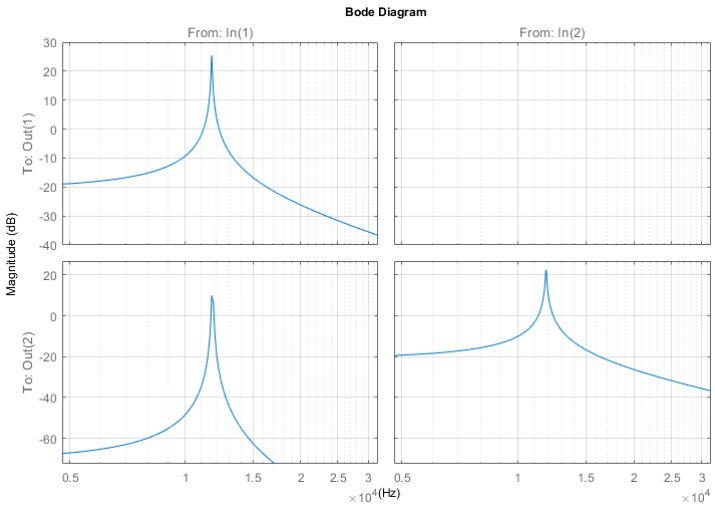



Figure 2.3 – Frequency response (magnitude) of the model of GYPRO3300 from its manufacturer.

Note that the gyroscope has been designed such that $G_{xx}$ and $G_{yy}$ are strong mechanical bandpass filters around their respective resonance frequency ($\omega_{0_x}$ and $\omega_{0_y}$). **This will be considered as true in the sequel of this report.**

In reality we cannot measure directly the displacements and actuate the modes directly with forces, as it has been assumed in [Fei and Yang, 2012] for instance. In the next paragraph we explain the instrumentation of the measurement and excitations.

► **Instrumentation for the excitation/measurement of both modes:** The gyroscope displacements are actuated and measured as described in [Saukoski, 2008]. The actuation (or excitation) and the measurement of the displacements of both modes are made by interdigital capacitances [5]. This interdigital capacitance is made by two combs placed face to face but shifted of one finger so that each finger of a comb faces a hole between two fingers of the other comb (interdigital placement). This way of placing the two combs (forming a double-comb) creates a capacitance between them, proportional to the distance between these combs.

5. Note that it exists other ways to actuate and measure the modes displacements. In [Yazdi et al., 1998] they consider piezo-electrical instrumentation for their MEMS gyroscope. We will not give further details as it is not our case.

On the gyroscope we place two double-combs for each mode: one used for the excitation and one use for the measurement of the displacement. For each double-comb, one comb is attached to the structure and the other one is attached to the mass. The Figure 2.4 illustrates the placement of the double-combs on the gyroscope. We will not detail here the physics behind this instrumentation, one can see the Appendix B for further details.

For the actuation, the double-combs are excited with a input voltage and it results in an electrostatic force proportional to the square of this input voltage. We will denote $V_{in_x}$ the input voltage in the drive mode and $V_{in_y}$ the input voltage in the sense mode.

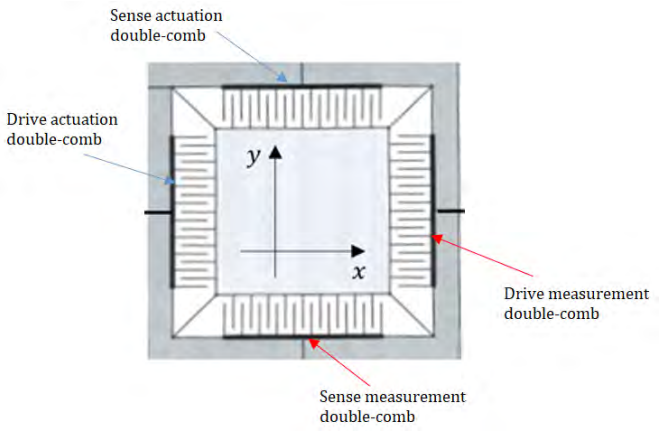



Figure 2.4 – Scheme of the MEMS gyroscope, instrumented with capacitive double-combs.

For the measurement, the capacitance of each measurement double-comb is directly proportional to the displacement of the concerned mode. We convert this capacitance with an operational amplifier circuit so that the output voltage of this circuit is an image of the capacitance and consequently of the displacement. These output voltages are amplified with a two-stage amplifier. We will denote $V_{out_x}$ the output voltage in the drive mode $V_{out_y}$ the output voltage in the sense mode from these amplifiers. In the sequel the new inputs are $V_{in_x}^2(t)$ and $V_{in_y}^2(t)$, while the new outputs are $V_{out_x}(t)$ and $V_{out_y}(t)$. We will denote $V_{in} = \begin{pmatrix} V_{in_x} & V_{in_y} \end{pmatrix}^T$, $V_{in}^2 = \begin{pmatrix} V_{in_x}^2 & V_{in_y}^2 \end{pmatrix}^T$ and $V_{out} = \begin{pmatrix} V_{out_x} & V_{out_y} \end{pmatrix}^T$. Then we obtain a new transfer function representation with these new inputs and outputs. We will keep abusively the notation $G$ for this transfer function representation in the sequel and it is now defined by the equation (2.11).

$$\begin{pmatrix} V_{out_x}(s) \\ V_{out_y}(s) \end{pmatrix} = \underbrace{\begin{pmatrix} G_{xx}(s) & G_{xy}(s) \\ G_{yx}(s) & G_{yy}(s) \end{pmatrix}}_{G(s)} \begin{pmatrix} V_{in_x}^{(2)}(s) \\ V_{in_y}^{(2)}(s) \end{pmatrix} \tag{2.11}$$

where $V_{in_x}^{(2)}(s)$ is the Laplace transform of the signal[6] $V_{in_x}^2(t)$ and $V_{in_y}^{(2)}(s)$ is the one of $V_{in_y}^2(t)$.

► **Temperature-dependency:** Due to its microscopic size, some parameters of the gyroscopes in both equations (2.8) and (2.9) depend strongly on the temperature (explained in [Guan et al., 2015] for the resonance frequency shift and [Xia et al., 2009] for the quality factor shift). More particularly the damping and stiffness coefficients of the beams that link the structure and the mass are affected by temperature variations. These variations of the mechanical properties are explained by the variation of the Young modulus of the material of these beams (silicon in our gyroscope) [Hopcroft et al., 2010].

6. The Laplace transform of the square of a signal is not equal to the square of the Laplace transform of the signal. Therefore we use the notation $V_{in_x}^{(2)}(s)$ to not have a confusion with the square operator.

The temperature can also affects the electronic implementation, i.e. here the electronic card Mobyle dynamics. It has been studied for instance in [Yongzhen et al., 2011]. The first results presented in this report do not take account of this effect. However it is a perspective for the improvement of the model as this effect has a strong impact on the MEMS dynamics, explained later in the paragraph 4.3.

► **Nonlinear working of the amplifiers:** These devices can have a nonlinear behaviors. As we are going to excite the gyroscope with sinusoidal excitation, the output signal will contain several multiples of the excitation frequency. It is the nonlinear harmonic distortion illustrated in the Figure 2.5. As we want to find a linear model we cannot model this effect in our approach. We can reduce it by decreasing the gain of the amplifier. Then in the sequel we will not consider this effect in our model.


Figure 2.5 – Illustration of the nonlinear distortion effect with $f$ the nonlinear function describing the amplifier behavior.

► **Discretization:** In reality we deal with sampled signals with a sample time denoted $T_s$. Then the obtained models can directly be obtained in the discrete-time domain. In that case we use the $\mathcal{Z}$ domain which described discrete-time system. Our model becomes the one in (2.12).

$$\begin{pmatrix} V_{out_x}(t) \\ V_{out_y}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} G_{xx}(z) & G_{xy}(z) \\ G_{yx}(z) & G_{yy}(z) \end{pmatrix}}_{G(z)} \begin{pmatrix} V^2_{in_x}(t) \\ V^2_{in_y}(t) \end{pmatrix} \tag{2.12}$$

where $z = e^{T_s s}$ is the shift operator [7]. As the signals are discretized, we need an anti-aliasing filter. The one of the card is composed of one first order filter with a cut-off frequency of 40kHz.

► **Noises:** The output voltages are noisy due to two major phenomena : flicker noise and mechanical-thermal noise [Saukoski, 2008]

- Flicker noise: this noise comes from the presence of active electronic components such as amplifiers. The power spectrum of this noise is proportional to $1/f$ where $f$ is the frequency (in Hz) in the low frequency range. After a particular frequency, called corner frequency, this noise is overshadowed by other noise sources.
- Mechanical-thermal noise: as the gyroscope is a dissipative system it is affected by thermal noise. By using the equipartition theorem of energy [Waterston and Strutt, 1892] and the Nyquist relation, the power spectral density of the noises on the drive and sense mode outputs ($i = \{x, y\}$) are given by $\Phi_{therm}(f) = 4k_b T d_{ii}$ with $k_b$ the Boltzmann constant, $T$ the absolute temperature and $d_{ii}$ the damping of the mode represented by the direction $i = \{x, y\}$. Note that it does not depend linearly on the temperature as it can be first interpreted from this expression because $d_{ii}$ depends also on $T$.

Noises are characterized by their density spectral power (DSP). It could be interesting to know the DSP of the MEMS gyroscope noise on the voltage outputs. We generally describe the noisy outputs of a system as a sum of the output signal from the transfer of the inputs and a noise signal. In our case

7. Note that normally it would have been correct to put the $\mathcal{Z}$ transform of the output and input vector. However we use the common abusive notation: $x(t+k) = z^k x(t)$ where $k$ is an integer.

we will denote $w_x(t)$ the noise on the drive output and $w_y(t)$ the noise on the sense output. Then to take account of the noise phenomena the model in (2.12) is completed as described in (2.13).

$$\begin{pmatrix} V_{out_x}(t) \\ V_{out_y}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} G_{xx}(z) & G_{xy}(z) \\ G_{yx}(z) & G_{yy}(s) \end{pmatrix}}_{G(z)} \begin{pmatrix} V^2_{in_x}(t) \\ V^2_{in_y}(t) \end{pmatrix} + \underbrace{\begin{pmatrix} w_x(t) \\ w_y(t) \end{pmatrix}}_{w(t)} \tag{2.13}$$

with $w(t) = \begin{pmatrix} w_x(t) & w_y(t) \end{pmatrix}^T$. We want to characterize the DSP of $w_x$ and $w_y$. It is simple to do it from the representation in (2.13) . Indeed we only need to measure the outputs without exciting the gyroscope ($V^2_{in} = 0$). We can identify the DSP of the noises from these data (not developed in this report). For the GYPRO3300 they are given in the Figure 2.6.

Both trends (flicker noise and mechanical-thermal noise) are visible. Then the we see the effect of the anti-aliasing filter after 10000Hz. What is important to see here is that the noise in the sense output is much more powerful than the one of the drive output. This has a strong consequence: the transfer from the drive mode input to the sense input is hidden by the noise even with the most powerful input that respects the input constraints explained later in the paragraph 3.1. Since we cannot identify $G_{yx}$, we will consider it equal to 0.

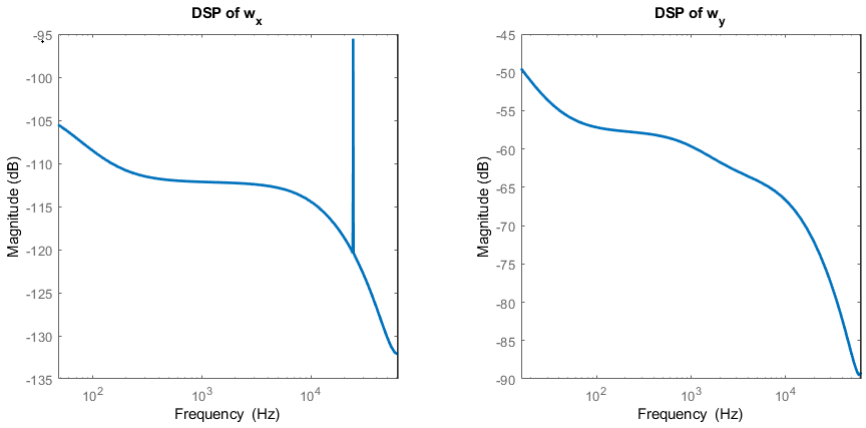



Figure 2.6 – DSP of the noises $w_x$ (left plot) and $w_y$ (right plot).

► **Parasite electrical cross-coupling of the modes:** As we use capacitance excitations and measurements, there are interactions between the electrical fields inside the gyroscope. This effect can couple the modes and it is not taken account in the model. Let us illustrate this phenomena in our case on the drive mode but the same conclusions hold for the sense mode. From prior knowledge of the manufacturer, we know that the drive resonance frequency is around $\omega_{0_x} = 2\pi f_{0_x}$ with $f_{0_x} = 11750$ Hz. For the sense mode, $f_{0_y} = 11876$ Hz.

We excite the drive mode with a sinusoidal input voltage at the half frequency of the gyroscope $\omega_{0_x}/2$ : $V_{in_x}(t) = a\cos(\omega_{0_x}/2t) + b$ with $a = b = 1$. These values are justified by the fact that we cannot excite with negative values for the input voltage on the electronic card Mobyle as it will be described in the subsection 3.1. The resulting external force on the drive mode $F_x$ is proportional to $V^2_{in_x}(t) = b^2 + a^2/2 + 2ab\cos(\omega_{0_x}/2t) + a^2\cos(\omega_{0_x}t)$. Then we excite two frequencies with this input: $\omega_{0_x}/2$ and $\omega_{0_x}$. If the MEMS has indeed been designed such that each mode is a strong bandpass filter around $\omega_{0_x}$ and $\omega_{0_y}$, the frequency $\omega_{0_x}/2$ should be filtered out [8] and the frequency $\omega_{0_x}$ should excite the drive mode resonance. The Figure 2.7 presents the FFT of the drive output $V_{out_x}(t)$ and sense output $V_{out_y}(t)$ from this sinusoidal excitation at the drive mode.

8. Indeed the gyroscope has been designed such that the diagonal transfer functions of $G$ are strong bandpass filters around their respective resonance frequency as shown in the Figure 2.3.

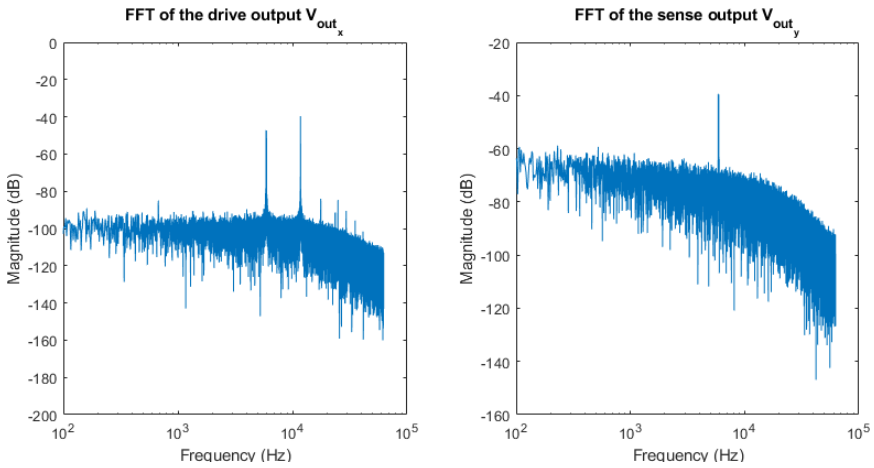



Figure 2.7 – FFT of the outputs $V_{out_x}$ (left plot) and $V_{out_y}$ (right plot) with a sinusoidal excitation at $\omega_{0_x}/2$.

We can see little peaks at $3\omega_{0_x}/2$, $4\omega_{0_x}/2 = 2\omega_{0_x}$, $5\omega_{0_x}/2$, etc in the FFT of $V_{out_x}$. This is the illustration of the nonlinear distortion previously discussed. We can also see a strong peak at $\omega_{0_x}$ : it is the effect of the resonance excitation on the drive mode. But there is also a strong peak at $\omega_{0_x}/2$ which cannot be explained from the mechanical dynamics of the gyroscope. This peak also appears in the FFT of $V_{out_y}$ at the same frequency. Let us try to characterize this effect, which is not modeled by the physics equations, with other experiments.

We do this experiment again but at the frequency $\omega_{0_x}/4$, then at $\omega_{0_x}/8$ and finally at $\omega_{0_x}/16$. For each experiment we analyze the gain of the sense and drive output FFT at the frequency of the sinus and at its double. Only the frequency of $V_{in_x}$ is excited by this unexplained phenomena while both frequencies of $V_{in_x}^2$ are not excited by this effect. We observe the same effect when exciting like that with $V_{in_y}$. It is in reality a parasite electrical effect that couples the modes (from one to another and itself), due to electrical field interaction.

We are going to model this effect with a transfer function matrix $E(z)$ whose input is $V_{in}$. Therefore, by considering that this effect is added with the effect of the mechanical excitation of the gyroscope, we can complete the model presented in (2.13) and we obtain the model in (2.14). It corresponds to the parallel structure given in the Figure 2.8.

$$\begin{pmatrix} V_{out_x}(t) \\ V_{out_y}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} G_{xx}(z) & G_{xy}(z) \\ G_{yx}(z) & G_{yy}(z) \end{pmatrix}}_{G(z)} \begin{pmatrix} V_{in_x}^2(t) \\ V_{in_y}^2(t) \end{pmatrix} + \underbrace{\begin{pmatrix} E_{xx}(z) & E_{xy}(z) \\ E_{yx}(z) & E_{yy}(z) \end{pmatrix}}_{E(z)} \begin{pmatrix} V_{in_x}(t) \\ V_{in_y}(t) \end{pmatrix} + \begin{pmatrix} w_x(t) \\ w_y(t) \end{pmatrix} \tag{2.14}$$

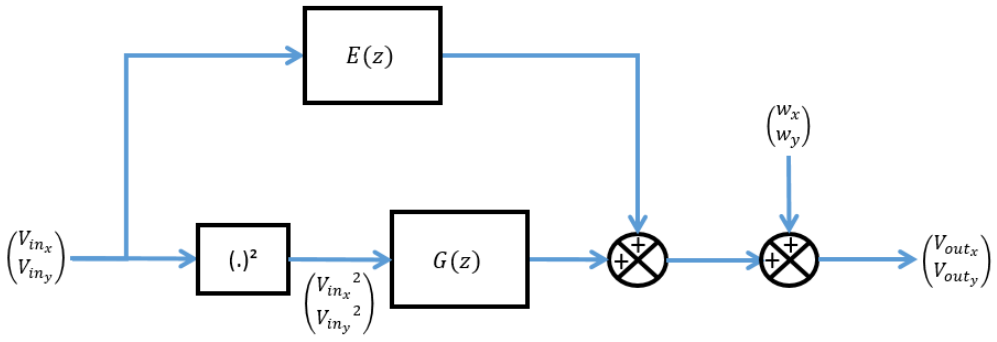



Figure 2.8 – Parallel model structure of the MEMS gyroscope.

This effect has been illustrated also in [Chen et al., 2003]. However in this article the electrostatic actuation and the parasite electrical coupling comes from the same input signal which is the electrical field [9]. Then both phenomena can be modeled in the same transfer function matrix by considering the exciting electric fields as inputs. In our case we cannot apply this method as the excitation of the modes is driven by $V_{in}^2$ and the parasite coupling is driven by $V_{in}$.

**Conclusion and motivation for our contribution in the MEMS gyroscope modeling:** In this section we studied the phenomena involved inside the MEMS gyroscope GYPRO3300 to model them. We started from the ideal case given in the paragraph 2.3, modeled by a simple non mechanically cross-coupled model. Then by considering all the nonidealities and the instrumentation of our gyroscope we had to complete the ideal model. We ended up with the model described by (2.14) which is the main contribution of this work during the first year in the MEMS gyroscope modeling literature survey. In the next section we give a method for the computation of $G$ and $E$.

# 3 Modeling of GYPRO3300 with a parallel structure identified with Prediction-Error-Method (PEM)

In this section we are going to identify the GYPRO3300. But first we need to generate data from the gyroscope. We will present the experimental setup and the faced problems in the next paragraph.

## 3.1 Experimental setup and faced problems on the electronic aspects

We have borrowed an electronic card with a MEMS gyroscope GYPRO3300 since March, which allowed us to perform experiments more easily. However the electronic card Mobyle was not programmed to perform identification experiments in the first place (open-loop experiments with user choice signals for instance). Indeed in the first architecture, the gyroscope is only excited by a mono-frequency sinusoidal signal and the output is demodulated with synchronous demodulation to retrieve the amplitude and the phase shift which do not correspond to our considered outputs. During the first months of the PhD thesis we had to reconfigurate the electronic card and correct some computing problems.

For the experiments some constraints have to be respected. Due to memory issue we cannot perform experiments lasting more than 20s. This is not problematic as we will only identify on short time sequences.

The input voltage must be positive. Indeed as the force is proportional to the square of the input voltage then it is not necessary to use the negative values of this voltage and the company developing the electronic card chose to restrict only on positive input voltages. This limitation must be taken account for our experiment. Then there is always a DC level on the input voltages $V_{in_x}$ and $V_{in_y}$.

We also faced a missing data issue, only for the output data, due to communication problem between the PC and the card. We have fixed this problem. Now that we are able to generate data we can explain our contribution on the modeling of this MEMS gyroscope.

In the next paragraph we present the prediction-error-method (PEM) used to identify both transfer function matrices. But first we present the possible issues with MIMO (Multiple Inputs Multiple Outputs) PEM identification

9. Indeed the electrostatic force is linear to the electric field and the parasite effect (which is an electrical field interaction) also takes as input the electrical field.

## 3.2 Motivation for the choice of a SISO-by-SISO approach

The general idea of all identification techniques dealing with transfer functions is to compute the parameters of the transfer function matrices after having specified the order of the numerator and denominator of each SISO transfer function.

We can try to identify all the transfer functions simultaneously: it is MIMO identification. However due to the number of transfer functions, the issue of input correlation in the MIMO framework [10] and complexity of the algorithm (local minimum issue) it is complex to try MIMO identification in a first place.

These issues motivated us to identify each transfer function in $E$ and $G$ in a SISO-by-SISO approach: we identify every transfer function separately from the others by exciting its corresponding input and measuring its corresponding output. In the next paragraph we present the method to identify a SISO model with input and output data with the prediction-error-method.

## 3.3 Introduction to prediction error method (PEM) identification in the SISO case [Ljung, 1999]

### 3.3.1 Presentation of the estimator of the PEM

Prediction-error identification is a data-based way to model a dynamic system, denoted $\mathcal{S}$. We will consider a SISO system with one input $u$ and one output $\Upsilon$. The identification process is based on experimental data from the true system $\mathcal{S}$. We suppose that we have $N$ consecutive data of the input and output of the system, sampled at the same sampling time $T_s$. Let us assume that the true system, represented in the scheme in the Figure 3.1 can be written in the following linear (discrete) form (3.1):

$$\mathcal{S}: \qquad \Upsilon(t) = F_0(z)u(t) + \nu(t) \tag{3.1}$$

with $F_0$ is a stable transfer function and $\nu(t)$ a noise signal.

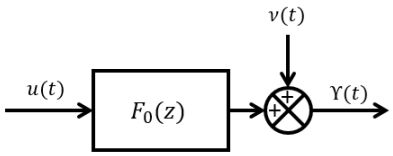



Figure 3.1 – Block scheme of the true system.

The input $u$ and the noise process $\nu$ are assumed to be independent: their cross-correlation function is equal to 0. This can allow us to extract from the data everything that is related to the input, independently from the noise.

The general purpose of system identification is to find a model that fits best the experimental data from the true system $\mathcal{S}$. This way of modeling is more accurate as it can model phenomena which are not taken account in the ordinary differential equations of the system but are visible within the data (neglected resonances in a flexible system such as cantilever beams for instance).

The first step is to choose a transfer function model structure $\mathcal{F} = \{F(z,\theta) \mid \theta \in \mathbb{R}^n\}$ where $\theta$ is the vector made up by the parameters of the transfer functions $F(z,\theta)$ with $n = \dim(\theta)$ and defined by

10. $V_{in}$ and $V_{in}^2$ might be correlated and this raises a problem in MIMO identification as described in [Mareels and Gevers, 1988].

the expression (3.2):

$$\tilde{\Upsilon}(t,\theta) = F(z,\theta)u(t) = \frac{B(z,\theta)}{A(z,\theta)}u(t) = z^{-n_k}\frac{\sum_{k=0}^{n_b} b_k z^{-k}}{1+\sum_{k=1}^{n_a} a_k z^{-k}}u(t) \tag{3.2}$$

where $\theta = \begin{pmatrix} a_1 & \cdots & a_{n_a} & b_0 & \cdots & b_{n_b} \end{pmatrix}^T$ and $n_b$, $n_a$ and $n_k$ are polynomial orders. This model structure is called Output-Error (OE). We will assume that the chosen-model structure (more precisely the chosen polynomial orders of the OE structure) contains $F_0$ (full-order assumption $F_0 \in \mathcal{F}$), i.e. $\exists \theta_0 \in \mathbb{R}^n$ such that $F(z,\theta_0) = F_0(z)$.

The idea of PEM identification is to compute the optimal $\theta$, denoted $\hat{\theta}_N$ such that we minimize at each time $t$ the difference $\Upsilon(t) - \tilde{\Upsilon}(t,\theta)$, called also prediction error and denoted $\epsilon(t,\theta)$. This is the so-called prediction error minimization identification (PEM). One common way to express the minimization of $\epsilon(t,\theta)$ at each time $t$ (for the $N$ samples) is to consider the minimization of the cost function made up by all the $N$ values of $\epsilon(t,\theta)$ in a least square formulation defined by the optimization (3.3):

$$\hat{\theta}_N = \arg\min_{\theta} \frac{1}{N}\sum_{t=1}^{N} \epsilon(t,\theta)^2 = \arg\min_{\theta} \frac{1}{N}\sum_{t=1}^{N} (\Upsilon(t) - \tilde{\Upsilon}(t,\theta))^2 \tag{3.3}$$

The minimization problem shown in (3.3) has not necessarily an unique solution. The only interesting solution from this minimization problem is the one linked to $\theta_0$. To obtain this uniqueness we need to have the persistently exciting input assumption or persistency [11]. There are two common input signals used in identification : filtered white noise and multisine. The persistently exciting input assumption is always true for filtered white noise and if we have at least $n/2$ sinusoids at different frequencies in the multisine then this assumption is also verified in the multisine case.

The solution of (3.3) changes from one experiment to another because of the random aspect of the white noise $\nu$ as explained in [Ljung, 1999]. It results in uncertainty set on the computed parameters. By modeling the noise $\nu(t)$ as a filtered white noise we can estimate the error $\theta_0 - \hat{\theta}_N$ as explained in [Ljung, 1999]. This error depends on the number of data $N$ and the input power spectrum $\Phi_u$ used for the identification. It will be discussed later in the subsection 4.4.

But how to be sure that we chose the right model order (full-order assumption) ? There exists some techniques to verify the model structure choice and the computed model. We present these verification aspects in the next paragraph.

### 3.3.2 Verification of the computed models

**Verification of the model structure:** The model structure will be validated, i.e. the full order assumption will be true, if there is no significant correlation between the prediction error computed at $\hat{\theta}_N$ ($\epsilon(t,\hat{\theta}_N)$) and the input signal $u$, more details in [Ljung, 1999].

**Verification of the computation of the model:** Another way to verify the model is to evaluate its efficiency to fit the output data. We generate other experimental data and compute the Best Fit indicator expressed in the equation (3.4):

$$\text{Best Fit} = \left(1 - \frac{||\Upsilon - \tilde{\Upsilon}||_2^{1/2}}{||\Upsilon||_2^{1/2}}\right) \times 100\% \tag{3.4}$$

11. For the SISO case it corresponds to a number of frequencies in the input power spectrum higher than the number of parameters to identify.

where $\Upsilon$ is the measured output of this second experiment, $\tilde{\Upsilon}(t) = F_0(z)u(t)$ and $||.||_2$ the $\mathcal{L}_2$ norm. Note that the Best Fit evaluates the fitting of the output with the model in (3.2) without taking account of the noise. A wrong Best Fit does not correspond to a wrong computed model if the noise is quite powerful compared to the excitation of $F_0$. Indeed by analyzing the expression of the Best Fit it can be linked to the ratio of the power spectrum of the noise to the power spectrum of the noisy output.

**How to identify in practice ?**

If the model structure is not verified, we have to change it. Then PEM identification is an iterative process presented by several steps:

— Step 1 : Generation of output data by excitation with persistently exciting inputs.
— Step 2 : Choice of the model structure (choice of polynomial orders).
— Step 3 : Computation of the parameters with optimization algorithm.
— Step 4 : Verification of the model structure and computed model. If not verified go back to the Step 2.

## 3.4 Black-box modeling of GYPRO3300 with prediction-error-method in the SISO configuration

This section presents the way and the results of the modeling of both transfer functions $E$ and $G$ with the PEM identification. In the next paragraph we present our experimental solution to model them.

### 3.4.1 Experiment choice for the modeling of $E$ and $G$ in a SISO framework

If we model with a SISO-by-SISO approach it means that we will do it transfer function by transfer function. But if, for instance, we want to identify $G_{xx}$ and then $E_{xx}$, we need to separate of the effect on the data used for the identification. This is true for all SISO transfer functions in $E$ and $G$.

It is better to do an experiment which corresponds to the operation of the system. In our case the control effort strategy on the drive mode, for instance, is to track a reference signal at $\omega_{0_x}$. So it would be better to excite around $\omega_{0_x}$ and $\omega_{0_y}$. Secondly the identification will be better with high Signal to Noise Ratio (SNR). This justifies also the fact that we want to excite the resonances of the gyroscope. To model all the transfer functions $G_{ij}$ and $E_{ij}$ with $(i,j) \in \{x,y\}$ the idea is to use multisine whose frequencies are around $\omega_{0_x}$ and $\omega_{0_y}$. Let us consider that we measure $V_{out_i}$ $(i = \{x,y\})$ and the input $V_{in_j}$ $(j = \{x,y\})$ is a multisine of the form $V_{in_j}(t) = V_{DC_j} + \sum_{k=1}^{n_j} A_{j_k} \cos(\omega_{j_k} t + \phi_{j_k})$, where $n_j$ is the number of sines, $\omega_{j_k}$ the frequencies of this multisine around $\omega_{0_x}$ and $\omega_{0_y}$, $A_{j_k}$ the amplitude and $\phi_{j_k}$ the randomly chosen phase shifts. The DC level $V_{DC_j}$ is necessary to keep a positive voltage as explained in the paragraph 3.1. With this input voltage, the output will be composed by the effect of $G_{ij}$ and also $E_{ij}$. As the input of $G_{ij}$ is $V_{in_j}^2$, it is also a multisine composed of sinusoids with:

- the same $n_j$ frequencies $\omega_{j_k}$ with $k \in [[1, n_j]]$ as in $V_{in_j}$, because of the presence of the DC level,
- $n_j$ frequencies of the form $2\omega_{j_k}$ with $k \in [[1, n_j]]$.
- $n_j(n_j-1)/2$ frequencies of the form $\omega_{j_k} + \omega_{j_l}$ with $(k,l) \in [[1, n_j]]^2$ and $k \neq l$,
- $n_j(n_j-1)/2$ frequencies of the form $\omega_{j_k} - \omega_{j_l}$ with $(k,l) \in [[1, n_j]]^2$ and $k \neq l$.

As the multisine frequencies $\omega_{j_k}$ are around $\omega_{0_x}$ and $\omega_{0_y}$ then $V_{in_j}^2$ will also have frequencies around $\omega_{0_x}$ and $\omega_{0_y}$. Other frequencies of $V_{in_j}^2$ are filtered. Therefore on the output $V_{out_i}$ the effects of $E_{ij}$ and $G_{ij}$ will not be distinguishable as they are at the same frequencies.

To be able to separate both effects on the output $V_{out_i}$ we excite the input $V_{in_j}$ with a multisine **around the half of resonance range**, i.e frequencies around $\omega_{0_x}/2$ and $\omega_{0_y}/2$. Therefore the frequencies $2\omega_{j_k}$ are around $\omega_{0_x}$ and $\omega_{0_y}$: then the gyroscope resonances will be excited and the effect will be seen in

the output $V_{out_i}$. The frequencies of $V_{in_j}^2$ which are not around $\omega_{0_x}$ are filtered. The effect of $E_{ij}$ will be concentrated around and in $[\omega_{0_x}/2,\ \omega_{0_y}/2]$. Therefore to get rid of the effect of $E_{ij}$ we filter the input $V_{in_j}^2$ and the output $V_{out_i}$ around the resonance frequency range $\omega_{0_x}$ and $\omega_{0_y}$ with a strong bandpass filter so that it filters out also the frequencies around $\omega_{0_x}/2$ and $\omega_{0_y}/2$ (and so the effect of $E_{ij}$ excitation). The new filtered input is denoted $\tilde{V}_{in_j}^2$ and the new filtered output is denoted $\tilde{V}_{out_i}$ and these new output and input signals will be our $\Upsilon$ and $u$ of the paragraph 3.3.1 for the identification of $G_{ij}$.

As we want to learn more about this parasite effect modeled by $E_{ij}$, we do another experiment in which we excite all the frequencies with a white noise process, like a Random Binary Sequence [12] (RBS) on input $V_{in_j}$. As we have previously computed $G_{ij}$ we can identify $E_{ij}$ by removing from the output $V_{out_i}$ the effect of the gyroscope excitation. This new output, denoted $\bar{V}_{out_i}(t)$, and expressed by $\bar{V}_{out_i}(t) = V_{out_i}(t) - G_{ij}(z)V_{in_j}(t)^2$ will be our $\Upsilon$ for the identification of $E_{ij}$ while $V_{in_j}$ will be the signal $u$.

**Experimental procedure:** To sum up, for all $(i,j) \in \{x,y\}$, we model each $G_{ij}$ and $E_{ij}$ in two steps:

- Excitation of the gyroscope with a multisine on $V_{in_j}$ at the half of the resonance range and measurement of $V_{out_i}$.
- Filtering of the input and output data with a strong bandpass filter in the resonance range: we obtain the new input $\tilde{V}_{in_j}^2$ and the new output $\tilde{V}_{out_i}$.
- Identification of $G_{ij}$ with $u = \tilde{V}_{in_j}^2$ and $\Upsilon = \tilde{V}_{out_i}$ with an OE model structure.
- Excitation with a RBS signal on $V_{in_j}$ and measurement of $V_{out_i}$.
- Removing of the mass excitation from the output data to isolate the parasite electrical effect: we obtain a new output signal $\bar{V}_{out_i}(t) = V_{out_i}(t) - G_{ij}(z)V_{in_j}(t)^2$.
- Identification of $E_{ij}$ with $u = V_{in_j}$ and $\Upsilon = \bar{V}_{out_i}$ with an OE model structure.

### 3.4.2 Results of the MIMO black-box identification with SISO-by-SISO approach

In this paragraph we give the results obtained for the identification of $G$ and $E$ with the approach described in the previous paragraph and the identification method explained in 3.3.1. We use an OE model structure and we will keep the notations $n_b$ and $n_a$ for the polynomial orders of that model structure defined in the expression (3.2). We chose $n_k = 1$. As explained in 2.4 it is not necessary to model $G_{yx}$ as this effect is hidden by the sense output noise $w_y$. We will consider that it is equal to 0.

**Verification of the PEM identification on $E$ and $G$:** For the choice of $n_a$ and $n_b$ we follow the procedure explained in the paragraph 3.3.2 for the choice of the model orders. The orders and the obtained Best Fit for each SISO transfer function of $G$ are given in the Table 1 and for $E$ in the Table 2.

| Transfer function $G$ | $n_b$ | $n_a$ | Best Fit on identification data | Best Fit on verification data |
|---|---|---|---|---|
| $G_{xx}$ | 2 | 2 | 68.38% | 59.46% |
| $G_{xy}$ | 1 | 2 | 0.76% | 1.39% |
| $G_{yx}$ | - | - | Noise level | Noise level |
| $G_{yy}$ | 2 | 2 | 93.88% | 91.42% |

Table 1 – Results of the PEM identification of $G$, obtained orders and Best Fit.

12. But note that here also a multisine is enough but it leads to results less accurate outside the frequencies of the multisine.

| Transfer function $E$ | $n_b$ | $n_f$ | Best Fit on identification data | Best Fit on verification data |
|---|---|---|---|---|
| $E_{xx}$ | 6 | 6 | 88.86% | 89.18% |
| $E_{xy}$ | 4 | 4 | 92.83% | 92.89% |
| $E_{yx}$ | 4 | 4 | 10.54% | 9.87% |
| $E_{yy}$ | 4 | 4 | 89.67% | 88.44% |

Table 2 – Results of the PEM identification of $E$, obtained orders and Best Fit.

For the residual analysis, the cross-correlations between the prediction error at the optimal $\hat{\theta}_N$ and the input for each transfer function are presented in the Appendix C.

From a first observation we can say that $G_{yy}$, $E_{xx}$, $E_{yy}$ and $E_{xy}$ are well identified, their Best Fit are very good. For $G_{xx}$ the wrong results of the Best Fit come from the resonance of $G_{xx}$ which has not been perfectly excited. Indeed as the mode is more lightly damped than in the sense mode (drive quality factor ten times higher than the one of the sense mode), exciting a little off the resonance $\omega_{0_x}$ will be poorly seen in the drive output $V_{out_x}$. Moreover we do not know perfectly this resonance, all the frequencies of the multisine $V_{in_x}^2$ might be a little off the resonance $\omega_{0_x}$. For $E_{yx}$ and $G_{xy}$ it is explained by the fact that the effect from the gyroscope excitation is partly hidden in the noise level.

**Bode diagram of computed G and E:** The Bode magnitude diagrams of the identified transfer function matrices are shown in the Figure 3.2 for $G$ and in the Figure 3.3 for $E$.

For $G$ the resonances peaks have been well identified. In $G_{xx}$ the peak is at 11749 Hz, corresponding to the experiment analysis and 11867 Hz for $G_{yy}$. A smaller resonance peak for $G_{xy}$ at 11810 has also been modeled. This model $G$ does not correspond to the model of the manufacturer, presented in the Figure 2.3. Their model comes only from physic equations and our model only from experimental data. Therefore our model is more reliable for a controller design.

For $E$ nearly all transfer functions present the same aspect (except for $E_{yx}$ which is not accurate due to the small effect of its contribution on the output data). There is a slope of +20 dB/dec which can illustrate a parasite capacitive coupling. Then around the resonance range the gain seems to be constant, something that has been observed by the electronic card developers. The anti-aliasing filter effect can be seen after the resonance range. What is important to see here is that the parasite electrical coupling is maximal at the resonance range (and at the half of it) which can cause several inaccuracies on the deduction of the angular rate $\hat{\Omega}$ if we do not take account of that parasite phenomena.

**Validation with a controller on the drive mode:** A controller has been designed from this model for the reference signal tracking on the drive mode by Fabrício Saggin with the $\mathcal{H}_\infty$ method, not explained in this report. He did not apply an input voltage on the sense mode. Then the concerned transfer function for the controller design is $G_{xx}$ and the control effort is $V_{in_x}$. He designed the controller based on the computed $G_{xx}$ and subtracts from the output $V_{out_x}$ the effect from the parasite electrical coupling such that the signal in the feedback loop compared to the reference is only the image of the mass excitation. This feedback signal is $V_{out_x}(t) - E_{xx}(z)V_{in_x}(t)$.

The results from the model and the experiment were the same and the error of the tracking of the reference signal was minimized by compensating with $E_{xx}$ ! Therefore our contribution can be validated. Indeed from these experiments with a controller we can validate our way to separate the gyroscope excitation and the parasite effect in two transfer functions (at least for the drive) and we can also validate the computation of $G_{xx}$ and $E_{xx}$.

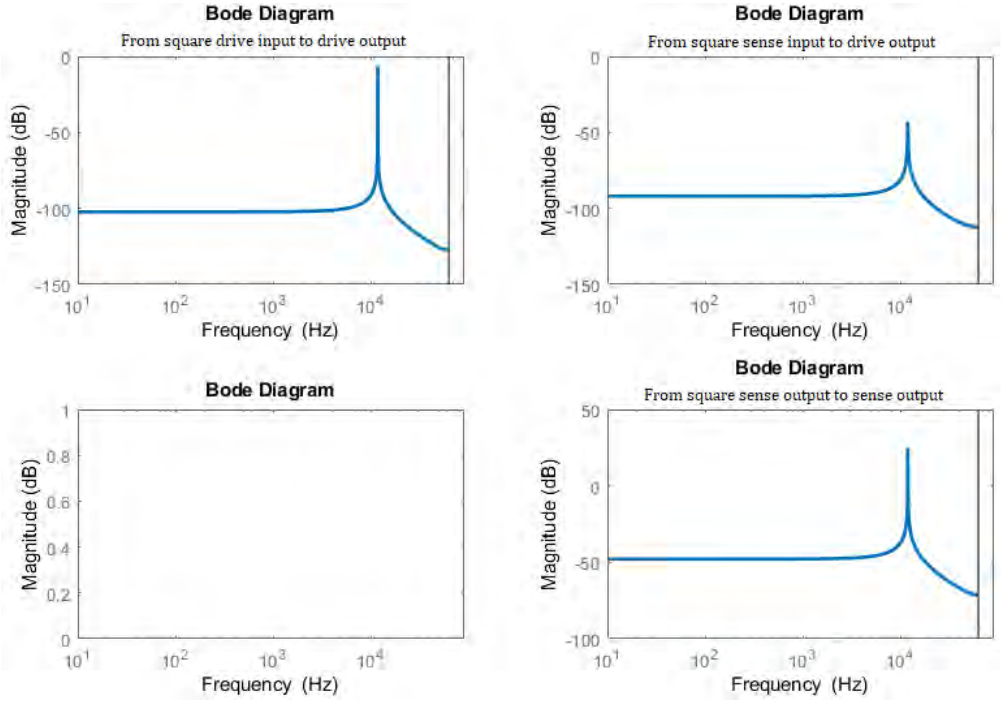



Figure 3.2 – Bode magnitude diagram of $G$ : $G_{xx}$ in the top-left plot, $G_{xy}$ in the top-right plot, $G_{yx}$ in the bottom-left plot and $G_{yy}$ in the bottom-right plot.

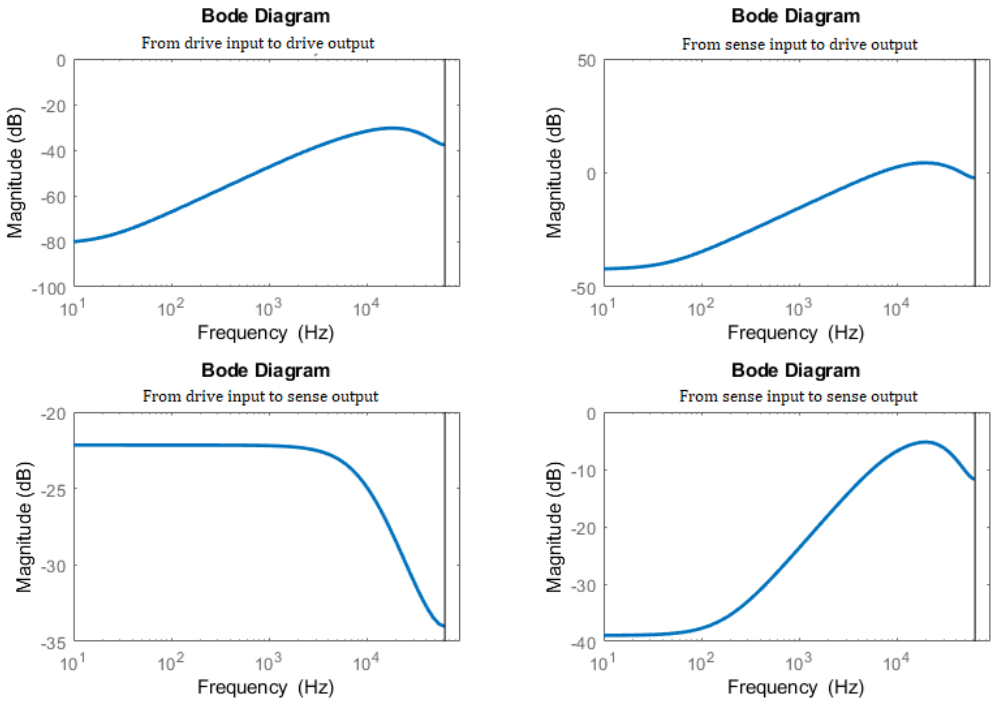



Figure 3.3 – Bode magnitude diagram of $E$ : $E_{xx}$ in the top-left plot, $E_{xy}$ in the top-right plot, $E_{yx}$ in the bottom-left plot and $E_{yy}$ in the bottom-right plot.

In the next section we give the main perspectives of the PhD thesis corresponding to improvements and problems faced with the LTI model.

# 4 Perspectives and roadmap of the PhD thesis

In this section we give the possible perspectives for this PhD thesis. These perspectives are directly linked to the computation approach of the model described in the previous section and the limitation of the obtained LTI model to describe other effects involved in the MEMS gyroscope.

## 4.1 MIMO modeling of the GYPRO3300: input correlation problem

Our approach to model the gyroscope was to do it SISO by SISO modeling as the algorithm is more efficient than in the MIMO one, less complex to identify [13] and we do not have input correlation issue. However for some control strategies both inputs in $V_{in}$ are excited in the same time and both outputs in $V_{out}$ are measured. Therefore it would be more accurate to model every transfer function in $E$ and $G$ with a MIMO approach. To circumvent the efficiency issue of the algorithm (local minimum problem) a good initial point for the computation will lead to the right result. Therefore we can use our previously computed transfer functions as initial point for the MIMO identification algorithm.

But to do so we need MIMO experimental data, i.e. all inputs are simultaneously excited and all outputs are simultaneously measured. In the MIMO configuration, the correlation between the inputs can lead to a wrong results. Indeed the property of the uniqueness of the computed $\hat{\theta}_N$ can be false with input correlation.

We want to identify the gyroscope with a multisine for $V_{in_x}$ and another one for $V_{in_y}$. If we put the sinusoids around $\omega_{0_x}/2$ and $\omega_{0_y}/2$ of each mode, then we can remove the frequencies of the input $V_{in}^2$ which are around the half of the resonance frequency (by filtering them). In this case we will not have common sinusoids and then no correlation between $V_{in}$ and $V_{in_x}^2$.

But it will be more appropriate to excite with sinusoids around $\omega_{0_x}$ and $\omega_{0_y}$. Indeed it is better to identify with the same experimental conditions used for the nominal operation of the gyroscope [14]. In the SISO-by-SISO approach we could not do that as we wanted to separate both effects, not possible with multisine around $\omega_{0_x}$ and $\omega_{0_y}$, to benefit from the strong bandpass filter effect of the gyroscope.

Now that we have computed SISO models we can try to identify every transfer function simultaneously with sinusoids around $\omega_{0_x}$ and $\omega_{0_y}$ by using the computed SISO models as good initial point for the algorithm. But the inputs $V_{in_x}^2$ and $V_{in_y}^2$ are also multisine with frequencies described in the paragraph 3.4.1. Then there are common sines between the inputs $V_{in_x}$ and $V_{in_x}^2$. More precisely all the sinusoids in the input $V_{in_x}$ are correlated with the ones of $V_{in_x}^2$. This is also the same for $V_{in_y}$ and $V_{in_y}^2$ and we cannot use a filtered version of these signals as both effects are around the exciting frequencies. This can have an effect on the consistency of the PEM estimator [15] in a MIMO identification. However this problem has not been so much studied in the literature [16] as all inputs can be freely chosen and generally we avoid to correlate them. A contribution in this domain could be a theoretic possible perspective of this PhD thesis.

Some first studies done during this first year of the PhD thesis have shown that it is possible to keep the consistency of the estimator for the parallel structure shown in the Figure 2.8 with multisine as long as we have a sufficient number of sines. We will not give further details as the theory has not

13. The number of transfer functions in MIMO identification can be high as in our case (8 transfer functions) and the cross-correlation analysis for the verification must be true **simultaneously** for all the transfer functions which make the choice of the polynomial orders more complex.

14. For the chosen control strategy for the drive mode during this first year, Fabricio Saggin chose to track a sinus at the the resonance frequency $\omega_{0_x}$. Then its control effort $V_{in_x}$ is around $\omega_{0_x}$ and not $\omega_{0_x}/2$.

15. Here we talk about the property of the uniqueness of the global minimum $\theta_0$.

16. There are some studies for subspace identification as in [Jansson and Wahlberg, 1998]. In [Mareels and Gevers, 1988] Gevers also studied the consistency of PEM estimator but this theory is not trivial to be applied easily in our case. However these articles can be a basis for the development of another simpler conditions for the persistency property.

been detailed yet. But it seems that a LTI MIMO identification is possible. However in this model we did not consider the angular rate. This is presented in the next paragraph.

## 4.2 Dependency of the model with the angular rate $\Omega$

During this first year we have computed a model of the gyroscope without angular rate. The aim of the gyroscope is the deduction of $\Omega$ so the model needs to explain accurately its effect on the dynamics. In the model from physical equations it is assumed to be a Coriolis contribution given in the Appendix A. We have to verify the validity of this expression by doing experiments with angular rate and studying the effect on the output $V_{out}$. If it is not satisfying then we will have to take account of the angular rate variations by computing a LPV (Linear Parameter Varying) black-box model.

There are different approaches to compute a LPV model. The local approach explained in the paper [Ghosh et al., 2018] is interesting. Indeed the idea is to compute LTI models for different angular rates. Then we interpolate these models (more precisely the parameters of those models) to obtain the LPV model. This will be one of the perspectives of this PhD thesis.

## 4.3 From LTI to LPV modeling for the temperature dependency

During the controller validation step a controller has been designed with the $\mathcal{H}_\infty$ method on this model by Fabrício Saggin and the results from the experiment and the model were the same the first day. When he did the experiment again the next week, the control effort saturated and the performances were not reached.

The phenomena that explains this behavior is the temperature dependency explained in the paragraph 2.4: the drive resonance frequency $\omega_{0_x}$ has shifted and as the mode is very slightly damped, the computed resonance frequency of the LTI model $\hat{\omega}_{0_x}$ is out of the real resonance range and is filtered. The controller has been designed to track a reference signal at $\hat{\omega}_{0_x}$, then the control effort saturates as this frequency is filtered.

In the set of specifications of the MEMS manufacturer the angular rate deduction performances must be reached for temperature between $-20°C$ to $80°C$. The first possibility to handle this is to consider the variations of the parameters as uncertainties on the parameters of the model at a nominal temperature. But from the described experiment with the controller and from the first results of Jorge Ayala on the robustness analysis the performances will not be obtained for all the possible values of the parameters coming from the temperature evolution. Then we need an adaptive controller, so we need a LPV model of the gyroscope to model the temperature dependency.

## 4.4 Uncertainties of the computed models with PEM identification

Due to noise presence on the output data when performing the identification experiments in 3.4.1, there are some uncertainties on the identified parameters resulting in a modeling error $\theta_0 - \hat{\theta}_N$. In PEM identification we can quantify this modeling error depending on $N$ and the input power spectrum $\Phi_u$ of the identification experiment[17]. To be able to do it we need first to model the noise $w$ as a filtered white noise : $w = H(z, \theta)e$ where $H$ is a monic ($H(0) = 1$), stable and inversely stable transfer function and $e$ a white noise vector. We have to consider a new $\theta$ which contains also the parameters of that noise model $H$.

If we go back to the control strategy of the MEMS gyroscope explained in the paragraph 2.2 for the deduction of the angular rate, we see that it depends on the computed parameters. Therefore as we have modeling errors on the parameters it results in errors in the deduction of the angular rate.

17. This relation is known and can be seen in [Ljung, 1999] for the SISO case and [Barenthin et al., 2008] for the MIMO case.

We can build a time-domain criterion featuring the error between the true angular rate $\Omega(t)$ and its deduction $\hat{\Omega}(t, \hat{\theta}_N)$ depending on the identified uncertain parameters. The criterion, denoted $\mathcal{V}(\hat{\theta}_N)$, can be expressed in a least-square formulation as presented in the (4.1).

$$\mathcal{V}(\hat{\theta}_N) = \frac{1}{N} \sum_{k=1}^{N} (\hat{\Omega}(\hat{\theta}_N, t) - \Omega_0(t))^2 \tag{4.1}$$

This error can be unsatisfying, for instance by exceeding the value of the set of specifications (which is a maximal error of 500 ppm for the manufacturer). Let say that we want to have $\mathcal{V}(\hat{\theta}_N) < \beta$ for different angular rate scenarii, where $\beta$ can come from the set of specifications. One perspective could be to determine the optimal experiment (the optimal $N$ and $\Phi_u$) such that the modeling error $\theta_0 - \hat{\theta}_N$ is small enough to guarantee $\mathcal{V}(\hat{\theta}_N) < \beta$. This is the aim of Experiment Design.

The problem in our case is the fact that $\hat{\Omega}(t, \hat{\theta}_N)$ is very complex. We can approximate this criterion with a Taylor second order approximation as presented in [Hjalmarsson, 2009] for complex systems and applied in [Forgione et al., 2015]. It results in a linear matrix inequality which involves $\Phi_u$. This can be a great contribution on the improvement of the accuracy of the MEMS gyroscope angular rate deduction with Experiment Design.

Another challenging problem is the input constraints in our case. As the inputs of the mechanical dynamics are te square of voltages, the optimal inputs must be positive as explained in the paragraph 3.1. This must be considered in the Experiment Design formulation and has not been studied so much in the literature survey.

## 4.5 Roadmap of the PhD thesis

To sum up briefly what has been discussed in this section we give here the roadmap of the PhD thesis for the next months in a chronological order:

- Input correlation study for PEM MIMO identification applied to multisine. This study can also lead to more general theoretical results which can contribute well in the literature.
- MIMO identification by using the model obtained with SISO-by-SISO approach as initialization of the algorithm. Verification of the prior knowledges on the computed model.
- Physical interpretation study of the black-box MIMO model.
- Theoretical analysis of the impact of computed parameter uncertainties on the angular rate deduction with the LTI MIMO model.
- First Experiment Design approach with second order approximation method for the minimization of uncertainties on the angular rate deduction.
- LPV modeling of the gyroscope with local approach. Reflection of an experimental setup to make vary the temperature and the angular rate for the LPV modeling.
- Experiment design on the LPV modeling for the minimization of the angular rate deduction uncertainties.

Note that, depending on the results of the two other PhD students with the model, some parts can be done again to improve or change the model.

# 5 Conclusion

In this report we discussed about a new way to model a MEMS gyroscope. After presenting the advantages and challenges around this type of gyroscope we study the different models. Our strategy was to start from the ideal model and then complete it by taking account of phenomena involved in the gyroscope (anisoelasticity, non proportional damping, capacitive instrumentation). However this more complete model cannot explain the noise power spectrum and the parasite electrical effect. Then we use a black-box approach with a new modeling structure (parallel model) with a SISO-by-SISO approach and by using PEM identification. The first results are encouraging and allow us to say that our model contribution is more accurate.

However it would be more accurate to do it in a MIMO configuration. To do it with a MIMO configuration we need to study the input correlation effect on the consistency of the PEM estimator as we face an input correlation issue in our problem.

The obtained LTI model does not consider the dependency of the dynamics with the angular rate. We need to verify the Coriolis expression explained in the Appendix A and if not satisfying, we need to model this effect with a black-box LPV approach.

The LTI model is not enough accurate as its parameters depend on the temperature evolution. The LPV approach can help us to describe the dependency with the temperature by interpolating LTI models computed at different temperatures (local approach).

Finally all computed LTI models have uncertainties due to the noise presence. In that case we need to study the impact of them in the angular rate deduction. If these uncertainties become too important we can minimize them by using Experiment Design. As the system is complex we can use a second-order approximation to simplify the computation.

# References

[One, 2018] (2018). Propositions d'emploi Post-doc dans le cadre du projet NEXT4MEMS, site internet onera.fr.

[Barenthin et al., 2008] Barenthin, M., Bombois, X., Hjalmarsson, H., and Scorletti, G. (2008). Identification for control of multivariable systems: Controller validation and experiment design via LMIs. *Automatica*, 44(12):3070–3078.

[Bombois and Scorletti, 2012] Bombois, X. and Scorletti, G. (2012). Design of least costly identification experiments : The main philosophy accompanied by illustrative examples. *Journal Européen des Systèmes Automatisés (JESA)*, 46(6-7):587–610.

[Chen et al., 2003] Chen, Y.-C., Hui, J., and M'Closkey, R. (2003). Closed-loop identification of a micro-sensor. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, volume 3, pages 2632–2637 Vol.3, Maui, Hawaii USA.

[Chen et al., 2005] Chen, Y.-C., M'Closkey, R., Tran, T., and Blaes, B. (2005). A control and signal Processing integrated circuit for the JPL-boeing micromachined gyroscopes. *IEEE Transactions on Control Systems Technology*, 13(2):286–300.

[Egretzberger et al., 2012] Egretzberger, M., Mair, F., and Kugi, A. (2012). Model-based control concepts for vibratory MEMS gyroscopes. *Mechatronics*, 22(3):241–250.

[Fei et al., 2010] Fei, J., Hua, M., and Xue, Y. (2010). A comparative study of adaptive control approaches for MEMS gyroscope. In *2010 IEEE International Conference on Control Applications*, pages 1856–1861, Yokohama, Japan.

[Fei and Yang, 2011] Fei, J. and Yang, Y. (2011). System Identification of MEMS Vibratory Gyroscope Sensor. *Mathematical Problems in Engineering*, 2011:1–12.

[Fei and Yang, 2012] Fei, J. and Yang, Y. (2012). Comparative study of system identification approaches for adaptive tracking of MEMS gyroscope. *International Journal of Robotics and Automation 2012*, 27(6).

[Forgione et al., 2015] Forgione, M., Bombois, X., and Van den Hof, P. (2015). Data-driven model improvement for model-based control. *Automatica*, 52:118–124.

[Ghosh et al., 2018] Ghosh, D., Bombois, X., Huillery, J., Scorletti, G., and Mercère, G. (2018). Optimal identification experiment design for LPV systems using the local approach. *Automatica*, 87:258 – 266.

[Guan et al., 2015] Guan, R., He, C., Liu, D., Zhao, Q., Yang, Z., and Yan, G. (2015). A temperature control system used for improving resonant frequency drift of MEMS gyroscopes. In *10th IEEE International Conference on Nano/Micro Engineered and Molecular Systems*, pages 397–400.

[Hjalmarsson, 2009] Hjalmarsson, H. (2009). System Identification of Complex and Structured Systems. *European Journal of Control*, 15(3):275–310.

[Hopcroft et al., 2010] Hopcroft, M. A., Nix, W. D., and Kenny, T. W. (2010). What is the Young's Modulus of Silicon? *Journal of Microelectromechanical Systems*, 19(2):229–238.

[Jansson and Wahlberg, 1998] Jansson, M. and Wahlberg, B. (1998). On Consistency of Subspace Methods for System Identification. *Automatica*, 34(12):1507–1519.

[Korniienko et al., 2017] Korniienko, A., Bombois, X., Scorletti, G., Dehaeze, T., and Bohoslavets, R. (2017). Rapport d'étude bibliographique et d'étude déterminant les fonctionnalités devant être intégrées à la plate-forme pour la mise en oeuvre des algorithmes de conception Projet NEXT4mems. Technical report, Laboratoire Ampère, École Centrale de Lyon.

[Ljung, 1999] Ljung, L. (1999). *System identification: theory for the user*. Prentice Hall information and system sciences series. Prentice Hall PTR, Upper Saddle River (NJ), second edition edition.

[Mareels and Gevers, 1988] Mareels, I. and Gevers, M. (1988). Persistency of Excitation Criteria for Linear, Multivariable, Time-Varying Systems. *Mathematics of Control, Signals, and Systems*, 1:203–226.

[M'Closkey et al., 1999] M'Closkey, R., Gibson, S., and Hui, J. (1999). System Identification of a MEMS Gyroscope. *Journal of Dynamic Systems, Measurement, and Control*, 123(2):201–210.

[Painter and Shkel, 2002] Painter, C. and Shkel, A. (2002). Identification of anisoelasticity for electrostatic trimming of rate integrating gyroscopes. *Smart Struct. Mater.*, 4700.

[Phani and Seshia, 2004] Phani, S. and Seshia, A. (2004). Identification of Anisoelasticity and Non-proportional Damping in MEMS Gyroscopes. *TechConnect Briefs*, 2(2004):343–346.

[Phani et al., 2006] Phani, S., Seshia, A., Palaniapan, M., Howe, R., and Yasaitis, J. (2006). Modal Coupling in Micromechanical Vibratory Rate Gyroscopes. *IEEE Sensors Journal*, 6(5):1144–1152.

[Saukoski, 2008] Saukoski, M. (2008). *System and circuit design for a capacitive MEMS gyroscope*. PhD thesis, Helsinki University of Technology, Department of Micro and Nanosciences.

[Waterston and Strutt, 1892] Waterston, J. and Strutt, J. W. (1892). On the physics of media that are composed of free and perfectly elastic molecules in a state of motion. *Phil. Trans. R. Soc. Lond. A*, 183:1–79.

[Xia et al., 2009] Xia, D., Chen, S., Wang, S., and Li, H. (2009). Temperature Effects and Compensation-Control Methods. *Sensors (Basel, Switzerland)*, 9(10):8349–8376.

[Yazdi et al., 1998] Yazdi, N., Ayazi, F., and Najafi, K. (1998). Micromachined inertial sensors. *Proceedings of the IEEE*, 86(8):1640–1659.

[Yongzhen et al., 2011] Yongzhen, F., Luo, B., and Wang, A. (2011). Analysis of temperature adaptability for frequency control loop for silicon micromechanical gyroscope. In *IEEE 2011 10th International Conference on Electronic Measurement Instruments*, volume 4, pages 346–349.

[Zhu Yu-Cal, 2007] Zhu Yu-Cal (2007). Black-box identification of mimo transfer functions: Asymptotic properties of prediction error models. *International Journal of Adaptive Control and Signal Processing*, 3(4):357–373.

# A Description of Coriolis effect

Let consider an object, with its inertial center denoted $M$, with a mass denoted $m$ moving into a direct orthonormal basis $(O, \overrightarrow{e}_x, \overrightarrow{e}_y, \overrightarrow{e}_z)$ [18] or $(O, x, y, z)$. This frame has a rotating motion with an angular rate vector $\overrightarrow{\Omega}$ along the $z-$axis. In this frame, the vectors composing the basis $(O, x, y, z)$ are constant. When the object $M$ moves in a reference frame which has not a non-uniformly motion [19] it is subjected to cinematic effects in addition to the external forces. One of them is the Coriolis force effect.

This force is applied only in the plane orthonormal to the angular rate vector $\overrightarrow{\Omega}$. By denoting $\overrightarrow{v} = \begin{pmatrix} \dot{x} & \dot{y} & \dot{z} \end{pmatrix}^T$ the velocity vector of $M$ then a force, denoted Coriolis, equal to $\overrightarrow{F}_{Coriolis} = 2m\overrightarrow{v} \wedge \overrightarrow{\Omega}$ is applied on the mass in the direction perpendicular to $\overrightarrow{\Omega}$ and $\overrightarrow{v}$. An example is given in the Figure A.1 when $\overrightarrow{\Omega}$ is along the $z-$axis.

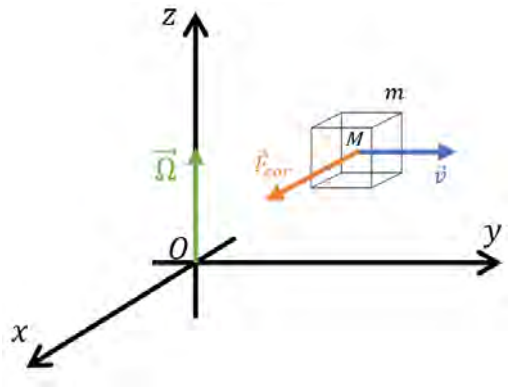



Figure A.1 – Illustration of the Coriolis effect.

The coordinates of the Coriolis force in this configuration are:

$$\overrightarrow{F}_{Coriolis} = \begin{pmatrix} F_{Coriolis_x} \\ F_{Coriolis_y} \\ F_{Coriolis_z} \end{pmatrix} = 2m \begin{pmatrix} \Omega\dot{y} \\ -\Omega\dot{x} \\ 0 \end{pmatrix}$$

.

# B Further details on the capacitive instrumentation

This Appendix is based on the instrumentation with double-combs illustrated in the Figure 2.4

We will denote $C^{act}$ the capacitance created by the actuation double-combs and $C^{mea}$ by the measurement double-combs. For the sake of simplicity of notation we will present the equations of the capacitance for both cases (actuation and measurement) as it is the same phenomena by using the notation $act/mea$ as index. For small motions of the mass (which is the case here) the capacitance $C$ depends linearly of the displacement of the mass, so $x(t)$ for the drive actuation double-combs and $y(t)$ for the sense actuation double-combs : $C^{act/mea}(t)_x = C_{0_x}^{act/mea} - \beta_x^{act/mea} x(t)$ and $C_y^{act/mea}(t) = C_{0_y}^{act/mea} - \beta_y^{act/mea} y(t)$ where $\beta_x^{act/mea}$, $\beta_y^{act/mea}$, $C_{0_x}^{act/mea}$ and $C_{0_y}^{act/mea}$ are constant coefficients, depending on the geometry and the electrical properties of the combs.

18. The vectors $\overrightarrow{e}_x$, $\overrightarrow{e}_y$ and $\overrightarrow{e}_z$ form a direct basis and $||\overrightarrow{e}_x|| = ||\overrightarrow{e}_y|| = ||\overrightarrow{e}_z|| = 1$.

19. Here we mean that the frame has not a constant speed straight-line motion. In other words the vector speed at the center of the frame is not constant.

If we apply a tension $V$ between two combs dedicated to the actuation, then the stored capacitive energy is given by $E(t) = \frac{1}{2}C^{act}(t)V^2(t)$, with $C^{act}(t)$ the value of the capacitance between the two combs placed face to face. We denote $V_{in_x}(t)$ the voltage applied to the drive actuation double-comb and $V_{in_y}(t)$ the one to the sense actuation double-comb. Then the external forces $F_x(t)$ and $F_y(t)$ are $F_i(t) = -\frac{dE_i}{dx}(t) = \beta_i^{act}V_{in_i}^2(t)$ with $i = \{x, y\}$. The forces do not depend on the position of the mass which is very interesting for control purposes.

Now let consider the drive and sense measurement double-comb. We know that $C^{mea}(t)_x = C_{0_x}^{mea} - \beta_x^{mea}x(t)$ and $C_y^{mea}(t) = C_{0_y}^{mea} - \beta_y^{mea}y(t)$. Then we can directly measure $x(t)$ and $y(t)$ with the measure of $C_x^{mea}(t)$ and $C_y^{mea}(t)$. This measurement is done by using a high-pass filter to remove the constant term. It converts the capacitance into a voltage that can be measured. Then this voltage is amplified by two stages of amplifier. We will denote $V_{out_x}$ this voltage output which is a linear image of $x$ for the drive mode and $V_{out_y}$ for the sense mode.

# C Cross-correlation results for the identification of $E$ and $G$ with PEM

This appendix presents the results on the cross-correlation between the input and the residual for the identification of $E$ and $G$ with PEM. For $G$, the one of $G_{xx}$ is presented in the Figure C.1, the one of $G_{xy}$ is presented in the Figure C.2 and the of $G_{yy}$ is presented in the Figure C.3. The one of $G_{yx}$ is not presented as there is no significant effect in the FFT of the output $V_{out_y}$ when we excite the drive mode.

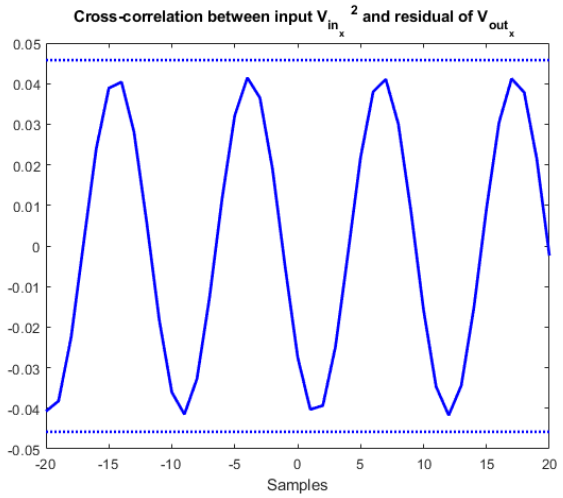



Figure C.1 – Cross-correlation between the input $V_{in_x}^2$ and the residual of $V_{out_x}$.

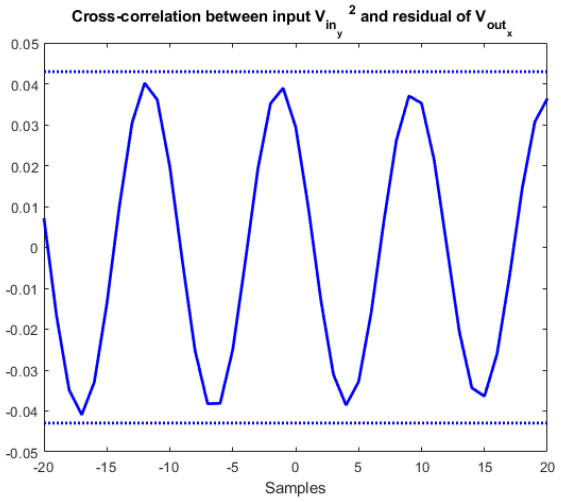



Figure C.2 – Cross-correlation between the input $V_{in_y}^2$ and the residual of $V_{out_x}$.

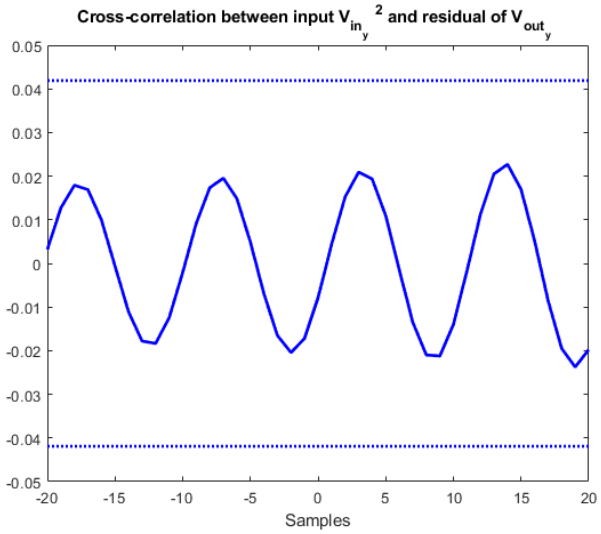



Figure C.3 – Cross-correlation between the input $V_{in_y}^2$ and the residual of $V_{out_y}$.

For $E$, the one of $E_{xx}$ is presented in the Figure C.4, the one of $E_{xy}$ is presented in the Figure C.5 and the of $E_{yy}$ is presented in the Figure C.6.

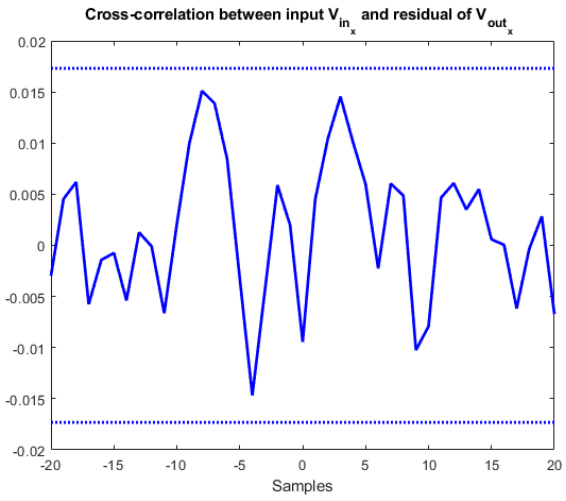



Figure C.4 – Cross-correlation between the input $V_{in_x}$ and the residual of $V_{out_x}$.

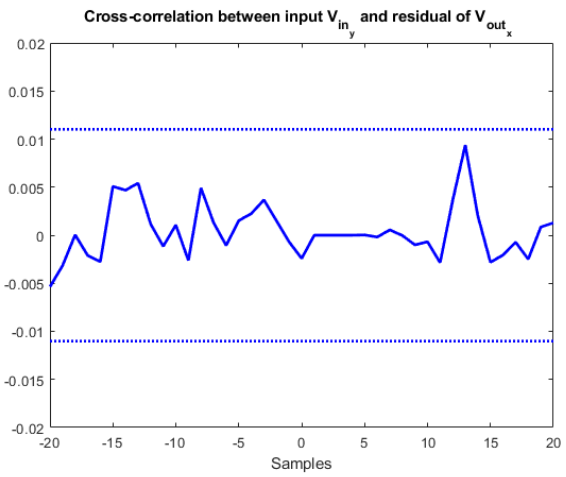



Figure C.5 – Cross-correlation between the input $V_{in_y}$ and the residual of $V_{out_x}$.

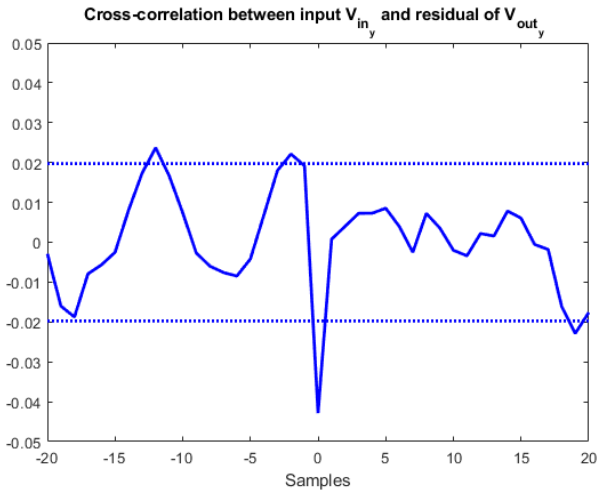



Figure C.6 – Cross-correlation between the input $V_{in_y}$ and the residual of $V_{out_y}$.

Université de Lyon
CNRS, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1

**Laboratoire Ampère**
Unité Mixte de Recherche du CNRS - UMR 5005
Génie Electrique, Automatique, Bio-ingénierie

# *Mémoire doctorant 1ère année*
# *2017 -2018*

| | |
|---|---|
| Nom - Prénom | **Errigo Florian** |
| email | **Florian.errigo@supergrid-institute.com** |
| Titre de la thèse | **« Convertisseurs de puissance avec stockage d'énergie intégré pour réseaux haute tension à courant continu »** |
| Directeur de thèse | **Venet Pascal** |
| Co- encadrants | **Sari Ali** |
| Dpt. de rattachement | **T1 : Systèmes et énergies sûrs** |
| Date début des travaux | **1/10/17** |
| Type de financement | **Salarié SuperGrid** |

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France
Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

# Table des matières

# Nomenclature

## Liste des acronymes

| | |
|---|---|
| **AC** | Courant Alternatif |
| **DC** | Courant Continu |
| **DAB** | Dual Active Bridge |
| **FC** | Convertisseur multicellulaire série (Flying Capacitor) |
| **HVDC** | High Voltage Direct Current |
| **IGBT** | Insulated Gate Bipolar Transistor |
| **LCC** | Line Commutated Converter |
| **MMC** | Convertisseur Modulaire Multiniveaux (Modular Multilevel Converter) |
| **NPC** | Convertisseur à neutre clampé (Neutral Point Clamped Converter) |
| **SM** | Sous-module |
| **VSC** | Voltage Source Converter |

# Liste des symboles

| | |
|---|---|
| $C$ | Capacité d'un sous module |
| $C_{mmc}$ | Capacité équivalente d'un convertisseur MMC |
| $E_{cinétique}$ | Energie cinétique |
| $E_{capactive}$ | Energie electrostatique |
| $H$ | Constant d'inertie |
| $H_{dc}$ | Constante d'inertie côté DC |
| $i$ | Indice pour le bras correspondant avec $i \in \{\ a,b,c\ \}$ |
| $I_{dc}$ | Courant DC |
| $i_{\boldsymbol{ac}}$ | Courant AC du réseau |
| $i_{\boldsymbol{es}}$ | Courant au sein du système de stockage de l'énergie |
| $i_{\boldsymbol{i}}^{\boldsymbol{u,l}}$ | Courant d'un demi-bras |
| $i_{\boldsymbol{sm}}^{\boldsymbol{u,l}}$ | Courant redressé au sein d'un sous module |
| $J$ | Moment d'inertie |
| $n$ | Nombre de sous modules insérés dans un demi-bras |
| $N$ | Nombre de sous modules par demi-bras |
| $P_e$ | Puissance électrique consommée |
| $P_m$ | Puissance produite |
| $S_n$ | Puissance apparente |
| $u, l$ | Indice pour le demi-bras supérieur et inférieur |
| $U_{sm}$ | Tension aux bornes d'un sous module |
| $V_{ac}$ | Tension simple AC |

| | |
|---|---|
| $V_{dc}$ | Tension DC |
| $v_{\boldsymbol{i}}^{\boldsymbol{u,l}}$ | Tension d'un demi bras |
| $\varphi$ | Déphasage entre la tension et le courant du réseau AC |
| $\omega$ | Vitesse angulaire |
| $\psi$ | Déphasage du courant circulatoire |

# Table des figures

# Chapitre 1

# Introduction générale

## 1.1 Contexte de l'évolution des réseaux électriques

Pendant de longues décennies, les réseaux électriques ont été considérés comme des systèmes matures. Cependant, depuis plusieurs années, ils sont sujets à de profondes mutations pour faire face à la demande croissante d'énergie électrique et garantir sa sécurité d'approvisionnement. En parallèle, la prise de conscience mondiale en vue de lutter contre le réchauffemment climatique a mené à des politiques énérgétiques avec des objectifs de réduction d'émissions des gaz à effet de serre conséquents [1]. Dans ce contexte, l'Europe s'est positionnée avec une approche progressive avec le développement des énergies renouvelables en vue de cette transition vers une économie décarbonée [2, 3, 4]. On pourra citer comme tournant majeur le Paquet Horizon 2020 visant à atteindre une consommation électrique de 20 % à partir de sources renouvelables d'ici à 2020 [3].

En parallèle, l'énergie est devenue une commodité commercialisable [5]. L'ouverture à la concurrence et la suppression des modèles de monopole verticalement intégrés se sont traduit par la séparation des activités de production, transport et de commercialisation [6]. De nouveaux producteurs peuvent s'installer librement et les consommateurs choisirent ouvertement leur fournisseur. A present l'électricité peut être échangée à tout instant au sein d'une même zone continentale entrainant inévitablement une augmentation du transit des puissances sur les lignes de transport. A l'origine, les infrastructures des réseaux de transport n'ont pas été dimensionnées pour des échanges à l'échelle internationale. La capacité des interconnexions entre pays se doit d'être augmentée. En Europe, la commision européenne a fixé comme objectif une augmentation de 15% des interconnexions d'ici à 2030 [7].

Le passage à un schéma de production décentralisé et au caractére incertain a mis en avant le besoin de restructuration des installations existantes vers des solutions plus flexibles et efficaces. Avec l'introduction de sources intermittentes, géographiquement dispersées, et l'augmentation des flux de puissance sur les lignes de transmission, les réseaux de transport électrique à courant alternatifs (AC) vont être sujet à des contraintes de fonctionnement de plus en plus denses. De nouveaux investissements sont nécessaires [8].

## 1.2 L'incitation aux infrastructures HVDC

En outre, la mise en place d'un marché de gros a conduit à une complexification des régimes d'approbation réglementaire et un cadre administratif plus astreignant vis à vis de la construction de nouvelles lignes, déjà coûteuses en temps. Mais le principal défi de cette transition est bien d'ordre technique. En plus de prendre en compte un besoin énergétique croissant, l'évolution des réseaux est dorénavant dictée par la localisation des nouvelles unités de production. Or les principaux gisements solaires et éoliens, représentants une grande part des énergies renouvelables, sont généralement éloignés des centres de consommation. Géographiquement, cela se traduit par un déplacement des centres de production vers les frontières et notamment les littoraux. Dans cette perspective, l'énergie électrique devra être transportée sur de plus longues distances. Face à ces changements, les limites de la technologie alternative apparaissent. La capacité de transport des liens AC est dépendante des phénomènes inductifs et capacitifs se produisant dans les lignes et les câbles. Une liaison câblée se retrouve rapidement limitée par son courant capacitif, responsable uniquement de pertes et ne véhiculant aucune puissance active, dès que sa longueur augmente, Fig. 1.1a. En dernier lieu, la connexion entre deux réseaux asynchrones n'est pas envisageable. De ce fait, le raccordement de sources offshores est inaccessible avec les technologies AC actuelles.

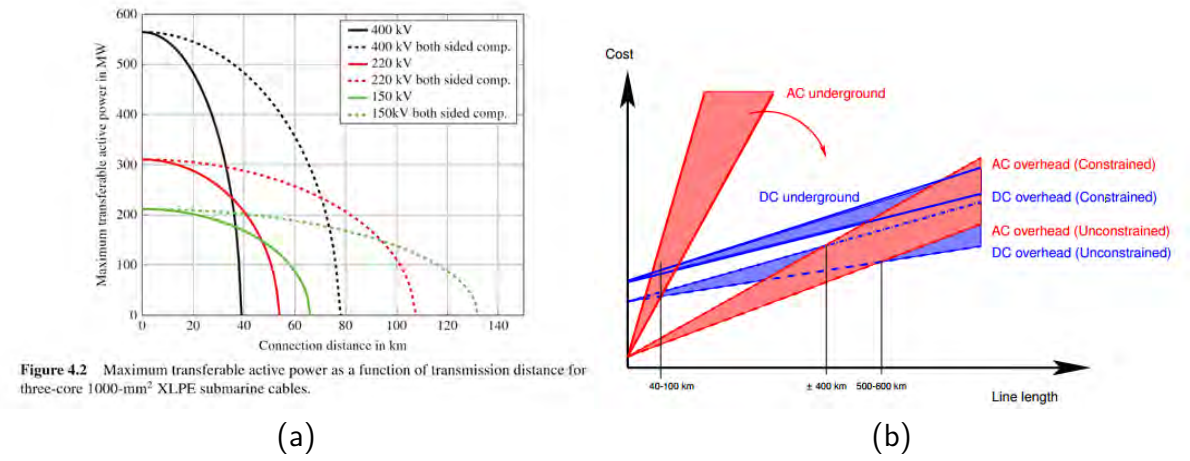



FIGURE 1.1 – (a) Puissance active transmissible en fonction de la distance pour un câble XLPE sous-marin à trois conducteurs (1000 mm$^2$) [9] (b) Comparaison économique entre un projet de lignes DC et AC en fonction de la distance de la ligne de transmission [9]

Pour faciliter cette nouvelle architecture, la technologie à courant continu sous haute tension (HVDC) est redevenue une solution plausible pour le transport de puissances considérables sous haute tension sur des distances élevées [10]. En théorie, la longueur des câbles en régime continu n'a pas de limites. D'un point de vue économique, le coût d'investissement d'une liaison HVDC est plus important du fait de la nécessité de stations de conversion, Fig. 1.1b. Cependant, le côut de transmission est plus faible compte tenu de pertes mineures en raison de l'absence de phénomènes tels que l'effet de peau ou la

production d'énergie réactive. En définitive, le désir de préservation de l'environnement et la réduction de l'impact visuel encourage à la construction de lignes à courant continu.

## 1.3 Le SuperGrid au coeur des réseaux de demain

Dans ce contexte, la mutualisation des systèmes offshores pourraient favoriser à la mise en oeuvre d'un vaste SuperGrid [9, 11, 12] qui viendrait renforcer les infrastructures AC. Cependant avant d'y arriver, il convient de s'assurer que l'on dispose de bases saines. Un des tournants majeurs dans les progrès remarquables des technologies HVDC a été l'évolution de l'électronique de puissance sous l'impulsion de l'avancée dans les technologies des semi-conducteurs [13]. On pourra citer comme acte fondateur les travaux du Dr. Lamm, considéré comme le père du HVDC, sur les diodes à vapeur de mercure qui a permis la commercialisation de la première liaison HVDC au monde [14]. A présent, les verrous technologiques liés à de tels ouvrages s'amenuisent et ouvrent la voie à de nouveaux réseaux combinant à la fois production centralisée\décentralisée et technologie AC\DC.

Aujourd'hui, la majorité des liaisons en services sont des liens point à point avec pour objectif d'augmenter le transfert d'énergie entre deux zones [15] ou faciliter l'accès aux sources renouvelables offshores [16, 17]. Tandis que les premiers réseaux multi-terminaux maillés commencent seulement à voir le jour en Chine [18, 19]. En parallèle, l'absence de standards a mené à des niveaux de tension distincts. Alors que la percée de l'électronique de puissance a permis l'utilisation de technologies de conversion différentes. On peut distinguer deux grandes familles.

— Les convertisseurs source de courant, *"Line commutated converter (LCC)"*, à la base des premières stations de conversion HVDC. Ils utilisent des thyristors comme élément de commutation. Par conséquent, cette topologie est dépendante d'un circuit extérieur afin d'assurer la fermeture à zéro de courant des composants de puissance, dans ce cas la tension du réseau AC. Ce dernier doit être *"fort"* pour garantir une activité sûre.
— En opposition, les convertisseurs à source de tension, *"Voltage Source Converter (VSC)"*, reposent sur des interrupteurs à commutation forcée commandables, Insulated Gate Bipolar Transistor (IGBT). Leur action est indépendante du comportement réseau.

Cette structure se retrouve d'autant plus avantageuse, comparée au système LCC, grâce à son contrôle indépendant des puissances active et réactive, une gestion plus aisée des flux de puissance ou encore un besoin en filtres moindre [9, 20]. Pour finir, sa capacité de démarrage sans présence de tension (Black Start) est un atout non négligeable. Elle s'impose ainsi comme la topologie de référence pour les futurs réseaux HVDC.

# Chapitre 2

# L'essor des convertisseurs modulaires multi-niveaux (MMC)

## 2.1 Les structures multi-niveaux conventionnelles

Sans envisager tous les aspects de la conversion DC-AC, on pourra citer comme structure historique "l'onduleur 2 niveaux". Toutefois, la faible tenue en tension des semi-conducteurs de puissance, de l'ordre du kilovolt, un taux de distorsion harmonique important et des pertes non négligeables restent des points bloquants pour des usages haute tension. Pour y pallier, l'une des premières solutions a été la mise en série de composants de puissance pour réaliser un seul et même interrupteur afin de disposer d'un calibre en tension plus important. Cependant, cette méthode s'est vite révélée insuffisante de part la difficulté de synchronisation de l'ensemble des commutations et le design de circuits de commande spécifiques.

Ainsi sont apparus les topologies multi-niveaux comme alternative à cette association directe d'IGBT. Elles se caractérisent par la mise en série de sources de tension élémentaires, réalisées à partir de condensateurs, fonctionnant à une fraction de la tension du bus DC globale. Il n'est plus nécessaire d'assurer des commutations synchrones et le stress sur les semi-conducteurs est réduit. Plusieurs niveaux de tension peuvent être obtenus. Une tension AC en forme de marche d'escalier, proche d'une sinusoïde, est générée minimisant la taille des filtres. On pourra mentionner comme principaux designs, les convertisseurs à neutre clampé (NPC) et multicellulaire série (FC) qui ont rapidement attirés l'attention [21, 22]. Néanmoins pour des applications haute tension, un nombre de niveaux élevé est difficilement atteignable compte tenu de la complexité des structures et le nombre de composants requis.

## 2.2 Les convertisseurs modulaires multi-niveaux

En 2001, les convertisseurs multi-niveaux ont connu une avancée importante avec la topologie proposée par R.Marquardt [23, 24]. Connu sous le nom de Modular Multilevel Converter (MMC), ce convertisseur modulaire multi-niveaux a l'avantage d'avoir une

structure totalement modulaire permettant d'atteindre des niveaux de tension plus élevés et d'améliorer la qualité des signaux de sortie.

### 2.2.1 Structure

Les convertisseurs triphasés MMC se composent de trois bras, divisés en un demi-bras supérieur et inférieur identiques (index {u,l}), Fig. 2.1. Ils se distinguent par la mise en série de N convertisseurs DC-AC élémentaires comme cellule de base, appelés sous-module (SM). A titre d'exemple, pas moins de 400 SMs sont utlisés par demi-bras pour une station de conversion dans le cadre de la liaison HVDC reliant la France et l'Espagne (2× 1000 MW, +\− 640 kV) [25]. La topologie demi-pont est la plus commune, composée de deux interupteurs et un condensateur autorisant deux niveaux de tension. Plusieurs alternatives ont été étudiées dans la littérature [26, 27]. On pourra mentionner, la topologie en pont complet avec comme avantage de générer une tension négative et limiter les courants de court-circuit. Ces SMs peuvent être contrôlés indépendamment les uns des autres. A l'aide d'une stratégie de modulation adéquate, il est possible de gérer le nombre de SMs insérés ou bypassés dans chaque demi-bras de telle manière à contrôler la tension générée mais aussi l'énergie stockée dans ces derniers à tout instant. Par suite, il est possible de réguler la tension de sortie, $V_{ac}$. Ce concept offre un nouveau degré de liberté à la fois en termes de contrôle, de modularité et de fiabilité.

Finalement, un bras du convertisseur MMC peut être considéré comme un générateur de tension continu autonome d'un point de vu du réseau DC. Une disparité de tension entre bras entraîne l'apparition d'un courant dit de circulation. Une inductance de bras, $L$ est ainsi connecté en série au sein de chaque demi-bras pour réduire ces courants et par ailleurs les courants de court-circuit lors d'un défaut.

### 2.2.2 Principe de fonctionnement

Un MMC permet de synthétiser une tension de sortie AC quasi-sinusoïdale en variant le pourcentage de SMs connectés au sein des demi-bras. En opposition, pour conserver une tension DC à une valeur fixe, la proportion de SMs connectés dans un bras doit rester constante. En considérant $n$ le nombre de SMs insérés dans le demi-bras inférieur, on aura à tout instant N-$n$ SMs actifs dans la partie supérieure du bras pour un MMC de N+1 niveaux.

En appliquant la loi des nœuds et la loi des mailles, il est possible de définir les relations suivantes au sein des demi-bras en régime établi, avec $\varphi$ le déphasage entre le courant et la tension du réseau. On considéra une unique phase sachant qu'un MMC triphasé correspond à l'association de trois systèmes monophasés identiques déphasés de 120 $^\circ$.

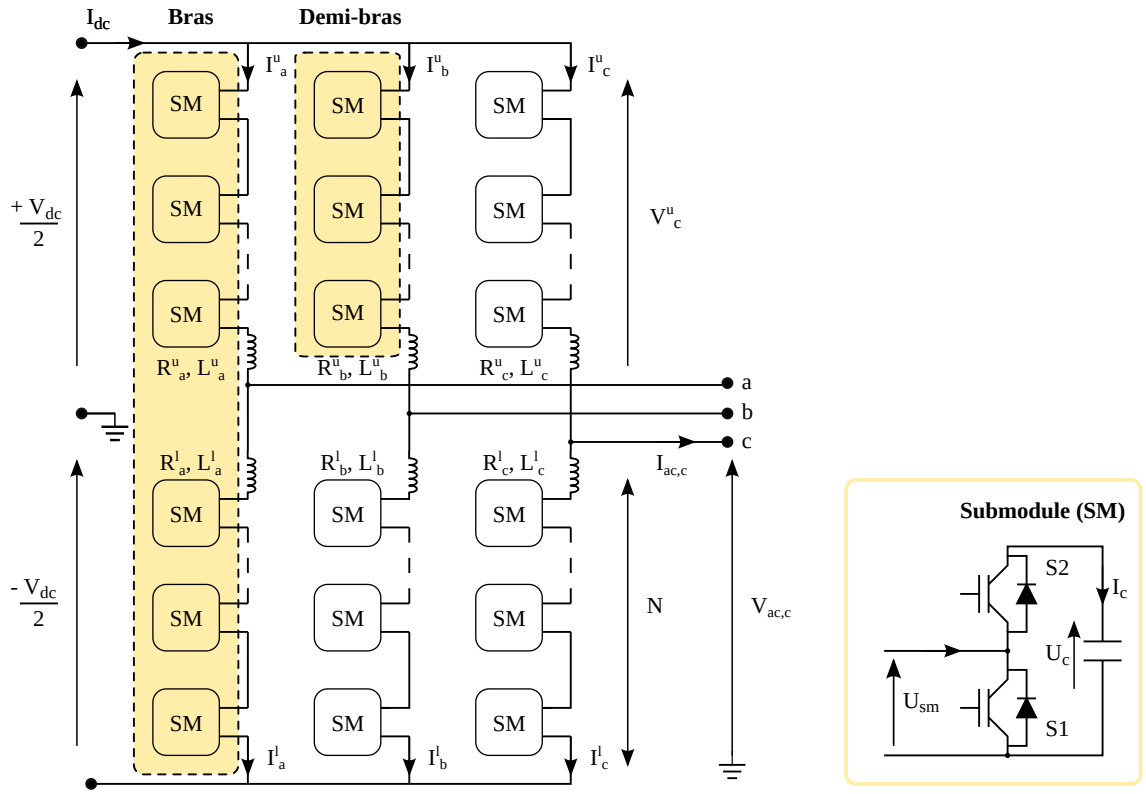



FIGURE 2.1 – Structure d'un convertisseur modulaire multi-niveaux (MMC)

$$i_a^u(t) = \frac{I_{dc}}{3} + \frac{i_{ac}\sqrt{2}sin(\omega t - \varphi)}{2} \qquad i_a^l(t) = \frac{I_{dc}}{3} - \frac{i_{ac}\sqrt{2}sin(\omega t - \varphi)}{2} \tag{2.1}$$

$$v_a^u(t) = \frac{V_{dc}}{2} - \frac{v_{ac}\sqrt{2}sin(\omega t)}{2} \qquad v_a^l(t) = \frac{V_{dc}}{2} + \frac{v_{ac}\sqrt{2}sin(\omega t)}{2} \tag{2.2}$$

Au regard de (2.1), (2.2), chaque demi-bras conduit une composante continue en provenance du réseau DC également distribuée, $\frac{I_{dc}}{3}$. A cela s'ajoute une composante alternative issue du réseau AC, équitablement répartie entre la partie supérieure et inférieure, $\frac{i_{ac}(t)}{2}$.

Finalement, une troisième composante est présente. Il s'agit d'un courant circulatoire entre bras résultant d'une différence de tension comme évoqué précédemment. Ce dernier est interne au convertisseur et n'affecte en aucun cas le système extérieur. Il comporte aussi une composante continue et alternative venant s'ajouter à (2.1), avec $\psi$ le déphasage associé.

$$i_{circ}(t) = i_{circ_{dc}} + i_{circ_{ac}}\sqrt{2}sin(\omega t + \psi) \tag{2.3}$$

### 2.2.3 Le contrôle de l'énergie, un nouveau degré de liberté

Les convertisseurs MMC se distinguent des topologies classiques par la redondance de cellules de stockage indépendantes de faible capacité pour garantir sa fonctionnalité. Toutefois, leur niveau d'énergie se doit d'être régulé sous peine d'entraîner une panne des composants des SMs ou un déséquilibre de puissance entre les réseaux AC, DC de part et d'autre du convertisseur. On peut introduire la dynamique interne du convertisseur au travers des puissances transitant au sein d'un demi-bras. En considérant le demi-bras supérieur, la puissance instantanée reçue s'écrit :

$$P_a^u(t) = v_a^u(t) i_a^u(t) \tag{2.4}$$

En utilisant les relations (2.1), (2.2) et sous l'hypothèse d'un régime permanent, la puissance moyenne dans un demi-bras est donnée par :

$$P_a^u = \frac{V_{dc} I_{dc}}{6} - \frac{V_{ac} I_{ac} cos(\varphi)}{4} - \frac{V_{ac} I_{circ_{ac}} cos(\psi)}{2} + \frac{V_{dc} I_{circ_{dc}}}{6} \tag{2.5}$$

Les deux premiers termes de la partie droite de l'équation ci-dessus représentent la différence de puissance entre les réseaux AC et DC y transitant. Elle nécessite d'être nulle pour préserver une marche stable. Les deux derniers termes restants traduisent la puissance échangée avec le demi-bras inférieure de la même phase et les deux autres bras du convertisseur. Ils permettent de réaliser ce que l'on appelle un équilibrage vertical et horizontal de l'énergie capacitive stockée au sein du convertisseur [28].

Il est dorénavant possible de gérer la dispersion de l'énergie totale stockée entre les demi-bras d'un MMC. On comprend de suite l'intérêt de cette structure flexible en matière de contrôle et de stockage distribué [29]. Ce dernier se retrouve accru s'il devient possible d'augmenter la capacité énergétique du convertisseur, en étant capable à la fois d'assurer son activité mais aussi de fournir des services système. Concrètement, cela revient à y ajouter des solutions de stockage de l'énergie en bénéficiant de la modularité du convertisseur via ses SMs. L'objectif de la présente thèse est d'étudier l'intégration d'une fonction de stockage au sein de cette topologie.

# Chapitre 3

# Vers un stockage distribué au sein des convertisseurs MMC

## 3.1 Favoriser l'émergence des réseaux de demain

### 3.1.1 Le besoin en services système rapides

En élargissant leur gamme de services, les MMC deviennent acteurs dans les échanges de flux de puissance. L'ajout d'une solution de stockage permettra de répondre aux désirs de services auxiliaires exprimés par les opérateurs réseaux pour faciliter les échanges et garantir la sécurité des installations. En parallèle, il est possible d'intervenir sur le marché de l'électricité [30]. Cependant, ces mécanismes requièrent des besoins en énergie et en puissance différents. En fonction de l'application désirée, il est primordial de définir les bénéfices d'une telle solution et de sa rentabilité sur le réseau. En effet, il n'apparaît pas viable de dimensionner un stockage, nécessitant une quantité d'énergie considérable, à la fois coûteux et volumineux. Les prestations de nivellement de charge ou d'arbitrage sur le marché, visant à acheter et stocker de l'électricité à bas prix et la revendre aux heures fortes, s'en retrouvent exclus tout comme la fourniture des réserves secondaires et tertiaires.

Dans un second temps, les énergies renouvelables se caractérisent par leur forte intermittence et faible prévisibilité. De part cette spécificité, il est de plus en plus difficile d'ajuster la production à la demande avec le risque de fragiliser l'équilibre du réseau. Par ailleurs, la création de liaisons HVDC se caractérise par l'utilisation d'électronique de puissance comme principale interface. Il s'en suit un découplage inhérent entre les deux réseaux AC, DC. Ces sources ne contribuent plus ou peu à l'inertie du système, rendant ce dernier plus vulnérable en cas de perturbations. Cette dernière est un paramètre essentiel dans le cadre de la stabilité en fréquence des réseaux traditionnels AC. L'inertie se caractérise par la capacité de réponse du réseau face à une variation de fréquence. Aujourd'hui, elle provient majoritairement de l'énergie stockée dans les masses tournantes des générateurs synchrones. La fréquence est utilisée comme un indicateur d'équilibre entre la production et la consommation d'électricité (3.1).

$$\frac{1}{2}J\frac{d\omega^2}{dt} = P_m - P_e \tag{3.1}$$

Si la demande en énergie électrique, $P_e$, du réseau devient trop importante par rapport à la production, $P_m$, la fréquence diminue et inversement. Ce gradient est d'autant plus accentué si l'inertie du système, $J$, est faible. Le risque inhérent est d'atteindre des excursions en fréquence inconciliable avec les exigences de sécurité des référentiels techniques des gestionnaires de réseaux. On caractérise généralement la réponse inertielle d'un système par sa constante, $H$, comme étant le rapport entre l'énergie stockée, $E_{cinétique}$, à vitesse nominale et la puissance nominale du système, $S_n$, (3.2).

$$H = \frac{E_{cinétique}}{S_n} = \frac{\frac{1}{2}J\omega_n^2}{S_n} \tag{3.2}$$

Dans la perspective de répondre au défi de l'insertion massive des énergies durables au profit de sources conventionnelles, de nombreuses études ont placé le réglage en fréquence et la problématique de la perte d'inertie comme axe de développement prioritaire pour assurer la disponibilité et la qualité des futurs réseaux électriques [31, 32, 33, 34]. C'est dans cette optique qu'on étudiera l'utilisation d'un système de stockage au sein d'un convertisseur modulaire multi-niveaux. De part leur contrôle flexible et rapide, comparé aux systèmes de production classiques, ils se retrouvent bien adaptés pour répondre à ce besoin de puissances important sur de courte durée. L'enjeu sera d'imaginer une solution pour réussir cette intégration.

Finalement, il convient de faire l'analogie avec un réseau DC. Dorénavant, le principal baromètre est la tension du bus, $V_{dc}$. Bien que le concept d'inertie n'est plus présent, la robustesse d'un réseau DC de résister à des fluctuations de tension est assurée par l'énergie stockée sous forme électrostatique dans les condensateurs des convertisseurs, $E_{capacitive}$, pour les systèmes VSC. De façon identique, l'équation dynamique qui caractérise un système DC est donnée par (3.3) et la notion de constante d'inertie peut être étendue (3.4).

$$\frac{1}{2}C_{mmc}\frac{dV_{dc}^2}{dt} = P_{ac} - P_{dc} \tag{3.3}$$

$$H_{dc} = \frac{E_{capacitive}}{S_n} = \frac{\frac{1}{2}C_{mmc}V_{dc}^2}{S_n} \tag{3.4}$$

Typiquement dans un réseau AC, la valeur de $H$ pour des générateurs conventionnels varie entre 2 et 10 sec [35] alors que dans un réseau DC, elle est de l'ordre de la dizaine de millisecondes [36]. Ceci s'explique principalement par une quantité d'énergie stockée plus faible. En conséquence, ces derniers sont plus à même à des risques d'instabilités. Ceci confirme de la nécessité de nouveaux services auxilaires dits rapides pour garantir

un fonctionnement stable des futurs systèmes électriques AC/DC. Bien que les besoins restent encore mal définis [37] et les standards en cours d'élaboration notamment côté DC [38].

## 3.2 Principe d'intégration général

Actuellement, la majorité des solutions de stockage stationnaires proposées sont orientées vers des systèmes de stockage, à base de batteries, massifs et volumineux avec un convertisseur dédié [39]. Bâtis sur la mise en série et parallèle de nombreuses cellules à faible tension, cette méthode est sujet à des problèmes de fiabilité et une difficulté dans la gestion des défauts en raison du manque de redondance. L'attractivité de ces travaux réside dans l'opportunité de distribuer ce stockage en plusieurs sous-systèmes standardisés en bénéficiant de l'architecture modulaire des convertisseur MMCs à l'aune notamment de la fiabilité du système. Le principe général d'intégration est représenté, Fig. 3.1.

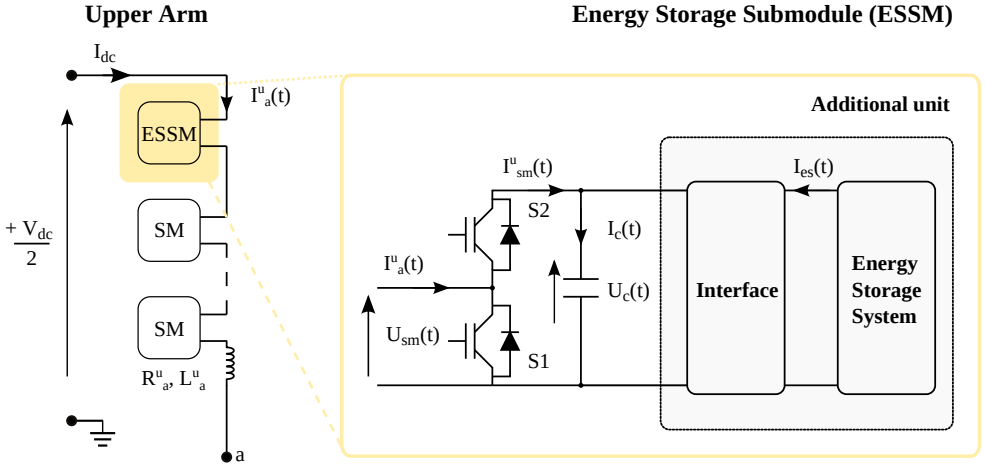



FIGURE 3.1 – Principe général d'intégration d'un système de stockage de l'énergie distribué au sein d'un convertisseur modulaire multi-niveaux (MMC)

## 3.3 Le dimensionnement d'une interface, une nécessité

En théorie, la méthode la plus aisée consisterait à connecter directement l'élément de stockage aux bornes de la capacité d'un SM. Toutefois en fonction de l'application et de la technologie de stockage choisie, plusieurs contraintes de dimensionnement sont à prendre en considérations réduisant le nombre de solutions disponibles. Dans ce contexte, l'insertion d'une interface additionnelle entre le SM et l'unité de stockage peut s'avérer

nécessaire. On pourra mentionner comme principaux arguments :

— *Contrôle imbriqué entre le SM et la solution de stockage de l'énergie*

En mettant en œuvre une connexion directe, la tension aux bornes du stockage de l'énergie est dépendante de celle du condensateur du SM. En raison du mode de fonctionnement du convertisseur, l'équilibrage de ce dernier respecte un processus précis pour garantir l'activité du convertisseur [40]. Conséquence, il devient difficile d'optimiser la profondeur de décharge de l'élément de stockage, limitant les opportunités en termes de contrôle.

— *La présence de composantes oscillatoires basses fréquences*

Comme évoqué précèdemment, les SMs d'un demi-bras sont introduits alternativement afin de reproduire une forme d'onde sinusoïdale. Par ailleurs, le courant dans un bras comporte au moins une composante alternative à la fréquence du réseau et une composante continue (2.1). Inévitablement, le courant circulant au sein d'un SM, $I_{sm}^{u}$, va comprendre des composantes oscillatoires basses fréquences au minimum à la fréquence et au double de la fréquence du réseau avec différentes amplitudes [41, 42]. Avec une connexion directe, ces dernières vont se répartir en fonction de l'impédance de branche des deux éléments avec le risque d'induire des pertes supplémentaires et d'affecter la durée de vie du stockage. Même si le lien de causalité entre oscillations et vieillisement prématuré reste encore flou [43, 44].

— *Un sur-dimensionnement du système de stockage*

Par ailleurs la plupart des technologies de stockage, comme les batteries ou les supercondensateurs, reposent sur des cellules de faible tension [45]. Alors que l'énergie stockée peut être dépendant de la tension, voir une fonction quadratique de la tension [46]. De ce fait, le nombre d'éléments requis pour tenir la tension imposée par la capacité d'un SM, de l'ordre du kilovolt [25], est conséquent avec la potentialité d'être surdimensionné en énergie. Finalement, la présence de composantes à faible fréquence va entraîner une augmentation de l'intensité du courant au sein du stockage avec le besoin d'augmenter le nombre de composants en parallèle afin de respecter leur limite en courant. Subséquemment, le risque est d'arriver à un système de stockage surdimensionné avec un coût non compétitif.

— *Les problématiques d'isolation*

En dernier lieu, l'ensemble des SMs d'un MMC sont référencés à un potentiel flottant. Cependant, durant chaque commutation, leur potentiel à la terre varie et des courants de fuite élevés peuvent apparaître. Similairement, la maintenance d'appareils à des tensions importantes n'est pas une tâche aisée. C'est pourquoi les problèmatiques d'isolation devront être abordées [47, 48, 49].

Au vue de ces observations, il est nécessaire de découpler le système de stockage pour garantir une intégration distribuée optimale, avançant le besoin d'une interface entre les deux entités.

## 3.4 Etat de l'art sur le stockage de l'énergie au sein des MMC

Dans un premier temps, l'objectif a été de réaliser un état de l'art sur les solutions évoquées dans la littérature ou proposées par les industriels [50, 51] pour pallier aux obstacles cités, Fig. 3.2. Avant d'étendre cette dernière aux exigences de notre application.

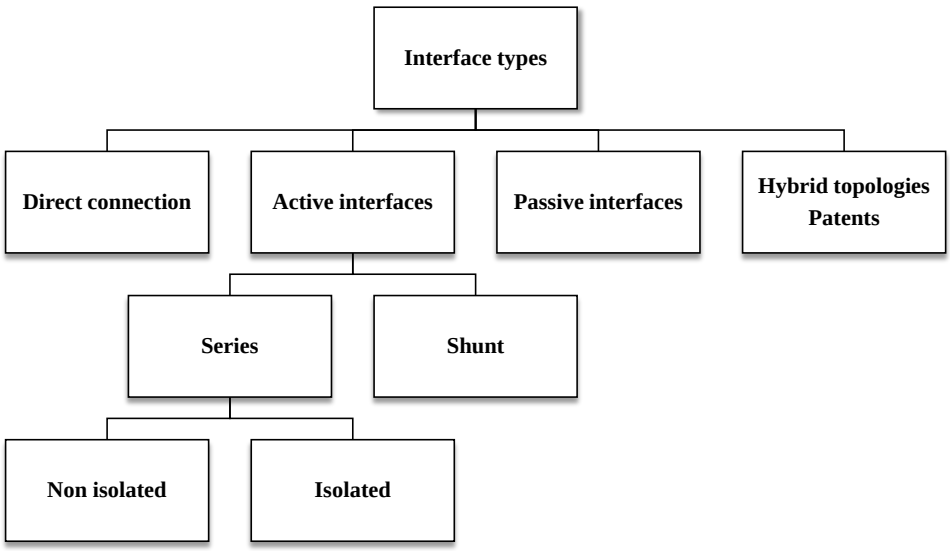



FIGURE 3.2 – Première classification des solutions répertoriées dans la littérature

### 3.4.1 Interfaces passives

Une des premières possibilités est l'utilisation de filtres passifs. La priorité est axée sur la suppression des composantes alternatives basses fréquences au sein du système de stockage. Ces méthodes sont reconnues comme étant simples, robustes et peu coûteuses. Toutefois, la conception d'un filtre peut vite s'avèrer fastidieux, volumineux, et induire des phénomènes de résonances non désirés. Les structures principalement utilisées sont des filtres passe-bas [52] ou résonants dimensionnés à une ou plusieurs fréquences spécifiques [53]. Bien que l"association des deux n'est pas à exclure [54]. Une dernière alternative est la combinaison avec des stratégies de contrôle afin de réduire les besoins en composants passifs [53]. En dernier lieu, ces topologies ne permettent pas de s'affranchir de la dépendance à la tension d'un SM et perdent de leur efficacité lorsque les paramètres du circuit évoluent. C'est pourquoi, les solutions actives à base de convertisseurs de puissance ont rapidement attirées l'attention.

### 3.4.2 Interfaces actives de type "shunt"

Pour une meilleure compensation des harmoniques générées par la dynamique des SMs, des filtres actifs ont été évoqués [54], Fig.3.3a. Néanmoins, l'intérêt de la présente méthode est aussi de limiter les ondulations de tension aux bornes de la capacité initiale d'un SM, et réduire son encombrement. Cette dernière représentant plus de 50% du volume d'un SM, élément de base d'un MMC [55]. Comme précédemment, le système de stockage ne peut être contrôlé individuellement ce qui constitue un inconvénient majeur.

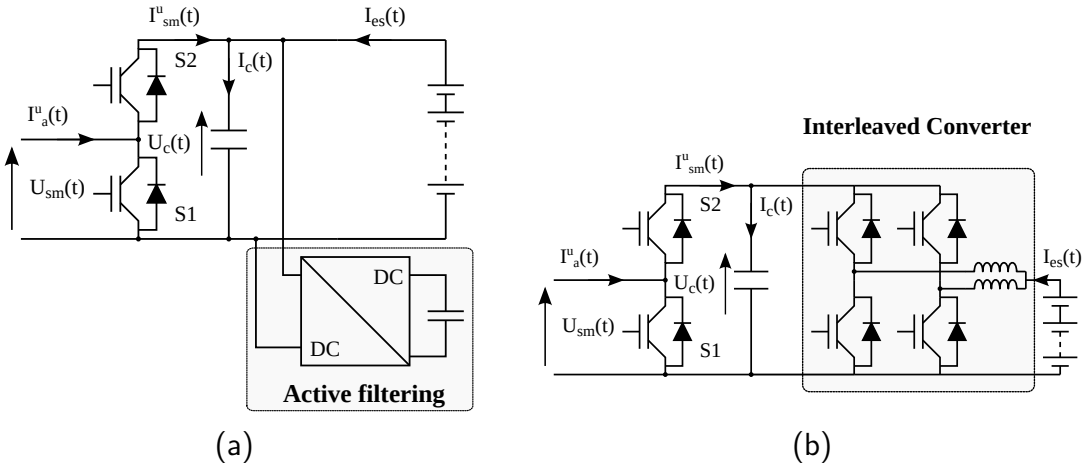



FIGURE 3.3 – (a) Interface active de type "shunt" (b) Interface active de type "série" - Convertisseur buck\boost entrelacé

### 3.4.3 Interfaces actives de type "series"

En implémentant un convertisseur en série, le contrôle des tensions offre un nouveau degré de liberté. A présent, la gestion énergétique entre les deux systèmes est parfaitement découplée et le dimensionnement du stockage s'en retrouve optimisé. Cependant, cette solution nécessite au minimum deux interrupteurs de puissance avec le risque d'induire des pertes supplémentaires ainsi que des problèmes de fiabilité. De plus, la commande se complexifie.

#### Interfaces actives non isolées

De nombreux articles suggèrent une topologie buck/boost classique reconnue pour sa simplicité [47, 54, 56, 57] malgré une inductance pouvant être volumineuse et un rendement limité. D'autant plus lorsqu'il s'agit de travailler avec un important rapport de conversion comme évoqué section 3.3.

Pour pallier à ces contraintes, des industriels ont proposé une topologie buck/boost entrelacée [50, 58]. Elle réside sur la mise en parallèle de cellules bidirectionnelles avec des signaux de commande déphasés, Fig.3.3b. L'objectif est de réduire l'ondulation totale du courant au sein du stockage par rapport à celles dans les inductances grâce au

processus d'entrelacement [59]. L'inconvénient principale est le doublement au minimum des composants nécessaires et la complexité de la commande.

Finalement, l'hypothèse de convertisseurs multi-niveaux a aussi été évoquée [60, 61]. Initialement utilisés pour permettre l'usage de semi-conducteurs avec un faible calibre en répartissant de manière homogène les contraintes en courant, tension sur plusieurs cellules de commutation, ici c'est l'opportunité de réduire le volume de l'inductance qui est recherchée à la faveur d'une fréquence apparente de commutation plus élevée vue par le filtre. Cependant la complexité de la solution est accrue tout comme le nombre de composants fortement multiplié.

**Interfaces actives isolées**

Pour des raisons de sécurité, il peut être nécessaire de disposer d'une isolation galvanique et le choix de convertisseurs DC-DC isolés est requis. La topologie la plus commune est le convertisseur Dual Active Bridge (DAB) introduit par [62], Fig.3.4. Elle se compose de deux onduleurs de tension connectés de part et d'autre d'un transformateur. L'inductance de fuite de ce dernier est utilisée pour le transfert de l'énergie. Cette structure s'est rapidement popularisée car elle fournit trois degrés de liberté additionnels : le rapport cyclique de chaque cellule de commutation et le déphasage entre les commandes. Grâce à cette flexibilité, ce convertisseur possède un assez bon rendement. En contrepartie, un transformateur et huit interrupteurs de puissance sont nécessaires bien qu'il autorise la réalisation de commutations douces pour modérer les pertes. De même, le transfert de puissance peut être limité par l'inductance de fuite.

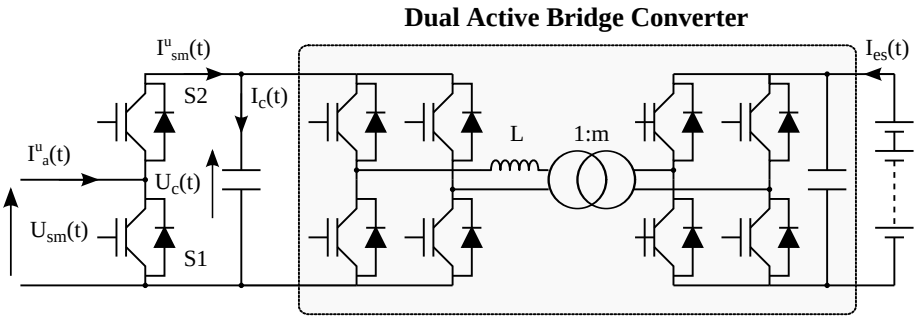



FIGURE 3.4 – Interfaces actives isolées - Dual Active Bridge Converter (DAB)

### 3.4.4 Topologies hybrides

En vue de répondre au défi d'intégration d'une solution de stockage de l'énergie, minimiser le nombre de composants, pallier les écarts haute\basse tension, de nouvelles idées ont été avancées. Dans [63], on propose de bénéficier de l'architecture modulaire du MMC. Un stockage commun à chaque SM situé à une même position au sein de chaque

bras est implémenté, Fig. 3.5. La capacité des trois SMs est reliée à un transformateur via un convertisseur en pont complet à quatre enroulements. La sortie de ce dernier est connectée à l'unité de stockage par un redresseur. L'objectif est à la fois d'obtenir une composante quasi-continue au sein du stockage et de garantir une isolation galvanique.

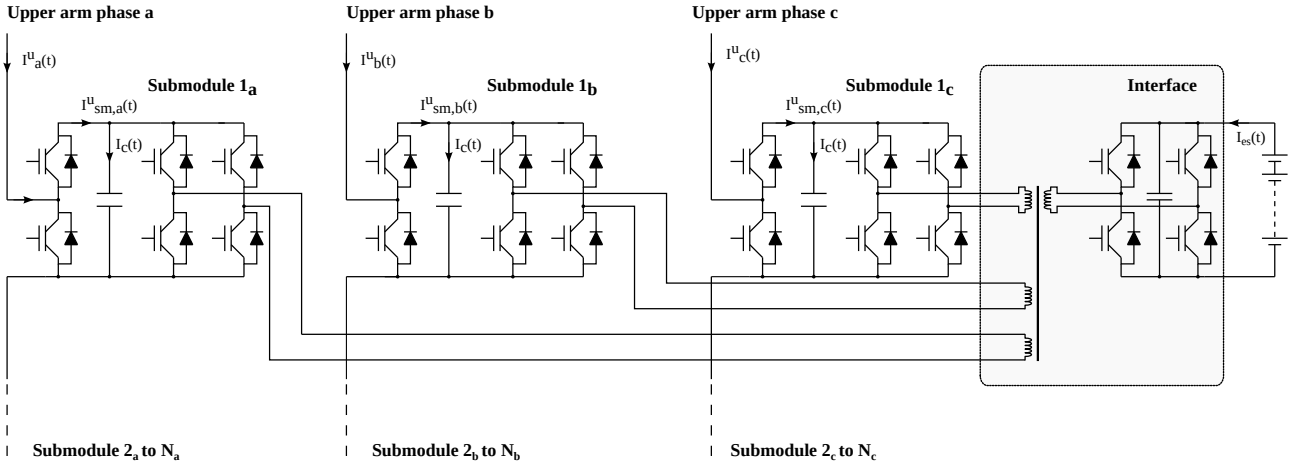



FIGURE 3.5 – Interface avec stockage commun à chaque SM situé à une même position au sein d'un bras [63]

Dans une perspective similaire, [64] a opté pour remplacer une partie des SMs d'un demi-bras par des convertisseurs multi-niveaux FC. Initialement prévu pour diminuer les exigences des capacités des SMs du MMC, sans affecter les formes d'ondes, il a aussi suggéré l'ajout d'un stockage commun aux bornes de ces nouvelles cellules. L'intérêt évoqué est la suppression des composantes basses fréquences grâce à la structure multi-niveaux. Toutefois, cette démarche est limitée aux applications moyenne tension car son efficacité diminue lorsque le nombre de SMs augmente.

Toujours dans l'optique d'éliminer les oscillations basses fréquences intrinséques au MMC, [65] se concentre davantage sur une solution propre au SM. Il envisage d'inclure en amont du système de stockage une source de tension contrôlable en opposition aux ondulations de tension aux bornes de la capacité d'un SM. Ainsi, l'élément de stockage n'est traversé que par une composante continue. Pour cela, il propose d'utiliser un convertisseur AC/DC couplé au secondaire d'un transformateur en série entre la capacité et l'élément de stockage, Fig. 3.6. Bien que cette solution offre de bonnes performances en termes de filtrage, la gestion énergétique du système de stockage n'est de nouveau plus indépendante du comportement du SM.

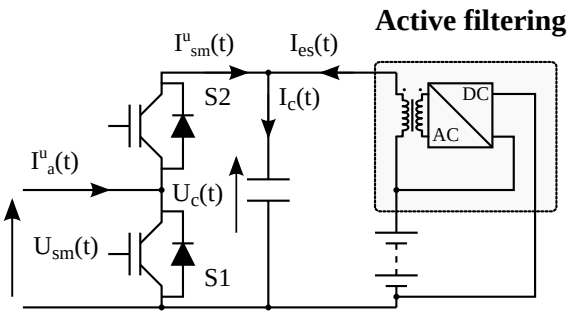



FIGURE 3.6 – Solution d'integration d'un système de stockage de l'énergie à l'aide d'une source de tension contrôlable proposée par [65]

### Bilan

Cette étude a souligné les nombreuses variantes pour intégrer une fonction de stockage. Grâce à la topologie modulaire du MMC, il est possible de distribuer le stockage dans un bras, un demi-bras ou bien dans l'ensemble des SMs. Au sein même de ces modules, il existe différentes possibilités pour réussir cette association. Il conviendra de définir le critère d'optimalité pour ce dimensionnement qui résultera d'un compromis entre plusieurs paramètres comme la fiabilité, le coût, la taille ou le rendement. Ce dernier dépendra obligatoirement de l'application finale et des services à fournir par le système de stockage.

Les premières études de ces derniers mois ont montré de la nécessité d'utiliser une solution basée sur un convertisseur DC-DC pour permettre une intégration optimale du stockage de l'énergie et de s'affranchir des contraintes de fonctionnement du convertisseur MMC. On pourra mentionner la réduction des ondulations de courant ou l'association d'une source haute tension avec un système basse tension. Néanmoins, des problématiques comme l'efficacité, la compacité et le coût devront être résolus tout comme la position idéale au sein du MMC. Une réflexion approfondie sur les convertisseurs adaptés à notre application est en cours de réalisation dans l'optique de proposer une topologie pour la suite de notre étude.

# Chapitre 4

# Conclusion et perspectives

## Conclusion

Ces premiers mois de thèse ont permis de rappeler les enjeux des réseaux électriques de demain et de l'intérêt économique, technique du transport de l'énergie électrique en HVDC notamment pour faciliter l'intégration des ressources renouvelables. On s'est attaché à étudier le fonctionnement et les différents constituants de cette technologie. Cette dernière est devenue attractive grâce au progrès de l'électronique de puissance, incarnée par les convertisseurs multi-niveaux. Bien que plusieurs topologies existent, les convertisseurs modulaires multi-niveaux (MMCs) semblent s'imposer comme la technologie privilégiée dans la perspective de réseaux maillés. On a ensuite analysé la structure complexe du MMC. L'étude de sa dynamique interne a permis de comprendre l'avantage de commander la distribution de l'énergie au sein du convertisseur. Initialement prévue pour garantir sa fonctionnalité, l'un des défis majeurs sera d'augmenter la capacité énergétique du convertisseur pour pallier aussi aux besoins en services système du réseau et devenir dorénavant acteur dans les échanges de flux de puissance. Dans cette thèse, on se propose d'étudier l'intégration de systèmes de stockage de l'énergie au sein de convertisseurs modulaires multiniveaux en vue de répondre aux challenges de résilience et de fiabilité des réseaux de transport et d'interconnexion à haute tension.

Dans un second temps, l'objectif est de discerner la valeur apportée par le système de stockage et de sa rentabilité d'un point de vue système. Par ailleurs, la diminution de l'inertie dans les réseaux, liée à l'insertion des énergies renouvelables, et le besoin en réglage en fréquence apparaissent comme un point critique dans le cadre de la sûreté des futurs systèmes électriques. De même, le problème se retrouve amplifié du côté DC avec une tension fortement sensible en cas de faibles perturbations. Pour l'instant, les travaux se sont portés essentiellement sur le dimensionnement d'un système de stockage en vue de pallier à ces problématiques.

Une analyse sur les contraintes d'intégration d'un stockage distribué et une étude bibliographique sur les solutions proposées dans la littérature ont été effectuées. En parallèle, un état de l'art sur les technologies de stockage de l'énergie a été mené. Ces parties ont identifié les verrous scientifiques à la mise en œuvre de ces travaux.

Enfin, une première approche comparative entre les différentes solutions possibles a été réalisée. Elle consistait à déterminer la distribution idéale au sein du convertisseur et l'interface la plus apte à répondre à nos besoins. La comparaison a tenu en compte du dimensionnement du système de stockage, des composants passifs et de la rentabilité des divers topologies au travers d'une évaluation économique. Elle a permis de confirmer de la nécessité d'associer un convertisseur pour faciliter l'intégration de systèmes de stockage d'énergie modulaires. De même, le coût du stockage devient indifférent de sa distribution à l'intérieur du convertisseur. Cependant, l'augmentation de la densité de puissance par sous-module, si le stockage n'est pas distribué de manière homogène, montre que la solution finale sera fortement influencée par le coût des convertisseurs élémentaires implémentés et des technologies de semi-conducteurs utilisées. Leur coût ne devra pas excéder l'économie effectuée sur la réduction du nombre de composants.

# Perspectives

Cette phase de travail a permis d'attester de l'intérêt d'intégrer une fonction de stockage dans les convertisseurs modulaires multi-niveaux en exploitation. Cependant, la réalisation d'un tel système exige un haut niveau de fiabilité et de durée de vie. De même que la rentabilité et les contraintes masses/volumes sont des questions cruciales.

Dans un premier temps, il sera judicieux de confirmer de la pertinence de cette solution en comparaison avec les stockages stationnaires avec convertisseur dédié actuelles. A court terme, le but sera de retenir une topologie pour la suite de notre étude.

En parallèle, l'identification des besoins du réseau mènera à la définition d'un profil de mission caractéristique à fournir par le stockage de l'énergie et à son dimensionnement optimisé. Une simulation de la solution proposée sera mise en œuvre. Cette partie sera suivie d'une validation expérimentale avec la réalisation d'un prototype à échelle réduite incluant le stockage, avec comme objectif d'aboutir à une simulation en temps réel du convertisseur avec stockage de l'énergie dans son environnement.

En dernier lieu, une analyse technico-économique plus poussée se devra d'être réalisée afin de déterminer de la profitabilité et de la viabilité d'une telle fonction.

Un échéancier prévisionnel des différentes tâches à réaliser pendant le déroulement de la thèse est fourni en annexe.

# Publications

Ces travaux de recherche ont déjà fait l'objet d'un premier article [42] présenté au cours de la conférence IEEE International Conference on Industrial Technology (ICIT) 2018 à Lyon. Un second article avec comité de lecture est en cours de réalisation pour cette seconde année. Il s'orientera principalement sur un état de l'art et la possibilité d'introduire des systèmes de stockage de l'énergie au sein des convertisseurs modulaires multi-niveaux. Plusieurs publications sont envisagées en fonction de la solution retenue ainsi que des résultats obtenus au cours des simulations et expérimentations.

# Bibliographie


[1] E. Commision. "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL The Road from Paris : assessing the implications of the Paris Agreement and accompanying the proposal for a Council decision on the signing, on behalf of the European Union, of the Paris agreement adopted under the United Nations Framework Convention on Climate Change". [Online]. Available : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016DC0110

[2] ——. "2020 Climate and Energy Package". [Online]. Available : https://ec.europa.eu/clima/policies/strategies/2020_en

[3] Directive 2001/77/EC of the European Parliament and of the Council of 27 september 2001 on the promotion of electricity produced from renewable energy sources in the internal electricity market. [Online]. Available : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32001L0077

[4] E. Commision. "2030 Climate and Energy Framework". [Online]. Available : https://ec.europa.eu/clima/policies/strategies/2030_en

[5] "Directive 96/92/CE du parlement européen et du conseil du 19 décembre 1996 concernant des règles communes pour le marché intérieur de l'électricité". [Online]. Available : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=celex%3A31996L0092

[6] D. S. Kirschen and G. Strbac, *Fundamentals of power system economics*. John Wiley & Sons, 2004.

[7] E. Commision. "Communication from the commission to the European parliament and the council - *European Energy Security Strategy*". [Online]. Available : https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52014DC0330

[8] ENTSO-E, "TYNDP 2018 Scenario Report - Main report," ENTSO-E, Tech. Rep., 2018.

[9] D. Van Hertem, O. Gomis-Bellmunt, and J. Liang, *HVDC Grids : For Offshore and Supergrid of the Future*, ser. IEEE Press Series on Power Engineering. Wiley, 2016. [Online]. Available : https://books.google.fr/books?id=cfWlCgAAQBAJ

[10] P. Fairley, "DC versus AC : The second war of currents has already begun [in my view]," *IEEE Power and energy magazine*, vol. 10, no. 6, pp. 104–103, 2012.

[11] [Online]. Available : http://www.bestpaths-project.eu/

[12] [Online]. Available : https://www.promotion-offshore.net/

[13] B. K. Bose, "Evaluation of modern power semiconductor devices and future trends of converters," *IEEE Transactions on Industry Applications*, vol. 28, no. 2, pp. 403–413, Mar 1992.

[14] U. Axelsson, A. Holm, C. Liljegren, K. Eriksson, and L. Weimers, "Gotland HVDC Light transmission-world's first commercial small scale DC transmission," in *Cired Conference*, vol. 32, 1999.

[15] L. Patricia, S. Silvia, and G. Sylvain, "New french-spanish vsc link [c/cd]," *CIGRE Session : International Council on Large Electric Systems. Paris, France : CIGRE*, pp. 1–14, 2012.

[16] B4.CIGRE. DolWin1 HVDC System. [Online]. Available : http://b4.cigre.org/Publications/Other-Documents/Compendium-of-all-HVDC-projects/DolWin1-HVDC-system

[17] ——. BorWin1 HVDC System. [Online]. Available : http://b4.cigre.org/Publications/Other-Documents/Compendium-of-all-HVDC-projects/BorWin-1-HVDC-system

[18] G. Bathurst and P. Bordignan, "Delivery of the nan'ao multi-terminal vsc-hvdc system," in *11th IET International Conference on AC and DC Power Transmission*, Feb 2015, pp. 1–6.

[19] G. Tang, Z. He, H. Pang, X. Huang, and X. p. Zhang, "Basic topology and key devices of the five-terminal dc grid," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 2, pp. 22–35, June 2015.

[20] O. E. Oni, I. E. Davidson, and K. N. I. Mbangula, "A review of lcc-hvdc and vsc-hvdc technologies and applications," in *2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC)*, June 2016, pp. 1–7.

[21] J. Rodriguez, J.-S. Lai, and F. Z. Peng, "Multilevel inverters : a survey of topologies, controls, and applications," *IEEE Transactions on Industrial Electronics*, vol. 49, no. 4, pp. 724–738, Aug 2002.

[22] H. Akagi, "Multilevel Converters : Fundamental Circuits and Systems," *Proceedings of the IEEE*, vol. 105, no. 11, pp. 2048–2065, Nov 2017.

[23] A. Lesnicar and R. Marquardt, "An innovative modular multilevel converter topology suitable for a wide power range," in *Power Tech Conference Proceedings, 2003 IEEE Bologna*, vol. 3. IEEE, 2003, pp. 6–pp.

[24] R. Marquardt, "Current rectification circuit for voltage source inverters with separate energy stores replaces phase blocks with energy storing capacitors," *German Patent (DE10103031A1)*, vol. 25, 2002.

[25] J. Peralta, H. Saad, S. Dennetière, J. Mahseredjian, and S. Nguefeu, "Detailed and averaged models for a 401-level MMC–HVDC system," *IEEE Transactions on Power Delivery*, vol. 27, no. 3, pp. 1501–1508, 2012.

[26] P. Ladoux, N. Serbia, P. Marino, and L. Rubino, "Comparative study of variant topologies for MMC," in *Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 2014 International Symposium on.* IEEE, 2014, pp. 659–664.

[27] A. Nami, J. Liang, F. Dijkhuizen, and G. D. Demetriades, "Modular Multilevel Converters for HVDC Applications : Review on Converter Cells and Functionalities," *IEEE Transactions on Power Electronics*, vol. 30, no. 1, pp. 18–36, Jan 2015.

[28] P. Münch, D. Görges, M. Izák, and S. Liu, "Integrated current control, energy control and energy balancing of Modular Multilevel Converters," in *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, Nov 2010, pp. 150–155.

[29] G. Henke and M.-M. Bakran, "Balancing of modular multilevel converters with unbalanced integration of energy storage devices," *2016 18th European Conference on Power Electronics and Applications (EPE'16 ECCE Europe)*, pp. 1–10, 2016.

[30] A. A. Akhil, G. Huff, A. B. Currier, B. C. Kaun, D. M. Rastler, S. B. Chen, A. L. Cotter, D. T. Bradshaw, and W. D. Gauntlett, *DOE/EPRI 2013 electricity storage handbook in collaboration with NRECA.* Sandia National Laboratories Albuquerque, NM, 2013.

[31] "Deliverable D1.1 Report on systemic issues," MIGRATE – Massive InteGRATion of power Electronic devices, Tech. Rep., 2016.

[32] "System Needs and Product Strategy," National Grid, Tech. Rep., 2017.

[33] ENTSO-E. "Nordic report Future system inertia". [Online]. Available : https://docs.entsoe.eu/id/dataset/nordic-report-future-system-inertia

[34] P. Tielens and D. Van Hertem, "The relevance of inertia in power systems," *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 999–1009, 2016.

[35] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control.* McGraw-hill New York, 1994, vol. 7.

[36] J. Beerten, O. Gomis-Bellmunt, X. Guillaud, J. Rimez, A. van der Meer, and D. V. Hertem, "Modeling and control of HVDC grids : A key challenge for the future power system," in *2014 Power Systems Computation Conference*, Aug 2014, pp. 1–21.

[37] R. H. Renner and D. Van Hertem, "Ancillary services in electric power systems with HVDC grids," *IET Generation, Transmission & Distribution*, vol. 9, no. 11, pp. 1179–1185, 2015.

[38] E. Commision. "Commission Regulation (EU) 2016/1447 of 26 August 2016 establishing a network code on requirements for grid connection of high voltage direct current systems and direct current-connected power park modules". [Online]. Available : https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R1447

[39] [Online]. Available : https://www.energystorageexchange.org/

[40] A. Antonopoulos, L. Angquist, and H.-P. Nee, "On dynamics and voltage control of the modular multilevel converter," in *Power Electronics and Applications, 2009. EPE'09. 13th European Conference on.* IEEE, 2009, pp. 1–10.

[41] K. Ilves, A. Antonopoulos, S. Norrga, and H.-P. Nee, "Steady-state analysis of interaction between harmonic components of arm and line quantities of modular multilevel converters," *IEEE transactions on power electronics*, vol. 27, no. 1, pp. 57–68, 2012.

[42] F. Errigo, P. Venet, L. Chedot, and A. Sari, "Optimal supercapacitor pack sizing for modular multilevel converter with integrated energy storage system," in *2018 IEEE International Conference on Industrial Technology (ICIT)*, Feb 2018, pp. 1760–1766.

[43] R. German, A. Sari, P. Venet, O. Briat, and J.-M. Vinassa, "Study on specific effects of high frequency ripple currents and temperature on supercapacitors ageing," *Microelectronics Reliability*, vol. 55, no. 9-10, pp. 2027–2031, 2015.

[44] S. Bala, T. Tengnér, P. Rosenfeld, and F. Delince, "The effect of low frequency current ripple on the performance of a Lithium Iron Phosphate (lfp) battery energy storage system," in *Energy Conversion Congress and Exposition (ECCE), 2012 IEEE.* IEEE, 2012, pp. 3485–3492.

[45] [Online]. Available : http://www.maxwell.com/products/ultracapacitors

[46] P. Venet, "Le stockage de l'énergie électrique par supercondensateurs : du composant au système," 2017.

[47] I. Trintis, S. Munk-Nielsen, and R. Teodorescu, "A new modular multilevel converter with integrated energy storage," in *IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society.* IEEE, 2011, pp. 1075–1080.

[48] S. Thomas, M. Stieneker, and R. W. De Doncker, "Development of a modular high-power converter system for battery energy storage systems," *EPE Journal*, vol. 23, no. 1, pp. 34–40, 2013.

[49] A. Christe, E. Coulinge, and D. Dujic, "Insulation coordination for a modular multilevel converter prototype," in *2016 18th European Conference on Power Electronics and Applications (EPE'16 ECCE Europe)*, Sept 2016, pp. 1–9.

[50] R. Alvarez, M. Pieschel, H. Gambach, and E. Spahic, "Modular multilevel converter with short-time power intensive electrical energy storage capability," in *2015 IEEE Electrical Power and Energy Conference (EPEC)*, Oct 2015, pp. 131–137.

[51] ABB. Dynapeaq ® Energy Storage System – A UK first. [Online]. Available : http://new.abb.com/facts/references/reference-dynapeaq---a-uk-first

[52] B. Novakovic and A. Nasiri, "Modular Multilevel Converter for Wind Energy Storage Applications," *IEEE Transactions on Industrial Electronics*, 2017.

[53] S. B. Wersland, A. B. Acharya, and L. E. Norum, "Integrating battery into MMC submodule using passive technique," in *2017 IEEE 18th Workshop on Control and Modeling for Power Electronics (COMPEL)*, July 2017, pp. 1–7.

[54] M. Vasiladiotis, "Modular Multilevel Converters with Integrated Split Battery Energy Storage," Ph.D. dissertation, STI, Lausanne, 2014.

[55] Z. Kong, X. Huang, Z. Wang, J. Xiong, and K. Zhang, "Active Power Decoupling for Submodules of a Modular Multilevel Converter," *IEEE Transactions on Power Electronics*, vol. 33, no. 1, pp. 125–136, Jan 2018.

[56] M. Schroeder, S. Henninger, J. Jaeger, A. Raš, H. Rubenbauer, and H. Leu, "Integration of batteries into a modular multilevel converter," in *2013 15th European Conference on Power Electronics and Applications (EPE)*, Sept 2013, pp. 1–12.

[57] T. Soong, "Modular Multilevel Converters with Integrated Energy Storage," phdthesis, University of Toronto, 2015.

[58] E. Spahic, S. Letzgus, G. Beck, G. Kuhn, and V. Hild, "Frequency Stabilizer in Transmission Systems," in *CIGRE-IEC 2016 Colloquium on EHV and UHV (AC and DC), Montreal, Canada*, 2016.

[59] P.-W. Lee, Y.-S. Lee, D. K. Cheng, and X.-C. Liu, "Steady-state analysis of an interleaved boost converter with coupled inductors," *IEEE Transactions on Industrial Electronics*, vol. 47, no. 4, pp. 787–795, 2000.

[60] M. Stojadinovic and J. Biela, "Comparison of high power non-isolated multilevel dc-dc converters for medium-voltage battery storage applications," in *Power Electronics and Applications (EPE'15 ECCE-Europe), 2015 17th European Conference on*. IEEE, 2015, pp. 1–10.

[61] A. Hillers, "Power Electronic Converter Systems for Modular Energy Storage Based on Split Batteries," Swiss Federal Office of Energy SFOE, Tech. Rep., 2016.

[62] R. W. A. A. D. Doncker, D. M. Divan, and M. H. Kheraluwala, "A three-phase soft-switched high-power-density DC/DC converter for high-power applications," *IEEE Transactions on Industry Applications*, vol. 27, no. 1, pp. 63–73, Jan 1991.

[63] L. J. Garces, R. Zhou, Z. Zhou, and D. Zhang, "System and method for integrating energy storage into modular power converter," Jun. 8 2017, US Patent App. 14/960,729.

[64] G. Konstantinou, J. Pou, D. Pagano, and S. Ceballos, "A hybrid modular multilevel converter with partial embedded energy storage," *Energies*, vol. 9, no. 12, p. 1012, 2016.

[65] T. Tengner and R. Alves, "Battery energy storage and power system," Aug. 23 2016, US Patent 9,425,681.

# Annexe A

# Calendrier prévisionnel

Phase 1
Phase 2
Phase 3
Phase 4
Phase 5
Phase 6
S1 Octobre 17 - Mars 18
S2 Avril 18 - Septembre 18
S3 Octobre 18- Mars 19
S4 Avril 19- Septembre 19
S5 Octobre 19- Mars 20
S6 Avril 20- Octobre 20
Oct Nov Dec Jan Fev Mar Av Mai Jn Jui Aou Sep Oct Nov Dec Jan Fev Mar Av Mai Jn Jui Aou Sep Oct Nov Dec Jan Fev Mar Av Mai Jn Jui Aou Sep
Phase 1 - Phase 2
Etat de l'art sur les convertisseurs pour réseau HVAC/HVDC
Etat de l'art sur le stockage de l'énergie
Etat de l'art sur l'introduction de solution de stockage
Etat de l'art sur le management de l'énergie
Définition du besoin
1ère approche - Methodologie de dimensionnement
Sélection des topologies les plus adaptées
1ère Etude - Rentabilité du système sur le marché de l'électricité
Etude des topologies choisies
Mise en œuvre d'une simulation
Proposition d'une topologie
Phase 3 - Phase 4
Etude détaillée de la topologie et de son interfce associée
Dimensionnement optimale du système de stockage
Etude du contrôle du système de stockage et de son interface associée
Simulation hors ligne
Etablissement du cahier des charges pour une version réduite
Dimensionnement du prototype
Réalisation du prototype
Phase 5 - Phase 6
Essais de la maquette
Simulation temps réel
Etude technico-économique approfondie
Tâches annexes
Rédaction du manuscrit
Prévisionnel
Réel

Université de Lyon
CNRS, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1

# Laboratoire Ampère

**Unité Mixte de Recherche du CNRS - UMR 5005**

Génie Electrique, Automatique, Bio-ingénierie

*Mémoire doctorant 1ère année*

*2017 -2018*

| | |
|---|---|
| Nom - Prénom | **GERAMIRAD-Hadiseh** |
| email | **Hadiseh.geramirad@supergrid-institute.com** |
| Titre de la thèse | **EMC study of gate drive for 3.3kV SiC MOSFET** |
| Directeur de thèse | **Christian Vollaire** |
| Co- encadrants | **Florent MOREL** |
| Dpt. de rattachement | **EE** |
| Date début des travaux | **01/10/2018** |
| Type de financement | **CIFRE-Super Grid Institute** |

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

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

## Abstract :

An approach for EMC study of gate drive in order to reduce the $di/dt$ and $dv/dt$ is introduced in this document. In this work it is trying to introduce a new approach that can increase the efficiency of power converter. The main aim of this thesis is to find a cost effective solution which can be applied to different power module with different specification along with solving the EMI problem. The results shown in this report show that by acting in limited number of elements in gate drive circuit, there is a reduction of common mode current.

## Contents

Table of Figures:

# 1. Background

Electromagnetic interference (EMI) is defined as any type of electromagnetic disruption that disturbs, degrades, changes, and limits the proper performance of electronic or electrical equipment [1]. Meeting EMI standards is known as an Electromagnetic compatibility (EMC) study. Proper function of electronic devices and equipment in the presence of the other electronic devices is the electromagnetic capability.

The new generation of switched mode power conversion systems generate EMI noise at each switching transient due to low capacitance of modern power semiconductors. This low capacitance makes the dream of power converter designers true to switch as fast as possible. This parasitic capacitance with the help of parasitic inductance of source limits the transients of power semiconductor.

Keeping in mind above information about EMI and power converter, in power converter gate drivers serves as an interface between power dies and microcontroller. Gate drivers convert and amplify the controller signal to drive the power semiconductor device and has its own role to protect the device against over shoot of current and surge voltage. Therefore, as an early stage design consideration, EMC problem should be issued and studied which can optimize the successful performance of the power converter.

The new generation of power semiconductor for fast switching and the need of new design of gate drive to control this type of power semiconductor are the main mindsets of this thesis. Emphasis is placed on the analysis of the commutation cell of the switching converter to show that it is possible to control the EMI of the device by handling some parameters of the driver.

# 2. General introduction

Technological advances in power electronics such as developments in SiC power devices cause relevant EMC challenges. Wide band gap material SiC now has great potential to replace Si as the dominant transistor topology [2] and in light of recent technology advances, commercial production of SiC MOSFET is feasible [3]. Low switching losses, reduction of cost are the main basis of replacing IGBTs by SiC MOSFETs, a comprehensive comparison showed the superior qualities of SiC MOSFET in [4].

There are some significant differences between SiC MOSFETs and IGBTs in a same range that act on gate drive design [2]. SiC MOSFETs typically have lower switching times (faster) and due to this fast switching they will reduce the switching losses in converters. The other difference due to the different physics of SiC MOSFETs and IGBTs is junction temperature. SiC MOSFET is tend to have less losses due to lower junction temperature than IGBTs which affect the efficiency of this device compare to IGBTs.

Si IGBTS normally are driving between -15 to +15 Voltages and SiC MOSFETS gate voltages are varying between -6 to +20 voltages [5]. Driving SiC MOSFET in 20 volts minimizes switching losses and improves the surge current for SiC devices. However current turn-off in SiC MOSFET is not comparable with the tail current turn-

off in IGBT since IGBT has its own snubbing during the transition. A higher ringing in combination with lower gate source voltage threshold make a great need of attention to MOSFET turn-off ringing current.

Using the advantage of fast switching ability of SiC MOSFET coming with aggregating EMI. So alongside of a dramatically increasing usage of SiC MOSFET, early stage works on SiC MOSFET gate drive must be done to meet the EMC challenges in front it.

## 3. Thesis objectives for the first year

Based on the aforementioned information, the objectives of first year are as follow:

- Provide quantitative understanding of switching behavior of SiC MOSFET.
- Overview of the current solutions to mitigate EMC problems with gate driver for SiC MOSFET.
- Choose a method that could reduce the EMI and first attempts to apply this method.
- To understand and improve the layout of current gate drive (Super Grid Institute prototype)
- Observe the level of the current overshoot and surge voltage in SiC MOSFET1.7 Cree Wolfspeed application
- Extract the parasitic element in layout of the power converter (BUCK converter → test platform of the SiC MOSFET)
- Model a equivalent circuit of test bench, as a first step it is planned to find a simple model which can provide a trustable behavior of device under test and complete the complexity of whole circuit by measuring the parasitic element that can affect directly the switching behavior.

## 4. Problem description and current solutions

MOSFETs usage as a switching part of switched mode power converter inherently increases ringing more due to the fact that there is no tail current which provides a great amount of damping as in IGBT gate drive circuits [5]. The analysis [6] has shown that in power switching mode applications based on SiC devices several issues such as oscillation which provoke electromagnetic interference (EMI) and overshoot require analysis and efficiency studies for achieving the maximum performance and extend the life of the devices. Considering the proper layout is a good practice to reduce the EMI in gate drive, control and minimizing ringing can be done to the acceptable level with the help of gate drive. Therefore, EMC problem addressed in this thesis are:

- Increased current and voltage overshoot due to the use of SiC devices
- The need of study of complicated coupling path and the role of parasitic elements in the layout of power converter

Therefore, it is trying to:

- Drive SiC MOSFET with less high frequency noise without sacrificing speed operation capability of SiC MOSFET. This clean switching should be done with less additional components to the converter in order to keep the cost effectiveness constraint.

## 5. Voltage and current over shoot in clamped inductive circuit

To investigate the EMI, it is meaningful to study the switching behavior. A double pulse test set up is a typical clamp inductive circuit to characterize the switching behavior of MOSFETs. This switched mode power converter with a half bridge configuration is shown in Figure 1.

The EMI noise caused by the hard switching are produced by parasitic elements like stray inductances which is in series with the device under test and the parasitic capacitances of the magnetic elements and parasitic capacitances of the ground connection. Those parasitic capacitances are in parallel with the parasitic capacitances of the MOSFET. Figure 1.b is presenting them. These parasitic elements produce voltage and current oscillation when $di/dt$ and $dv/dt$ expose to them.

Low side MOSFET Q_L is the controlled switch and the MOSFET Q_H is the freewheeling switch. During the turn on of the Q_L, the high side MOSFET is always turned off and its body diode conducts the load current. Switching behavior of the SiC MOSFET can be explained by charging and discharging of internal capacitors.

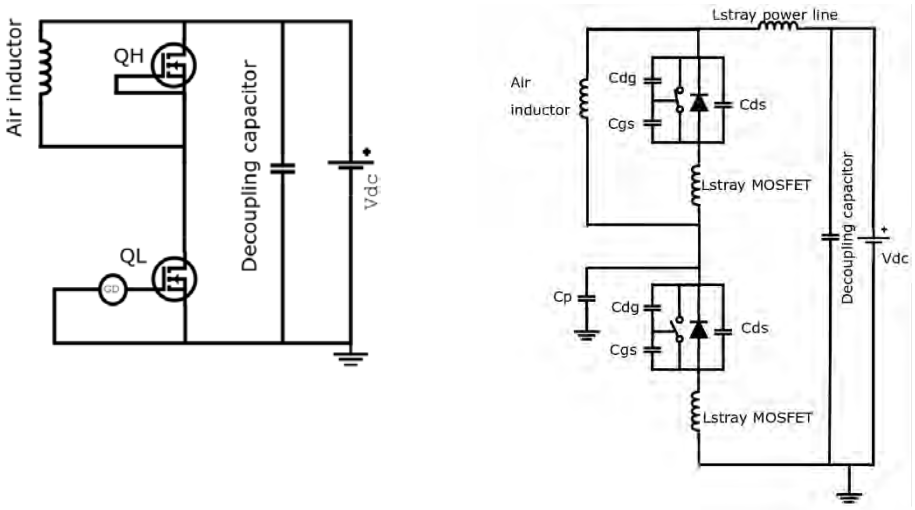



*Figure 1: Clamp inductive circuit (a)double pulse test (b)double pulse test considering parasitic elements*

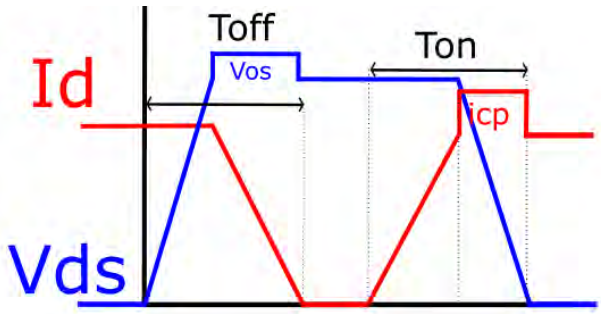



*Figure 2: Current and voltage spikes in hard switching*

In turn off process the drain source voltage $V_{ds}$ increases to maximum switching voltage, then the drain current $i_d$ starts reduce to zero and therefore the diode current which is the current through the stray inductors in the circuit start to increase the $V_{Dos}$, see (Eq.1).

$$V_{Dos} = L_{stray} \frac{di_{Lstray}}{dt} \quad \text{(Eq.1)}$$

This voltage will be added to the switch voltage as it is illustrated in
Figure 2. On the other hand, when the switch turns on, the switch current increases firstly. The switch voltage starts to decrease and $\frac{di}{dt}$ is exposed to parasitic capacitances in parallel with diode and the parasitic capacitances to the ground. This current can be calculated by (Eq.2)

$$i_{C_p} = C_p \frac{dv_{c_p}}{dt} \quad \text{(Eq.2)}$$

This current will be added to the switching current and will make the current spikes as it is shown in Figure 2.

## 6. Recent research to overcome EMI problem for MOSFET

1. Increasing the gate resistance

Increasing the gate resistance is the conventional approach to reduce voltage/current spikes by changing charging and discharging time of the capacitors $C_{gs}$and $C_{ds}$. It is widely known that the gate resistor variation of the power devices can control the switching losses behavior and the oscillation. However, a tradeoff between these two parameters is required since increasing the $R_g$ causes a dramatic increase of switching losses. At the given amount of resistance the challenge is to get the moderate oscillation concerning the minimal losses which needs a proper adjustment between efficiency and EMI noise in a power converter [6].

2. Sunbber circuit

Using snubber circuits is another conventional alternative. In fact, in many power converters still using snubber circuits is the solution to damp overshoot of current and voltage oscillation. Although efficient snubber can remove the oscillation but it can particularly reduce the efficiency of the system [7].

3. Resonance circuit

In the last years, new solutions have been presented to reduce the losses while they are controlling the $di/dt$ and $dv/dt$. Different forms of resonance circuit gate drives were presented a few years ago as a solution for Si and SiC MOSFETs at high frequency operation, regarding of the low cost of resonance gate drive. They are efficient solutions in case of losses also.

Two types of resonant gate drive have been presented which resonant voltage/current across a series or parallel resonant circuit named parallel or series resonant gate drive. This technique has been researched extensively for MOSFETs to reduce the needed gate charge at high switching frequencies and usually it is applicable by means of inductor. However, the gate voltage rise time is considerably slowed down by current source gate drive [8], [9].

In series type resonant gate driver circuit with reverse blocking diodes, gate charge is supplied with the help of inductive charging at turn on state (Figure 3). The resonant current (series resonant), charges and discharges $C_{iss}$through the resonant inductor, cause a reduction of the output voltage and the power consumption in the gate drive circuit. It is important to notice that regulating gate source voltage is difficult since it depends on the resistance in resonant circuit.

In parallel resonance (Figure 3), the circuit recovers the energy stored in $C_{iss}$ to the dc power supply, when discharging of $C_{iss}$. The dc power supply for the gate drive circuit has to provide a certain amount of energy during charging period which is bigger than the energy stored in $C_{iss}$. Although the excessive energy is theoretically recovered to the dc power supply during discharging period, it leads to increase the power consumption in a practical implementation [10].

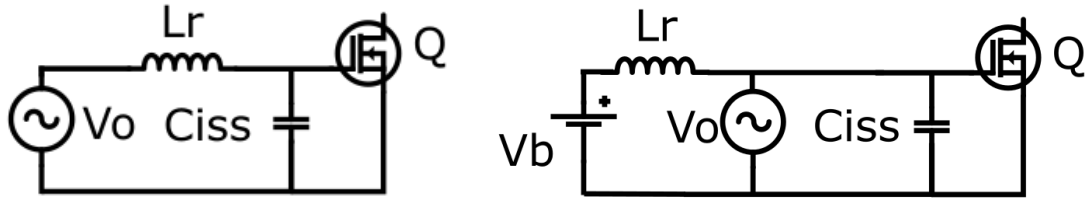



*Figure 3: (a)Series resonance (b)Parallel resonance*

In [11], to overcome the limitation of the losses, they tried to transfer a constant and defined peak gate voltage. It has been done with the help of voltage feedback to the gate current path. It is tried to continuously adjust the energy supplied to the resonant circuit and keep the gate voltage constant and consequently reduced the drain voltage fall time at turn on state which resulted to reduce the losses.

Hideaki Fujita [9] proposes a series resonant gate drive circuit turning the gate charge polarity of the MOSFET by using the resonant current through a series resonant circuit. This circuit consists of MOSFETs and a resonant inductor, and the inductor and the input capacitance in the driven MOSFET form a simple series resonant circuit. The theoretical analysis showed that the proposed circuit makes it possible to reduce the power consumption by a factor of ten, compared with a conventional gate-drive circuit. The experimental results obtained from the case study which is the half bridge inverter illustrated a good conversion efficiency as high as 99%.

Above presented RGD[1] circuit are designed only for specific applications and sometimes may not operate for others. As instance, a certain design approach must be taken for the Synchronous buck converter circuit. Since the circuit has two devices that are switching, the applicable RGD circuits are designed to adapt the need for those type of circuit. Coupled inductors can be used to drive both switches using one magnetic core. Zhiliang Zhang proposed resonant gate drive with two complementary drive signals to drive two MOSFETs in one leg with the adjusted dead time, more simple circuit and cost effective [12]. Proposed gate drive is a dual channel isolated RGD solution suitable for the DC-DC converter which recover the energy of gate with better efficiency compare to conventional resonant gate drive. It can drive two MOSFETs as a dual channel driver and at the end it can drive with negative gate voltage which avoids false triggering. The negative gate drive voltage ensures high reliability of the turn off state to prevent the high ratio of voltage variation in comparison with conventional resonant gate drive. Similar to this isolated dual resonant gate drive is presented in [13].

These drivers are simple solutions and low cost but have a fixed profile and the variable parameters are limited.

4. Active gate drive

Active gate drive has been introduced mostly to control the switching transition of the Si MOSFETs/IGBTs in low frequency [14], [15] which may be adopted to be used in SiC MOSFETs [16], [17].AGD[2] techniques control the switching by controlling the gate circuit of the MOSFET/IGBT and consequently does not require any modification in the original topology of power converter neither in using of passive neither active devices. The principle is to change the slope profile of the gate source voltage. Recent researchers on Active Gate Drive classified them in total to three groups of control: gate voltage control, gate current control, gate resistance control.

- Controlling the gate resistance

  In controlling the gate resistance, the impedance of the charging and discharging path of the gate is increasing during the $di/dt$ oscillation. This increased impedance decrease the $di/dt$ therefore the

[1] Resonant gate drive
[2] Active Gate Drive

voltage overshoot will be reduced. This impedance in the next step will be reduced in turn off state or high $dv/dt$ state to decrease the $dv/dt$ which cause less losses also. In [18] this method has been applied for IGBT with an increase in switching losses in turn on state and successful supersession of $di/dt$. In 2017, Alejandro Paredes, released new result of this method applied to SiC MOSFET [19]. The control loop was made of two switches in parallel for both turn on and turn off state to control the gate resistance with different values. The gate resistance will be changed based on the reference value of $di/dt$ and $dv/dt$ set for the control loop.

- Controlling the gate current

  In AGD by controlling the current, the strategy is to control the gate during the turn off or voltage rise time with maximum current. This maximum exposed gate current contracts the voltage rise time and decreases the losses respectively. As instance, in [20], this method has been applied to IGBT and the authors believe that this method can be applied for SiC MOSFET. In the proposed gate drive for MOSFET, the control is independent of the slopes of the drain voltage and current. It can be done only by shaping the gate current, similar to this research is done for SiC MOSFET devices [21]. To summarize this method briefly it can be said: the goal is to charge the input capacitance as fast as possible which speed up the variation in drain source voltage by supplying gate-current pulses at the beginning of the Miller plateau. Considering in this method the variation of the drain current will not change. This injected current has different methods to be applied to gate drive circuit which can be seen in referenced articles in this section. Although $di/dt$ is controlled but the above mentioned methods were not successful in controlling the $dv/dt$.

- Controlling the gate voltage

  The concept of this method also is similar to other types of AGD. Which are practically can be done in closed loop control. In this control method, by using a voltage generator it is possible to shape the waveform of the gate source voltage. In an iterative control loop, a reference voltage is compared with drain source voltage and act as a lever in the circuit of gate drive to inject the voltage to the gate source voltage. for example, in [22], this method is applied by injecting the transitional gate voltage in two different stages, for turn on and turn off state, although the method was successful to decrease the EMI emissions in buck converter but there is no interest in losses.

  In fact, the negative point in current method is the time consumption, this feedback control takes time to act in the circuit of gate drive. As it is already mentioned in the introduction, the advantage of SiC MOSFET is fast switching and consequence of this method is in contrary with the aim of using SiC MOSFET in power converter as a switching part. This problem has been addressed in [23]. In this research, the difference between IGBT and MOSFET was carefully considered and a delay compensator was provided additionally in feedback control loop to overcome the challenge of small switching transition of SiC MOSFET. Considerable reduction of $di/dt$ can be observed while the challenge of losses still exist in this research.

5. Inductive boost control

Inductive feedback or inductive boost control of gate drive, is a new gate drive EMC study topic which does not have profound discussion yet. At the first overview of this method, Ogata in [24], did a brief study in influence of the mutual inductance between the gate drive loop and power loop. In this work, besides the layout of the power converter, the effect of coupling between control loop and power loop in overshoot of voltage has been discussed precisely. Considering in this article, only the turn off state was debated, the result has been revealed that by negative coupling a reasonable tradeoff between losses and EMI emissions can be achieved. Similar to this concept, in [25] and [26], it is tried to actualized this coupling and take the advantage of it. In these two papers, only a transformer added to the circuit of power converter to transfer energy between the power path and the gate control path while all the typical design of gate drive remains without modification. Results, illustrated that this coupling could compensate the influence of the parasitic elements which are changing the switching behavior. Inductive coupling has been applied to boost the gate current in this article for several switches with different specification and it is shown that although in some cases it could not decrease the overshoot of current and voltage, but keeping the same level of noise it could reduce the losses significantly.

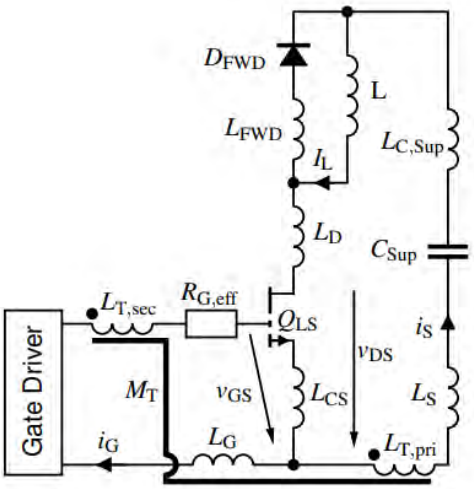



*Figure 4: Double pulse set up with inductive coupling [26]*

## 7. Method going to be applied

By overviewing the current solution on gate drive EMC mitigation, it can be summarized in following lines:

- The conventional method including the gate resistance and resonant gate drive although works in decreasing the overshoot but not for improving the efficiency. The number of parameters that can act on design of gate drive are limited and for each application it should change based on the specification and requirement of the application. For example, changing from buck to synchronous buck converter needs new design of the gate drive.

- In case of closed loop control, changing the gate resistance seems to be effective but this method has the lack of adjustability of $di/dt$ and $dv/dt$ in different operating condition because in the control loop discrete values of resistance should be used.

- Controlling the $di/dt$ and $dv/dt$ in control loop at the same time which can prevent losses to increase, until now has not been applied in SiC MOSFET due to the significant delay time producing by the control loop. This increasing of transition time is not negligible in SiC MOSFET.

- Considering all cited research needs a certain change in a circuit of gate drive and needs separate constraint for turn on and turn off the MOSFET, inductive boost has the advantage of simplicity to modify the circuit. Another advantage is cost consideration in design of the gate drive circuit, by considering that control block and sensor add more expense in closed loop control application, inductive coupling with less expense can be apply.

- The new approach of inductive boost gate drive, needs to be studied extensively to pass the challenge of soft switching, synchronous buck converter, practical installation in power module which make the great motivation to be studied further.

Inductive Boost gate driver operation: The coupling part introduces a certain amount of energy from the source connection due to the variation of the source current to gate drive path during rise and fall times. Depending on the direction of the coupling, this current variation makes positive or negative voltage which result to higher or lower gate current.

Based on the goal of the designer this negative or positive voltage, can make the switching of the MOSFET faster or slower. Figure 5 Illustrates the steps of the inductive feedback in changing the gate source voltage profile. When the gate drive is in on state, gate source voltage starts to increase up to threshold voltage, which charge the input capacitance of the SiC MOSFET. MOSFET starts to turn on, therefore, the source current start to change which create the voltage through the coupling. Depending on the coupling this voltage can be negative or positive. This voltage changes the voltage waveform of the gate source voltage therefore, gate current increase or decrease consequently. In a case that current increases, it boosts the turn on of the MOSFET and when it decreases, it can slow down the switching rate.

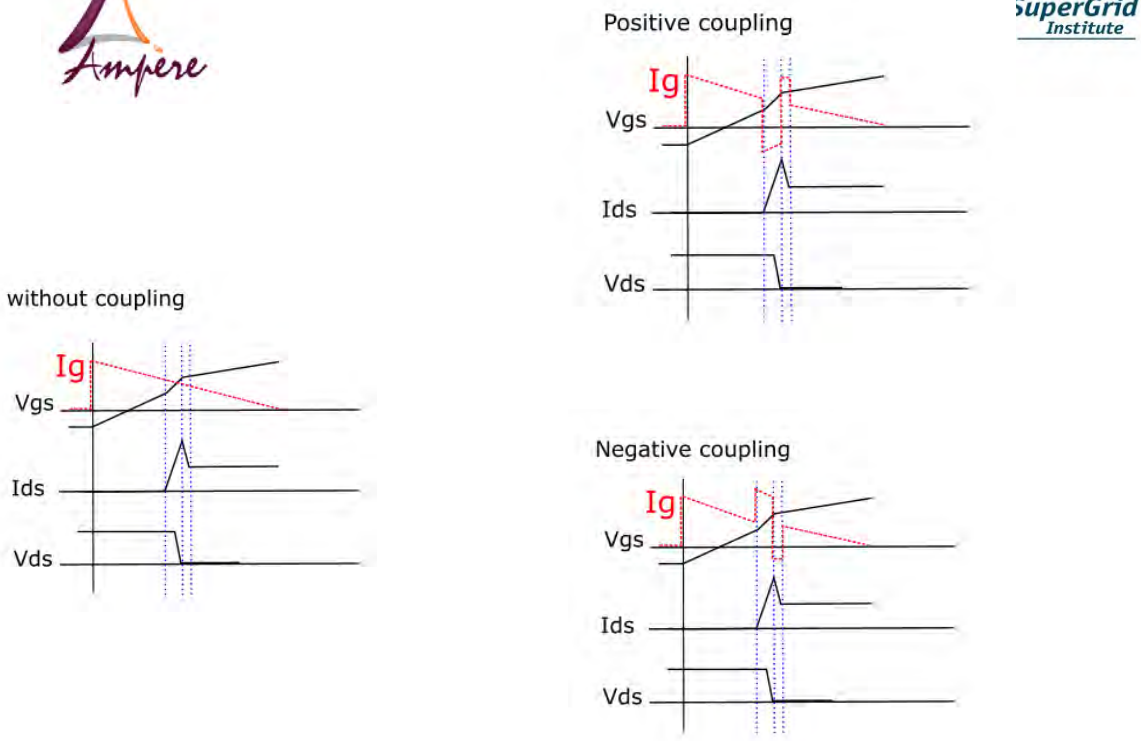



*Figure 5: Inductive feedback operation*

## 8. Double pulse test simulation and experimental result:

The significantly faster switching of the SiC MOSFET module requires a more comprehensive understanding of the effect of the parasitic elements to achieve a successful EMI model. All physical circuit has stray inductance caused by bond wires, board traces. In clamped inductive circuit, the load should be much bigger than stray inductance in order to not make any damping on the resonant circuit. In complicated circuits and geometry, recognizing and measuring these inductances are difficult also. The voltage drop across these inductances is expressed in Eq.3.

$$V = L \times \frac{di}{dt} \qquad \text{(Eq.3)}$$

Considering the "rule of thumb" in EMC design, it is possible to consider $\frac{10nh}{cm}$ for physical connection, to estimate the value of parasitic inductance in test bench where it is difficult to measure, and for the rest of the system, it was tried by impedance meter to estimate the value of the parasitic inductors and capacitances on double pulse experimental platform(see Figure 6).

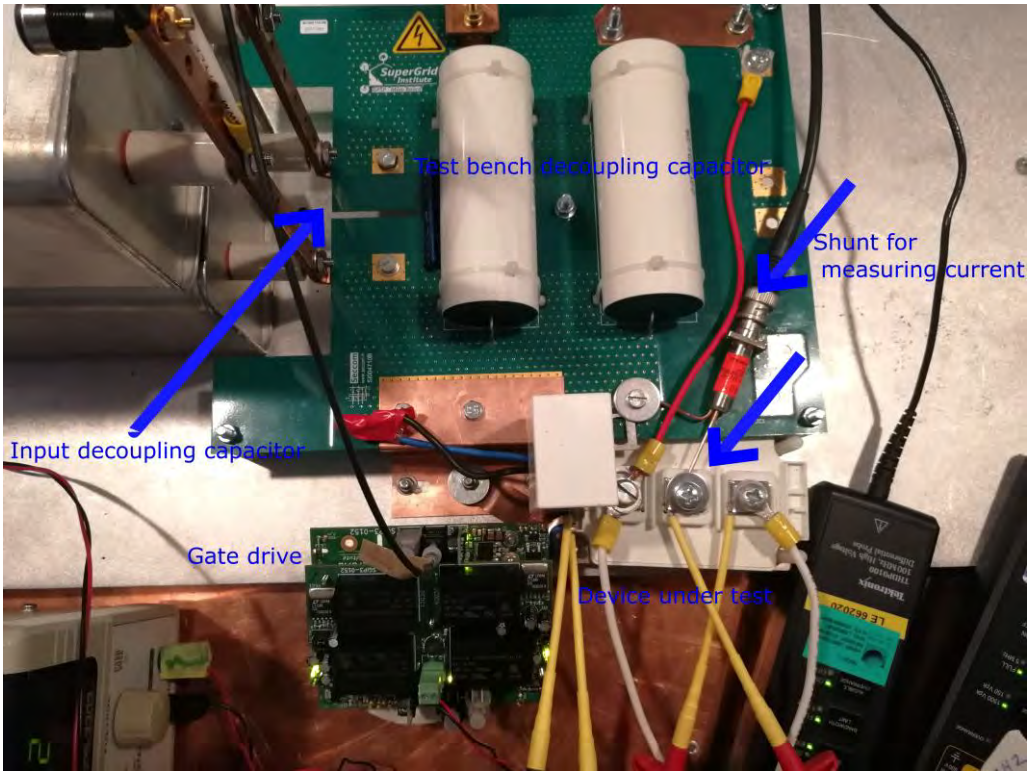



*Figure 6: Double pulse set up*

One of the most critical parameter in SiC MOSFET application is the overshoot of current. Based on the datasheet of MOSFET, it should be ensured during the manipulation that current will not pass its maximum. The overshoot of voltage also is due to the resonant circuit formed by the output capacitance of the module and stray inductance exist between the module and the link capacitance. The voltage overshoot can be seen when the low side MOSFET is on, while the other MOSFET is carrying the freewheeling current (see Figure 7). At the beginning, both MOSFETs are off, the freewheeling current from the load is biased forward through the diode of the high side MOSFET, which cause a small and negative voltage across the load, in fact the magnitude of this voltage is equal to voltage drop of the diode.

Back to the situation when the low side MOSFET is on, the high side MOSFET is bias forwarding the freewheeling current. This causes a short circuit to be formed at the moment of turn-on. As it can be seen in Figure 7 current begins to flow from the link capacitance. The current through the upper diode is (Eq.4):

$$I_D = I_{freewheel} - I_{dclink} \quad \text{(Eq.4)}$$

This current will continue until it reaches this point where $I_{freewheel} = I_{link}$. At this state, the diode become reversed biased and introduces a capacitive load to circuit. This capacitive is the sum of the reverse capacitance of the upper diode and the output capacitance $C_{oss}$of the lower power MOSFET.

While the voltage across the load is increasing, current will flow through the load as it is shown in Figure 8 , the DC link current will separate between the load and DC link which will charge the output capacitance. Based on the above information, resonant circuit constitute of $Coss$ and stray inductance will be made up.

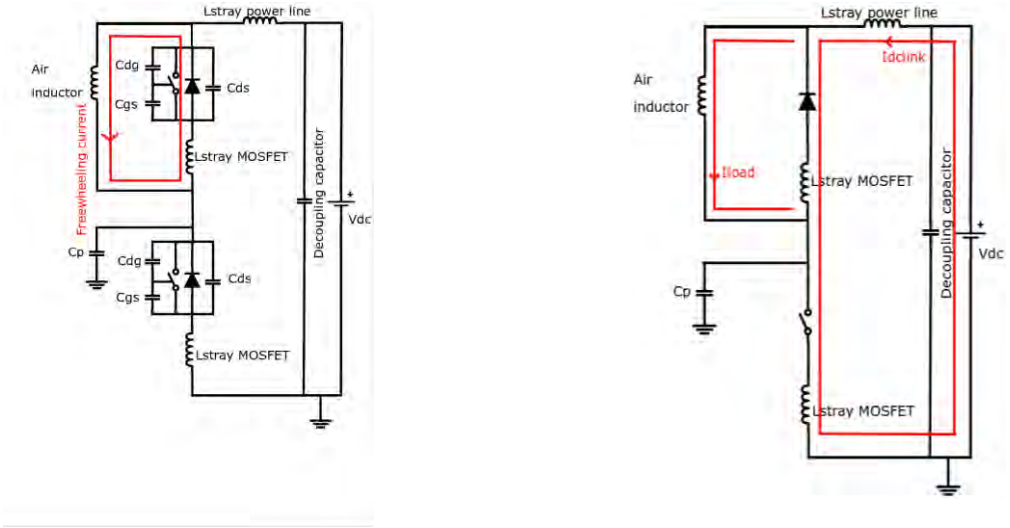



*Figure 7: current in (a) clamp inductive equivalent circuit (b)Freewheeling equivalent circuit*

Besides this stray inductor and parasitic capacitances that make the resonant circuit and respectively the oscillation, there is an internal resistor of power MOSFET, and other resistive part of the circuit. These resistance damped the oscillations and an overshoot across the $C_{oss}$. The resonance frequency is made up of:

$$f_n = \frac{1}{2\pi\sqrt{LC}} \tag{5}$$

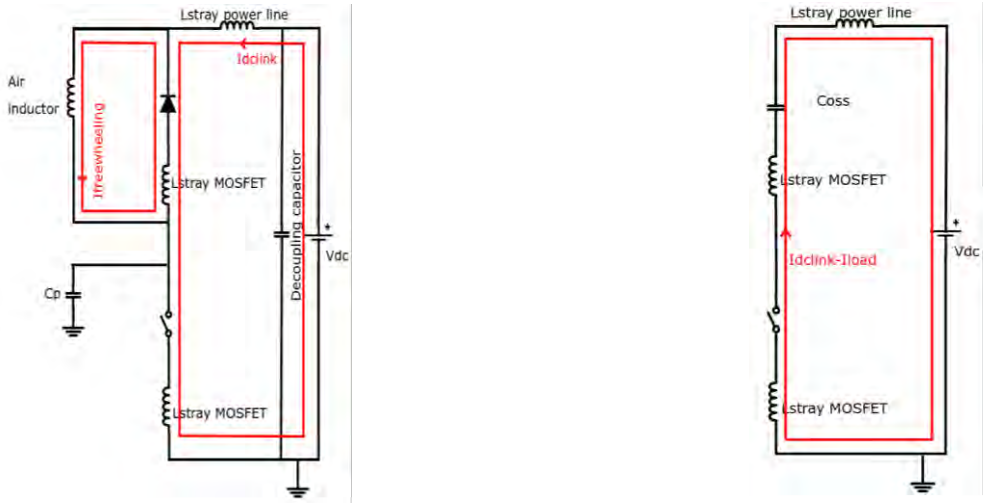



*Figure 8: current path(a) commutation (b) current over shoot*

## 8.1. Double pulse test measurement

Two pulse inductive test has been done on the power module CAS300M17BM2 to evaluate the performance of the test setup. At the first test, it has been done in 1200V input and 64A load current with 0.2 Ω external gate resistance in order to fasten the switching time and observe the switching behavior. The certain amount of ringing can be seen in this measurement. It is possible to reduce the amount of ringing by slowing down the switching speed but this in turn increases the amount of switching loss which is clear by changing the slop of voltage variation in measurement by external gate resistance equal to 5 Ω (Figure 9 and Figure 10).

One of the key advantages of the SiC MOSFET is fast switching speed and it is possible to cancel out this key advantage by slowing the switching speed down too much. Recognizing that there will always be some amount of ringing present therefore an engineering tradeoff needs to be made to ensure that voltage overshoot does not damage the device while keeping the switching speed advantage.

By comparing the waveform resulted from simulation in Figure 10, it can be seen that there is a certain amount of ringing in steady state with the resonance frequency of 12 MHz which differs from the simulation. This difference comes from the following points:

The most probable explanation is that achieving comprehensive reasonable model with all parasitic elements is extremely complex and time-consuming. However, this task can be simplified by creating a circuit that simulates conditions of behavior of MOSFET. The second, at these low inductance values, there is always amount of uncertainty caused by the repeatability of the impedance meter test fixture as well as other factor in impedance meter like calibration. At the end, considering output capacitance of the power module which is typically proportional to the current rating of the power MOSFET, has specific value. Due to nonlinear behavior of $C_{oss}$ which is varying significantly based on the $V_{ds}$, a simple first order analysis is difficult.

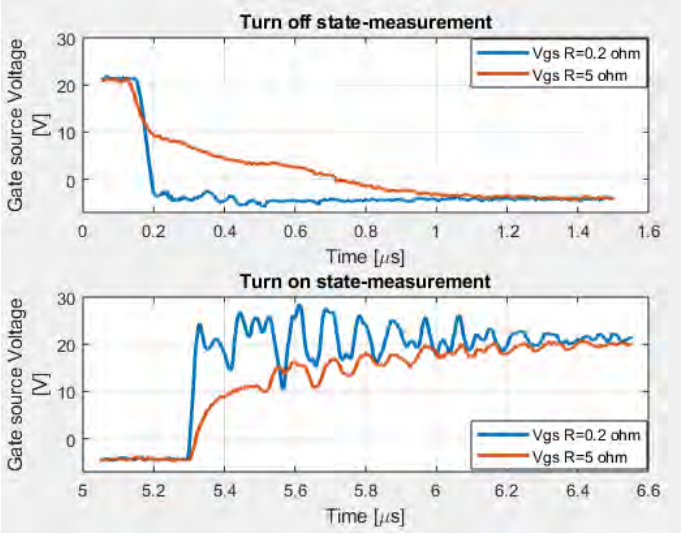



*Figure 9: Gate source voltage profile*

In EMC studies, the trial and error method is one of the well-known method which can help the designer to find a proper model that represent the similar behavior to the real application, in this thesis, this method has been applying until finding the proper model which can be adjusted for the subjected solution.

At the first year of this work, considering the previous EMC studies in SuperGrid institute in DC-DC converter, EMI modeling has started with Ansys simplorer platform. In this software the model of power module has been provided which revealed some inaccuracy that can be improved in following months by comparing with experiments. However, in simulation, finding a reasonable simplified circuit to analyze resonant behavior is difficult to achieve and is a long term goal. Detailed resonant equivalent circuit needs to study energy storage in the module to predict the value of the capacitance in order to have same order of magnitude in rise time and fall time. (Complete comparison is provided in ANNEX1).

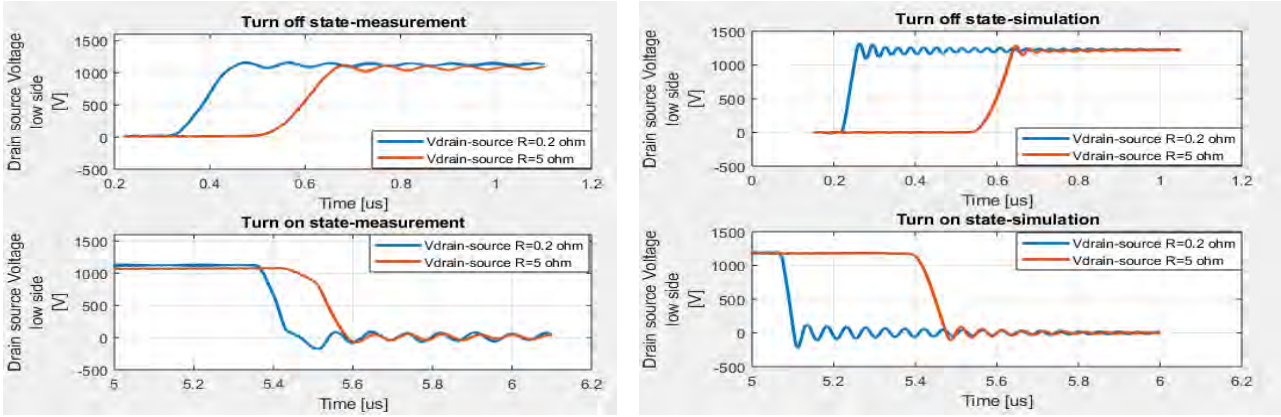



*Figure 10: Drain source voltage*

## 8.2. Inductive coupling:

Considering the model that can show the influence of the parasitic elements and the switching behavior of the MOSFET, first possible method has been applied. As it is explained in section 5, the coupling between the source current and the gate path has been done (Figure 11). This coupling has been done by mutual inductance with absolute inductor of 10nH for the primary and 100nH in secondary side. Considering that two types of direct and indirect coupling has been applied also, a parametric study of the value of coupling factor has been done in order to find the optimal point. That is why the selected value for positive coupling is +0.5 and negative coupling factor is -0.4. As it can be seen in Figure 12, negative coupling can reduce the current overshoot significantly while keeping the voltage overshoot in turn off state still moderate. This reduction of current overshoot in negative coupling is along with reduction of voltage slop. It can be seen in Figure 13, how it can changes the power in turn on and turn off state. In other side, there is a considerable reduction of losses which comes with positive coupling. An alternative to reduce the overshoot and keep the decreasing attitude of losses in positive coupling can be increasing the gate resistance (Figure 14,Figure 15).

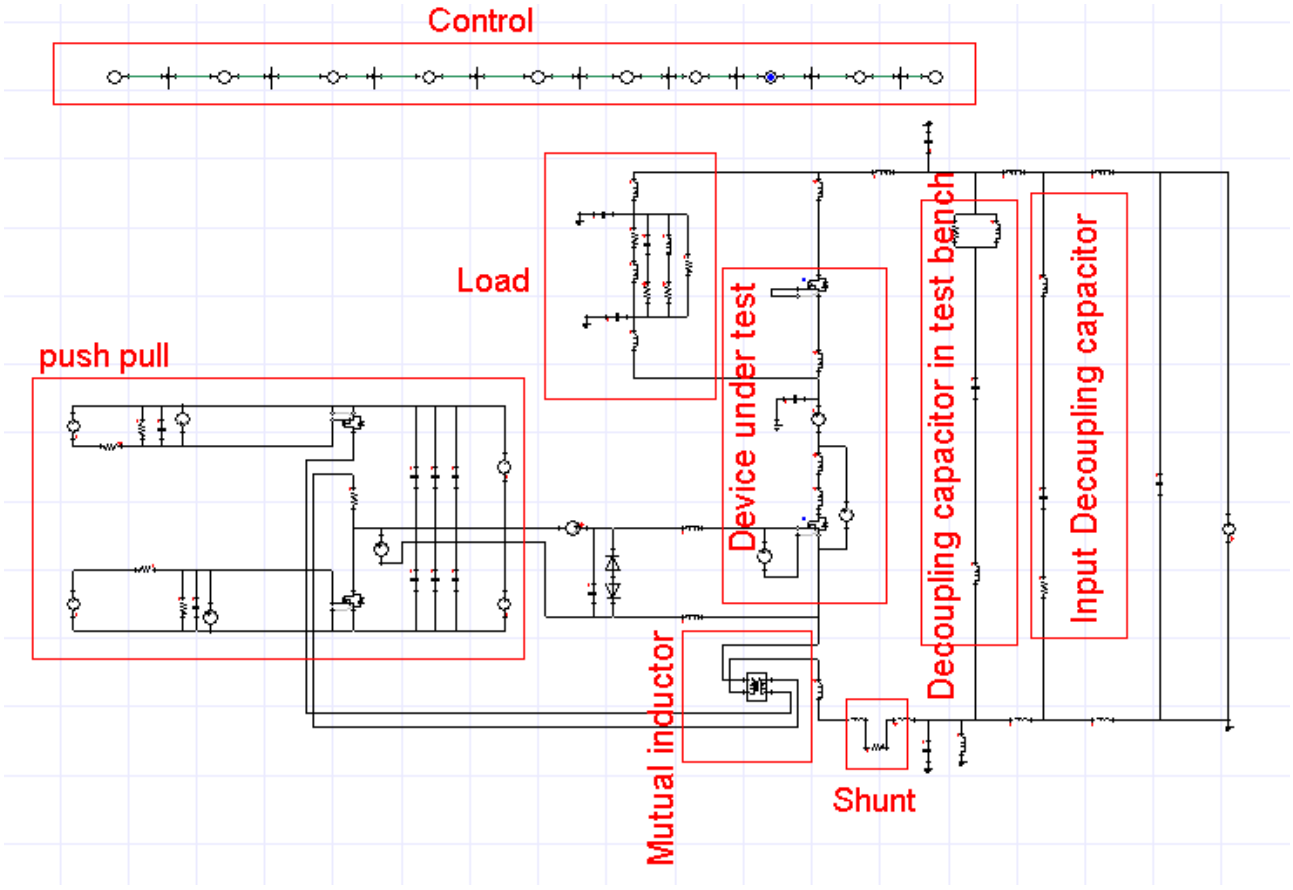



*Figure 11: Simulation of inductive coupling*

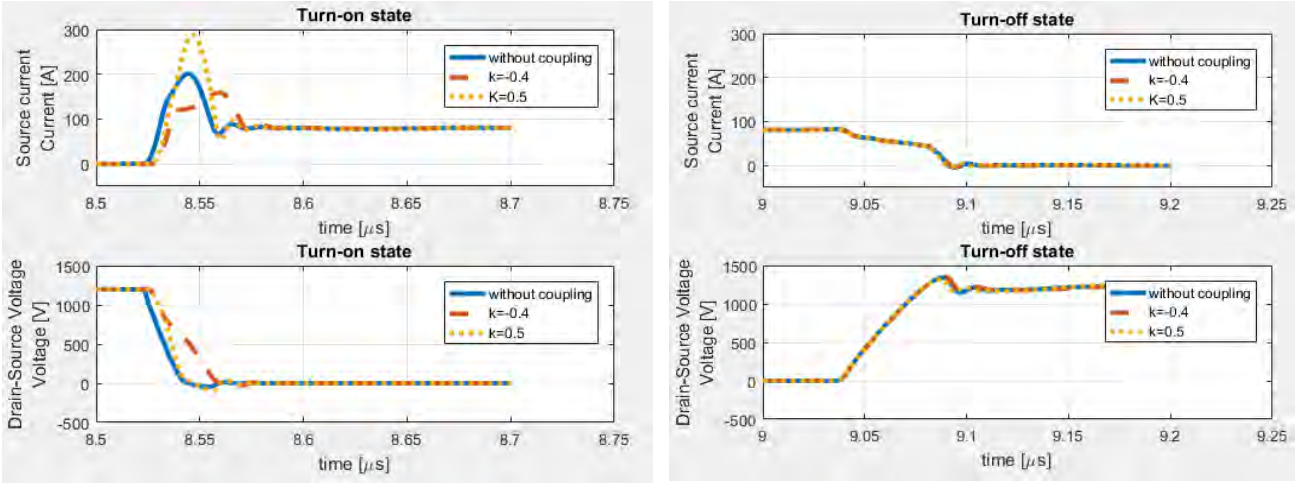



*Figure 12: switching behavior with inductive coupling*

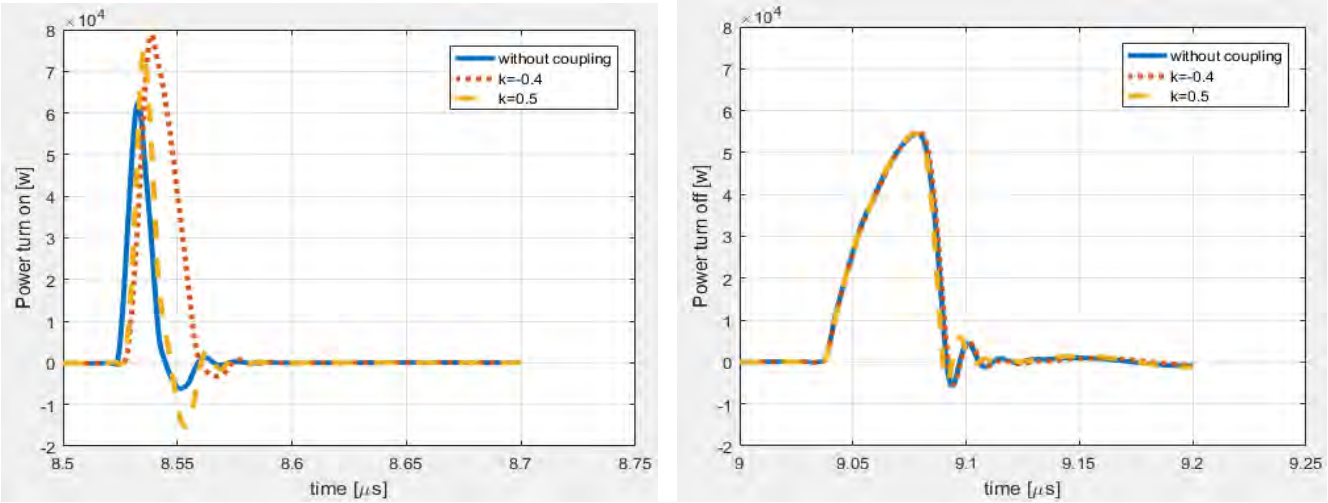



*Figure 13: Power variation in inductive coupling*

by comparing switching losses calculation, it can be seen the turn on switching losses decreased around 15% by positive coupling while it increased around 10% by negative coupling. In turn off switching losses has the variation around 5%.

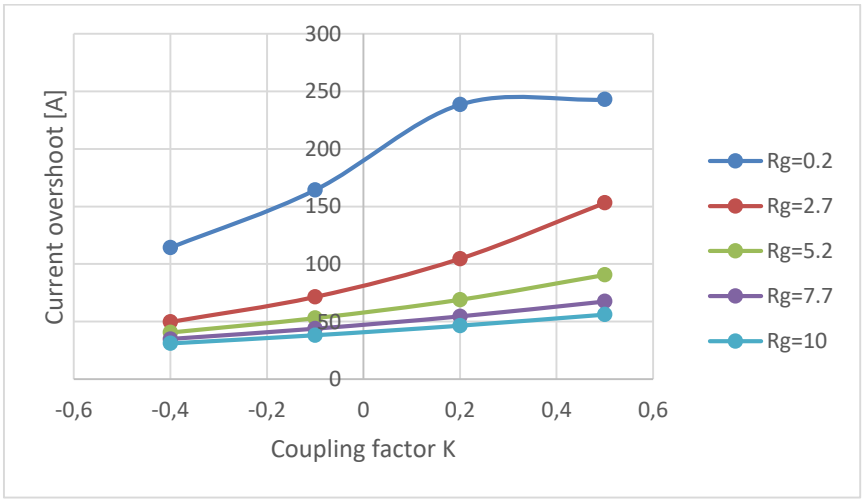



*Figure 14: Parametric study of current overshoot*

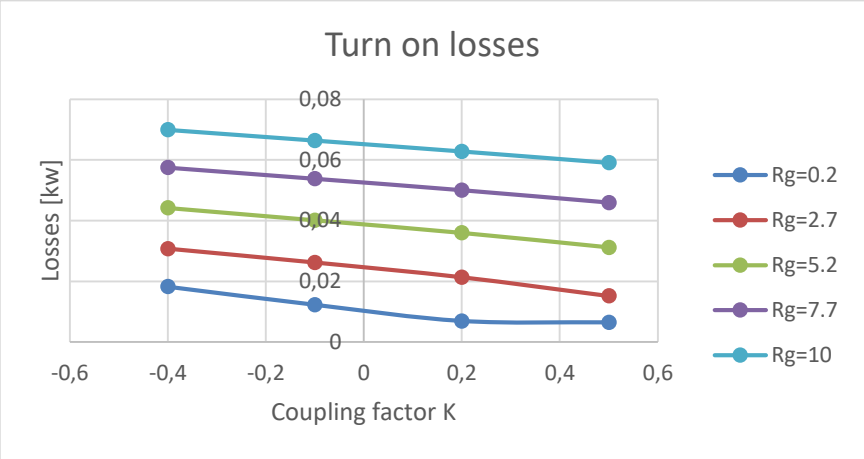



*Figure 15: Turn on losses variation based on coupling factor and gate resistance*

## 9. Conclusion and perspectives:

In this report, the nature of SiC MOSFET which is the basis of this thesis has been explained. In addition, parameters and elements which affect and generate current and voltage spikes are presented. One of the key element of EMC study is to provide the EMI model, and here it is tried to explain the steps to get the proper model and indeed following stages that should be carried out to find the optimized one. Depending on the application, there are some solutions to reduce the EMI in power converter. In this thesis, not only controlling the $di/dt$ in turn-on and the $dv/dt$ in turn- off transitions but also keep low switching losses with fixed gate resistance is priority, therefore the goal is to find a solution-considering the cost and losses- can moderate oscillation of voltage and current.

The work has been started based on the first prototype of Super-Grid institute. A Gate drive for high power converter which is applied for IGBT. In power converter gate drive is an interface between control command and power part. Considering EMC studies should be an early stage of design, in this thesis, it is trying to act on a gate drive as one of the elements to reduce EMI in power converter. Besides reducing the overshoot of source current improving the efficiency of power converter with the help of gate drive is included in to do list for next two years. The main interest is to keep the gate drive circuit as simple as possible regarding that is capable to be applied to different power module. At the end, by shaping the voltage waveform slope, it is targeted to reduce the Common Mode current. Common mode current investigating result in clamped inductive circuit has been shown in ANNEX1. To get to those objectives there are some challenges in front that should be solved in following months:

- The equivalent circuit uses fixed values of impedances to simplify the modeling process. However, in actual situations, the component values (inductance, capacitance, and resistance) are varying based on the frequency, voltage and current. Therefore, in different conditions the result is different from the measurement, however in this application the proper equivalent circuit of power converter which can

provide similar behavior depends significantly to the model of power module; it is planned to test the equivalent circuit in different software in order to find the closest behavior. This task has been started from the middle of first year of thesis and it is planned to be finished until the middle of second year.

Until now it has been tried in Simplorer and in next six months it is going to be tested in LTspice, by the middle of the second year, a comprehensive comparison between two generated model will be done in order to find the final equivalent circuit.

- First measurement has shown a great effect of parasitic inductor in switching loop, for the next year, it is planned to find a proper platform for the connection of gate drive to the power module since the final goal is to have a gate drive capable for switching the different power module.
- One important difference of presented buck converter equivalent circuit is that parasitic capacitance to the grounded is included while in most of the EMI studies of gate drive, this element has not been considered. An extensive input common mode current study has been started and will be continued in following year. The aim is to find out the attitude of Common Mode current in this application based on different voltage waveform. (First result in ANNEX1)
- To apply inductive coupling method, there is a great need of optimized coupling, it can be done with a transformer or a coil, in both case, it should be designed carefully in order to prevent to add minimum stray inductor to the switching loop. This stage of thesis is planned to start from the beginning of the second year.
- The selected method has not been tried yet neither in soft switching, neither in synchronous buck converter. The idea is to try the method in second and third year of thesis in synchronous converter, considering the short circuit protection and proper isolation in half bridge configuration should be considered.
- During the second six months of the second year and first half of the third year, a new prototype will be tested and all the EMI study will be applied for the final report at the end of third year.

***References***


[1] M. I. Montrose, EMC and the printed circuit board: design, theory, and layout made simple, vol. 6, John Wiley & Sons, 2004.

[2] L. F. S. Alves, P. Lefranc, P.-O. Jeannin and B. Sarrazin, "Review on SiC-MOSFET Devices and Associated Gate Drivers".

[3] A. Stefanskyi, Ł. Starzak and A. Napieralski, "Review of commercial SiC MOSFET models: Topologies and equations," in *Mixed Design of Integrated Circuits and Systems, 2017 MIXDES-24th International Conference*, 2017.

[4] M. K. Das, "SiC MOSFET module replaces up to 3x higher current Si IGBT modules in voltage source inverter application," *Bodo's Power Systems*, vol. 22, p. 24, 2013.

[5] P. Bogónez-Franco and J. B. Sendra, "EMI comparison between Si and SiC technology in a boost converter," in *Electromagnetic Compatibility (EMC EUROPE), 2012 International Symposium on*, 2012.

[6] J. Wang, H. S.-h. Chung and R. T.-h. Li, "Characterization and experimental assessment of the effects of parasitic elements on the MOSFET switching performance," *IEEE Transactions on Power Electronics*, vol. 28, pp. 573-590, 2013.

[7] C. U-Yaisom, W. Khanngern and S. Nitta, "The study and analysis of the conducted EMI suppression on power MOSFET using passive snubber circuits," in *Electromagnetic Compatibility, 2002 3rd International Symposium on*, 2002.

[8] B. Arntzen and D. Maksimovic, "Switched-capacitor DC/DC converters with resonant gate drive," *IEEE Transactions on Power Electronics*, vol. 13, pp. 892-902, 1998.

[9] H. Fujita, "A resonant gate-drive circuit capable of high-frequency and high-efficiency operation," *IEEE transactions on Power Electronics*, vol. 25, pp. 962-969, 2010.

[10] Z. Yang, S. Ye and Y.-F. Liu, "A new dual channel resonant gate drive circuit for synchronous rectifiers," in *Applied Power Electronics Conference and Exposition, 2006. APEC'06. Twenty-First Annual IEEE*, 2006.

[11] P. Anthony, N. McNeill and D. Holliday, "High-speed resonant gate driver with controlled peak gate voltage for silicon carbide MOSFETs," *IEEE Transactions on Industry Applications*, vol. 50, pp. 573-583, 2014.

[12] Z. Zhang, F.-F. Li and Y.-F. Liu, "A high-frequency dual-channel isolated resonant gate driver with low gate drive loss for ZVS full-bridge converters," *IEEE Transactions on Power Electronics*, vol. 29, pp. 3077-3090, 2014.

[13] Z. Yang, S. Ye and Y.-F. Liu, "A new dual-channel resonant gate drive circuit for low gate drive loss and low switching loss," *IEEE Transactions on Power Electronics*, vol. 23, pp. 1574-1583, 2008.

[14] J. W. K. Y. Lobsiger and J. W. Kolar, "Closed-Loop di/dt and dv/dt IGBT Gate Drive Concepts," *ed. ETH Zurich, Switzerland: Power Electronic Systems Laboratory*, 2013.

[15] Y. Lobsiger and J. W. Kolar, "Closed-loop IGBT gate drive featuring highly dynamic di/dt and dv/dt control," in *Energy Conversion Congress and Exposition (ECCE), 2012 IEEE*, 2012.

[16] Z. Zhang, F. Wang, L. M. Tolbert and B. J. Blalock, "Active gate driver for crosstalk suppression of SiC devices in a phase-leg configuration," *IEEE Transactions on Power Electronics*, vol. 29, pp. 1986-1997, 2014.

[17] K. Yamaguchi, Y. Sasaki and T. Imakubo, "Low loss and low noise gate driver for SiC-MOSFET with gate boost circuit," in *Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE*, 2014.

[18] S. Takizawa, S. Igarashi and K. Kuroki, "A new di/dt control gate drive circuit for IGBTs to reduce EMI noise and switching losses," in *Power Electronics Specialists Conference, 1998. PESC 98 Record. 29th Annual IEEE*, 1998.

[19] A. Paredes, H. Ghorbani, V. Sala, E. Fernandez and L. Romeral, "A new active gate driver for improving the switching performance of SiC MOSFET," in *Applied Power Electronics Conference and Exposition (APEC), 2017 IEEE*, 2017.

[20] S. Musumeci, A. Raciti, A. Testa, A. Galluzzo and M. Melito, "Switching-behavior improvement of insulated gate-controlled devices," *IEEE Transactions on Power Electronics*, vol. 12, pp. 645-653, 1997.

[21] B. Wittig and F. W. Fuchs, "Analysis and comparison of turn-off active gate control methods for low-voltage power MOSFETs with high current ratings," *IEEE Transactions on Power Electronics*, vol. 27, pp. 1632-1640, 2012.

[22] N. Idir, R. Bausiere and J. J. Franchaud, "Active gate voltage control of turn-on di/dt and turn-off dv/dt in insulated gate transistors," *IEEE Transactions on Power Electronics*, vol. 21, pp. 849-855, 2006.

[23] H. Riazmontazer, A. Rahnamaee, A. Mojab, S. Mehrnami, S. K. Mazumder and M. Zefran, "Closed-loop control of switching transition of SiC MOSFETs," in *Applied Power Electronics Conference and Exposition (APEC), 2015 IEEE*, 2015.

[24] K. Ogata and K. Wada, "Influence of induced voltage noise on switching characteristics for a power converter circuit," in *URSI Asia-Pacific Radio Science Conference (URSI AP-RASC)*, 2016.

[25] M. Ebli, M. Wattenberg and M. Pfost, "A gate driver approach enabling switching loss reduction for hard-switching applications," in *Power Electronics and Drive Systems (PEDS), 2017 IEEE 12th International Conference on*, 2017.

[26] M. Ebli and M. Pfost, "A novel gate driver approach using inductive feedback to increase the switching speed of power semiconductor devices," in *Power Electronics*

*and Applications (EPE'17 ECCE Europe), 2017 19th European Conference on*, 2017.

[27] D. a. I. D. Vries, "A resonant power MOSFET/IGBT gate driver," in *Applied Power Electronics Conference and Exposition, 2002. APEC 2002. Seventeenth Annual IEEE*, 2002.

ANNEX.1

Simulating the equivalent circuit of double pulse test has been done with two different model of power module CAS300M17BM2, the average model and the dynamic model. It should be noticed that all the results presented here in simulation and experimental, are at the 1200V switching voltage and 64A load current unless it will be noted on the figure. The load charge has been done with two different pulse with the 80us as a first pulse and 5us as a second pulse since the absolute inductor of air inductor is around 1.5mH.

The rise time and fall time based on the IEC 60747-8-4, CEI 60747-8-4 standard has been calculated approximately between 10% and 90% of Gate-Source voltage. gate source voltage variation is between -5 to 21 Volt therefore the rise time and fall time was measured between about 19 to -4.5 volts. By looking at Figure 16, Figure 17and Figure 18, it can be seen that the different resonant circuit between simulation and experimental made a different ringing in current and voltage. Although around 10% the simulation in rise time is longer than experiments.

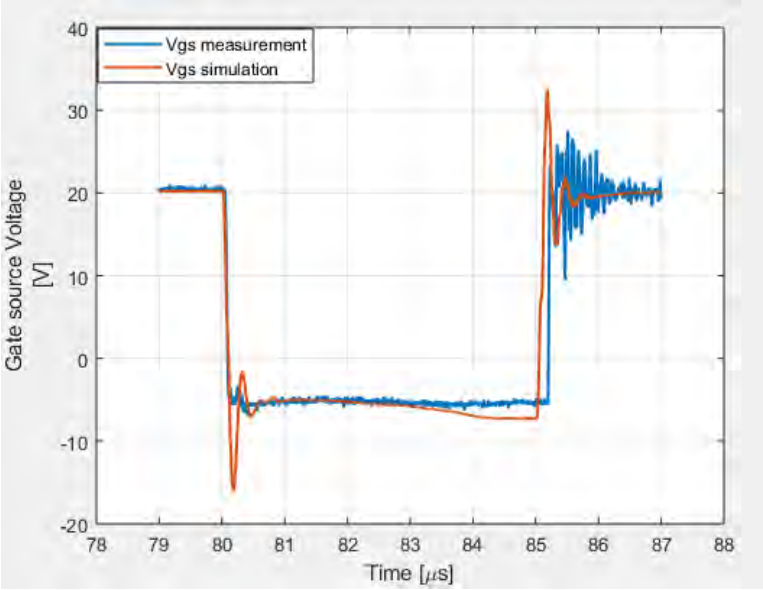



*Figure 16: Vgs comparison in simulation and experimental*

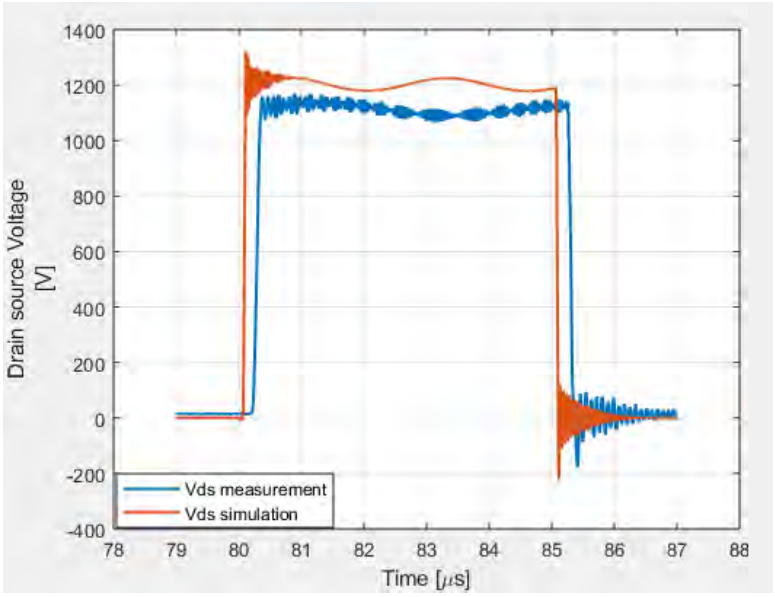



*Figure 17: Vds comparison in simulation and experimental*

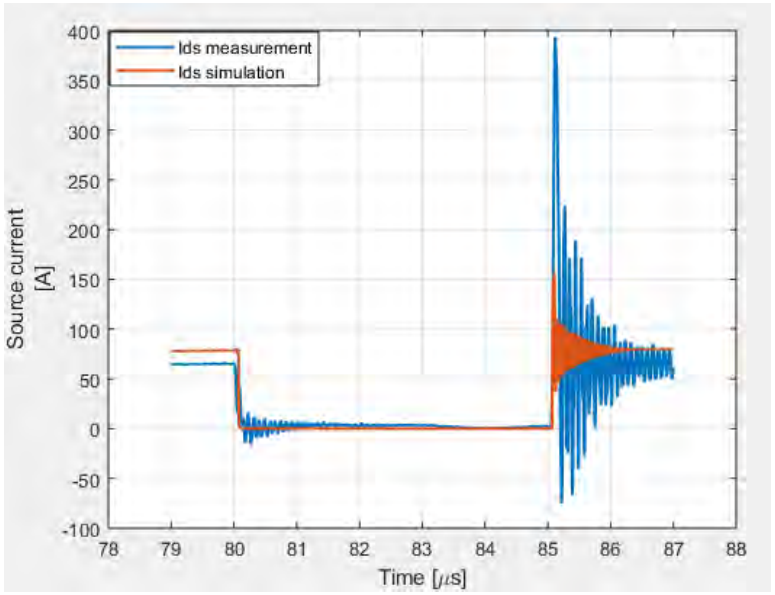



*Figure 18: Ids comparison in simulation and experimental*

In Figure 18, a big amount of current overshoot can be seen with the gate resistance of 0.2Ω in experimental result. Although a total amount of stray inductor that cause ringing in simulation and experimental set up is relatively equal, the provided model of power module in Simplorer does not show a compatible behavior.
It was tried to change the gate resistance of gate drive to 5Ω to monitor voltage and current variation. By calculating the rise time and fall time, the increase of slope from 100ns to 1us can be seen by increasing the

gate resistance (see Figure 19). This slowing down the switching speed results in more cleaning gate source voltage indeed. This effect is not evident in drain source voltage(Figure 21) but it clearly decreased the current overshoot(Figure 20). The important point is to consider how much this slowing down the switching speed can change the losses in Sic MOSFET as well as Common mode current. In turn off there is 16 % of decreased losses while as it was expected it increase the losses in turn-on state around 10%. In other words, if current overshoot and losses considered in a same frame. The tradeoff was done by increasing the losses up to 10% in turn on, lead to decrease current overshoot up to 60%.

| | Losses off-state (kw) | Losses on-state (kw) |
|---|---|---|
| Rg=0.2 Ω | 0.0029 | 0.0131 |
| Rg=5 Ω | 0.0018 | 0.0143 |

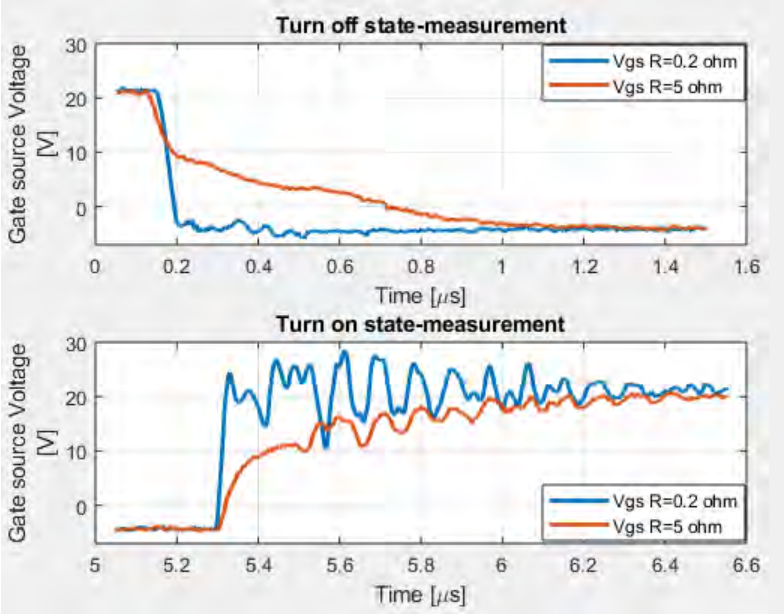



*Figure 19:Vgs comparison with different gate resistance*

Increasing the gate resistance changes the rise time of the source current and multiplication of this current to voltage make an apparent change in losses. By probing the rate of change of voltage and current it can be understood that this increasing the gate resistance, make the charging and discharging of the input and output capacitance longer. This change of rate of voltage variation decrease the input Common Mode Current as it can be seen in Figure 22.

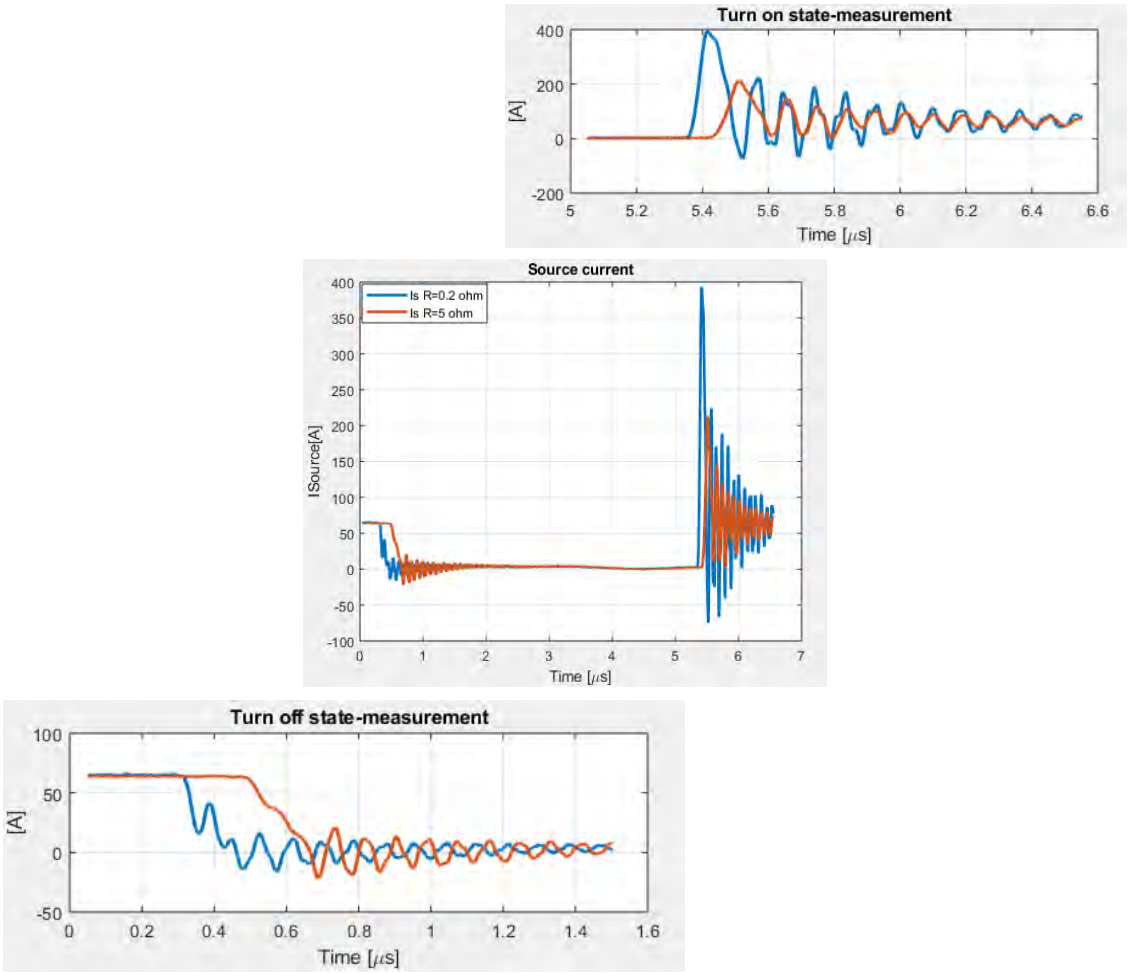



*Figure 20: source curent oscillation*

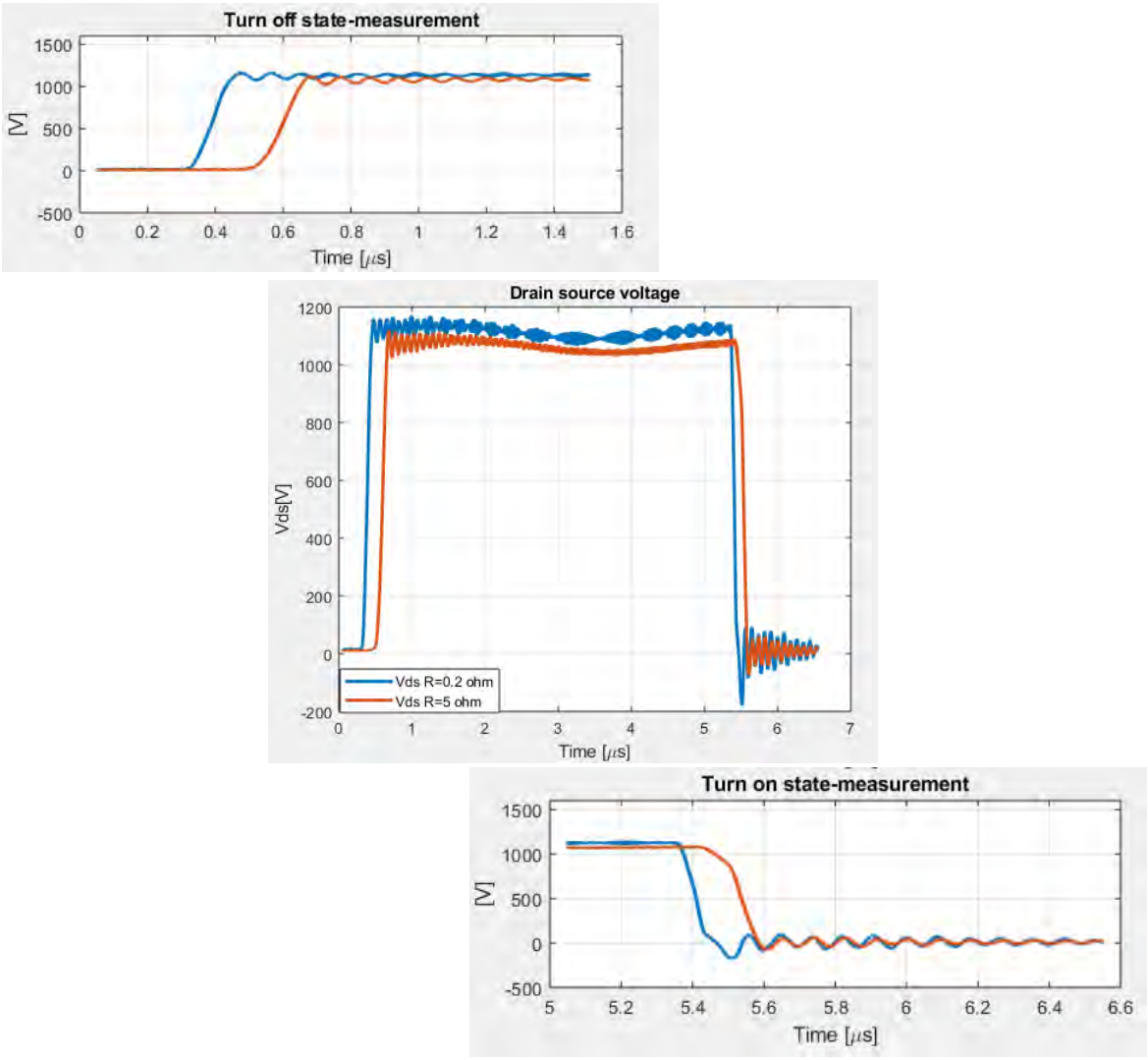



*Figure 21: Drain source voltage oscillation*

Changing the rate of change in voltage has a direct impact on the Common mode current. Since the value of the parasitic elements are constant during the switching behavior of the SiC MOSFET, it can be seen that this voltage profile change can reduce the common mode current. The input common mode current of the supply of the gate drive has been measured in order to see the influence of this voltage variation for two different cases (Figure 23, Figure 24).

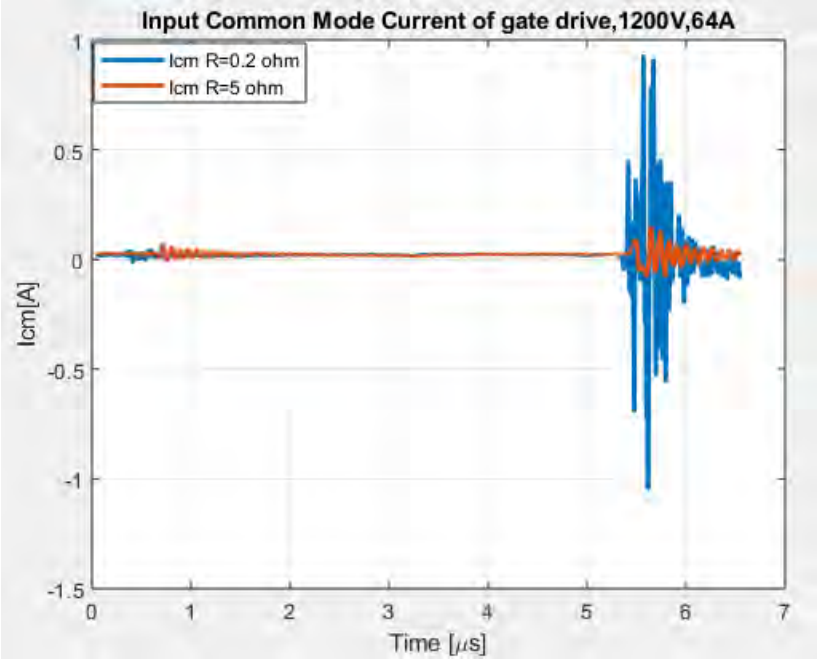
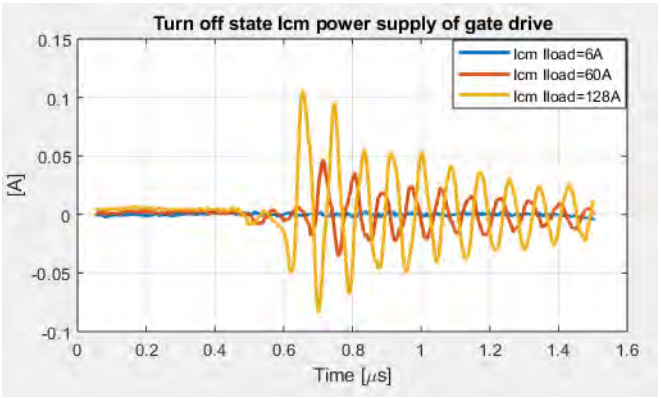
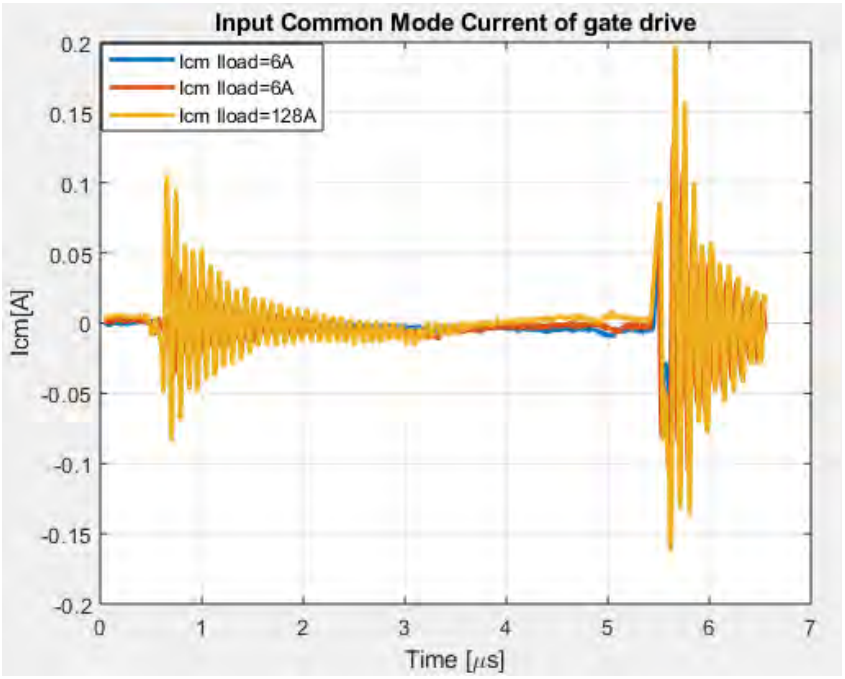



*Figure 22: Input common mode current*



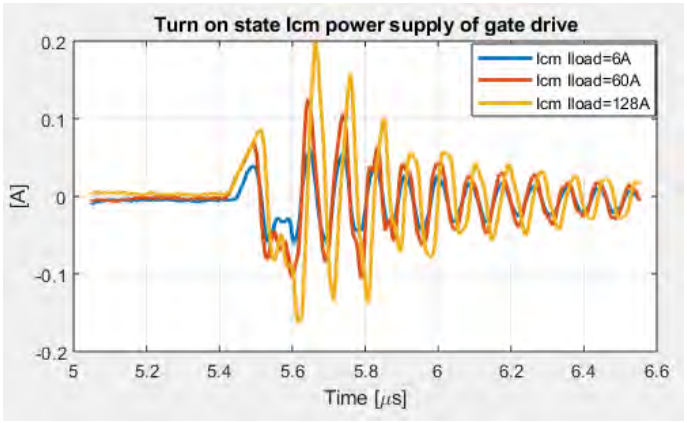




*Figure 23: Common mode current for different load current*

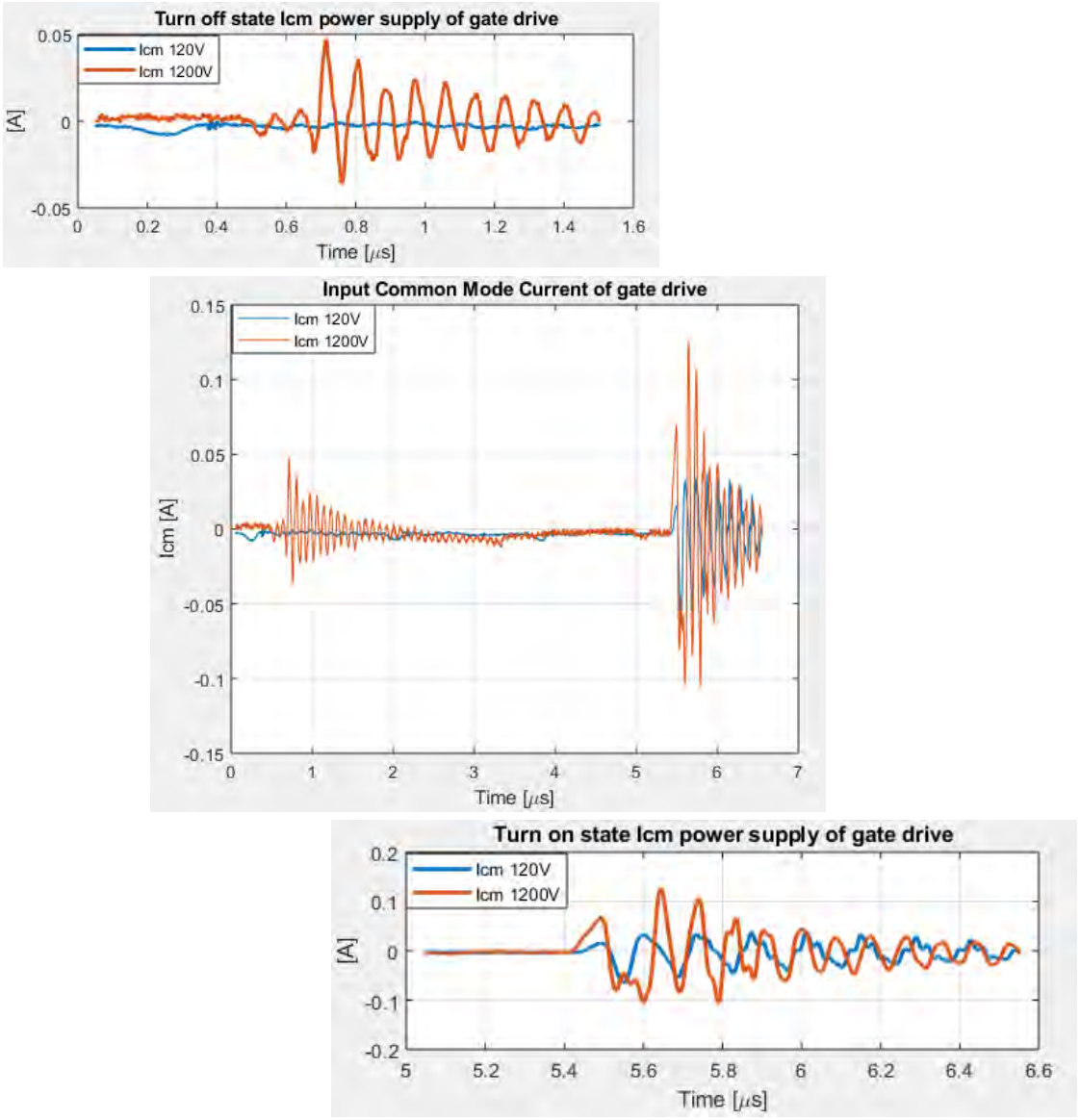



*Figure 24: Common mode current for different voltage*

From above results it can be said that in future gate drive should be tested with more different value of resistance to find the tendency of losses based on the external gate resistance in order to find the suitable tradeoff between losses and external gate resistance while the goal is to overcome EMI emissions.

Université de Lyon
CNRS, Ecole Centrale Lyon, INSA Lyon, Université Claude Bernard Lyon 1

**Laboratoire Ampère**

**Unité Mixte de Recherche du CNRS - UMR 5005**

Génie Electrique, Automatique, Bio-ingénierie

# *Mémoire doctorant 1ère année*

# *2017 -2018*

| | |
|---|---|
| Nom - Prénom | **Kircher Alexandre** |
| email | **alexandre.kircher@ec-lyon.fr** |
| Titre de la thèse | **Estimation résiliente** |
| Directeur de thèse | **Bako Laurent** |
| Co- encadrants | **Benallouch Mohamed, Blanco Eric** |
| Dpt. de rattachement | **Electronique, Electrotechnique, Automatique** |
| Date début des travaux | **1er octobre 2017** |
| Type de financement | **Bourse académique** |

Laboratoire Ampère – Ecole Centrale de Lyon – 36, avenue Guy de Collongue - 69134 Ecully cedex – France
Tél : +33 (0) 4 72 18 60 99 Fax : +33 (0) 4 78 43 37 17 http://www.ampere-lab.fr

## Résumé du Rapport

Dans le domaine de l'estimation des systèmes, les exigences en matière de performances et de sécurité nous amènent à prendre en compte de plus en plus de types de perturbation. Cependant, si des estimateurs d'état, tels que l'estimateur de KALMAN développé à partir des années soixante, permettent de gérer efficacement des bruits gaussiens présents au sein d'un système, ils gèrent bien moins efficacement des perturbations soudaines et échappant à une modélisation sous forme de signal aléatoire, tels que les défaillances de capteur par exemple.

Le présent rapport détaille donc les activités effectuées au cours de ma première année de thèse et durant laquelle nous avons cherché à mettre en place un estimateur optimal étant capable de gérer efficacement ces *nouveaux* types de perturbation. Il expose tout d'abord le principe de l'estimation d'état, formule le problème auquel cette dernière répond puis présente un état de l'art de l'estimation qui nous permet de préciser le cahier des charges pour notre futur estimateur. Dans un deuxième temps, une structure répondant à ce cahier des charges est proposée et nous discuterons de sa mise en œuvre pour des problématiques hors-ligne et en ligne. Dans les deux cas, des éléments de l'analyse de ces performances seront présentés, pour finir par un bilan sur ce qu'il reste à effectuer quant à l'analyse de cet estimateur ainsi que sur les activités envisagées pour les deux prochaines années.

## Abstract

Regarding the estimation of systems, one of the key elements to improve the performance and versatility of observers is to take into account more and more types of disturbancies. However, disturbancies which cannot be modeled by random signals – such as sensor failures – are more often than not beyond the scope of classical observers, such as Kalman estimator, which are typically designed to handle gaussian noises.

The following report deals with the proposition of an optimisation-based observer architecture in order to handle that *new* type of disturbancies. In a first part, we will discuss the problem and the aim of state estimation, and then a review of the state of the art will enable us to give a more precise definition of the goals of our observer. Secondly, the architecture itself will be presented and as we discuss its application to online and offline problematics, we will provide several analysis results and leads to assess its theoretical performances. We will finally discuss the remaining points yet to investigate and the global plan regarding our work for the two following years.

# Table des matières



# Table des figures

## Nomenclature

| | |
|---|---|
| $\mathbb{R}$ | ensemble des réels |
| $\mathbb{R}^+$ | ensemble des réels positifs |
| $\mathbb{K}^{a\times b}$ | ensemble des matrices avec $a$ lignes et $b$ colonnes à valeurs dans le corps $\mathbb{K}$ |
| $\mathbb{K}^a$ | ensemble des vecteurs colonnes à $a$ éléments à valeurs dans $\mathbb{K}$ |
| $\mathcal{S}_a^+(\mathbb{K})$ | ensemble des matrices définies positives à $a$ lignes à valeurs $\mathbb{K}$ |
| $\mathcal{K}_\infty$ | ensemble des fonctions $g$ de $\mathbb{R}^+$ dans $\mathbb{R}+$ strictement croissantes telles que $g(0)=0$ et $g(\lambda)\to\infty$ quand $\lambda\to\infty$ |
| $\mathrm{Id}_a$ | matrice identité (carrée) à $a$ lignes |
| $t\in\mathbb{N}$ | variable de temps *discrète* |
| $a_t$ | Valeur d'une variable $a$ à l'instant $t$ |
| $x_t\in\mathbb{R}^n$ | Vrai vecteur d'état (à l'instant $t$) |
| $\hat{x}_{t\|k}$ | Estimation du vecteur d'état (à l'instant $t$) pour des mesures jusqu'à l'instant $k$ |
| $X_t\in\mathbb{R}^{n\times t+1}$ | Vraie trajectoire du système jusqu'à $t$, c'est-à-dire $X_t=\begin{pmatrix}x_0 & x_1 & \cdots & x_t\end{pmatrix}$ |
| $\hat{X}_t$ | Trajectoire estimée du système jusqu'à $t$, c'est-à-dire $\hat{X}_t=\begin{pmatrix}\hat{x}_{0\|t} & \hat{x}_{1\|t} & \cdots & \hat{x}_{t\|t}\end{pmatrix}$ |
| $Z_t\in\mathbb{R}^{n\times t+1}$ | Trajectoire quelconque, variable de décision de $V_t$ avec $Z_t=\begin{pmatrix}z_0 & z_1 & \cdots & z_t\end{pmatrix}$ |
| $V_t(:\mathbb{R}^{n\times t+1}\to\mathbb{R}^+)$ | Fonction coût |
| $\phi_t(:\mathbb{R}^n\to\mathbb{R}^+)$ | Fonction coût convexe s'appliquant sur la dynamique du système à l'instant $t$ |
| $\psi_t(:\mathbb{R}^m\to\mathbb{R}^+)$ | Fonction coût convexe s'appliquant à la sortie du système à l'instant $t>0$ |
| $V_t^*(:\mathbb{R}^n\to\mathbb{R}^+)$ | Fonction coût en ligne |

**Représentation d'état linéaire temps-variant (LTV) en temps discret :**

$$\begin{cases} x_{t+1} &= A_t x_t + B_t u_t + \omega_t \\ y_t &= C_t x_t + D_t u_t + \nu_t \end{cases} \quad \text{avec} \quad \begin{cases} x_t\in\mathbb{R}^n & \text{Vecteur d'état} \\ u_t\in\mathbb{R}^{n_u} & \text{Entrée connue du système} \\ A_t\in\mathbb{R}^{n\times n} & \text{Matrice d'état à l'instant } t \\ B_t\in\mathbb{R}^{n\times n_u} & \text{Matrice de commande à l'instant } t \\ \omega_t\in\mathbb{R}^n & \text{Perturbation dynamique inconnue} \\ y_t\in\mathbb{R}^{n_y} & \text{Vecteur de sortie du système} \\ C_t\in\mathbb{R}^{n_y\times n} & \text{Matrice d'observation à l'instant } t \\ D_t\in\mathbb{R}^{n_y\times n_y} & \text{Matrice d'action directe à l'instant } t \\ \nu_t\in\mathbb{R}^{n_y} & \text{Perturbation inconnue en sortie} \end{cases}$$

## Lexique

| | |
|---|---|
| Estimation en ligne : | Problème où on estime l'état du système au fur et à mesure que les mesures arrivent. |
| Estimation hors-ligne : | Problème où l'on a accès à toutes les mesures entre un instant initial $t_0$ et un instant final *fixé* $t_f$ pour estimer l'état du système sur ce même intervalle. |
| Perturbation impulsive : | Événement soudain et imprévisible ayant lieu au sein d'un système et pouvant mettre en péril son intégrité. |
| Perturbation continue : | perturbation inhérente au système et toujours présente en son sein |

# 1 Introduction

Dans certains systèmes complexes, tels que les réseaux de distribution électrique ou hydraulique, il paraît irréalisable de placer des capteurs à tous les points auxquels les opérateurs de ces systèmes aimeraient récupérer de l'information : une méthode classique pour pallier à ce problème est de placer un nombre raisonnable de capteurs dans le réseau et de reconstruire l'information aux points non mesurés à l'aide d'un objet mathématique appelé un « estimateur ». Cependant, il devient de plus en plus crucial de connaître précisément l'état de ces systèmes :

— Dans le cas des réseaux électriques, l'arrivée de nouvelles sources d'électricité demande une surveillance encore plus accrue. En effet, ces nouvelles sources dîtes « à énergie renouvelable » se caractérisent par des constantes de temps beaucoup plus petites que les sources classiques : une éolienne, du fait du vent, peut ne pas produire la même quantité d'électricité d'une heure à l'autre. En conséquence, pour pouvoir assurer au mieux la stabilité du réseau, il est nécessaire d'avoir une bien meilleure résolution temporelle pour repérer les variations de tension dues à ces nouvelles sources.

— Dans le cas des réseaux hydrauliques, pour empêcher au maximum le gaspillage de l'eau, il faut pouvoir détecter le plus rapidement possible des problèmes d'étanchéité ou de fuite, ce qui demande une connaissance de l'état du réseau toujours plus fine.

De plus, parallèlement à ces besoins croissants en matière de performance d'estimation, la question de la fiabilité de cette estimation se pose : en effet, les performances se doivent d'être assurées même en présence de dysfonctionnements. Du fait du rapport privilégié qui existe entre l'estimation et les mesures du système qui lui sont rapportées, c'est souvent au niveau des ces dernières que les dysfonctionnements sont considérés dans la littérature [KA13] [Mis+17].

Parmi les dysfonctionnements envisagés, les plus répandus sont la défaillance capteur et l'attaque d'un intervenant tiers dans la liaison estimateur/capteur : en effet, de plus en plus d'estimateurs sont mis en œuvre sur des systèmes numériques qui peuvent ainsi être la cible de cyber-attaques. L'objectif de la thèse est donc de mettre en place une stratégie pour permettre d'estimer un système même en présence de tels dysfonctionnements dont la façon dont d'être modélisés dépasse le cadre classique de l'estimation.

Le présent rapport a pour but de rendre compte des points majeurs sur lesquels j'ai pu travailler au cours de ma première année de thèse. Dans une première partie, nous aborderons d'un point de vue très général les problématiques d'estimation résiliente pour parvenir à la formulation d'objectifs que nous avons établis pour cette thèse. Dans une deuxième partie, nous verrons la direction que nous avons considérée pour remplir ces objectifs, ce qui passe par la définition d'une classe d'estimateur issue de la théorie de l'optimisation et pour laquelle il nous est nécessaire de mener une analyse approfondie des performances : seront donc présentés le principe de cet estimateur, les cas qu'il sera capable de traiter et l'état d'avancement de l'analyse que nous menons sur ses performances. Enfin, nous conclurons sur un bilan des points effectués en cours d'année, ceux en traitement et ceux qu'il restera à considérer durant les deux prochaines années.

# 2 Mise en contexte

Afin d'expliciter au mieux la problématique du sujet et nos objectifs, il paraît indispensable de définir ce que nous entendons par « estimation résiliente ».

Le but de cette partie est donc de développer, dans un premier temps, l'intérêt de l'estimation d'état, ce qui va nous amener à la formulation du problème global que l'on va chercher à résoudre. Ensuite, au vu de l'état de l'art, nous viendrons à définir plus précisément les objectifs de notre étude.

## 2.1 Problématique de l'estimation d'état

### 2.1.1 Intérêt de l'estimation d'état

Dans le domaine du contrôle des systèmes, réaliser des objectifs fixés exige une bonne connaissance de ces dits systèmes. Cette dernière s'obtient alors par le biais de l'identification, domaine de l'automatique qui permet de trouver un modèle qui décrit *suffisamment le système* par rapport à nos exigences. Cette connaissance permet alors, par exemple, de synthétiser un correcteur permettant d'asservir le système.

Une des stratégies de commande classique, appelée « commande à retour d'état », exige notamment d'avoir une connaissance sur l'état dans lequel se trouve le système à un instant $t$ afin d'adapter la commande. Dans le cas idéal, les informations sur l'état du système nécessaires au contrôle de ce dernier sont accessibles à la mesure : la mise en œuvre du correcteur consiste alors *simplement* à mesurer les grandeurs qui nous intéressent à l'aide de capteurs pour pouvoir ensuite les lui envoyer. Mais nous ne sommes pas toujours dans ce cas, et il existe de nombreuses applications dans lesquelles on ne peut accéder à ces données intéressantes. Cela peut être dû à deux choses :

— La donnée intéressante n'est pas physiquement obtensible dans des conditions satisfaisantes, c'est-à-dire qu'il n'existe pas de capteur respectant le cahier des charges d'acquisition que l'on souhaite.

— La quantité de données intéressantes est trop grande par rapport aux moyens alloués à leur acquisition.

Ce constat se généralise alors pour tous les systèmes dont on souhaite extraire de l'information, que ce soit pour du contrôle ou pour d'autres utilisations comme la surveillance de santé des systèmes.

Pour pallier à ce problème, l'idée est de *reconstruire* l'information qui n'est pas mesurée : c'est ce qu'on appelle l'*estimation*. Une approche consisterait alors à simuler le système à partir de son modèle *seul*. Cela pose cependant le problème de l'accumulation des erreurs dues notamment à la modélisation. En effet, un modèle sera toujours une représentation imparfaite d'un système du fait de l'impossibilité de prendre en compte l'ensemble des processus physiques qui ont lieu en son sein, et ce parfois avec des échelles de temps très différentes. Au vu des fonctions du système qu'on cherche à modéliser et des horizons de temps sur lesquels on souhaite le faire, la négligence de certains phénomènes physiques est donc une étape incontournable de l'identification. Le risque d'une simulation est alors, à la manière d'une boucle ouverte, d'accumuler les erreurs dues à la modélisation, ce qui entraînerait une divergence croissante entre la simulation et le système réel au cours du temps.

L'enjeu principal de l'estimation d'état est donc le suivant :

*Comment mettre en œuvre notre connaissance d'un système afin de reconstruire l'état interne de ce dernier ?*

### 2.1.2 Principe de mise en œuvre de l'estimation d'état

Comme indiqué précédemment, la simulation d'un système par un modèle qui en est décorrélé n'est pas une façon viable de faire de l'estimation. Une façon plus élaborée de procéder est de mettre en lien un modèle du système avec des mesures prises dans le vrai système : comme dans le cas d'un correcteur en boucle fermée, on apporte au modèle des mesures, qui sont des informations partielles de l'état du vrai système, afin qu'il reconstruise la totalité de l'état. On nomme alors *estimateur* ou *observateur* le dispositif (théorique ou pratique) se servant des entrées et des sorties du systèmes afin de reconstruire des données internes et non mesurables sur son état.

La structure d'un estimateur va évidemment dépendre de la structure du modèle qui est utilisée. Dans le cas de l'estimation d'état, la structure de modèle utilisée est celle d'une représentation d'état. Pour notre étude, nous nous restreindrons aux systèmes Linéaires-Temps Variant (LTV) en temps discret de la forme

$$\begin{cases} x_{t+1} &= A_t x_t + B_t u_t + \omega_t \\ y_t &= C_t x_t + D_t u_t + \nu_t \end{cases} \quad \text{avec } x_0 \in \mathbb{R}^n \text{ état initial du système} \tag{2.1}$$

où

| | |
|---|---|
| $x_t \in \mathbb{R}^n$ | Vecteur d'état |
| $u_t \in \mathbb{R}^{n_u}$ | Entrée connue du système |
| $A_t \in \mathbb{R}^{n \times n}$ | Matrice d'état à l'instant $t$ |
| $B_t \in \mathbb{R}^{n \times n_u}$ | Matrice de commande à l'instant $t$ |
| $\omega_t \in \mathbb{R}^n$ | Perturbation dynamique inconnue |
| $y_t \in \mathbb{R}^{n_y}$ | Vecteur de sortie du système |
| $C_t \in \mathbb{R}^{n_y \times n}$ | Matrice d'observation à l'instant $t$ |
| $D_t \in \mathbb{R}^{n_y \times n_u}$ | Matrice d'action directe à l'instant $t$ |
| $\nu_t \in \mathbb{R}^{n_y}$ | Perturbation inconnue en sortie |

La première équation du modèle décrit la dynamique du système et comment l'état évolue au cours du temps : au sein de ce rapport, nous l'appellerons donc *équation dynamique* ou *équation d'état* du système. La deuxième exprime comment les sorties (ou mesures) du systèmes sont obtenues : nous l'appellerons *équation d'observation* du système dans ce rapport.

D'un point de vu général, ce type de modélisation sert à relier les entrées $u_t$ du système à ses sorties $y_t$ par le biais de variables intermédiaires contenues dans le vecteur d'état $x_t$ : elle est donc tout à fait appropriée à notre problématique. De plus, les quantités $\omega_t$ et $\nu_t$ servent à modéliser certaines imperfections du modèle par rapport à la réalité :

— $\omega_t$ peut modéliser des phénomènes internes ou externes extrêmement complexes intervenant dans la dynamique du système.

— $\nu_t$ peut correspondre à un bruit de capteur venant parasiter les mesures.

Ce sont des inconnues du système qui rendent l'estimation plus compliquée : l'estimation d'état consiste donc à obtenir une estimation $\hat{x}_t$ du vecteur d'état $x_t$ du système à partir de ses entrées et de ses sorties et ce malgré la non-connaissance des perturbations $\omega_t$ et $\nu_t$. La figure 1 présente un schéma-bloc résumant ce principe.

Il est à noter que dans la plupart des cas, le signal $u_t$ est connu à l'avance par l'opérateur : pour plus de simplicité, nous développerons le cas $u_t = 0$ pour tout $t$. Le système considéré sera alors de la forme

$$\begin{cases} x_{t+1} &= A_t x_t + \omega_t \\ y_t &= C_t x_t + \nu_t \end{cases} \quad \text{avec } x_0 \in \mathbb{R}^n \tag{2.2}$$

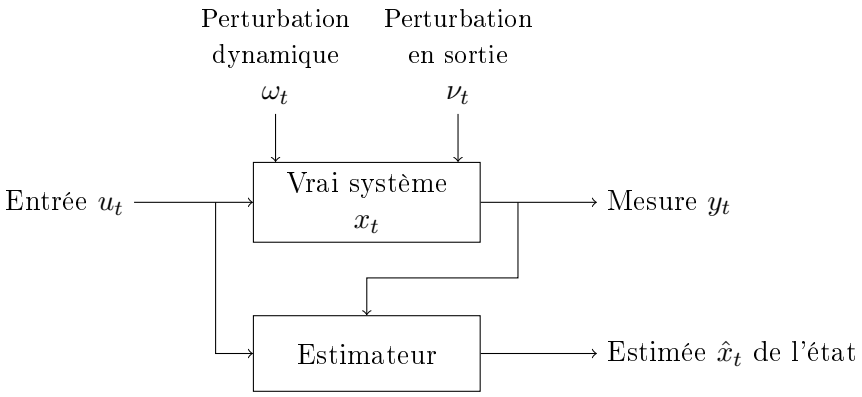



FIGURE 1 – Schéma-bloc du principe de l'estimation d'état

Comme on peut le voir dans le modèle (2.2), l'état ne s'exprime dans les sorties d'un système qu'au travers de la matrice d'observation $C_t$. Or, on sait par exemple qu'une matrice n'est pas forcément injective, c'est à dire que $x_1$ différent de $x_2$ n'implique pas forcément $C_t x_1$ différent de $C_t x_2$. Il paraît donc possible de confondre deux états différents si l'on considère uniquement les mesures du système : c'est pourtant spécifiquement le problème auquel on a affaire en estimation d'état. A partir de ce constat, comment s'assurer que l'on est capable de différencier deux états du système à partir de l'observation de ses sorties uniquement ? Pour répondre à cette question, il est nécessaire d'introduire la notion *d'observabilité* d'un système :

**Définition 2.1 (*Observabilité d'un système [Oga09]*)**

*Un système est dit complètement observable si tout état initial $x_0 \in \mathbb{R}^n$ peut-être déterminé à partir de l'observation des mesures $y_t$ du système sur un horizon de temps fini $0 \leq t \leq t_1$.*

Intuitivement, on peut comprendre la notion d'observabilité comme une propriété de distinguabilité : est-ce que l'observation des mesures $y_t$ permet de faire la différence entre deux états initiaux $x_0$ et $x'_0$ différents ? Dans le cas des systèmes temps-invariants (LTI) en temps discret et en l'absence de bruit, l'observabilité possède un critère de vérification simple. En effet, considérons le système

$$\begin{cases} x_{t+1} &= Ax_t \\ y_t &= Cx_t \end{cases} \tag{2.3}$$

avec deux états initiaux $x_0$ et $x'_0$ distincts. Le système est alors observable si et seulement si pour tout $x_0$ et $x_0$ dans $\mathbb{R}^n$ avec $x_0 \neq x'_0$, il existe un $t_1$ tel que

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{t_1} \end{pmatrix} \neq \begin{pmatrix} y'_0 \\ y'_1 \\ \vdots \\ y'_{t_1} \end{pmatrix}$$

où pour tout $t$, $y_t$ et $y'_t$ sont les sorties du systèmes à l'instant $t$ pour les états initiaux $x_0$ et $x'_0$ respectivement : si un tel $t_1$ n'existait pas, alors pour tout $t$, on aurait $y_t = y'_t$ et il ne serait pas possible de déterminer les états initiaux $x_0$ et $x'_0$ à partir de l'observation des mesures du système.

Or, on sait que pour tout $t$,

$$y_t = CA^t x_0 \text{ et } y'_t = CA^t x'_0$$

donc on en déduit que

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{t_1} \end{pmatrix} = \mathcal{O}_{t_1}(A,C)x_0 \text{ où } \mathcal{O}_{t_1}(A,C) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{t_1} \end{pmatrix}$$

et de même, $\begin{pmatrix} y'_0 & y'_1 & \cdots & y'_{t_1} \end{pmatrix}^\top = \mathcal{O}_{t_1}(A,C)x'_0$.

Finalement, on obtient que le système est observable si et seulement si pour tout $x_0$, $x'_0$ dans $\mathbb{R}^n$ distincts, il existe $t_1$ tel que

$$\mathcal{O}_{t_1}(A,C)(x_0 - x'_0) \neq 0 \quad \text{soit} \quad (x_0 - x'_0) \notin \mathrm{Ker}(\mathcal{O}_{t_1}(A,C))$$

Cette condition est assurée si et seulement si pour tout $t_1$, la matrice $\mathcal{O}_{t_1}(A,C)$ est injective, c'est-à-dire $\mathrm{Ker}(\mathcal{O}_{t_1}(A,C)) = \{0\}$, ce qui équivaut à ce qu'elle soit de rang plein. Or, on peut montrer que cela équivaut à ce que seule la matrice

$$\mathcal{O}_{n-1}(A,C) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}$$

soit de rang plein : on appelle cette matrice *la matrice d'observabilité* du système, puisque l'étude de son rang permet d'évaluer l'observabilité du système, et on parle très souvent de l'observabilité d'un couple $(A,C)$ pour désigner celle du système tout entier du fait que seules ces deux matrices interviennent dans la matrice d'observabilité.

Dans le cas LTV, un raisonnement analogue peut être mené : cependant, les matrices $A_t$ et $C_t$ changeant à chaque instant, il n'est pas possible de s'arrêter à une taille donnée : ainsi, on dira que le système LTV est **observable sur l'horizon de temps** $[t_1; t_2]$ si la matrice

$$\begin{pmatrix} C_{t_1} \\ C_{t_1+1}A_{t_1} \\ C_{t_1+2}A_{t_1+1}A_{t_1} \\ \vdots \\ C_{t_2}A_{t_2-1}A_{t_2-2}\dots A_{t_1} \end{pmatrix}$$

est de rang plein.

Cette propriété est donc une notion centrale de l'estimation d'état, car elle va bien souvent définir une condition nécessaire pour pouvoir estimer le vecteur d'état du système.

## 2.2 Estimateurs d'état classiques

### 2.2.1 Estimation sans bruit : estimateur de Luenberger

Mathématicien américain de l'université de Stanford, David Luenberger publie en 1964 un article [Lue64] concernant l'observation de l'état des systèmes dynamiques linéaires. Dans le cas d'un système admettant une représentation d'état Linéaire Temps-Invariante (LTI) en temps discret

$$\begin{cases} x_{t+1} &= Ax_t \\ y_t &= Cx_t \end{cases} \tag{2.4}$$

Par rapport au système dont le modèle est défini par (2.1), nous sommes ici dans un cas moins général, puisque temps-invariant d'une part, et ne prenant pas de bruit en compte d'autre part. Le principe à l'origine de l'article de Luenberger est de construire un deuxième système ayant pour entrée les mesures du système, c'est-à-dire d'équation dynamique

$$z_{t+1} = Gz_t + Hy_t, \tag{2.5}$$

et dont le vecteur d'état $z_t$ est une combinaison linéaire de $x$, c'est-à-dire qu'il existe une matrice $T$ telle que $z_t = Tx_t + a_t$, avec $a_t$ une certaine quantité qu'il serait intéressant d'avoir faible, voire tendant vers le vecteur nul. Contrairement au premier système, ce deuxième n'a aucune réalité physique, les matrices $G$ et $H$ sont donc avant tout des degrés de liberté sur lequel l'opérateur peut jouer.

Le résultat principal obtenu par Luenberger concerne les conditions d'existence de $T$ :

**Theorème 2.1 (*Existence de T [Lue64]*)**

*Une matrice $T$ telle que*

$$\forall t \geq 0, \quad z_t = Tx_t + G^t(z_0 - Tx_0) \tag{2.6}$$

*existe et est unique si et seulement si les matrices $A$ et $G$ n'ont aucune valeur propre en commun.*
*$T$ est alors donnée par la résolution de l'équation $TA - GT = HC$.*

**Démonstration.** Si $T$, telle que (2.6), existe et est unique, alors en remplaçant $z_t$ par $Tx_t + G^t(z_0 - Tx_0)$ dans (2.5), on obtient

$$Tx_{t+1} = GTx_t + Hy_t$$

Or, en remplaçant $x_{t+1}$ et $y_t$ par leur expression issue de (2.4), on obtient alors

$$\begin{aligned} & TAx_t = GTx_t + HCx_t \\ \Leftrightarrow \quad & (TA - GT - HC)x_t = 0 \end{aligned}$$

Cette équation devant être vraie *a priori* pour tout $x_t$, on en déduit qu'il faut que

$$TA - GT = HC$$

Cette équation est appelée *équation de* Sylvester, et admet une unique solution si et seulement si $A$ et $G$ n'ont aucune valeur propre en commun [HJ12] : puisque $T$ est unique, alors $A$ et $G$ n'ont aucune valeur propre commune.

Réciproquement, si $A$ et $G$ n'ont aucune valeur propre en commun, alors on définit $T$ comme l'unique solution de l'équation $TA - GT = HC$. En remplaçant $HC$ par $TA - GT$ dans (2.5), on obtient alors

$$\begin{aligned} & z_{t+1} = Gz_t + (TA - GT)x_t \\ \Leftrightarrow \quad & z_{t+1} - Tx_{t+1} = G(z_t - Tx_t) \end{aligned}$$

puisque $x_{t+1} = Ax_t$. Cela permet, par récurrence immédiate, d'aboutir à la relation (2.6). □

Ce théorème démontre uniquement l'existence de ce qu'on appelle un estimateur réduit de Luenberger : pour que cette structure soit réellement un estimateur d'état, il est nécessaire que $T$ soit inversible afin de pouvoir obtenir $x$ à partir de $z$. Il existe alors un théorème permettant d'assurer l'inversibilité de $T$.

**Theorème 2.2 (*Inversibilité de T [Lue64]*)**

*Supposons que $A$ et $G$ n'ont aucune valeur propre en commun. Si la paire $(A, C)$ est observable et la*

*paire $(F, G)$ est commandable*, alors l'unique solution $T$ de l'équation $TA - GT = HC$ est inversible.*

Si on multiplie par $T^{-1}$ à gauche et à droite dans (2.6), on obtient

$$x_t = T^{-1}\left(z_t - G^t(z_0 - Tx_0)\right)$$

Or, le terme $(z_0 - Tx_0)$ n'est la plupart du temps pas calculable : en effet, si $z_0$ est un degré de liberté de l'estimateur, il n'est pas possible d'avoir accès à la valeur exacte de l'état initial $x_0$ du vrai système. Cependant, $G$ est également un degré de liberté de l'estimateur : en particulier, la seule contrainte qu'elle doit vérifier pour assurer l'existence de $T$ et d'avoir des valeurs propres différentes de celles de $A$. On peut donc la choisir stable, c'est-à-dire avec toutes ses valeurs propres de module strictement inférieur à 1, ce qui implique que le terme $G^t$ tend vers la matrice nulle quand $t$ tend vers $+\infty$. On en déduit alors que $T^{-1}z_t$ tend asymptotiquement vers $x_t$.

L'estimation $\hat{x}_t$ qu'on va alors prendre du vrai vecteur $x_t$ est

$$\hat{x}_t = T^{-1}z_t$$

Pour des raisons d'implémentation, il est alors intéressant de chercher une relation de récurrence liant $\hat{x}_{t+1}$ et $\hat{x}_t$ : ainsi,

$$\begin{aligned}\hat{x}_{t+1} &= T^{-1}z_{t+1} \\ &= T^{-1}(GT\hat{x}_t + Hy_t)\end{aligned}$$

Or, si $TA - GT = HC$, alors $T^{-1}GT = A - T^{-1}HC$ : on en déduit donc

$$\hat{x}_{t+1} = A\hat{x}_t + L(y_t - \hat{y}_t) \tag{2.7}$$

avec $\hat{y}_t = C\hat{x}_t$ et $L = T^{-1}H$. La structure de l'estimateur apparaît ici comme étant celle d'un système en boucle fermé prenant pour entrée $y_t$ et pour matrice de commande $L$. La figure 2 présente un schéma-bloc de l'estimateur de Luenberger.

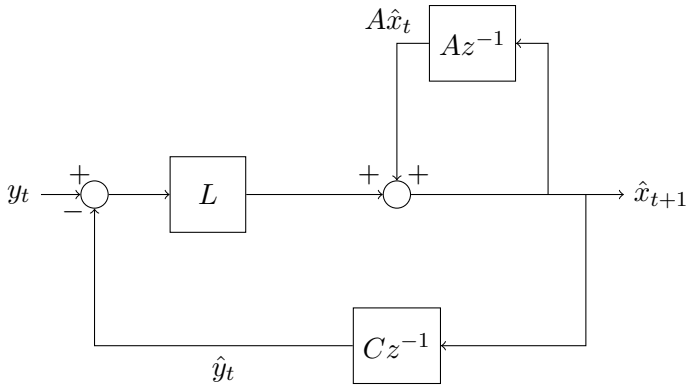



Figure 2 – Schéma-bloc du principe de l'estimateur de Luenberger

La matrice $L$ est donc une matrice particulière, souvent appelée le **gain de l'estimateur**. Comme dans le cas du contrôle, cette matrice est celle qui va corriger la trajectoire de l'estimateur pour la faire coller au plus proche de la vraie trajectoire du système. De plus, si l'on a pu discuter précédemment sur la

∗. La *commandabilité* est une notion complémentaire de l'observabilité qui s'assure qu'on peut amener un système à tous les états possibles en un temps fini. Dans le cas des systèmes LTI, la paire $(F, G)$ est commandable si et seulement si la paire $(F^\top, G^\top)$ est observable.

disposition des valeurs propres de $G$, en pratique, ce sont celles de $L$ qui sont définies par l'opérateur. En effet, comme $T$ est inversible, il est à noter que $G$ et $A-LC$ possèdent les mêmes valeurs propres puisque $A-LC = T^{-1}GT$. On peut donc choisir les valeurs propres de $L$, par la méthode du placement de pôle par exemple, de telle sorte à ce que les valeurs propres de $A-LC$ soient toutes de module strictement inférieur à 1, ce qui permet d'affirmer que l'estimateur converge asymptotiquement vers le vrai état. Par ailleurs, plus les valeurs propres de $A-LC$ sont de modules faibles, plus l'estimateur converge rapidement.

Initialement développés pour les systèmes LTI, de nombreux chercheurs se sont intéressés à l'extension des estimateurs de Luenberger à d'autres systèmes, notamment les systèmes linéaires temps-variant [Oro+18] ou encore le cas encore plus général des systèmes non-linéaires [Afr+17]. Toutefois, même dans ses extensions, l'estimateur de Luenberger possède deux problèmes principaux :

— Aucune notion d'optimalité n'est impliquée dans le choix du gain d'estimateur $L$ : comment être sûr que ce choix est le plus approprié vis-à-vis du cahier des charges fixé par l'opérateur ?

— Aucun bruit n'est pris en compte dans la synthèse de l'estimateur, ce qui peut paraître utopique et assez éloignée de la réalité. Nous n'avons donc aucune garantie théorique quant aux performances de l'estimateur en présence de bruit.

C'est en raison de ces deux défauts principaux que l'on en vient à s'intéresser à l'estimateur des moindres carrés qui synthétise un estimateur par rapport à un critère bien définie et en prenant en compte du bruit.

### 2.2.2 Estimateur des moindres carrés

Pour l'estimateur des moindres carrés, le système considéré est le système LTI en temps discret (bien que les développements qui vont suivre s'étendent facilement aux LTV) suivant

$$\begin{cases} x_{t+1} &= Ax_t \\ y_t &= Cx_t + v_t \end{cases} \quad \text{avec } (A,C) \text{ observable} \tag{2.8}$$

On peut constater l'introduction d'une perturbation en sortie du système : du fait de sa présence, il est nécessaire d'avoir une plus grande exigence quant aux performances de notre estimateur. L'idée est donc d'introduire une fonction coût $\mathcal{C}_t$ de $\mathbb{R}^{n\times nt}$ dans $\mathbb{R}^+$ et de chercher à la minimiser afin de trouver la *meilleure* estimation au regard de ce critère ce critère peut porter sur tout ou une partie de la trajectoire du système, c'est à dire la matrice

$$X_t = \begin{pmatrix} x_0 & x_1 & \cdots & x_t \end{pmatrix},$$

si bien que l'estimation qu'on va obtenir ne sera pas uniquement celle de l'état $x_t$ à l'instant $t$ mais celle de toute la trajectoire $X_t$. L'estimée $\hat{X}_t$ est alors définie par

$$\hat{X}_t \in \min_{Z_t} \mathcal{C}_t(Z_t)$$

Dans le système tel que formulé dans (2.8), on constate que pour tout $t \geq 0$, on a

$$x_t = A^t x_0. \tag{2.9}$$

ce qui entraîne que le vecteur de toutes les mesures du système de 0 à $t$ s'écrit

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_t \end{pmatrix} = \mathcal{O}_t x_0 + \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_t \end{pmatrix} \text{ avec } \mathcal{O}_t = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^t \end{pmatrix} \tag{2.10}$$

Cela permet également de poser la fonction coût $\mathcal{C}'(x_0) = \mathcal{C}(\text{traj}(x_0))$ où $\text{traj}(x_0) = \begin{pmatrix} x_0 & Ax_0 & \cdots & A^t x_0 \end{pmatrix}$. Les deux fonctions auront alors le même minimum, et cela revient au même de minimiser $\mathcal{C}_t$ par rapport à toute la trajectoire ou de minimiser $\mathcal{C}'_t$ par rapport à l'état initial puisque la détermination de ce dernier suffit pour reconstruire entièrement la trajectoire.

En l'occurrence, l'estimateur des moindres carrés consiste à utiliser une fonction coût quadratique de la forme

$$\mathcal{C}'_t(z) = \sum_{k=0}^{t} \|y_k - CA^t z\|^2_{R_k} \tag{2.11}$$

où $z$ est la variable de décision du problème d'optimisation et $\|\cdot\|_{R_k}$ une norme telle que pour tout $z$ dans $\mathbb{R}^n$,

$$\|z\|_{R_k} = z^\top R_k^{-1} z \text{ avec } R_k \text{ matrice symétrique définie positive}$$

Le choix de la matrice $R_t$ a évidemment son importance et dépend du bruit $v_t$ : en effet, dans le cas d'un bruit modélisé par une variable aléatoire, si on la connaît, on utilise la covariance de la variable aléatoire associée au bruit(voire annexe A.1). Cela entraîne que le critère normalise chaque sortie par rapport à la contribution du bruit $v_t$ à cette sortie : ainsi, dans le cas d'une sortie où la variance du bruit est grande, la norme va appliquer un faible coefficient (égal à l'inverse de cette variance), et inversement pour une sortie où le bruit a une variance faible. Le critère fait donc en sorte de ne privilégier aucune sortie et de normaliser leur contribution.

En s'inspirant de la formule (2.10), le critère $C'_t$ peut en réalité se réécrire

$$\mathcal{C}'_t(z) = \|Y_t - \mathcal{O}_t(A,C)z\|^2_{\mathcal{R}_t} \quad \text{avec} \quad Y_t = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_t \end{pmatrix} \quad \text{et} \quad \mathcal{R}_t = \text{diag}(R_0, R_1, \ldots, R_t) \tag{2.12}$$

où $\mathcal{R}_t$ désigne une matrice diagonale par blocs ayant les matrices $R_k$ de 0 à $t$ sur sa diagonale. Cette forme est très connue dans la littérature (comme par exemple [KSH00]), et est un cas particulier pour lequel il est possible d'avoir une expression analytique du minimum. Dans le cas où le système est observable, l'état initial estimé vaut donc

$$\hat{x}_{0|t} = \left(\mathcal{O}_t^\top \mathcal{R}_t^{-1} \mathcal{O}_t\right)^{-1} \mathcal{O}_t^\top \mathcal{R}_t^{-1} Y_t \tag{2.13}$$

Cette formulation est intéressante, mais nécessite de mettre à jour $\mathcal{O}_t$, $\mathcal{R}_t$ et $Y_t$ à chaque instant pour calculer $\hat{x}_{t|0}$. Hors, leur dimension croît en fonction du temps, ce qui est problématique du point de vue des ressources allouées au calcul. Une façon de contourner cela est alors de mettre en place une formule récursive pour calculer $\hat{x}_{t+1}$ en fonction de $\hat{x}_t$.

Pour trouver une telle relation, on va développer la formule de (2.13) pour faire réapparaître $A$, $C$ et les $R_t$ : ainsi, on obtient

$$\hat{x}_{0|t} = \left(\sum_{k=0}^{t} (CA^k)^\top R_k^{-1} CA^k\right)^{-1} \sum_{k=0}^{t} (CA^k)^\top R_k^{-1} y_k.$$

En posant

$$\begin{cases} F_t = \displaystyle\sum_{k=0}^{t} (CA^k)^\top R_k^{-1} CA^k \\ V_t = \displaystyle\sum_{k=0}^{t} (CA^k)^\top R_k^{-1} y_k \end{cases},$$

on en déduit les relations de récurrence

$$\begin{cases} F_t = F_{t-1} + (CA^t)^\top R_t^{-1} CA^t \\ V_t = V_{t-1} + (CA^t)^\top R_t^{-1} y_t \end{cases}.$$

Cela permet enfin d'écrire

$$\hat{x}_{0|t} = F_t^{-1} V_t \tag{2.14}$$

$$= \left(F_{t-1} + (CA^t)^\top R_t^{-1} CA^t\right)^{-1} \left(V_{t-1} + (CA^t)^\top R_t^{-1} y_t\right) \tag{2.15}$$

Or, le lemme d'inversion matricielle, aussi appelé formule de Sherman-Morrison-Woodbury [HJ12], permet d'écrire

$$\left(F_{t-1} + (CA^t)^\top R_t^{-1} CA^t\right)^{-1} = F_{t-1}^{-1} - F_{t-1}^{-1}(CA^t)^\top \left(R_t + CA^t F_{t-1}^{-1}(CA^t)^\top\right)^{-1} CA^t F_{t-1}^{-1}$$

ce qui nous donne

$$\begin{aligned} \hat{x}_{0|t} = \hat{x}_{0|t-1} &- F_{t-1}^{-1}(CA^t)^\top \left(R_t + CA^t F_{t-1}^{-1}(CA^t)^\top\right)^{-1} \hat{y}_{t|t-1} \\ &+ \left[F_{t-1}^{-1} - F_{t-1}^{-1}(CA^t)^\top \left(R_t + CA^t F_{t-1}^{-1}(CA^t)^\top\right)^{-1} CA^t F_{t-1}^{-1}\right] (CA^t)^\top R_t^{-1} y_t \end{aligned}$$

puisque $\hat{x}_{0|t-1} = F_{t-1}^{-1} V_{t-1}$ et que d'après (2.9), la mesure estimée $\hat{y}_t$ s'écrit

$$\hat{y}_t = C\hat{x}_t = CA^t \hat{x}_{0|t-1}$$

Il est alors possible de montrer que

$$\hat{x}_t = A^t \hat{x}_{0|t} = A\hat{x}_{t-1} + A^t F_{t-1}^{-1}(CA^t)^\top \left(R_t + CA^t F_{t-1}^{-1}(CA^t)^\top\right)^{-1} (y_t - \hat{y}_{t|t-1}) \tag{2.16}$$

Cette expression n'est pas pratique, puisqu'elle comprend des puissances de la matrice $A$ qui peuvent être extrêmement lourdes à calculer au fur et à mesure que le temps croît. Cependant, en posant

$$P'_t = A^t F_{t-1}^{-1} (A^t)^\top$$

on montre que l'expression (2.16) équivaut à

$$\hat{x}_t = A\hat{x}_{t-1} + P_{t+1} C^\top R_{t+1}^{-1} (y_{t+1} - \hat{y}_{t+1|t}) \tag{2.17}$$

avec

$$\begin{cases} P'_t = AP_{t-1}A^\top \\ P_t = \left((P'_{t-1})^{-1} + C^\top R_t^{-1} C\right)^{-1} \end{cases} \tag{2.18}$$

Cette forme est alors une forme récursive, puisque l'on obtient des paramètres permettant de calculer $\hat{x}_t$ directement à partir de ceux permettant de calculer $\hat{x}_{t-1}$. De plus, poser

$$L_t = \left((P'_{t-1})^{-1} + C^\top R_t^{-1} C\right)^{-1}$$

permet de se rendre compte que l'estimateur des moindres carrés a en réalité une structure d'estimateur de Luenberger avec un gain d'estimateur temps-variant égal à $L_t$.

Nous sommes donc capable d'obtenir un estimateur optimal par rapport au critère que nous nous sommes fixés. Mais comment gérer le cas où l'on voit également apparaître des perturbations dans l'équation dynamique du système ?

### 2.2.3 Estimateur de Kalman

En 1960, KALMAN [Kal60] propose une structure permettant de réaliser de l'estimation dans des systèmes discrets possédant une représentation d'état linéaire temps-variante (LTV). Si dans le cas historique présenté dans son article, KALMAN s'intéresse au cas purement stochastique (c'est à dire sans entrée déterministe), le cas le plus intéressant pour nous est celui d'une représentation d'état LTV temps discret de la forme

$$\left\{ \begin{array}{rcl} x_{t+1} & = & A_t x_t + w_t \\ y_t & = & C_t x_t + v_t \end{array} \right. \tag{2.19}$$

où $v_t$ et $w_t$ sont des bruits blancs gaussiens indépendants entre eux (voir Annexe A.1). Du fait de la présence de ces bruits, les vecteurs $x_t$ et $y_t$ sont également aléatoires. Connaissant des mesures du système jusqu'à l'instant $t$, l'estimateur de KALMAN a alors pour principe de produire l'estimation $\hat{x}_t$ du vecteur d'état $x_t$ qui minimise la trace de la matrice de covariance de l'erreur d'estimation, c'est-à-dire que

$$\hat{x}_t = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \ \operatorname{tr}(Cov[(x_t - z)(x_t - z)^\top]) \tag{2.20}$$

La matrice $P_t$ définie par

$$P_t = Cov[(x_t - \hat{x}_t)(x_t - \hat{x}_t)^\top]$$

va alors jouer un rôle important dans la synthèse de l'estimée. Bien que pouvant être regroupées en une seule étape, l'estimation de KALMAN se déroule traditionnellement en deux étapes :

— **Prédiction :** à l'instant $t$, la meilleure estimation (au sens du critère (2.20)) du vecteur d'état $x$ à l'instant $t+1$ que l'on peut faire est

$$\hat{x}_{t+1|t} = A_t \hat{x}_t$$

puisque l'action de $w_t$ ne peut être anticipée. Il est également possible de montrer que la matrice de covariance de cette prédiction est

$$P_{t+1|t} = A_t P_t A_t^\top + Q_t$$

où $Q_t$ est la matrice de covariance du bruit $w_t$.

— **Mise à jour :** on prend maintenant connaissance d'une nouvelle mesure du système $y_{t+1}$. L'estimée prédite $x_{t+1|t}$ va donc être mise à jour par un terme correctif dépendant de la différence entre la vraie mesure $y_{t+1}$ et la mesure prédite $\hat{y}_{t+1|t} = C_{t+1}\hat{x}_{t+1|t}$, ce qui donne

$$\hat{x}_{t+1} = \hat{x}_{t+1|t} + P_{t+1} C_{t+1}^\top R_{t+1}^{-1} (y_{t+1} - \hat{y}_{t+1|t})$$

avec

$$P_{t+1} = \left( P_{t+1|t}^{-1} + C_{t+1}^\top R_{t+1}^{-1} C_{t+1} \right)^{-1}$$

où $R_t$ est la matrice de covariance du bruit $v_t$.

Si l'on regroupe l'estimateur de KALMAN en une seule étape, on obtient alors

$$\left\{ \begin{array}{l} \hat{x}_{t+1} = A_t \hat{x}_t + P_{t+1} C_{t+1}^\top R_{t+1}^{-1} (y_{t+1} - C_{t+1} A_t \hat{x}_t) \\ P_{t+1} = \left( \left( A_t P_t A_t^\top + Q_t \right)^{-1} + C_{t+1}^\top R_{t+1}^{-1} C_{t+1} \right)^{-1} \end{array} \right. \tag{2.21}$$

Non seulement on constate que l'estimateur de KALMAN possède une structure d'estimateur de LUENBERGER avec un gain temps-valant $L_t = P_{t+1} C_{t+1}^\top R_{t+1}$. mais on remarque également que l'estimateur obtenu par la méthode des moindres carrés est un cas particulier de celui de KALMAN : en effet, pour $Q_t = 0$, on constate que $P'_t$ dans le cas de l'estimateur des moindres carrés est égal à $P_{t|t-1}$ dans le cas de KALMAN,

et que les deux $P_t$ sont égale. L'estimateur de KALMAN peut donc être considéré comme un estimateur des moindres carrés prenant en compte un bruit au niveau de la dynamique du système.

Cette structure d'estimation connut un essor presque immédiat, notamment dans le domaine de l'aérospatial, où les ingénieurs de la NASA ont été les premiers à l'implémenter dans le cadre du programme Apollo dès 1961 [MS85]. De nombreux scientifiques se sont penchés sur l'estimateur de Kalman, certains s'intéressant à ses performances, comme DEYST et PRICE [DP68] qui se sont penchés sur sa stabilité asymptotique, d'autres à des moyens d'étendre son domaine d'application : c'est ainsi qu'en 1961, aidé de BUCY, KALMAN [KB61] publia un article pour étendre ses travaux aux systèmes linéaires continus. Une théorie, intitulée *Filtre de* KALMAN *étendu*, existe pour traiter le cas des systèmes non linéaires, mais cette dernière ne donne aucune garantie d'optimalité quant à l'estimée obtenue [JU04].

Cependant, le type de bruit pris en compte par les estimateurs présentés ne peut être une représentation fidèle d'un dysfonctionnement tel qu'une défaillance capteur : en effet, ce genre de problème n'arrivant que soudainement, il ne peut être modélisé par une bruit obéissant à une densité de probabilité comme ceux gérés par un estimateur de KALMAN. Il est donc nécessaire de s'intéresser plus spécifiquement à des estimateurs cherchant à prendre en compte ce genre de dysfonctionnement.

## 2.3 Estimation résiliente

Une question revenant souvent dans les problématiques d'estimation est celle du domaine de performance de l'estimateur : quelles caractéristiques le système, et le cas échéant sa modélisation, doivent-ils vérifier pour s'assurer que la structure d'estimation choisie ait les performances attendues ?

De manière générale, on va vouloir qu'un estimateur conserve son niveau de performance en présence d'une certaine quantité de bruit, qui peut très bien modéliser des erreurs de modélisation du système ou de vrais phénomènes physiques non désirés : lorsque c'est le cas, on peut dire d'un tel estimateur qu'il est robuste ou résilient. Cette partie a donc pour objectif d'expliquer la différence entendue entre robustesse et résilience, de justifier le choix du terme « résilient » et de présenter quelques propositions d'estimateurs résilients.

### 2.3.1 Préambule : nuance entre robustesse et résilience

« Robustesse » et « résilience » sont deux termes qui font référence à une résistance de l'estimateur vis-à-vis de perturbations pouvant mettre à mal ses performances. En fonction de leur occurrence, nous avons décidé de les séparer en deux catégories :

— **Perturbations continues :** Certaines perturbations sont inhérentes à un système réel et sont constamment présentes au cours du temps. Elles peuvent correspondre à de nombreux phénomènes, comme un bruit au niveau des capteurs, ou bien l'écart entre le modèle et le système. Du fait de leur caractère ordinaire, ces perturbations ne sont donc pas une menace pour son fonctionnement, mais peuvent cependant diminuer les performances de l'estimateur associé.

— **Perturbations impulsives :** Au sein d'un système peuvent survenir des *événements soudains et parfois brefs* compromettant l'intégrité du système : par exemple, l'apparition d'un court-circuit dans un circuit électrique change drastiquement sa topologie et sort complètement des hypothèses du modèle qu'on a pu formuler au préalable. Existent aussi les attaques provenant d'opérateurs extérieurs et visant à endommager le système.

Pour expliquer la différence entre robustesse et résilience, nous nous sommes donc intéressés à l'utilisation de ces termes dans la littérature et pour quels types de perturbations :

— L'estimation d'état **robuste** est un concept abordé dans la littérature depuis au moins les années 70 [SH77] et vise principalement à synthétiser des estimateurs résistants aux incertitudes ordinaires. L'estimateur de Kalman est d'ailleurs déjà un estimateur robuste, puisqu'il prend en compte des perturbations continues, à savoir des bruits présents au niveau de la dynamique du système et des capteurs, contrairement à l'estimateur de Luenberger.

— L'estimation d'état **résiliente** (ou **sécurisée**) est un concept beaucoup plus récent apparu dans les années 2010 [Paj+14] suite à l'intérêt grandissant pour les systèmes cyber-physiques, ou CPS (*Cyber-Physical Systems*). Ce genre de système étant géré par ordinateur, la sécurité vis-à-vis des attaques informatiques, considérées comme étant une propriété que leurs estimateurs doivent posséder.

Les définitions que nous avons décidé de retenir sont donc les suivantes :

**Robustesse** capacité d'une estimateur à être résistant aux perturbations continues.

**Résilience** capacité d'un estimateur à rejeter les perturbations impulsives pour garder une estimée fiable.

Puisque notre but principal est la mise en œuvre l'estimation centralisée d'un système décentralisé qui peut alors être vulnérable à de nombreuses défaillances impulsives, le terme *résilient* est donc celui retenu. Évidemment, il est toutefois pertinent de parler de la robustesse d'un estimateur résilient, puisque nous pourrons être amenés à considérer des systèmes présentant à la fois des perturbations continues et impulsives.

### 2.3.2 État de l'art sur l'estimation résiliente

En matière d'estimation résiliente, les perturbations impulsives considérées sont souvent des attaques ayant lieu au niveau des capteurs ou des actionneurs : les travaux dans le domaine considèrent donc souvent des perturbations impulsives à ces deux niveaux. Par ailleurs, il est intéressant de réfléchir à la façon dont sont modélisés ces perturbations : en effet, leur caractère empêche toute mise en place d'un modèle décrivant leur valeur, que ce soit d'un point de vue déterministe ou statistique. Par conséquent, les bruits impulsifs seront presque toujours considérés comme étant non bornés. Nous partons cependant du principe que des hypothèses peuvent être faites sur leur fréquence d'occurrence, temporellement ou spatialement : en effet, dans le cas d'une défaillance de capteur, on peut par exemple supposer que les capteurs sont suffisamment fiables pour que peu d'entre eux arrêtent de fonctionner simultanément.

Dans l'article de Fawzi, Tabuada et Diggavi [FTD14], on considère ainsi le cas d'un système LTI de la forme

$$\begin{cases} x_{t+1} &= Ax_t + B(u_t + w_t) \\ y_t &= Cx_t + f_t \end{cases} \tag{2.22}$$

avec des perturbations impulsives $f_t$ et $w_t$ non bornées mais rares. Par ailleurs, il est supposé que l'ensemble des capteurs et des actionneurs sur lesquels $w_t$ et $f_t$ peuvent respectivement agir est fixe au cours du temps. Le problème présenté dans l'article est un problème hors-ligne, c'est-à-dire que possédant $T$ mesures du système, ils cherchent à retrouver l'état initial $x_0$ du système à l'aide d'un estimateur. Les résultats principaux sont :

— Les conditions sous lesquels cet estimateur est capable de retrouver l'état initial à partir de $T$ mesures sachant qu'un nombre donné d'entre elles sont corrompues.

— Une démarche pour synthétiser cet estimateur de manière optimale

Si ces résultats sont très intéressants d'un point de vue théorique, l'hypothèse d'une absence de bruits continus au sein du système paraît forte. De plus, l'obtention de l'estimateur par une synthèse optimale est une démarche adaptée à une problématique hors-ligne, cette dernière ne peut être viable dans le cadre d'une implémentation en ligne.

En 2016, [ST16] reviennent sur ce problème en prenant une formulation similaire en ne considérant cette fois-ci pas de perturbations impulsive au niveau de l'entrée du système. De plus, les hypothèses sur le bruit impulsif ne sont pas les mêmes : il n'est ici plus question d'avoir un ensemble constant de capteurs pouvant être attaqués. Tous peuvent être potentiellement attaqués, mais il existe un nombre maximum de capteurs pouvant être attaqués simultanément. Enfin, le problème ne se fait plus exactement en hors-ligne, mais sur un horizon de temps glissant : en effet, la problématique soulevé dans l'article est de reconstruire une version retardée du vecteur d'état $x_{t-\tau+1}$ (avec $\tau$ le retard ainsi que les attaques $f_{t-\tau+1}$,$f_{t-\tau+2}$,...,$f_t$ à partir des $\tau$ dernières mesures $y_{t-\tau+1}$,$y_{t-\tau+2}$,...,$y_t$. Afin de réaliser cela, les auteurs proposent une synthèse optimal d'un estimateur de Luenberger par la minimisation d'un critère : cependant, ce problème n'étant pas bien défini, il peut mener à plusieurs estimations différentes possibles. Ils discutent alors des hypothèses devant vérifier le système afin que le problème ait une unique solution.

Bien que le problème à horizon glissant soit plus adapté à une implémentation en ligne de l'estimateur, le fait que ce soit non pas $x_t$ mais une version retardée de $x_t$ qui soit reconstruite à l'instant $t$ pose des problèmes pour la commande à retour d'état, puisque si le retard est trop grand, le système aura pu potentiellement changer complètement d'état. De plus, l'absence de bruits continus reste une situation sans doute trop idéale.

Ainsi,Mishra [Mis+17] publie en 2017 un article prenant cette fois-ci en compte d'éventuels perturbations continues dans un système de la forme

$$\begin{cases} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + v_t + f_t \end{cases} \tag{2.23}$$

avec $w$ et $v$ des bruits blancs gaussiens centrés indépendants entre eux. Le bruit impulsif combine alors les hypothèses des deux articles précédents : il ne peut agir que sur un nombre fixe de capteurs et les capteurs sur lesquels il agit ne varient pas dans le temps. L'objectif est alors de bien estimer le système sur un horizon de temps compris entre $t_1$ et $t_1 + N - 1$ avec $N$ entier.

La stratégie utilisée est de reconstruire l'état à partir d'un estimateur de Kalman se basant sur un ensemble de mesures non corrompues : pour cela, l'auteur propose dans un premier temps d'étudier un indicateur d'attaque pour un ensemble donné de capteurs. Un algorithme est alors proposé : il estime l'état du système de $t = 0$ jusqu'à $t_1 - 1$ puis comparer la différence entre les vraies mesures $y_t$ et les mesures prédites $\hat{y}_t$ (sans nouvel apport d'information) pour tout $t$ entre $t_1$ et $t_1 + N - 1$. Cette différence est alors comparée à la valeur moyenne attendue, et si la différence est plus grande, l'algorithme renvoie une valeur.

L'idée est alors de chercher un ensemble des capteurs non touchés par le bruit impulsif afin d'avoir une bonne estimation à l'aide d'un estimateur de Kalman réalisé avec les mesures provenant de ces capteurs. Des discussions sont alors menées dans l'article afin de s'assurer que l'indicateur renvoyé par l'algorithme est suffisamment fiable (en jouant notamment sur la taille de la fenêtre de prédiction $N$). Cependant, il est évident que cette implémentation pose des problèmes de besoins en calculs. Tout d'abord, cette solution n'est réalisable qu'en hors-ligne, puisque si $t_1$ grandit, l'algorithme va mettre de plus en plus de temps à s'effectuer. De plus, il est prévu que l'algorithme ait lieu plusieurs fois jusqu'à trouver le bon ensemble de capteurs. Même si les auteurs discutent d'une méthode pour trouver plus rapidement cet ensemble, l'algorithme reste extrêmement lourd en ressources calculatoires.

## 2.4 Objectifs

On cherche à mettre en œuvre un estimateur sur une unité de calcul centralisant des mesures provenant de capteurs délocalisés : du fait de la communication entre capteurs et unité de calcul, cet estimateur est

vulnérables à de nombreuses perturbations pouvant être modélisées par des bruit impulsifs, que ce soit des attaques, des défaillances de capteur, etc. Du fait de la nature numérique du calculateur, l'estimateur devra être formulé **en temps discret**, et par conséquent le modèle de notre système aussi. La forme générale du modèle sera alors de la forme

$$\begin{cases} x_{t+1} &= A_t x_t + B_t u_t + \omega_t \\ y_t &= C_t x_t + D_t u_t + \nu_t \end{cases} \tag{2.24}$$

où la nature exacte des perturbations $\omega$ et $\nu$ n'est pas supposée explicitement : cependant, pour toute perturbation quelconque, on peut toujours la scinder en deux parties, une continue et une impulsive. Du fait de la vulnérabilité de l'estimateur aux attaque de capteurs, on va en particulier souvent considérer que la perturbation $\nu_t$ se décompose en deux parties

$$\nu_t = v_t + f_t$$

avec $v$ perturbation continue et $f$ perturbation impulsive. $v$ pourra être modélisé par un signal aléatoire, ou on pourra supposer la connaissance d'une borne supérieure pour ce bruit. $f$ sera un bruit non borné et admettant une seule hypothèse, sur la rareté spatiale et temporelle de ses occurrences, qu'il conviendra d'expliciter pour mettre en place des critères qui assurent les performances de notre estimateur. De plus, du fait de la décomposition de $\nu$ (qui n'est pas unique), on peut poser $f_t = 0$ lorsque le capteur n'est pas attaqué. Les caractéristiques que devra réunir l'estimateur sont alors les suivantes

- Sa mise en œuvre devra assurer une certaine optimalité.
- Si l'objectif final est d'implémenter le problème en ligne avec un coût de synthèse et de calcul raisonnable, nous envisagerons aussi le cas hors-ligne.
- L'estimateur devra se montrer robuste aux perturbations continues qui peuvent être présentes dans la dynamique du système ou au niveau de ses mesures.
- Il devra garder un bon niveau de performance en présence de bruit impulsif.

Au vu de l'étude bibliographique présentée précédemment, aucun article ne répond complètement à ce cahier des charges. Nous allons donc essayer de définir un nouvel estimateur afin d'y répondre au mieux.

# 3 Résilience aux bruits impulsifs : méthode d'estimation optimale

Comme présenté dans la sous-partie 2.2, il est possible de définir une estimation comme la minimisation d'un critère qui retranscrit nos attentes vis-à-vis de l'estimation. Les critères étudiés jusque là étant principalement composés de fonctions quadratiques, nous allons essayer dans cette partie de proposer un nouveau critère pour l'estimation.

## 3.1 Formulation du problème d'optimisation

On se place dans le cas d'un système LTV en temps discret de représentation d'état 2.2.

Comme pour l'estimateur des moindres carrés, il paraît normal de partir sur une fonction coût qui prend en compte toute la trajectoire $X_t = \begin{pmatrix} x_0 & x_1 & \cdots & x_t \end{pmatrix}$ du système. Du fait de la présence des bruits à la fois au niveau de la dynamique et des mesures du système, le critère devra donc comporter ces éléments en son sein. Ainsi, on définit le critère suivant

$$V_t(Z_t) = \chi(z_0 - \mu_0) + \sum_{k=0}^{t-1} \phi_k(z_{k+1} - A_k z_k - B_k u_k) + \sum_{k=1}^{t} \psi_k(y_k - C_k z_k) \tag{3.1}$$

avec $Z_t = \begin{pmatrix} z_0 & z_1 & \cdots & z_t \end{pmatrix} \in \mathbb{R}^{n \times t+1}$ où pour tout $k$, $z_k$ est un vecteur de $\mathbb{R}^n$, et $\phi_k : \mathbb{R}^n \to \mathbb{R}^+$ et $\psi_k : \mathbb{R}^{n_y} \to \mathbb{R}^+$ sont des fonctions coût convexes. $Z_t$, dans $\mathbb{R}^{n \times t}$, est donc la variable de décision de la fonction coût objectif $V_t$ et représente une trajectoire possible du système.

On peut noter la présence d'un terme optionnel servant à évaluer la distance entre l'état initial de la trajectoire $Z_t$ et un certain vecteur $\mu_0$. En effet s'il est généralement impossible de connaître avec certitude l'état initial $x_0$ d'un système, il est par exemple possible de connaître un ensemble borné de $\mathbb{R}^n$ dans lequel ce dernier est censé se trouver. $\mu_0$, pouvant être pris au hasard dans cette zone pour ce cas, sert donc en réalité à traduire cette connaissance. Cependant, tout le terme $\chi(z_0 - \mu_0)$ peut éventuellement être omis dans le cas où aucune connaissance sur l'état initial du système n'est supposée.

Nous avons supposé que $u$ était une entrée connue du système. Pour des soucis de clarté, nous considèrerons le cas du système libre c'est-à-dire $u_t = 0$ pour tout $t$, ce qui permet de simplifier l'expression de $V_t$ en

$$V_t(Z_t) = \chi(z_0 - \mu_0) + \sum_{k=0}^{t-1} \phi_k(z_{k+1} - A_k z_k) + \sum_{k=1}^{t} \psi_k(y_k - C_k y_k). \tag{3.2}$$

On définit finalement la trajectoire estimée $\hat{X}_t$ comme la solution du problème d'optimisation

$$\text{Trouver } \hat{X}_t \in \underset{Z_t}{\operatorname{argmin}}\ V_t(Z_t). \tag{3.3}$$

Cela peut s'interpréter comme : $\hat{X}_t$ est la trajectoire la plus probable parmi toutes les trajectoires hypothétiques $Z_t$.

Pour tout $t$, la résolution directe du problème d'optimisation entraîne la rééstimation de l'état du système à partir de $t = 0$. Par conséquent, on peut écrire $\hat{X}_t$ sous la forme

$$\hat{X}_t = \begin{pmatrix} \hat{x}_{0|t} & \hat{x}_{1|t} & \cdots & \hat{x}_t \end{pmatrix} \tag{3.4}$$

où, pour tout $k$ entre 0 et $t$, $\hat{x}_{k|t}$ est une estimation de $x_k$ prenant en compte les mesures reçues jusqu'à l'instant $t$.

Se pose alors de savoir si le problème est bien posé : le problème d'optimisation admet-il toujours une solution, et si oui, est-elle unique ? Tout d'abord, une hypothèse de convexité a été faite sur les fonctions coût qui composent $V_t$ : on en déduit que $V_t$ est une fonction convexe. Cela signifie que s'il existe un minimum local de $V_t$, alors ce dernier est un minimum global. Cependant, cela ne suffit pas à garantir l'existence d'un minimum. Pour cela, il conviendra de vérifier que la fonction $V_t$ est coercive, c'est-à-dire que

$$\lim_{\|Z_t\| \to +\infty} V_t(Z_t) = +\infty$$

Une fonction coercive se comporte de la même manière qu'une norme, tendant vers plus l'infini quand la norme de son argument tend vers plus l'infini. De plus, on en déduit qu'un minimum de $V_t$ existe toujours et qu'il est global : cependant, si le minimum est unique, l'ensemble des arguments minimisants possibles, c'est-à-dire l'ensemble $\operatorname{argmin}_{Z_t} V_t(Z_t)$ peut ne pas être réduit à singleton : cependant, du fait de la coercivité, cet ensemble arguments minimisant doit être borné.

De manière évidente, les performances de l'estimateur sont directement dépendantes des fonctions $\chi$, $\phi_k$ et $\psi_k$ (dont le choix, en présence de bruit impulsifs, sera discuté en sous-partie 3.2). Un des enjeux principaux va être d'énoncer des conditions que doit vérifier le système ainsi que les fonctions coût pour que l'estimateur ainsi défini produise une estimation suffisamment proche du vrai état : pour quantifier cette proximité, on va souvent s'intéresser à l'erreur d'estimation qui peut s'écrire

$$E_t = \begin{pmatrix} \hat{x}_{0|t} - x_0 & \hat{x}_{1|t} - x_1 & \dots & \hat{x}_t - x_t \end{pmatrix} = \begin{pmatrix} e_{0|t} & e_{1|t} & \dots & e_t \end{pmatrix}$$

et l'objectif sera d'en trouver une borne supérieure, voire d'en connaître la limite quand $t$ tend vers l'infini.

Enfin, il est important de noter que la résolution du problème d'optimisation (3.3) n'est pas forcément adaptée en pratique. En effet, si l'on est à $t$ fixé (**problèmes hors-ligne**), alors la mise en œuvre est avant tout un problème de faisabilité : les seules limites pour obtenir la meilleure estimée au sens de notre fonction coût sont les limites de l'optimisation convexe. Dans le cas où $t$ n'est pas fixe et peut potentiellement tendre vers l'infini (**problème en ligne**), un problème de ressources calculatoires se pose également. Tel qu'il est posé, le problème d'optimisation exige de recalculer l'ensemble de la trajectoire qui croît à chaque $t$. La croissance de $t$ entraîne alors une difficulté croissante à obtenir les estimées du fait de les tailles des matrices en jeu. Il est donc nécessaire de trouver une implémentatio plus adaptée au cas en ligne.

Dans la réalité, les états ayant eu lieu longtemps auparavant ne jouent pas de rôle dans la valeur de l'état à l'instant présent : c'est pourquoi nous pourrons être amenés à considérer une autre forme de la fonction coût $V_t$

$$V_t(Z_t) = \chi(z_0 - \mu_0) + \sum_{k=0}^{t-1} \lambda^{t-1-k} \phi_k(z_{k+1} - A_k z_k) + \sum_{k=1}^{t} \lambda^{t-k} \psi_k(y_k - C_k y_k). \tag{3.5}$$

avec $\lambda$ un coefficient réel compris entre 0 et 1 appelé **facteur d'oubli** : la forme de $V_t$ est faite de telle sorte à pondérer les fonctions coût $\phi_k$ et $\psi_k$ par un coefficient inversement proportionnel à leur ancienneté. Plus l'instant $k$ sera ancien, moins $\phi_k$ et $\psi_k$ auront de poids dans $V_t$, avec pour convention que $\phi_t$ et $\psi_t$ ont une pondération égale à 1. Pour faire la distinction entre les deux formes de $V_t$, nous utiliserons les termes « fonction coût avec/sans facteur d'oubli ».

Dans la suite de cette partie, nous aborderons la question de quelles fonctions coûts mettre en place pour traiter le cas des perturbations impulsives puis nous donnerons des éléments sur l'analyse de cet

estimateur en hors-ligne et en ligne pour le cas des bruits continus. Enfin, nous parlerons de la suite de l'analyse et des problématiques de mise en œuvre.

## 3.2 Fonctions coûts pour la gestion des perturbations impulsives

Un bruit impulsif n'étant présent que dans l'équation de mesure du système (2.2) étudié, c'est uniquement au niveau des $\psi_k$ que ce dernier va intervenir dans l'expression (3.2) de $V_t$. Si l'objectif final est de trouver des propriétés que les fonctions coûts doivent vérifier sans forcément expliciter leur forme exacte, il peut-être intéressant de voir quelles fonctions coûts ont déjà été envisagées pour traiter ce type de problème.

Dans cette partie, nous allons donc nous focaliser sur la somme

$$\sum_{k=1}^{t} \psi_k(y_t - C_t z_t) \tag{3.6}$$

en supposant également que la perturbation de mesure $\nu$ correspond uniquement à une perturbation impulsive $f$, c'est-à-dire que pour tout $k$, $\nu_k = f_k$. Une interprétation possible d'un bruit impulsif est de dire qu'il est creux : en effet, si on considère la matrice

$$\begin{pmatrix} f_1 & f_2 & \cdots & f_t \end{pmatrix}$$

alors celle-ci doit être creuse du fait que ses lignes et ses colonnes sont creuses pour représenter la rareté spatiale et temporelle du bruit. De plus, on sait que pour le vrai état, on a

$$y_t - C_t x_t = f_t$$

Ainsi, lorsqu'on cherche à minimiser (3.6) par rapport à $Z_t$, cela équivaut à minimiser

$$\sum_{k=1}^{t} \psi_k(m_t) \tag{3.7}$$

par rapport à

$$M_t = \begin{pmatrix} m_0 & m_1 & \cdots & m_t \end{pmatrix}$$

avec pour tout $k$, $m_k \in \mathcal{Z}_t$ où

$$\mathcal{Z}_t = \{m \in \mathbb{R}^{n_y} \setminus \exists z \in \mathbb{R}^n \text{ tel que } m = y_t - C_t z\}$$

Cette formulation est évidemment moins pratique à mettre en œuvre mais permet de comprendre une qualité que doivent avoir les fonctions $\psi_k$ : elles doivent encourager la parcimonie, c'est-à-dire que leur minimisation doit naturellement privilégier des arguments minimisants contenant beaucoup de zéros. En effet, on veut se rapprocher au mieux de $F_t$, la matrice $M_t$ solution du problème d'optimisation (3.7) doit être la plus creuse possible.

Une première idée serait de prendre la norme $\ell_0$ définie par

$$\forall m \in \mathbb{R}^{n_y}, \qquad \|m\|_{l_0} = \text{card}(\text{supp}(m))$$

où $\text{supp}(m)$ désigne le support d'un vecteur $m$,c'est-à-dire l'ensemble de ses éléments non nuls, et $\text{card}(\mathcal{E})$ désigne la cardinalité d'un ensemble $\mathcal{E}$, à savoir son nombre d'éléments. Cette norme est donc égale au

nombre de composantes non nulle du vecteur : minimiser par rapport à cette dernière permettrait de s'assurer d'avoir le bruit simulé le plus creux possible.

Le problème principal de cette norme est qu'elle n'est pas convexe, ce qui empêche toute utilisation de la théorie de l'optimisation convexe. Cependant, et comme cela est présenté par Candes, Wakin et Boyd [CWB07], la norme $\ell_1$ définie par

$$\forall m = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_{n_y} \end{pmatrix} \in \mathbb{R}^{n_y}, \qquad \|m\|_{\ell_1} = \sum_{i=0}^{n_y} |\mu_i|$$

encourage également la parcimonie tout en étant convexe, et ce bien plus que la norme euclidienne (aussi appelée norme $\ell_2$) définie par

$$\forall m = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_{n_y} \end{pmatrix} \in \mathbb{R}^{n_y}, \qquad \|m\|_{\ell_2} = \sqrt{\sum_{i=0}^{n_y} \mu_i^2}.$$

Un exemple de cette parcimonie est développé en Annexe A.2. Toutefois, ce n'est peut-être pas le seul type de fonction coûts convexe le permettant, c'est pourquoi nous cherchons avant tout à dégager des propriétés que les fonctions coût doivent vérifier plutôt que des formes explicites de ces fonctions coût.

## 3.3 Cas hors-ligne : présentation et analyse

Dans cette sous-partie, on se place donc dans le cas où, à $t$ fixé, on estime la totalité de la trajectoire $X_t$ de 0 à $t$

$$\hat{X}_t = \begin{pmatrix} \hat{x}_{0|t} & \hat{x}_{1|t} & \dots & \hat{x}_{t|t} \end{pmatrix}$$

Comme on se place dans un cas où toute la trajectoire du système nous intéresse, nous utiliserons la formulation (3.2) de $V_t$ sans facteur d'oubli et sans connaissance sur l'état initial du système

$$V_t(Z_t) = \sum_{k=0}^{t-1} \phi_k(z_{k+1} - A_k z_k) + \sum_{k=1}^{t} \psi_k(y_k - C_k z_k).$$

On cherche alors à trouver une borne supérieure de la norme du vecteur d'erreur d'estimation

$$E_t = \begin{pmatrix} \hat{x}_{0|t} - x_0 & \hat{x}_{1|t} - x_1 & \dots & \hat{x}_t - x_t \end{pmatrix} = \begin{pmatrix} e_{0|t} & e_{1|t} & \dots & e_t \end{pmatrix}$$

afin de s'assurer que l'estimation est toujours à une distance bornée de la vraie trajectoire.

En hors ligne, sauf mention contraire, $e_i$ signifiera toujours $e_{i|t}$.

Afin de mener l'analyse en hors ligne, on réintroduit tout d'abord la définition d'une fonction de classe $\mathcal{K}_\infty$, classe très utilisée dans le domaine de l'automatique [Kha01] :

**Définition 3.1 (*Fonction de classe $\mathcal{K}_\infty$ [Kha01]*)**

*Une fonction $g : \mathbb{R}^+ \to \mathbb{R}^+$ est dite de classe $\mathcal{K}_\infty$ si et seulement si :*

1. $g(0) = 0$
2. *$g$ est strictement croissante*

3. $\lim_{\lambda \to +\infty} g(\lambda) = +\infty$

A partir de cette définition, nous sommes en mesure de définir une classe de fonction :

**Définition 3.2 (*Fonction de classe $\mathcal{N}^a$ et de classe $\mathcal{N}^a_\gamma$*)**

*Une fonction $\xi : \mathbb{R}^a \to \mathbb{R}^+$ est dite de classe $\mathcal{N}^a$ si elle vérifie les points suivants :*

1. *$\xi$ est continue définie positive.*
2. *$\xi$ est une fonction paire*
3. *il existe une fonction $q : \mathbb{R}^+ \to \mathbb{R}^+$ de classe $\mathcal{K}_\infty$ telle que pour tout $\lambda$ dans $\mathbb{R}$ et tout $z$ dans $\mathbb{R}^a$,*

$$\xi(z) \geq q\left(\frac{1}{|\lambda|}\right)\xi(\lambda z) \tag{3.8}$$

*Si, de plus, il existe un réel positif $\gamma > 0$ tel que pour tout $z_1$, $z_2$ dans $\mathbb{R}^a$*

$$\xi(z_1 - z_2) \geq \gamma\xi(z_1) - \xi(z_2) \tag{3.9}$$

*on dit que $\xi$ est de classe $\mathcal{N}^a_\gamma$.*

Comme $q$ n'est pas nécessairement unique, il peut être intéressant de préciser la fonction avec laquelle $\xi$ vérifie (3.8) : on précisera donc parfois qu'une fonction $\xi$ est « $\mathcal{N}^a_\gamma$ pour la fonction $q$ ».

L'introduction de ces deux nouvelles classes de fonctions a pour but de généraliser les normes habituellement utilisées pour faire de l'estimation optimale. En particulier, la norme $\ell_2$ est convexe, paire, et continue. De plus, toute norme vérifie la propriété d'homogénéité

$$\|x\| = \frac{1}{|\lambda|}\|\lambda x\|$$

dont (3.8) est une généralisation. Enfin, toutes les normes vérifient par définition l'inégalité triangulaire : or, la relation (3.9) en est une version généralisée car pour $\gamma = 1$, on retombe sur l'inégalité triangulaire originale.

Nous allons donc dorénavant supposer que $\phi_k$ et $\psi_k$ appartiennent à cette classe : plus précisément, on suppose que pour tout $k$, il existe $\alpha_k > 0$, $\beta_k > 0$, ainsi que $g_k$ et $h_k$ de classe $\mathcal{K}_\infty$ tels que $\phi_k$ est de classe $\mathcal{N}^n_{\alpha_k}$ pour $g_t$ et $\psi_k$ est de classe $\mathcal{N}^{n_y}_{\beta_k}$ pour $h_k$.

Une propriété intéressante de ces classes est qu'il existe une inégalité liant les fonctions de classe $\mathcal{N}^a$ à des normes :

**Lemme 3.1**

*Soit $\xi$ une fonction de classe $\mathcal{N}^a$ pour $q \in \mathcal{K}_\infty$. On a alors*

$$\forall z \in \mathbb{R}^a, \qquad \xi(z) \geq Dq(\|z\|) \tag{3.10}$$

*avec $D = \min_{\|z\|=1} \xi(z)$.*

**Démonstration.** On considère l'ensemble $\mathcal{S}$ défini par

$$\mathcal{S} = \{z \in \mathbb{R}^a \backslash \quad \|z\| = 1\} \tag{3.11}$$

$\mathcal{S}$ est compact (fermé borné) donc les extremums de la fonction $\xi$ continue sur $\mathcal{S}$ sont atteints, *i.e.* il existe $z^\star$ tel que

$$\forall z \in \mathcal{S}, \qquad \xi(z) \geq \xi(z^\star) \tag{3.12}$$

Posons $D = \xi(z^\star)$. On sait que pour tout $z$ dans $\mathbb{R}^a$, $z/\|z\|$ appartient à $\mathcal{S}$. Or, d'après (3.8),

$$\xi(z) \geq q(\|z\|)\xi\left(\frac{z}{\|z\|}\right) \tag{3.13}$$

ce qui nous donne directement le résultat (3.10). □

L'introduction de ces classes et du lemme qui en découle nous a permis de développer le résultat suivant :

**Theorème 3.1 (*Borne supérieure de la norme de l'erreur d'estimation*)**

*Soit un système défini en (2.2). Si pour tout $k$, $\phi_k$ est de classe $\mathcal{N}^n_{\alpha_k}$ pour $g_t$ et $\psi_k$ est de classe $\mathcal{N}^{n_y}_{\beta_k}$ pour $h_k$, alors la norme de l'erreur d'estimation de l'estimée obtenue par la résolution du problème (3.3) est bornée par*

$$\|E_t\| \leq q_t^{-1}\left(\frac{2b_t}{D}\right) \tag{3.14}$$

*avec*

$$b_t = \sum_{k=0}^{t-1} \phi_k(\omega_k) + \sum_{k=1}^{t} \psi_k(\nu_k) \tag{3.15}$$

$$D = \min_{\|Z_t\|=1} \left\{ \sum_{k=0}^{t-1} \alpha_k \phi_k(z_{k+1} - A_k z_k) + \sum_{k=1}^{t} \beta_k \psi_k(C_k z_k) \right\} \tag{3.16}$$

$$\forall \lambda \in \mathbb{R},\ q_t(\lambda) = \min_{k \in [0;t]} \{g_k(\lambda), h_k(\lambda)\} \tag{3.17}$$

Une démonstration de cette borne est présentée en Annexe A.3.

Cette relation est particulièrement intéressante pour montrer qu'en présence de bruits bornés uniquement, l'erreur d'estimation sera toujours bornée *a priori* : En effet, si on considère, par exemple, que les perturbations $\omega$ et $\nu$ sont bornées, alors la quantité à droite de l'inégalité va être bornée (à $t$ fixé). De plus, cette relation est valable pour n'importe quelle norme : ainsi, pour l'étudier, il pourra être intéressant de considérer plusieurs normes. Cependant, on constate que $D$ dépend de la norme choisie.

Cependant, il reste encore à étudier plus en détail cette borne, notamment pour savoir si elle très large ou plutôt resserrée en pratique. Par ailleurs, nous travaillons pour adapter cette analyse au bruits impulsifs.

## 3.4 Cas en ligne : utilisation du *Forward Dynamic Programming*

Dans cette sous-section, on considère le problème en ligne : par conséquent, on cherche à estimer l'état du système parallèlement à l'obtention des mesures par les capteurs. Comme nous l'avons vu précédemment, la minimisation direct de la fonction coût $V_t$ n'est pas adaptée, car la complexité de ce problème d'optimisation tend vers l'infini au fur et à mesure que le temps avance.

Pour contourner ce problème, l'objectif est de le reformuler afin de trouver une fonction coût qui ne dépend que de l'état qu'on cherche à estimer : pour cela, on va faire appel à la méthode du *Forward Dynamic Programming*.

Par ailleurs, comme c'est l'état à l'instant $t$ qui nous intéresse tout particulièrement, nous allons considérer le problème (3.3) pour $V_t$ avec facteur d'oubli mais avec une connaissance de l'état initial du système, soit

$$V_t(Z_t) = \chi(z_0 - \mu_0) + \sum_{k=0}^{t-1} \lambda^{t-1-k} \phi_k(z_{k+1} - A_k z_k) + \sum_{k=1}^{t} \lambda^{t-k} \psi_k(y_k - C_k y_k).$$

### 3.4.1 Principe du *Forward Dynamic Programming*

La programmation dynamique (ou *Dynamic Programming* en anglais) est une théorie proposée notamment par Richard Bellman [Bel54] dans les années cinquante pour résoudre des problèmes d'optimisation. Le principe fondamental est de remplacer un problème d'optimisation par des sous-problèmes de complexité moindre.

Suivant l'idée d'obtenir une fonction coût ne dépendant que d'un vecteur, on définit une fonction coût en ligne $V_t^*$ telle que

$$V_t^*(z) = \min_{\substack{Z_t \\ z_t = z}} V_t(Z_t). \tag{3.18}$$

$V_t^*(z)$ est la plus petite valeur de $V_t(Z_t)$ sur l'ensemble des trajectoires $Z_t$ *aboutissant en* $z$ *à l'instant* $t$. Ainsi, lorsque $V_t^*$ est minimisée par rapport à $z$,

$$\min_z V_t^*(z) = \min_z \left\{ \min_{\substack{Z_t \\ z_t = z}} V_t(Z_t) \right\} = \min_{Z_t} V_t(Z_t) \tag{3.19}$$

du fait que l'ordre des variables de minimisation n'a pas d'importance. Ainsi, on en déduit que $V_t$ et $V_t^*$ possèdent le même minimum, et celui-ci existe et est unique du fait que $V_t$ est convexe coercive. Il existe donc au moins un état minimisant $V_t^*$ : cet état correspond alors à l'état présent dans une des trajectoires $\hat{X}_t$ minimisant $V_t$ (on rappelle que $\operatorname{argmin}\{V_t\}$ n'est pas forcément réduit à un singleton).

Ainsi, nous pouvons écrire une première propriété de $V_t^*$ :

**Propriété 1.** Soit un système (2.2) avec $V_t$ et $V_t^*$ définies par (3.2) et (3.18) respectivement. $z$ appartient à $\operatorname{argmin}_z V_t^*(z)$ si et seulement s'il existe une trajectoire dans $\operatorname{argmin}_{Z_t} V_t(Z_t)$ telle que $z_t = z$.

Cette propriété montre tout l'intérêt de la fonction coût $V_t^*$ : en effet, on va maintenant considérer le problème

$$\text{Trouver } \hat{x}_t \in \operatorname*{argmin}_z V_t^*(z) \tag{3.20}$$

comme notre problème en ligne principal. D'après la propriété 1, si on résout le problème (3.20) et obtient $\hat{x}_t$, alors il existe une trajectoire $\hat{X}_t$ aboutissant en $\hat{x}_t$ à l'instant $t$ et solution du problème 3.3. Cela va donc dans le bon sens, puisque l'on obtient la même estimée mais à partir de la minimisation d'une fonction coût ne portant que sur un seul vecteur d'état.

Cependant, ce constat n'est pas suffisant : en effet, si le problème a été résolu en apparence, il n'empêche que la définition de $V_t^*$ contient elle aussi un problème d'optimisation, et qui lui admet un nombre de variables de décision qui croît avec le temps. L'intérêt du *Forward Dynamic Programming* réside en réalité dans l'existence d'une relation de récurrence liant $V_t^*$ et $V_{t-1}^*$, intitulée *équation de Bellman* :

**Theorème 3.2 (*Equation de* Bellman)**

*Soit $V_t^*$ définie comme en (3.18). Alors pour tout $t > 0$,*

$$V_t^*(z) = \min_s \left\{ \lambda V_{t-1}^*(s) + \phi_{t-1}(z - A_{t-1}s) \right\} + \psi_t(y_t - C_t z)) \tag{3.21}$$

Une démonstration de cette relation de récurrence est présentée en Annexe A.4 : on constate par ailleurs que le cas sans facteur d'oubli se déduit à partir de ce théorème en prenant $\lambda = 1$.

Grâce à cette relation de récurrence, on voit que le problème (3.20) semble plus faisable que le problème (3.3), puisqu'il demande la résolution de deux problèmes d'optimisation, chacun portant sur $n$ variables de décision. On constate donc que le nombre de variables de décision n'augmente pas avec le temps.

En particulier, cette forme particulière permet de retrouver l'estimateur de Kalman dans le cas où les fonctions coûts $\phi_k$ et $\psi_k$ sont des formes quadratiques.

**Theorème 3.3 (*Forward Dynamic Programming avec fonctions coûts quadratique*)**

*Soit un système LTV (2.2). Considérons le critère (3.5) avec $\lambda = 1$ et pour tout $k$, $\chi$, $\phi_k$ et $\psi_k$ définis par*

$$\forall z \in \mathbb{R}^n, \quad \chi(z) = \frac{1}{2} z^\top S^{-1} z \tag{3.22}$$

$$\forall z \in \mathbb{R}^n, \quad \phi_k(z) = \frac{1}{2} z^\top Q_k^{-1} z \tag{3.23}$$

$$\forall z \in \mathbb{R}^{n_y}, \quad \psi_k(z) = \frac{1}{2} z^\top R_k^{-1} z. \tag{3.24}$$

*avec $S$, $Q_k$ dans $\mathcal{S}_n^+(\mathbb{R})$ et $R_k$ dans et $\mathcal{S}_{n_y}^+(\mathbb{R})$ respectivement. La fonction $V_t^*$ définie par (3.18) peut alors s'écrire sous la forme*

$$V_t^*(z) = \frac{1}{2}(z - \hat{x}_t)^\top P_t^{-1}(z - \hat{x}_t) + r_t$$

*avec*

$$\hat{x}_{t+1} = A_t \hat{x}_t + P_{t+1} C_{t+1}^\top R_{t+1}^{-1}(y_{t+1} - C_{t+1} A_t \hat{x}_t) \tag{3.25}$$

$$P_{t+1} = \left( (Q_t + A_t P_t A_t^\top)^{-1} + C_{t+1}^\top R_{t+1}^{-1} C_{t+1} \right)^{-1} \tag{3.26}$$

*et $r_t$ une quantité définie indépendante de $z$. Par ailleurs, $\hat{x}_t$ est l'unique solution du problème (3.20) pour tout $t$.*

Une démonstration de ce théorème est présenté en Annexe A.5.

On a donc l'assurance de manipuler une classe d'estimateurs optimaux qui englobe les estimateurs de Kalman. Cependant, il reste un dernier obstacle à l'implémentation. En effet, pour être exploitée, la relation (3.21) requiert implicitement une expression analytique de $V_t^*$ dépendant à la fois de $z$ et de $t$. Or, ceci ne peut être garanti avec le peu d'hypothèses qui ont été faites sur $\chi$, $\phi_k$ et $\psi_k$. Le *Forward Dynamic Programming* aura en quelque sorte *inversé le problème* : nous sommes passés d'un problème non implémentable en ligne mais avec une expression claire de la fonction coût à un problème conceptuellement facilement implémentable mais dont la fonction coût est purement théorique.

### 3.4.2 Piste d'analyse

Notre piste principale d'analyse est d'utiliser la fonction $V_t^*$ pour borner l'erreur d'estimation $e_t = x_t - \hat{x}_t$ : en effet, on a l'intuition que lorsque le problème est « bien posé », la différence entre $V_t^*(x_t)$ et $V_t^*(\hat{x}_t)$ est censée refléter la distance entre $x_t$ et $\hat{x}_t$, c'est-à-dire l'erreur d'estimation.

La stratégie se décompose en deux parties :

1. Trouver une borne supérieure pour $V_t^*(x_t) - V_t^*(\hat{x}_t)$
2. Trouver un lien entre la norme de l'erreur d'estimation et $V_t^*(x_t) - V_t^*(\hat{x}_t)$ : idéalement, on aimerait trouver $q$ de classe $\mathcal{K}_\infty$ telle que

$$\|e_t\| \leq q\left(V_t^*(x_t) - V_t^*(\hat{x}_t)\right) \tag{3.27}$$

Nous sommes déjà parvenus à réaliser le premier point, et le développement mathématique correspondant est développé en annexe A.6. Pour le deuxième point, notre idée est de d'appliquer le lemme 3.1 à la fonction $\mathcal{G}_t$ définie par

$$\mathcal{G}_t(e) = V_t^*(\hat{x}_t + e) - V_t^*(\hat{x}_t)$$

Cette fonction a comme propriété intéressante que pour $e = e_t$,

$$\mathcal{G}_t = V_t^*(x_t) - V_t^*(\hat{x}_t)$$

Le principe serait alors d'appliquer le lemme 3.1 : cela permettrait d'obtenir une inégalité du même type que (3.27). Cependant, nous n'avons pas encore réussi à démontrer l'appartenance de $\mathcal{G}_t$ à la classe $\mathcal{N}^n$.

C'est donc le point bloquant actuellement de cette démonstration. Dans le cas où les deux points seraient démontrés, leur mise en commun permettrait de prouver que l'erreur de l'estimation est bornée en présence de bruits bornés, puisque s'il existait $M \in \mathbb{R}^+$ tel que

$$\phi_k(w_k) + \psi_{k+1}(v_{k+1}) < M \tag{3.28}$$

on aurait alors

$$\mathcal{G}_{t+1}(e_{t+1}) \leq \lambda^{t+1}\mathcal{G}_0(e_0) + M\frac{1-\lambda^{t+1}}{1-\lambda} \tag{3.29}$$

On voit bien que dans le cas où le facteur d'oubli est strictement inférieur à 1,la partie droite de l'inégalité va tendre vers une constante quand $t$ tend vers l'infini.

## 3.5 Suite de l'étude

De nombreuses choses demeurent en cours quant à l'analyse de cette structure d'estimateur optimal :

— Dans le cas hors-ligne, une première borne supérieure pour l'erreur d'estimation a été trouvée. Cette borne est constante en présence de bruits bornés, ce qui montre que dans ce cas, l'erreur d'estimation sera bornée *a priori*. Cependant, il est encore nécessaire de l'étudier, ainsi que d'éventuellement chercher à l'améliorer. De plus, il reste encore à démontrer l'efficacité de l'estimateur en présence de bruit impulsif : de nombreuses pistes sont envisagées, mais aucun résultat notable n'a été obtenu au moment de l'écriture de ce rapport.

— Dans le cas en ligne, aucun résultat n'a encore été trouvé concernant une potentielle borne de l'erreur d'estimation : des pistes sont envisagées, notamment celle présentée dans la partie 3.4.2. Pour parvenir à relier la norme de $e_t$ à la valeur de $\mathcal{G}_t(e_t)$ il est envisagé d'utiliser un lemme ayant été démontré dans l'analyse hors-ligne (voir Annexe A.3 Lemme 3.1), mais son utilisation n'est pas immédiate.

— Afin de faire une étude complète de la structure d'estimateur, il est nécessaire de coupler cette analyse théorique avec des simulations permettant d'étudier plus quantitativement les performances de l'estimateur. Certains essais ont déjà été effectués, mais sans plan d'expérience permettant une analyse numérique pertinente.

# 4 Conclusion - Perspectives

Deux axes ont donc été développés simultanément durant cette première année de thèse. Partie intégrante d'une démarche scientifique qui se veut approfondie, l'étude bibliographique, présentée dans la partie 2, a permis de préciser le contexte scientifique dans lequel se situe le sujet de thèse. En expliquant l'intérêt de faire de l'estimation d'état et en présentant des résultats classiques issus de la littérature, nous avons pu définir précisément les objectifs de notre démarche.

Le deuxième axe, présenté dans la partie 3, est le développement d'une structure d'estimation optimale, nouvelle par la généralité de ses fonctions coûts. De nombreux points restent évidemment à explorer, que ce soit du point de vue de l'analyse même des performances de l'estimateur ou encore des améliorations pouvant y être apportées. Cependant, notre travail avance à un rythme convenable, et nous nous sommes fixés jusqu'à fin décembre pour effectuer l'ensemble des tâches énoncées dans la partie 3.5. Deux articles au moins sont prévus : le premier portera sur l'adaptation du problème en hors ligne à un problème à horizon glissant, et le deuxième devrait compiler l'ensemble de notre analyse sur l'estimateur une fois que celle-ci sera finalisée.

Par ailleurs, pour la suite de notre travail, nous aborderons le problème de la planification de capteurs : dans le cas d'un estimateur implémenté sur un calculateur détaché du système, il reste toutefois nécessaire de lui faire parvenir les mesures des différents capteurs. Or, le débit d'information se retrouve contraint par les moyens de transmission utilisés pour relier capteurs et calculateur : bien souvent, on se retrouve donc avec trop d'information à transmettre par rapport au temps disponible. En partant du principe que l'architecture ne peut être changée, l'enjeu va être de réussir à maintenir un bon niveau de performance en utilisant un nombre limité de capteurs : la « planification de capteurs » désigne alors le fait de choisir quelle mesure récupérer à quel moment. De plus, dans le cas où la bande-passante entre le système et le calculateur est suffisamment bien dimensionnée, la planification de capteur peut également avoir son importance : en effet, l'objectif pourrait alors être d'obtenir le même niveau de performance avec moins de capteurs utilisés. Ainsi, on pourrait garder le même nombre de capteurs tout en en utilisant moins, ce qui améliorerait la redondance (le fait d'avoir des capteurs prêts à être utilisés en cas de défaillance des capteurs actifs) et par conséquent la sécurité, qui est une thématique de plus en plus cruciale dans les réseaux par exemple.

Enfin, vous pourrez trouver en Annexe A.7 un planning prévisionnel donnant les grandes lignes de mon travail de recherche ainsi que les différentes échéances que je devrai respecter pour les deux années à venir.

# Références


[Afr+17] C. Afri, V. Andrieu, L. Bako et P. Dufour. "State and Parameter Estimation: A Nonlinear Luenberger Observer Approach". In : *IEEE Transactions on Automatic Control* 62.2 (2017), p. 973–980. doi : 10.1109/TAC.2016.2566804.

[Bel54] R. Bellman. "The theory of dynamic programming". In : *Bulletin of the American Mathematical Society* 60.6 (1954), p. 503–515. url : https://projecteuclid.org/euclid.bams/1183519147.

[Bla17] E. Blanco. "Introduction au filtrage optimal (Polycopié de cours)". In : (2017).

[CWB07] Emmanuel J. Candes, Michael B. Wakin et Stephen P. Boyd. "Enhancing Sparsity by Reweighted L1 Minimization". In : *arXiv:0711.1612 [math, stat]* (2007). arXiv : 0711.1612.

[DP68] J. Deyst et C. Price. "Conditions for asymptotic stability of the discrete minimum-variance linear estimator". In : *IEEE Transactions on Automatic Control* 13.6 (1968), p. 702–705. doi : 10.1109/TAC.1968.1099024.

[FTD14] H. Fawzi, P. Tabuada et S. Diggavi. "Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks". In : *IEEE Transactions on Automatic Control* 59.6 (2014), p. 1454–1467. doi : 10.1109/TAC.2014.2303233.

[HJ12] R. A. Horn et C. R. Johnson. *Matrix Analysis*. 2ème édition. Cambridge University Press, 2012.

[HS65] E. Hewitt et K. Stromberg. *Real and Abstract Analysis: A modern treatment of the theory of functions of a real variable*. Springer-Verlag, 1965. url : //www.springer.com/us/book/9783540780182.

[JU04] S. J. Julier et J. K. Uhlmann. "Unscented filtering and nonlinear estimation". In : *Proceedings of the IEEE* 92.3 (2004), p. 401–422. doi : 10.1109/JPROC.2003.823141.

[KA13] M. Korkali et A. Abur. "Robust Fault Location Using Least-Absolute-Value Estimator". In : *IEEE Transactions on Power Systems* 28.4 (2013), p. 4384–4392. doi : 10.1109/TPWRS.2013.2264535.

[Kal60] R. E. Kalman. "A new approach to linear filtering and prediction problems". In : *Journal of basic Engineering* 82.1 (1960), p. 35–45. doi : 10.1115/1.3662552.

[KB61] R. E. Kalman et R. S. Bucy. "New Results in Linear Filtering and Prediction Theory". In : *Journal of Basic Engineering* 83.1 (1961), p. 95–108. doi : 10.1115/1.3658902.

[Kha01] H. K. Khalil. *Nonlinear Systems*. Pearson, 2001.

[KSH00] T. Kailath, A. H. Sayed et B. Hassibi. *Linear Estimation*. Prentice Hall, 2000. url : https://infoscience.epfl.ch/record/233814.

[Lue64] D. G. Luenberger. "Observing the State of a Linear System". In : *IEEE Transactions on Military Electronics* 8.2 (1964), p. 74–80. doi : 10.1109/TME.1964.4323124.

[Mis+17] S. Mishra, Y. Shoukry, N. Karamchandani, S. N. Diggavi et P. Tabuada. "Secure State Estimation Against Sensor Attacks in the Presence of Noise". In : *IEEE Transactions on Control of Network Systems* 4.1 (2017), p. 49–59. doi : 10.1109/TCNS.2016.2606880.

[MS85] L. A. Mcgee et S. F. Schmidt. *Discovery of the Kalman filter as a practical tool for aerospace and industry*. 1985. url : https://ntrs.nasa.gov/search.jsp?R=19860003843.

[Oga09] K. Ogata. *Modern Control Engineering*. Pearson, 2009.

[Oro+18] J. O. Orozco-López, C. E. Castañeda, A. Rodríguez-Herrero, G. García-Sáez et E. Hernando. "Linear Time-Varying Luenberger Observer Applied to Diabetes". In : *IEEE Access* 6 (2018), p. 23612–23625. doi : 10.1109/ACCESS.2018.2825989.

[Paj+14] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee et G. J. Pappas. "Robustness of attack-resilient state estimators". In : *2014 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*. 2014 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS). 2014, p. 163–174. doi : 10.1109/ICCPS.2014.6843720.

[SH77] A. Sebald et A. Haddad. "Robust state estimation in uncertain systems: Combined detection-estimation with incremental MSE criterion". In : *IEEE Transactions on Automatic Control* 22.5 (1977), p. 821–825. doi : 10.1109/TAC.1977.1101602.

[ST16] Y. Shoukry et P. Tabuada. "Event-Triggered State Observers for Sparse Sensor Noise/Attacks". In : *IEEE Transactions on Automatic Control* 61.8 (2016), p. 2079–2091. doi : 10.1109/TAC.2015.2492159.

# A Annexes

## A.1 Rappel sur les signaux aléatoires

Le but de cette annexe est de donner rapidement les définitions des différents objets mathématiques liés aux signaux aléatoires et pouvant être utilisés au sein de ce rapport. Une grande partie de ce qui suit est inspiré du polycopié d'E. Blanco pour le cours de deuxième année à l'Ecole Centrale de Lyon intitulé « Estimation et Transmission de l'information » [Bla17].

Un signal aléatoire $x$ est un signal dont l'évolution est modélisée par une variable aléatoire $X$ dépendant du temps et dont la réalisation dictera la valeur du signal. Ainsi, pour tout $t$, $x(t)$ sera une réalisation d'une variable aléatoire $X(t)$ : cette modélisation ne donne donc pas qu'une seule valeur possible pour le signal mais une famille de signaux qui sont tous des réalisations temporelles possibles de $X$.

**Fonctions de répartitions et densités de probabilité :** pour décrire une variable aléatoire $X$ dépendant du temps, il est nécessaire d'avoir sa fonction de répartition $F_X$ définie sur $\mathbb{R}\times\mathbb{R}$ par

$$F_X(x,\tau) = P(X(\tau)\leq x)$$

où $P(X(\tau)\leq x)$ est la probabilité qu'une réalisation de la variable aléatoire $X(\tau)$ soit inférieure ou égale à $x$. Si la fonction $F_X(\cdot,t)$ admet une infinité de valeurs et que l'intégrale de sa dérivée sur $\mathbb{R}$ est égale à 1 (d'après le théorème de Lebesgue[†]), il est alors également possible de définir sur $\mathbb{R}\times\mathbb{R}$ une densité de probabilité $p_X$ de $X$ telle que

$$P(a\leq X(\tau)\leq b) = \int_a^b p_X(x,\tau)\mathrm{d}x$$

Il est à noter que $a$ et $b$ peuvent éventuellement être égaux à l'infini, et on retrouve bien

$$\int_{-\infty}^{+\infty} p_X(x,\tau)\mathrm{d}x = 1$$

Dans le cas de deux variables aléatoires $X$ et $Y$, on peut définir une loi conjointe dont la fonction de répartition s'écrit

$$F_{XY}(x,y,\tau_x,\tau_y) = P(X(\tau_x)\leq x, Y(\tau_y)\leq y)$$

où $P(X(\tau_x)\leq x, Y(\tau_y)\leq y)$ est la probabilité qu'une réalisation de $X(\tau_x)$ soit inférieure ou égale à $x$ *et* qu'une réalisation de $Y(\tau_y)$ soit inféireure ou égale à $y$. Si $F_{XY}(\cdot,\cdot,\tau_x,\tau_y)$ admet une infinité de valeur, on peut définir une densité de probabilité conjointe $p_{XY}$ telle que

$$F_{XY}(a,b,\tau_x,\tau_y) = \int_{-\infty}^{a}\int_{-\infty}^{b} p_{XY}(x,y,\tau_x,\tau_y)\mathrm{d}y\mathrm{d}x$$

†. E. Hewitt et K. Stromberg. *Real and Abstract Analysis: A modern treatment of the theory of functions of a real variable*. Springer-Verlag, 1965. URL : //www.springer.com/us/book/9783540780182.

**Espérance, Variance, Intercorrélation et Covariance :** On peut définir l'espérance $E[X(\tau)]$ d'une variable aléatoire $X$ à l'instant $\tau$ par la relation

$$E[X(\tau)] = \int_{\mathbb{R}} x p_X(x, \tau) \mathrm{d}x,$$

et sa variance par

$$Var[X(\tau)] = E[(X(\tau) - E[X(\tau)])^2].$$

Pour deux variables aléatoires $X$ et $Y$, on peut définir leur interocrrélation (auto-corrélation si $X = Y$) par

$$\Gamma_{XY}(\tau_x, \tau_y) = E[X(\tau_x)Y(\tau_y)]$$

et leur covariance mutuelle par

$$Cov[X(\tau_x, Y(\tau_y)] = E[(X(\tau_x) - E[X(\tau_x)])(Y(\tau_y) - E[Y(\tau_y)])].$$

Enfin, pour deux vecteurs $X(t) = \big(X_1(t), X_2(t), \cdots, X_n(t)\big)^\top$ et $Y(t) = \big(Y_1(t), Y_2(t), \cdots, Y_m(t)\big)^\top$ de variables aléatoires, on peut finalement définir la matrice de covariance comme la matrice $P_{XY}(t)$ de taille $n \times m$ telle que

$$P_{XY}(t) = Cov[XY^\top]$$

où l'opérateur $Cov[\cdot]$ s'applique terme à terme à la matrice $XY^\top$.

**Indépendance :** deux variables aléatoires $X$ et $Y$ sont dites indépendantes si la fonction de répartition de leur loi conjointe vérifie

$$\forall \tau_x, \tau_y \in \mathbb{R}, \qquad F_{XY}(x, y, \tau_x, \tau_y) = F_X(x, \tau_x) F_Y(x, \tau_y)$$

Intuitivement, cela s'interprète comme le fait que les deux processus aléatoires n'influent pas l'un sur l'autre, et que connaître la réalisation de l'un ne donne pas d'information sur la réalisation de l'autre.

**Blancheur :** Un signal aléatoire $x$ est dit blanc lorsque

1. $\forall t \in \mathbb{R}$, $E[X(t)] = 0$
2. $\forall t \in \mathbb{R}$, $Var[X(t)] = \sigma^2$
3. $\forall t_1, t_2 \in \mathbb{R}$, $\Gamma_X(t_1, t_2) = \Gamma_X(0, 0)\delta_{t_1, t_2}$

où *sigma* est une constante positive et $\delta$ est le symbole de Kronecker. On suppose donc que le signal est stationnaire (l'espérance et la variance ne dépendent pas du temps) et que les différentes variables aléatoires $X(t)$ n'ont aucun lien entre elles.

**Variable aléatoire gaussienne :** Un variable aléatoire gaussienne est une variable pour laquelle la densité de probabilité s'écrit

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

où $\mu$ est la moyenne statistique/espérance de la variable et $\sigma^2$ sa variance.

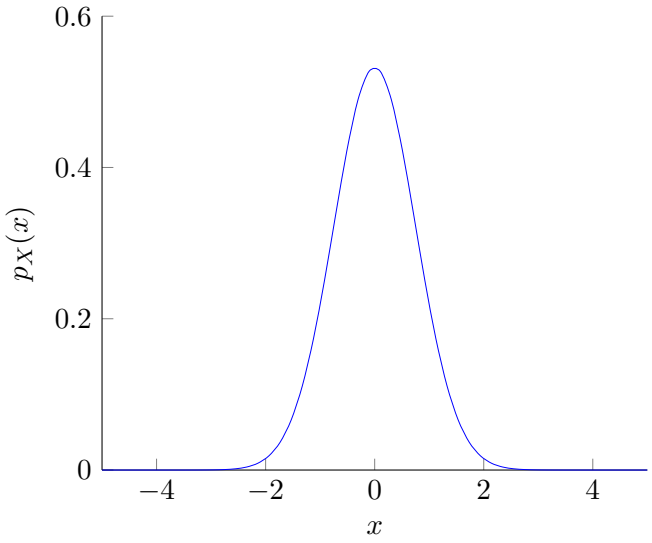



FIGURE 3 – Densité d'une variable aléatoire de moyenne nulle et de variance $0.75^2$

## A.2 Exemple montrant la parcimonie de $\ell_1$

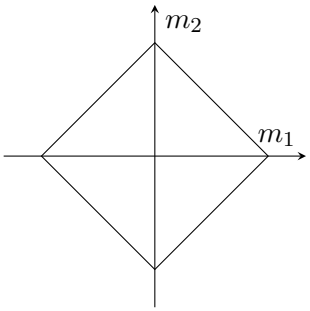
Dans cette annexe, on essaie de mettre en évidence le fait que la norme $\ell_1$ a tendance à encourager la parcimonie des solutions d'un problème d'optimisation au travers d'un exemple simple. On considère ainsi les problèmes d'optimisation sur $\mathbb{R}^2$

$$\underset{m=(m_1,m_2)\in\mathbb{R}^2}{\operatorname{argmin}} \|m\|_{\ell_1} \text{ avec } m_1 = am_2 + b \tag{P1}$$

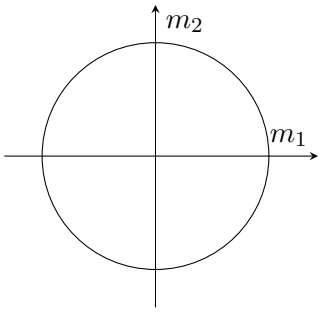
et

$$\underset{m=(m_1,m_2)\in\mathbb{R}^2}{\operatorname{argmin}} \|m\|_{\ell_2} \text{ avec } m_1 = am_2 + b \tag{P2}$$

Les problèmes consistent donc à trouver le point sur la droite d'équation $y = ax + b$ ayant la plus petite norme ($\ell_1$ ou $\ell_2$ suivant le problème). La figure 4 présente une comparaison de l'allure des lignes de niveau des normes, c'est-à-dire les courbes sur lesquelles tous les points ont la même norme. Ainsi, on constate que les lignes de niveau de $\ell_1$ sont des carrés avec les coins sur les axes, tandis que les lignes de niveau de $\ell_2$ sont des cercles. Les lignes de niveau ont leur importance, car les solutions des problèmes




(a) Ligne de niveau de la norme $\ell_1$




(b) Ligne de niveau de la norme $\ell_2$

FIGURE 4 – Comparaison des lignes de niveau pour les deux normes

$(P1)$ et $(P2)$ sont les points de la droite d'équation $y = ax + b$ intersectant la ligne de niveau associée à la plus petite valeur. Un exemple de résolution est présenté sur la figure 5.

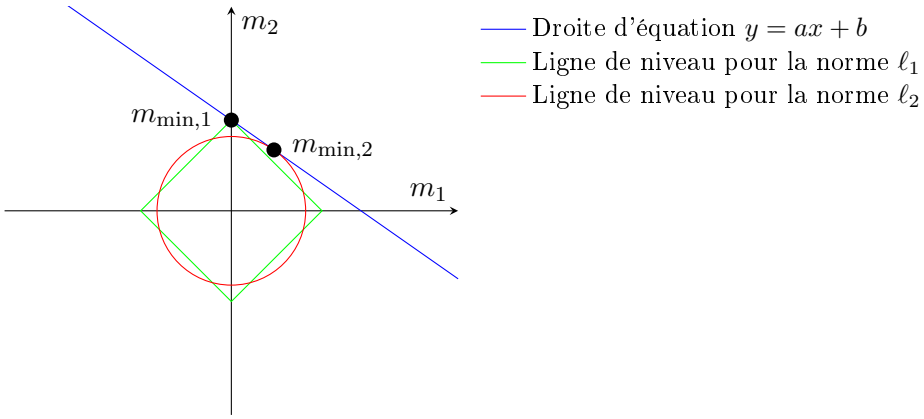



FIGURE 5 – Exemple de solutions pour les problèmes $(P1)$ et $(P2)$

Comme les lignes de niveau associées à $\ell_2$ sont des cercles centrés en 0, la droite est tangente à la ligne de niveau passant par la solution $m_{\min,2}$ de $(P2)$, ce qui ne présage pas d'une potentielle parcimonie dans l'argument minimisant $(P2)$. Les lignes de niveau de la norme $\ell_1$, ne peuvent intersecter la droite d'équation $y = ax + b$ qu'au niveau d'un de ses coins. Cela veut donc dire que dans tous les cas, la solution $m_{\min,1}$ de $(P1)$ sera creuse car une de ses deux coordonnées sera nulle.

## A.3 Démonstration du Théorème 3.1

Pour démontrer ce théorème, l'idée est de partir de la définition de $\hat{X}$, à savoir

$$\hat{X}_t \in \underset{Z_t}{\operatorname{argmin}}\{V_t(Z_t)\},$$

ce qui nous permet d'écrire

$$V_t(\hat{X}_t) \leq V_t(X_t)$$

En reprenant l'expression (3.2) de $V_t$, on a alors

$$V_t(\hat{X}_t) \leq V_t(X_t) \Leftrightarrow \sum_{k=0}^{t-1} \phi_k(\hat{x}_{k+1} - A_k\hat{x}_k) + \sum_{k=0}^{t} \psi_k(y_k - C_k\hat{x}_k) \tag{A.1}$$

$$\leq \sum_{k=0}^{t-1} \phi_k(x_{k+1} - A_k x_k) + \sum_{k=0}^{t} \psi_k(y_k - C_k x_k) \tag{A.2}$$

Or, si on applique l'inégalité triangulaire (3.9) à $\phi_k$ et à $\psi_k$ pour tout $k$, alors on peut écrire

$$\phi_k(\hat{x}_{k+1} - A_k\hat{x}_k) = \phi_k(e_{k+1} - A_k e_k + x_{k+1} - A_k x_k) \geq \alpha_k\phi_k(e_{k+1} - A_k e_k) - \phi_k(A_{k+1} - A_k e_k) \tag{A.3}$$

$$\psi_k(y_k - C_k\hat{x}_k) = \psi_k(y_k - C_k x_k - C_k e_k) \qquad \geq \beta_k\psi_k(C_k e_k) - \psi_k(y_k - C_k x_k) \tag{A.4}$$

De plus, on sait que les expressions $z_{k+1} - A_k z_k$ et $y_k - C_k z_k$ ont une valeur particulière lorsque $Z_t = X_t$. En l'occurrence, on a

$$x_{k+1} - A_k x_k = \omega_k \tag{A.5}$$

$$y_k - C_k x_k = \nu_k \tag{A.6}$$

Par conséquent, en reportant les inégalités (A.3) et (A.4) dans la partie gauche (A.1) de l'inégalité précédente, on obtient

$$\sum_{k=0}^{t-1}\phi_k(\hat{x}_{k+1}-A_k\hat{x}_k)+\sum_{k=1}^{t}\psi_k(y_k-C_k\hat{x}_k)\geq\sum_{k=0}^{t-1}\alpha_k\phi_k(e_{k+1}-A_ke_k)+\sum_{k=1}^{t}\beta_k\psi_k(C_ke_k) \tag{A.7}$$

$$-\left(\sum_{k=0}^{t-1}\phi_k(x_{k+1}-A_kx_k)+\sum_{k=1}^{t}\psi_k(y_k-C_kx_k)\right) \tag{A.8}$$

En combinant cette inégalité avec les formules (A.5) et (A.6) pour les injecter dans l'inégalit (A.1)/(A.2), on obtient finalement que

$$V_t(\hat{X}_t)\leq V_t(X_t)\Rightarrow\sum_{k=0}^{t-1}\alpha_k\phi_k(e_{k+1}-A_ke_k)+\sum_{k=1}^{t}\beta_k\psi_k(C_ke_k) \tag{A.9}$$

$$\leq 2\left(\psi_0(x_0-\mu_0)+\sum_{k=0}^{t-1}\phi_k(\omega_k)+\sum_{k=1}^{t}\psi_k(\nu_k)\right) \tag{A.10}$$

On pose

$$F_t(E_t)=\sum_{k=0}^{t-1}\alpha_k\phi_k(e_{k+1}-A_ke_k)+\sum_{k=1}^{t}\beta_k\psi_k(C_ke_k) \tag{A.11}$$

$F_t$ est une fonction de $\mathbb{R}^{n\times t}$ dans $\mathbb{R}^+$ : cependant, sans perte de généralité, on peut considérer $F_t$ comme étant une fonction sur un vecteur de $\mathbb{R}^{(n+1)t}$, puisqu'il suffit de vectoriser les matrices, c'est-à-dire de les réarranger sous forme de vecteur colonne.

L'idée est alors d'appliquer le lemme 3.1 à $F_t$. Pour cela, il est maintenant question de montrer que $F_t$ est bien de classe $\mathcal{K}^{(n+1)t}$ :

— $F_t$ est continue car composée uniquement de fonctions continues

— $F_t$ est positive en tant que somme de fonctions positive. De plus, si $F_t(E_t)=0$, alors chaque terme de la somme qui compose $F_t(E_t)$ doit être nul puisqu'ils sont tous positifs. Ainsi, on a

$$\begin{cases}\forall k, & \phi_k(e_{k+1}-A_ke_k)=0\\ \forall k & \psi_k(C_ke_k)=0\end{cases} \tag{A.12}$$

en particulier, pour $k=0$, on constate que $\psi_0(C_0e_0)=0$, ce qui implique que $C_0e_0=0$ puisque par définitions, les $\phi_k$ et les $\psi_k$ ne s'annulent qu'en 0. On remarque également que $\phi_0(e_1-A_0e_0)=0$ implique $e_1=A_0e_0$. Par récurrence immédiate, on se rend compte que pour que $F_t(E)$ soit égal à zéro, il faut que

$$\begin{pmatrix}C_0\\ C_1A_0\\ C_2A_1A_0\\ \vdots\\ C_tA_{t-1}A_{t-2}\dots A_0\end{pmatrix}e_0=0$$

La matrice d'observabilité sur l'horizon $[0;t]$ (voir la sous-partie 2.1.2) apparaît donc naturellement, et cette equation implique $e_0=0$ (et par conséquent $E_t=0$) puisque la matrice est de rang plein du fait que le système est observable sur l'horizon $[0;t$.

— On sait déjà que toutes les fonctions $\phi_k$ et $\psi_k$ vérifient une version généralisée de l'homogénéité (3.8) : on a donc déjà

$$\forall k \geq 0,\ \forall \lambda \in \mathbb{R},\ \begin{cases} \alpha_k \phi_k(e_{k+1} - A_k e_k) \geq g_k\left(\frac{1}{|\lambda|}\right) \phi_k(\lambda e_{k+1} - A_k(\lambda e_k)) \\ \beta_k \psi_k(C_k e_k) \geq h_k\left(\frac{1}{|\lambda|}\right) \psi_k(C_k(\lambda e_k)) \end{cases}$$

Pour montrer que $F_t$ vérifie l'inégalité (3.8), nous allons montrer que si une fonction quelconque $\xi$ a pour structure

$$\xi(z) = \sum_{i \in I} \xi_i(z) \tag{A.13}$$

avec $I$ de cardinalité finie et pour tout $i$, $\xi_i$ vérifiant l'inégalité (3.8) pour une fonction $q_i$, alors elle vérifie (3.8). Or, il est évident qu'à $\lambda$ fixé, on a

$$\xi_i(z) \geq \left(\min_{i \in I} q_i\left(\frac{1}{|\lambda|}\right)\right) \xi_i(\lambda z)$$

Par conséquent, en posant pour tout $\lambda \in \mathbb{R}$,

$$q(\lambda) = \min_{i \in I} q_i(\lambda), \tag{A.14}$$

On a

$$F(z) = \sum_{i \in I} F_i(z) \geq \sum_{i \in I} q\left(\frac{1}{|\lambda|}\right) F_i(\lambda z) = q\left(\frac{1}{|\lambda|}\right) F(\lambda z)$$

donc $\xi$ vérifie (3.8), et par conséquent $F_t$ également.

Par application du Lemme 3.1, on en déduit que

$$\forall E_t \in \mathbb{R}^{n \times t}, \qquad F_t(E_t) \geq D q_t(\|E_t\|)$$

avec

$$D = \min_{\|Z_t\|=1} F_t(Z_t)$$
$$\forall \lambda \in \mathbb{R},\ q_t(\lambda) = \min_{k \in [0; t-1]} \{g_k(\lambda), h_{k+1}(\lambda)\}$$

L'utilisation de cette inégalité avec l'inégalité (A.9)/(A.10)permet finalement écrire

$$D q_t(\|E\|) \leq 2\left(\sum_{k=0}^{t-1} \phi_k(\omega_k) + \sum_{k=1}^{t} \psi_k(\nu_k)\right) \tag{A.15}$$

$$\Leftrightarrow \|E\| \leq q_t^{-1}\left(\frac{2\left(\sum_{k=0}^{t-1} \phi_k(\omega_k) + \sum_{k=1}^{t} \psi_k(\nu_k)\right)}{D}\right) \tag{A.16}$$

Enfin, poser

$$b_t = \sum_{k=0}^{t-1} \phi_k(\omega_k) + \sum_{k=1}^{t} \psi_k(\nu_k)$$

permet d'achever la démonstration. □

## A.4 Démonstration du Théorème 3.2

On cherche à montrer que

$$V_t^*(z) = \min_s \left\{\lambda V_{t-1}^*(s) + \phi_{t-1}(z - A_{t-1}s)\right\} + \psi_t(y_t - C_t z))$$

D'après la définition de $V_t^*$, on a

$$V_t^*(z) = \min_{Z_t} \{V_t(Z_t) \text{ avec } z_t = z\} \tag{A.17}$$

$$= \min_e \left\{ \min_{Z_t} \{V_t(Z_t) \text{ avec } z_t = z, z_{t-1} = s\} \right\}, \tag{A.18}$$

puisque l'ordre des variables par rapport auxquelles on minimise n'a pas d'importance.

D'après la définition (3.2) de $V_t$, on obtient la relation de récurrence

$$V_t(Z_t) = V_{t-1}(Z_{t-1}) + \phi_{t-1}(z_t - A_{t-1}z_{t-1}) + \psi_t(y_t - C_t z_t) \text{ avec } Z_t = \begin{pmatrix} Z_{t-1} & z_t \end{pmatrix}, \tag{A.19}$$

ce qui nous permet de déduire

$$V_t^*(z) = \min_e \left\{ \min_{Z_{t-1}} \{V_{t-1}(Z_{t-1}) + \phi_{t-1}(z - A_{t-1}s) + \psi_t(y_t - C_t z) \text{ avec } z_{t-1} = s\} \right\} \tag{A.20}$$

car $Z_t$ n'est plus dans l'expression à minimiser : cette dernière peut donc être effectuée par rapport à $Z_{t-1}$ et la condition sur $z_t$ disparait. Enfin, le terme $\phi_{t-1}(z - A_{t-1}s)$ ne dépend pas de $Z_{t-1}$ donc on peut le sortir de la première minimisation, et le terme $\psi_t(y_t - C_t z)$ ne dépend ni de $Z_{t-1}$ ni de $s$ donc il peut être sorti des deux minimisations :

$$V_t^*(z) = \min_s \left\{ \min_{Z_{t-1}} \{V_{t-1}(Z_{t-1}) \text{ avec } z_{t-1} = s\} + \phi_{t-1}(z - A_{t-1}s) + \psi_t(y_t - C_t z) \right\} \tag{A.21}$$

$$= \min_s \left\{V_{t-1}^*(s) + \phi_{t-1}(\alpha - A_{t-1}s)\right\} + \psi_t(y_t - C_t z) \tag{A.22}$$

ce qui finit de démontrer le théorème.

□

## A.5 Démonstration du Théorème 3.3

Dans le cas où nous sommes, $V_t$ s'écrit

$$\begin{aligned} \forall Z_t \in \mathbb{R}^{n \times t+1}, \quad V_t(Z_t) = &\frac{1}{2}(z_0 - \mu_0)^\top S^{-1}(z_0 - \mu_0) \\ &+ \sum_{k=0}^{t-1} \frac{1}{2}(z_{k+1} - A_k z_k)^\top Q_k^{-1}(z_{k+1} - A_k z_k) \\ &+ \sum_{k=1}^{t} \frac{1}{2}(y_k - C_k z_k)^\top R_k^{-1}(y_k - C_k z_k), \end{aligned}$$

Nous allons procéder par récurrence sur $t$. Pour $t = 0$, on a

$$V_0^*(z) = V_0(z) = \frac{1}{2}(z - \mu_0)^\top S^{-1}(z - \mu_0). \tag{A.23}$$

En posant $P_0 = S$, $\hat{x}_0 = \mu_0$ et $r_0 = 0$, $V_0^*$ peut alors être réécrit sous la forme

$$V_0^*(z) = \frac{1}{2}(z - \hat{x}_0)^\top P_0^{-1}(z - \hat{x}_0) + r_0 \tag{A.24}$$

ce qui est conforme au théorème.

Supposons maintenant que le terme est vrai jusqu'au rang $t$. A ce moment là, la fonction coût $V_{t+1}$ va s'écrire, d'après l'équation de Bellman

$$\begin{aligned}
V_{t+1}^*(z) = & \min_s \{V_t^*(s) + \phi_t(z - A_t s)\} + \psi_{t+1}(y_{t+1} - C_{t+1} z) \\
= & \min_s \left\{ \frac{1}{2}(s - \hat{x}_t)^\top P_t^{-1}(s - \hat{x}_t) + r_t + \frac{1}{2}(z - A_t s)^\top Q_t^{-1}(z - A_t s) \right\} \\
& + \frac{1}{2}(y_{t+1} - C_{t+1} z)^\top R_{t+1}^{-1}(y_{t+1} - C_t z) \\
= & \frac{1}{2}(s_{\min} - \hat{x}_t)^\top P_t^{-1}(s_{\min} - \hat{x}_t) + r_t + \frac{1}{2}(z - A_t s_{\min})^\top Q_t^{-1}(z - A_t s_{\min}) \\
& + \frac{1}{2}(y_{t+1} - C_{t+1} z)^\top R_{t+1}^{-1}(y_{t+1} - C_t z)
\end{aligned}$$

où $s_{\min}$ est obtenu en différentiant la fonction à minimiser par rapport à $s$ : $s_{\min}$ vaut alors

$$s_{\min} = M_t^{-1}\left(P_t^{-1}\hat{x}_t + A_t^\top Q_t^{-1}\alpha\right) \text{avec } M_t = P_t^{-1} + A_t^\top Q_t^{-1} A_t. \tag{A.25}$$

En remplaçant $e_{\min}$ par son expression, $V_{t+1}^*$ peut être réarrangé sous la forme

$$V_{t+1}^*(z) = \frac{1}{2} z^\top \Delta_t z + \rho_t^\top z + \theta_t \tag{A.26}$$

où $\Delta_t$, $\rho_t$ et $\theta_t$ sont égaux à

$$\Delta_t = \left(Q_t + A_t P_t A_t^\top\right)^{-1} + C_{t+1}^\top R_{t+1}^{-1} C_{t+1} \tag{A.27}$$

$$\rho_t = -Q_t^{-1} A_t M_t^{-1} P_t^{-1} \hat{x}_t - C_{t+1}^\top R_{t+1}^{-1} y_{t+1} \tag{A.28}$$

$$\theta_t = r_t + \frac{1}{2} y_{t+1}^\top R_{t+1}^{-1} y_{t+1} + \frac{1}{2}\hat{x}_t^\top A_t^\top \left(Q_t + A_t P_t A_t^\top\right)^{-1} A_t \hat{x}_t. \tag{A.29}$$

Ensuite, en utilisant la formule

$$\forall z_1, z_2 \in \mathbb{R}^n, \forall \in \mathcal{S}_n^+, \qquad \frac{1}{2} x^\top Q x + z^\top x = \frac{1}{2}(x + Q^{-1} z)^\top Q (x + Q^{-1} z) - \frac{1}{2} z^\top Q^{-1} z,$$

on obtient

$$V_{t+1}^*(z) = \frac{1}{2}(z + \Delta_{t+1}^{-1}\rho_t)^\top \Delta_t^{-1}(z + \Delta_{t+1}^{-1}\rho_t) + r_t - \rho_t^\top \Delta_{t+1}^{-1}\rho_t + \theta_t \tag{A.30}$$

Si l'on pose

$$P_{t+1} = \Delta_t^{-1} \tag{A.31}$$

$$\hat{x}_{t+1} = \Delta_{t+1}^{-1}\rho_t \tag{A.32}$$

$$r_{t+1} = r_t - \rho_t^\top \Delta_{t+1}^{-1}\rho_t + \theta_t, \tag{A.33}$$

on peut alors écrire

$$V_{t+1}^*(z) = (z - \hat{x}_{t+1})^\top P_{t+1}^{-1}(z - \hat{x}_{t+1}) + r_{t+1} \tag{A.34}$$

On constate que $r_t$ est bien une quantité indépendant de $z$, et $P_{t+1}$ vérifie la définition générale (3.26) de $P_t$. Concernant $\hat{x}_{t+1}$, on a

$$\hat{x}_{t+1} = P_{t+1}^{-1} Q_t^{-1} A_t M_t^{-1} P_t^{-1} \hat{x}_t + P_{t+1}^{-1} C_{t+1}^\top R_{t+1}^{-1} y_{t+1} \tag{A.35}$$

On peut finalement montrer que

$$P_{t+1}^{-1} Q_t^{-1} A_t M_t^{-1} P_t^{-1} = A_t - P_{t+1}^{-1} C_{t+1}^\top R_{t+1}^{-1} C_{t+1} A_t$$

ce qui achève la démonstration.

□

## A.6 Éléments mathématiques de la piste d'analyse en ligne

Pour trouver une borne supérieure de $V_t^*(x_t) - V_t^*(\hat{x}_t)$, l'idée est de s'intéresser à son évolution lorsque le temps croît. En utilisant le formalisme présenté dans la partie 3.4.2, c'est-à-dire avec $\mathcal{G}_t(e_t) = V_t^*(x_t) - V_t^*(\hat{x}_t)$ on a tout d'abord que

$$\begin{aligned}
\mathcal{G}_t(e_t) &= V_{t+1}^*(x_{t+1}) - V_{t+1}^*(\hat{x}_{t+1}) && \text{(A.36)}\\
&= \min_s \{\lambda V_t^*(s) + \phi_t(x_{t+1} - A_t s)\} + \psi_{t+1}(y_{t+1} - C_{t+1} x_{t+1}) && \text{(A.37)}\\
&\quad - \Big(\min_s \{\lambda V_t^*(s) + \phi_t(\hat{x}_{t+1} - A_t s)\} + \psi_{t+1}(y_{t+1} - C_{t+1}\hat{x}_{t+1})\Big) && \text{(A.38)}
\end{aligned}$$

d'après l'équation de Bellman

Or, par définition du minimum, la valeur d'une fonction pour un argument quelconque est toujours supérieure ou égale à son minimum, d'où

$$\min_s \{\lambda V_t^*(s) + \phi_t(x_{t+1} - A_t s)\} \leq \lambda V_t^*(\hat{x}_t) + \phi_t(x_{t+1} - A_t \hat{x}_t). \tag{A.39}$$

De plus, si on prend $s_{\min}$ dans $\operatorname{argmin}_s V_t^*(s)$, c'est-à-dire

$$\min_s \{\lambda V_t^*(s) + \phi_t(\hat{x}_{t+1} - A_t s\} = \lambda V_t^*(s_{\min}) + \phi_t(\hat{x}_{t+1} - A_t s_{\min}), \tag{A.40}$$

alors comme

$$V_t^*(\hat{x}_t) \leq V_t^*(s_{\min}) \tag{A.41}$$

du fait que par définition, $\hat{x}_t$ minimise la fonction $V_t^*$, on obtient finalement

$$\min_s \{\lambda V_t^*(s) + \phi_t(\hat{x}_{t+1} - A_t s)\} \geq \lambda V_t^*(\hat{x}_t) + \phi_t(\hat{x}_{t+1} - A_t s_{\min}) \tag{A.42}$$

En injectant (A.39) et (A.42) dans (A.36), on obtient une *inégalité de récurrence*

$$\begin{aligned}
\mathcal{G}_{t+1}(e_{t+1}) \leq \lambda \mathcal{G}_t(e_t) &+ \phi_t(x_{t+1} - A_t x_t) + \psi_{t+1}(y_{t+1} - C_{t+1} x_{t+1}) && \text{(A.43)}\\
&- \phi_t(\hat{x}_{t+1} - A_t s_{\min}) - \psi_{t+1}(y_{t+1} - C_{t+1}\hat{x}_{t+1}) && \text{(A.44)}
\end{aligned}$$

Tous les termes présents dans la partie droite de l'inégalité ne sont pas facilement interprétable : cependant,on sait par exemple que

$$\begin{cases} x_{t+1} - A_t x_t = \omega_t \\ y_{t+1} - C_{t+1} x_{t+1} = \nu_{t+1} \end{cases}$$

ce qui, injecté dans (A.43) et en enlevant les termes négatifs de la droite de l'inégalité (ce qui a pour conséquence d'élargir la borne supérieure), permet d'obtenir

$$\mathcal{G}_{t+1}(e_{t+1}) \leq \lambda \mathcal{G}_t(e_t) + \phi_t(\omega_t) + \psi_{t+1}(\nu_{t+1}) \tag{A.45}$$

Par récurrence immédiate, on déduit alors que

$$\mathcal{G}_{t+1}(e_{t+1}) \leq \lambda^{t+1} \mathcal{G}_0(e_0) + \sum_{k=0}^{t} \lambda^{t-k} \left( \phi_k(\omega_k) + \psi_{k+1}(\nu_{k+1}) \right) \tag{A.46}$$

## A.7 Planning prévisionnel pour les deux prochaines années

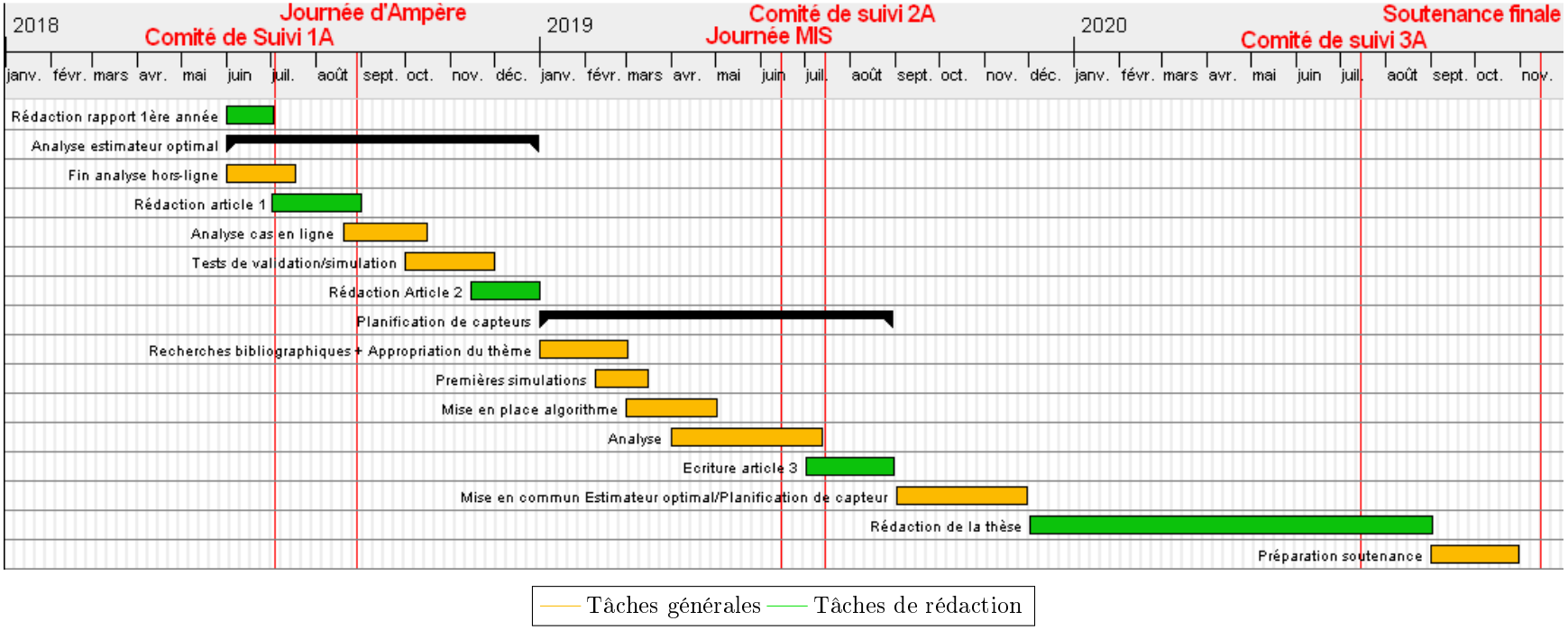



FIGURE 6 – Planning prévisionnel pour la suite de la thèse